LEVERAGING GENOTYPES IMPUTATION AND POLYGENIC RISK SCORES IN MALARIA SUSCEPTIBILITY

A Masters Dissertation

By

Peter Opiyo Kimathi

Division of Human Genetics

Department of Pathology

Faculty of Health Sciences (FHS)

Email Address: kmtpet002@myuct.ac.za

Supervisor: Prof. Emile R. Chimusa

Email Address: emile.chimusa@uct.ac.za

December 2019

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

DECLARATION

I, **Kimathi Peter**, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: December 13, 2019

Acknowledgment

Firstly, I would like to thank by supervisor Prof. Emile R. Chimusa for his motivation, immense knowledge and timely feedback. His continuous guidance and support have been invaluable throughout this dissertation process.

My sincere thanks goes to the Developing Excellence in Leadership and Genetics Training for Malaria Elimination in sub-Saharan Africa (DELGEME) for the financial support and trainings in various areas of bioinformatics.

I'm also grateful to my fellow colleagues in Prof. Emile's research group, and the UCT Human Genetic team for stimulating discussions and all the fun during the two years.

Finally, I thank my family members, friends and relatives for their spiritual and moral support.

Related Publications

1. Peter Opiyo Kimathi, Jacquiline W. Mugo, and Prof. Emile Chimusa
   A Review of Polygenic Risk Scores Methods. Under review 2019.
2. Peter Opiyo Kimathi, Jacquiline W. Mugo, Wonderful T. Choga1, Kabongo E., Francis E. Agamah, Gaston K. Mazandu, and Prof. Emile Chimusa
   Evaluation of Genotypes Imputation methods performance in African population. Under review 2019.
3. Wonderful T. Choga1, Kabongo E. Ntumba1, Peter O. Kimathi, Francis E. Agamah, Gaston K. Mazandu, Emile R. Chimusa
   Progress on Functional Genome-wide Association Studies: Applications, Benefits and Limits. Under review 2019.

List of Figures

List of Tables

Abstract

Background

Over the past few years, Genome Wide Association Studies (GWAS) have identified thousands of genetic variants that are associated with a wide range of complex traits, and have provided valuable insights as far as their genetic architectures are concerned. In malaria studies too, GWAS has been successful and a number of genetic variants have been identified. Despite the success, the complete aetiology of malaria, and many complex traits in general, remains poorly understood. A key concern is that the missing heritability remains too large, with some of the variants identified in some populations failing to replicate using independent study populations. Indeed comparable sources have revealed that the statistical power of association studies can be improved either via genotypes imputation approaches or by treating the whole genome of an individual as a risk predictor using Polygenic Risk Scores (PRS). However, imputation remains at modest in Africa populations with few (or no) studies (study) have evaluated the potential of imputation tools in African populations. On the other hand, although the utility of PRS has been shown in other studies, it has neither been assessed in African population nor applied in an infectious disease, like malaria.

Methodology

We evaluated the performance of five popular genotypes imputation methods (IMPUTE4, minimac 4, IMPUTE2, minimac3 and BEAGLE4) using case control datasets that mimics African populations, European populations and the admixed populations simulated with FractalSIM. We assessed imputation performance based on internal imputation quality metrics and the genotypes concordance. We applied the best imputation tool based on the assessment results to impute raw genotypes data of severe malaria case control studies from MalariaGEN of three African populations: Kenya, The Gambia and Malawi. Similarly, we obtained summary statistics of the same datasets, and imputed the summary statistics with ImpG. We performed an association on the imputed raw genotypes, and compared the association results with that of ImpG based imputation. Additionally, we performed meta-analysis with METASOFT, and compared the meta-analysis result of ImpG based imputation and that from imputed raw genotypes associations. Finally, we assessed five PRS methods (PRSice, LDpred(p+t), PRSoS, PLINK and PRScS) in predicting genetic

risk in African population, and applied the best PRS method to predict the genetic risk of severe malaria.

Results

IMPUTE2 recorded the best performance based on imputation accuracy and concordance for the African (accuracy=80.21% and concordance=99.2%) and the admixed samples (accuracy=69.46% and concordance=90.92%) for variants with MAF>0.05. Other tools recorded similar accuracy and concordance although BEAGLE 4 recorded the lowest concordance and accuracy across all the African and admixed datasets. For the real genotypes data, no SNP attained the genome wide significant threshold of 5.0×10^{-8} *for Malawi and the Gambia datasets. However, for the Kenyan dataset, 9 SNPs on chromosome 11 were significantly associated with severe malaria. 3 of these SNPS were located on the HBG2 genes and the remaining 6 had not been reviewed. No SNP attained the genome wide threshold for the ImpG imputed summary statistics for all the populations. For IMPUTE2 based meta-analysis, only one SNP rs12295158 located on the HBB region was significant across all the meta-analysis model (with P-value of* 2.88×10^{-12} *for fixed (FE),* 2.88×10^{-12} *random (RE) and* 9.64×10^{-12} *binary effect (BE) respectively). On the other hand ImpG based meta-analysis, two SNPs were signicant across all the meta-analysis model (rs183731078 located on RFX3 with P-values of* $8.40 \times 10^{-9}, 8.40 \times 10^{-9}, 4.47 \times 10^{-8}$ *for FE, RE and BE respectively, and rs8096513 located on DLGAP1* $1.43 \times 10^{-9}, 1.43 \times 10^{-9}, 1.01 \times 10^{-8}$ *with P-value for FE, RE and BE respectively). Pathway enrichment and analysis of these genes revealed that both of these genes are associated with malaria. Finally, for the PRS, PRSoS recorded the best performance based on Nargalkerke's* R^2 *(0.01736) and area under curve (AUC) (0.511). Other PRS methods recorded slightly similar results with PLINK recording the least. The odds of having severe malaria was estimated as 2.869, and a unit change of PRS scores was associated with -5.143 change in odds of having severe malaria with P-value of 0.0193 at* $\alpha = 0.05$*. However, the scores could only explain 1.28% of the phenotypic variance.*

Conclusion

Our results provide foundation for future studies in genetics, especially in African population, where the best performing imputation tool remains a mystery. Moreover, our results have demonstrated the potential of application of PRS in infectious diseases.

Contents

Chapter 1

Introduction

1.1 Overview

Malaria is one of the oldest infectious diseases that continues to be among the leading causes of morbidity and mortality across the world. Characteristics symptoms of malaria were first described by Neu ching, the Chinese medical canon [1], and were first edited by emperor Huang Ti, and can be traced back to as early as 2700 BC. It was also documented by most of early civilized societies like Greek, Roman, Indian and Arabs [1]. The characteristic symptoms are manifested inform of fevers, chills, general malaise, headaches, nausea and vomiting, body aches among others [2].

Malaria infection in human is caused by a protozoan parasite of genus *Plasmodium* [3]. It is transmitted to human being by a bite from a female anopheles mosquito [2]. Five known species of plasmodium parasites cause malaria in human i.e *Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale, Plasmodium malariae and Plasmodium knowlesi* [4, 5]. Of the five known species of Plasmodium genus, *Plasmodium vivax* is the most widespread and *Plasmodium falciparum* is the most virulent, and is the cause of high morbidity and mortality, especially in Africa [4, 6].

Despite the fact that malaria can be prevented by mosquito vector controls through use of mosquito nets, removal of breeding sites of mosquitoes and use of insecticides, among other effective case managements that have been advocated by World Health Organization (WHO), it is only drugs that can cure an infection from malaria [1]. While malaria continue to claim more lives with death toll rising every year, attempts to reduce this burden in most cases are thwarted by the fact that malaria parasites keep on developing resistance to the anti malarial drugs that are being invented [4, 6]. This has presented a serious health challenge and has consequently made its elimination and control very difficult [7]. Although efforts have been made to come up

with an effective vaccine that can prevent an infection from *falciparum malaria*, which is the most common and the most dangerous form of malaria in Africa, none has so far materialized [3]

One challenge in eliminating malaria is the understanding complex interaction that emanates from the parasite and the host [3]. These complex interactions have made the organisms to evolve strategies that have resulted into signatures of selections being exerted on the parasite genomes by the human host, and parasite too exerting footprints of malaria selection on the human genomes [3]. Other factors that adds more complexities to these interactions include environmental factors, parasite virulence, transmission intensity, different genetic background of the host among others [8]. It is therefore very important to understand the molecular basis or mechanism of these interactions, and a comprehensive understanding of the host and parasite factors [3]. Such understanding can inform the development of effective antimalarial drugs, effective vaccines and can also be applied in the risk prediction strategies [9].

On the other hand, Genome Wide Association Studies (GWAS) has been applied in the identification of variants associated with susceptibility or risks in complex traits, and for understanding the genetic architecture of complex traits [9]. Similarly, in malaria studies too, GWAS has been successful and a number of genes that confer susceptibility and protection against malaria have been identified [10]. Nevertheless, the overall contribution of the human genetics factors on malaria risk remains very low. Mackinnon et al. [11] for example used a pedigree data and estimated the overall host genetic factors to be approximately 25% [11]. This is similar to a recent study in Tanzania which reported that approximately 22% of the total variation in malaria risk can be explained by the genetics factors of the host [12]. While cognizant of the fact that it is difficult to generalize the overall estimate of the host genetic factors that contribute to a disease like malaria that involves the interaction of both genetic and environmental factors, which are inseparable in most study designs [11], and with the implicated variants being population or location specific [12] , it is clear that these estimates are still very low. Furthermore, most of these associations never replicated in different populations and several conflicting findings have been reported as well [13, 14]. By the fact most of the studies have failed to replicate or showed conflicting results implies that either the current studies are underpowered and much of the variation is still unexplained or missing [10], or the study approaches that have been applied so far do not address the overall picture. For this reason, more research is still needed to unravel these unexplained genetic variations. The question is then whether several association signals across the genome, whose effect sizes are too small to attain the stringent genome wide threshold could explain the remaining source of the genetic variation [15] or can more power be gained by including more variants into the study using different approaches?

One way of utilizing the variants that have small effect sizes, that in most cases do not meet the stringent GWAS threshold, is by treating the whole genome as a risk predictor [16], a concept that is well implemented in polygenic risk scoring methods [16, 17, 18] and linear mixed models [15]. Yang et al. for example used linear mixed model software (GCTA [15]) and illustrated that much of heritability of a trait can be accounted for by evaluating the effect of all SNPs [15]. Similarly, Ripke et al. [19] applied PRS in schizophrenia and showed that a large number of markers, actually up-to half of the total markers, were jointly associated with schizophrenia as opposed to the initial study which only identified a few of the variants. Indeed both polygenic score and linear models have been used to infer a wide range of complex traits to the extent that it is now generally accepted that all complex traits are polygenic [16]. Malaria being a complex trait [11, 13] trait is therefore no exception.

Another way of accounting for the remaining unexplained genetic variation in a disease like malaria is through imputation approaches. By including the variants that are not typed into the study, imputation increases the likelihood of finding the true causal signal of a disease [20]. Moreover, imputation can combine several studies through meta-analysis thus increasing the power [20]. Although imputation has been applied in genetics studies of African populations, up-to date no study has evaluated and hence recommended the most appropriate tool that can accurately handle African populations, that is known to have high genetic diversity and sub-structure.

Here in this project, we will focus on how the power of GWAS can be improved: first via imputation approaches and second using Polygenic Risk Score methods, in estimating the risk of malaria in African populations. We therefore evaluate the potential of genotypes imputation tools and present a recommendation of the most accurate tool that will guide future studies, particularly for the African populations. We moreover review and evaluate the polygenic risk score methods and finally, we apply the best PRS method in genetic prediction of malaria.

1.2 Key Concepts in Genetic Variation

1.2.1 Polymorphism

Polymorphisms are the variations in the human DNA sequences. These variations are what makes individuals unique and also determines an individual risk or susceptibility to a given trait or disease. The most common form of variations are the Single Nucleotide Polymorphisms (SNPs), that are popularly referred to as modern unit of genetic variation [21]. These variations

occur at specific loci of the genome, and can occur during DNA replication in which a single base pair may be left out, replaced or a single base pair might be added. In most cases, SNPs have two alleles and the frequency of a SNP in a population is given in terms of the frequency of the less common allele. This frequency is referred as the minor allele frequency [21].

1.2.2 Hardy Weinberg Equilibrium

Hardy Weinberg Equilibrium (HWE) describes how alleles/genotypes frequencies in a given population are inherited from generation to generation. It assumes that in the absence of migration, natural selection, mutation, and assortative mating the genotypes frequency tend to remain constant in a diploid population [22, 23]. Therefore, the genotypes frequencies under HWE can be estimated from the allele frequencies [22, 23]. It is very important to check if the observed genotypes of the target sample deviates from HWE [24]. Deviations from HWE raises a signal and could imply either genotyping error or problems with the population structure or it could also indicate the presence of selection [24]. Samples that deviates from HWE must be corrected or excluded from further analysis to avoid spurious associations [21].

To illustrate the concept of HWE, let N be the number of unrelated subjects of diploid populations. Suppose 0 is the reference allele and 1 is the alternate allele. Denote the frequency of 0 as f_0 and that of 1 as f_1. By diploid, there are $2N$ alleles so that the frequency of $f_0 = \frac{2N_{00}+N_{01}}{2N} = p$ and, $f_1 = 1 - f_0 = 1 - p = q$. For the first generation,

$$p + q = 1 \tag{1.1}$$

And for the second and subsequent generations,

$$p^2 + 2pq + q^2 = 1. \tag{1.2}$$

Thus by HWE, the expected genotype frequencies of $00, 01$ and 11 are given by the following.

$$f_{00} = p^2$$
$$f_{01} = 2pq$$
$$f_{11} = q^2$$

Observed genotype frequencies and the expected genotype frequencies from HWE can be compared using Pearson Chi-Square test for goodness of fit, Fisher exact test among others [24]. However, Fisher exact test is more preferred since it does not rely on the chi-square approximation, which can be poor when the genotypes counts are low [25]

Results from HWE can well be interpreted using qq plot of the log of p-values. Deviations from the $y = x$ line signifies deviations from HWE.

1.2.3 Linkage of Disequilibrium

Linkage of Disequilibrium, abbreviated as LD, is the correlation of alleles at one SNP to alleles at another SNP within a population [22]. Typically, it describes the degree with which allele at one SNP is inherited or correlated with an allele at another SNP within a population [26]. Thus, LD is used by the geneticist to describe changes in genetic variation within a population over time [27].

In a population of fixed size, LD decays over time due to random mating. This in particular leads to repeated random recombination of events which breaks away segments of contiguous stretches of chromosomes from founder generation sequentially until all the alleles in the contiguous chromosome segments are independent or are in linkage equilibrium [21, 27]. The rate of LD decay depends on population size, number of founder chromosomes in the population and the number of generations the population has existed [26].

Different populations have different patterns of LD [21]. African descent population in particular have smaller regions of contiguous chromosomal stretches hence smaller regions of LD [22]. This is due to the fact that African population is the most ancestral thus has undergone more recombination of events than any other population [21].

LD can be quantified by many measures although $D^{'}$ and r^2, which have been showed to be highly correlated, are commonly used [21]. However, LD is mostly reported in terms of r^2 as explained by [21]. To illustrate how $D^{'}$ and r^2 measures LD, let AB and ab be two SNPs at two loci. Denote the frequency of AB as f_{AB} and f_{ab}, f_{Ab}, f_{aB} to be the frequency of ab, Ab and aB respectively. Then,

$$D^{'} = \begin{cases} \frac{f_{AB}f_{ab} - f_{Ab}f_{aB}}{\min(f_A f_b, f_a f_B)} & \text{if } f_{AB}f_{ab} - f_{Ab}f_{aB} > 0 \\ \frac{f_{AB}f_{ab} - f_{Ab}f_{aB}}{\min(f_A f_B, f_a f_b)} & \text{if } f_{AB}f_{ab} - f_{Ab}f_{aB} < 0 \end{cases} \tag{1.3}$$

$D^{'} = 0$ implies complete linkage of equilibrium between two SNPs/markers [21]. This signifies frequent combinations between two SNPs and statistical independence under HWE. $D^{'} = 1$ means complete LD thus indicates no recombination between the two markers.

$$r^2 = \frac{(f_{AB}f_{ab} - f_{Ab}f_{aB})^2}{f_A f_B f_a f_b} \tag{1.4}$$

The interpretation of r^2 is similar to the interpretation of $D^{'}$ [21].

When two SNPs are in LD, then only one SNP needs to be genotyped to capture the variation at nearby sites in the genome, hence preventing genotyping of SNPs that are redundant [21].

The SNPs that are selected to capture the variations in the nearby regions of LD are called **tag SNPs**. It is important to note that tag SNPs selected for a given population may not work well for another population since LD is population specific [21]. The tag SNPs are particularly useful in the design of custom genotype array for a specific population.

1.2.4 Genome Wide Association Studies

Overview

Genome Wide Association Studies (GWAS) are studies that examine SNPs or genetic variations of thousands of individuals to identify the genetic variants which are associated with a particular trait or risk of a disease [27]. GWAS can be conducted at either family-based level or at a population-based level or both. However, the most common study design in GWAS are those that involve case/control study design. The main objective of GWAS is to identify the genetic variants that cause common diseases, and to understand how these genetic variants contribute to a particular trait or risk of a disease, with the assumption of common disease common variant hypothesis [21]. The understanding can be applied in predicting those who are at risk in a population [21, 27].

GWAS is actually the most efficient way to discover the genetic variants associated with a disease when we have very little molecular knowledge of the disease and has been applied in developing better treatment and prevention strategies, using results from the GWAS findings [21, 24]. Association analysis in GWAS is done by comparing the frequencies of genotypes or alleles between cases and controls [21, 24]. There are many strategies for performing association analysis in GWAS. However, Single SNP scan, which sequentially examines each SNP with the null hypothesis that there is no association, is more popular hence most commonly used [24]. GWAS association tests, that are used especially if the phenotypes are quantitative includes linear regression, analysis of variance, and for binary traits, logistic regression is preferred [24].

Despite successfully identifying thousands of SNPs that are associated with complex traits or diseases, most of results from GWAS have only accounted for a small proportion of the phenotypic variation. Below, we discuss current challenges in GWAS particularly in African populations.

Current Challenges in GWAS

Advancement in technology has made it possible to measure millions of SNPs at a cheaper cost, thus making studies utilizing GWAS common. This has indeed revolutionized the field of genetics with thousands of alleles that influence disease risk being discovered and replicated

across studies. However, despite the success, just a few of the loci identified via GWAS are associated with large or moderate disease risk, with some of the well known disease risk being missed by these associations. As a matter of fact, the genetic variance explained by GWAS findings is disappointingly low which has raised questions as far as their relevance or utility in risk assessment is concerned, and application in personal genetics, or their suitability in genetics testing [28]. Below, we give current challenges in GWAS studies, especially in a resource constrained continent like Africa.

African studies have lower levels linkage of disequilibrium, hence require denser genotyping arrays than the ones that are currently available. For example, based on HapMap from a single ethnic groups, it is estimated that 1.5 million SNPs from African populations have the same statistical power as 0.6 millions SNPs of European study [10]. This implies that larger number of SNPs are needed to achieve adequate statistical power across different ethnic groups in Africa [10].

Despite the fact that large sample sizes are possible to obtain, the greatest limitation is obtaining well characterized samples [29]. African samples, in particular, are characterized by high levels of ethnic diversity. This again, may lead to false positive association due to population structure. For example, Jallow et al. examined the possibility of false positive findings using 402,814 SNPs from 958 cases and 1,382 controls from The Gambian [10]. Interestingly, they found a high false positive association test statistics of $\lambda = 1.23$ in the raw data, which reduced to $\lambda = 1.07$ after accounting for the self reported ethnicity [10]. Additionally, variation of haplotype structure (which is common in African populations) among groups may reduce the power of association especially when data across multiple study sites are combined [10]. For example, band et el. [30] showed how the patterns of association of HbS differs across different populations in Africa.

Another challenge is handling large datasets [31]. GWAS datasets are typically very huge and are computationally intensive to analyze [31]. Thus, successful GWAS analysis require powerful computers with very many processors and high storage capacity. Although efficient strategies using distributed architectures, like clusters, cloud based systems, and super computers, have been proposed, these resources are very expensive and are not within reach to some of the institutions in the developing countries [32], like in Africa. This therefore makes it impossible for them to conduct such studies.

Disappointing results from GWAS also presents a very big challenge to researchers. So far, variants that have been identified to be significantly associated with a trait have weak relative risk ranging from 1.0 to 1.2, which is too weak to be applied in genetics testing [33]. Moreover,

missing heritability is still very large for most of the traits. In height for example, which is known to be highly heritable trait, 40 variants were found to be significantly associated with height for tens of thousands of individuals. However, these loci accounted for only 5% of the phenotypic variation , which is far much below the theoretical estimate of heritability, which is approximately 80% [34]. This therefore, raises questions about the utility of GWAS as disease classifier and their usefulness in risk assessment in personal genetics [35]

Finally, another challenge in GWAS is that results from GWAS cannot be directly translated into an immediate clinical use. As Du et al. [31] put it, that the cost of conducting quality GWAS is very high and the end result will actually attract high impact journal however translation to immediate clinical use is still lacking. This therefore raises the question whether GWAS researches are just for publication or health oriented or the chip companies are intended beneficiaries for GWAS [31].

1.3 Overview of Polygenic Risk Scores (PRS)

Polygenic Risk Score (PRS) analysis is one of the most popular and current methods for summarizing the effect of a collection of variants [36]. Using a single score for each individual in the study, PRS sums the risk alleles corresponding to a given trait or disease for an individual, weighted by the estimates of the effect sizes from GWAS published result or meta-analysis [17]. Typically, PRS summarizes the whole genome wide data of each individual in the study into a single variable which can then be tested for the association with a trait of interest, and can used as a measure of an individual liability or tendencies to a trait or phenotype of interest [36].

Polygenic methods uses two datasets; the first dataset is referred to as the training or the discovery dataset [16]. This dataset is used to estimate the effect sizes for each SNP, which are then selected according to some P-Value threshold [16]. The second dataset is called the target dataset, from which the PRS are calculated for each subject by calculating the weighted sum of the risk alleles using only the SNPS selected from the training sample [16].

To illustrate how PRS works, let y to be a vector of phenotypes of n unrelated or independent individuals in the target sample and G to be an $n \times q$ matrix of genotypes, where G_i is a vector of genotypes dosages for individual i in the target sample.

PRS of an individual is defined by the following [18].

$$\text{PRS}_i = \sum_{j=1}^{M} \hat{\beta}_j G_{ij} \tag{1.5}$$

Where $\hat{\beta}_j$ in 1.5 is the effect size estimate for SNP j from the training sample and G_{ij} is the genotype dosage of the target sample, $j \in \{1, \ldots, M\}$ and $i \in \{1, \ldots N\}$ and N is the number of individuals in the training sample.

There are several PRS methods but the basic idea is the same. **Figure 1.1** illustrates how PRS scores can be obtained and applied in association analysis.

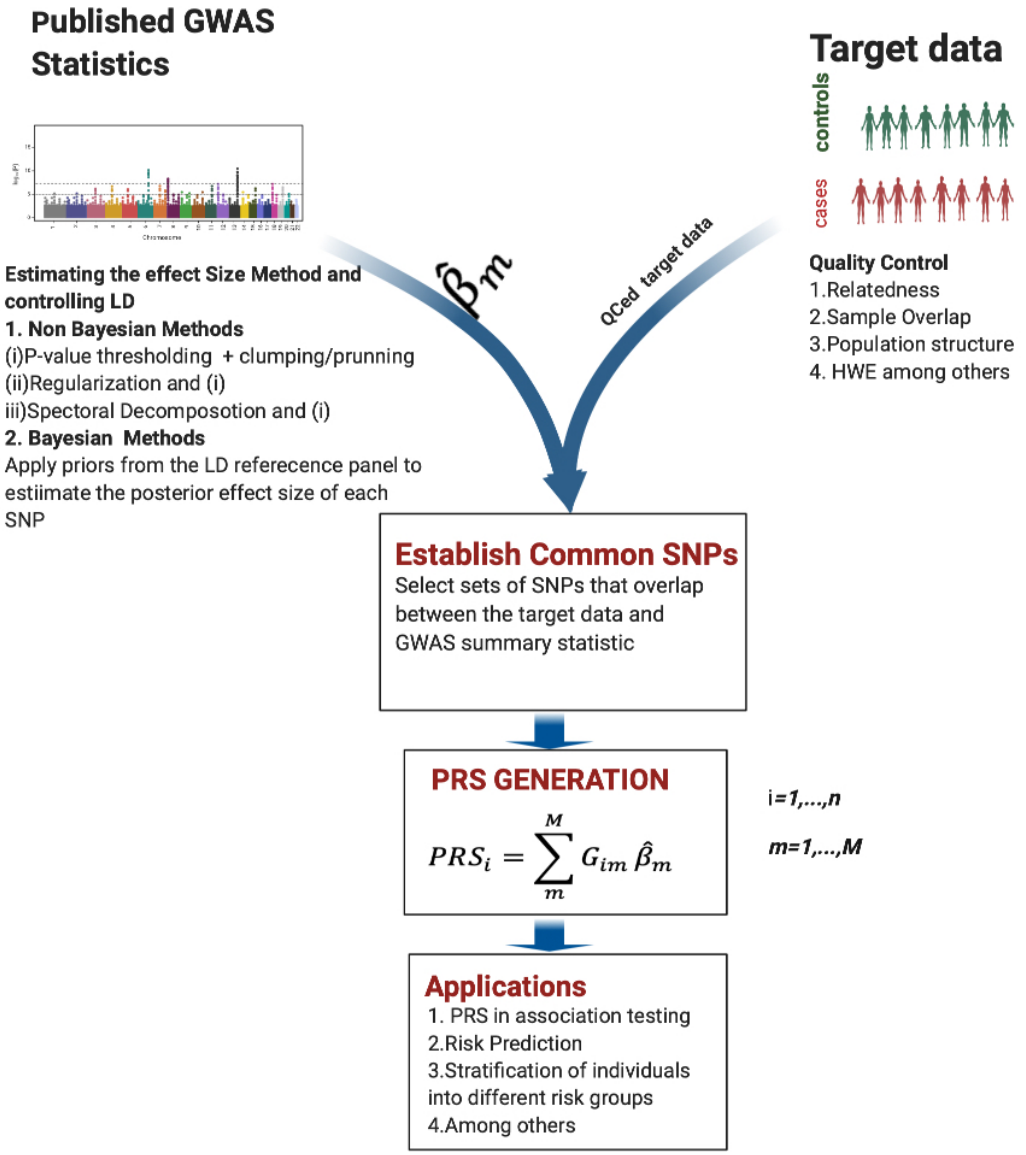


Figure 1.1: *An illustration of how PRS scores are obtained and applied in the association analysis.*

Here in this thesis, We will give a brief review of some of the popular PRS methods, discussing various aspects of their implementation.

PRS has been applied in studies like Schizophrenia, cancer, continuous traits like height among others [19, 37, 38]. Despite the fact that evidence from malaria GWAS suggests that malaria could be polygenic [10]. To date, no literature has been published that illustrates the potential of PRS application in malaria, and in infectious diseases in general. By the fact that genetic

factors identified to date only explain a small proportion of the malaria risk, and that most of the variants identified had effect sizes that did not meet the stringent GWAS threshold implies that more insights can be gained by evaluating the overall effect of all the variants in each individual. Indeed, comparable sources have suggested the existence of polygenic inheritance in malaria in some of the early publications [11, 13]. In the preceding chapters, we will discuss malaria and the genetic susceptibility, and attempt to answer the question whether malaria is polygenic or not. We will finally attempt to predict the risk of malaria using the PRS.

1.4 Malaria and The Human Genetic Susceptibility

1.4.1 Overview

Early findings suggests that in Africa alone, *P. falciparum* records approximately 500 million cases [39], killing in the order of a million African children every year [10, 40]. *P. falciparum* is considered the deadliest due to the fact that it can attack all red blood cells of all ages, while the other species like *P. vivax* and *P. Ovale* for instance can only invade the young red blood cells, and *P. malaria* on the other hand can only attack aging red blood cells [41].

If the interventions for Malaria controls are correctly implemented, then malaria is not only treatable but preventable as well [42]. However, to prevent malaria, we need to understand different life stages of the *plasmodium* parasite and their mechanism of adaptation in different hosts. This understanding in particular facilitated the development of different vector control programs that were applied in the 20th century and were very instrumental in controlling malaria [5]. In the next section, we give an overview of plasmodium parasite life cycle, that was very instrumental in controlling malaria in 20^{th} century. Nevertheless, to fully eliminate malaria, we need to understand the genetic basis of severe malaria. It is estimated that human genetic factors accounts for approximately 25% of the risk of severe malaria in areas where *p. falciparum* malaria infection is common. Additionally, hemoglobin S (*HbS*) is known as the strongest determinant of malaria risk and contributes approximately 2% of the total malaria variation [10]. Understanding these factors can provide valuable insights of molecular mechanisms of pathogenesis and protective immunity that can be applied in the development effective treatment options and vaccines [10, 30].

1.4.2 Plasmodium Parasite Life Cycle

Plasmodium has a complex life cycle with two distinct phases of life that involves two hosts, vertebrate and the invertebrate [2, 43]. The first phase is the asexual cycle, which takes place inside the human host and the second phase is the sexual phase which takes place inside the mosquito (invertebrate) [43].

Asexual phase begins when an infected mosquito, with the infection acquired from the previous blood meal, inject sporozoites into the human host during the blood meal [2, 39]. The injected sporozoites takes approximately 30-60 minutes to reach the liver via the blood stream [39]. Inside the human host, the parasite undergoes two stages; the exo-erythrocytic stage and the erythrocytic stage of the cycle [2]. Exo-erythrocytic stage takes place in the liver during which the sporozoites multiplies asexually [2, 5]. During this time, the sporozoites undergo many cell divisions producing 10,000 to 30,000 descendants for each sporozoite [39]. These descendants sporozoites matures in the hepatocytes to become schizonts [2, 5, 39]. Typically, the cell division for the sporozoites and the maturity of the descendants of the sporozoites into schizonts lasts for about 6 to 15 days and after which the schizonts bursts or raptures releasing thousands of merozoites into the blood stream for circulation [39]. This marks the end of exoerythrocytic stage, and the beginning of the erythrocytic stage [39]. The merozoites then invade the red blood cells, mature and multiply asexually producing descendants merozoites [2]. The descendants merozoites then develop into immature trophozoites and finally into mature trophozoites, which then develop into schizonts [2]. Erythrocytic stage lasts for about 44 -72 hours and terminates when the infected red blood cell burst releasing the merozoites [2]. The released merozoites can either begin the erythrocytic cycle again by infecting the new red blood cells or can develop into gametocyte, then infect the mosquito during the blood meal [2].

Sexual cycle, also known as sporogonic cycle, similarly begins during the blood meal when a feeding mosquito takes blood from an infected person [2, 39]. During the blood meal, the male gametocyte (known as microgametocyte) or the female gametocyte (macrogametocyte) may be injested, or both [2]. Inside the mosquito, microgametocye and macrogametocyte matures to become microgametes and macrogametes [2]. Microgamete then fertilizes the macrogamete in the midgut of the mosquito forming what is known as a zygote, which then matures to become ookinete [2, 39]. The ookinete develops into oocyst while invading the midgut of the mosquito at the same time. Finally, the oocyst matures and then raptures to release sporozoites, which can make their way to the human host through blood meal hence completing the life cycle [2, 39]. **Figure 1.2** illustrates the schematic life circle of the plasmodium parasite that causes malaria in

human.

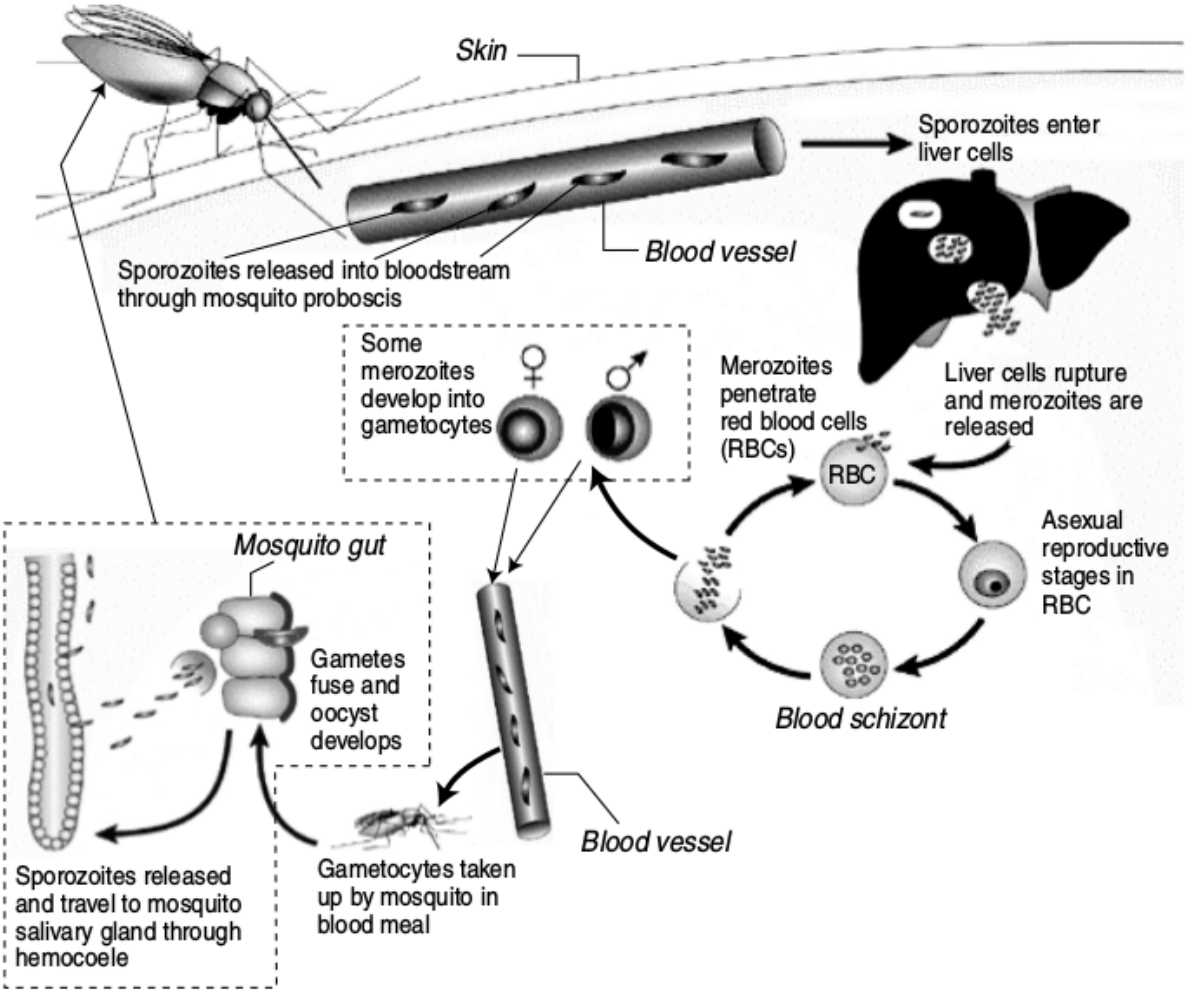


Figure 1.2: *schematic life circle of the plasmodium parasite that causes malaria in human [39].*

1.4.3 Human Genetic Susceptibility and Malaria

Studies suggest that malaria prevalence and incidence increased between 5,000 to 10,000 years ago [13]. It is during this period that there was massive expansion of agriculture, forest clearing and animal domestication, that are said to have enhanced the success of anopheles mosquito [1]. Human haplotypes studies have also shown that alleles that offer protection against malaria have undergone selection the same time frame [13]. Consequently, genetic mutations that offer protection against malaria varies in different populations due to difference in environmental factors, population structure, transmission intensity, effects of specific variant on a given host genotypes, and many more [3].

Malaria, the strongest force of the evolutionary selection on the human genome, is known to be the source of hematological disorders like hemoglobinopathies (*HbS*, *HbC* and *HbE*), quantitative

hemoglobinopathies (like the thalassamias), membrane mutations (like spherocytosis, elliptocytosis, ovalocytosis) and enzymopathies (like glucose-6-phosphate dehydrogenase deficiency (*GD6P*)) [3, 11]. These genetics disorders have risen to high frequencies in malaria endemic regions and thus suggests that malaria selective pressure is very strong [3]. A typical example is thelassaemia, which causes mild anaemia and is common in Mediterraneans, reached and maintained the high frequency due to selective advantage against *P. falciparum* and not as a result high mutation rate [44]. Moreover, these disorders vary with different populations having independent evolutionary response to malaria as can be witnessed from both global and local malaria levels [3, 13].

Table 1.1: *Common red blood cell variants that affect resistance to malaria [13].*

Gene	Protein	Function	Reported Genetic Associations with Malaria
FY	Duffy antigen	Chemokine receptor	FY*O allele completely protects against P. vivax infection.
G6PD	Glucose-6-phosphatase dehydogenase	Enzyme that protects against oxidative stress	G6PD deficiency protects against severe malaria.
GYPA	Glycophorin A	Sialoglycoprotein	GYPA-deficient erythrocytes are resistant to invasion by P. falciparum.
GYPB	Glycophorin B	Sialoglycoprotein	GYPB-deficient erythrocytes are resistant to invasion by P. falciparum.
GYPC	Glycophorin C	Sialoglycoprotein	GYPC-deficient erythrocytes are resistant to invasion by P. falciparum.
HBA	α-Globin	Component of hemoglobin	α^{+}Thalassemia protects against severe malaria but appears to enhance mild malaria episodes in some environments.
HBB	β-Globin	Component of hemoglobin	HbS and HbC alleles protect against severe malaria. HbE allele reduces parasite invasion.
HP	Haptoglobin	Hemoglobin-binding protein present in plasma (not erythrocyte)	Haptoglobin 1-1 genotype is associated with susceptibility to severe malaria in Sudan and Ghana.
SCL4A1	CD233, erythrocyte band 3 protein	Chloride/bicarbonate exchanger	Deletion causes ovalocytosis but protects against cerebral malaria.

Infectious diseases in general are considered as complex traits, that involves the interplay of host genetic factors, environmental factors and the pathogen genome [45]. As such, approaches

that works in the identification of variants that underlies complex trait can also be applied in infectious diseases. Prior to GWAS, linkage studies and candidate gene studies were applied in the identification of variants that are involved in the predisposition of complex disease [45], malaria inclusive. However,these studies had limited power, and in most cases the associations that were identified never replicated [45].

1.4.4 Malaria-Specific GWAS

The first GWAS malaria in Africa was done with a study data from The Gambia in 2009, using 958 cases and 1,382 controls with 402,814 SNPs after quality control. A total of 139 SNPs were identified to be nominally associated with malaria at a threshold of P-value $< 10^{-4}$, of which six had a P-value of $P < 10^{-6}$ [10]. However, only one SNP rs11036238 $OR = 0.61$, P-value $= 3.9 \times 10^{-7}$, located on the *HBB* region, replicated in an independent sample. The replication samples had 1,087 cases and 2,376 controls. Surprisingly, even the authentic loci like the *ABO* gene, *G6PD, HBA1-HBA2* also failed to reach the level of association even after genotyping of these regions. However, the authors argued that some of these variants have reached fixation in the Gambian population, while others are rare in the Gambia [10] hence could not attain the significance level. Additionally, low tagging efficiency by the Affymetrix array was also implicated as the reason why some of the known loci failed to attain the association threshold. Interestingly, imputation resulted to an increase in power as was observed from the *HbS* variant that had a P-value of 3.9×10^{-7} in the initial GWAS before imputation and a P-value of 4.5×10^{-14} after imputation. This study, despite highlighting the potential of imputation, they did not apply a guided choice for the imputation tool, a gap that we present in the current study. The study moreover recommended that different methodological approaches for GWAS in Africa. Genotypes imputation is the most cost effective approach of improving the power of GWAS, even as researchers are working on different approaches to entangle Africa population GWAS data, which present a lot of challenges during the GWAS analysis.

Another notable GWAS study on malaria was done by Timmann et al. [46] that identified two novel resistance loci for severe malaria using Ghanian study sample. Although this study moderately identified 41 loci that were associated with severe malaria, only four loci had a threshold of $P < 5 \times 10^{-8}$, of which two were novel and other two had been identified by other studies [13]. The two novel loci were *ATP2B4* on chromosome 1q32.1 under recessive model and *MARVELD3* on chromosome 16q22.2 using additive model. Other loci are summarized in the table below. Recently, Ravenhall et al. [47] conducted a GWAS in the North Easter part of

Tanzania using 449 cases and 465 typed at 15.2 million SNPs. This study again, despite small sample size, discovered novel loci that had not been identified before. Among the novel regions identified includes *IL-23R* and *IL-12RBR2, KLHL3* among othes [47].

In other GWAS studies, Band et al. [30] did a meta-analysis of African population from Kenya, The Gambia and Malawi. The study had a total of 5,425 cases and 6,891 controls with 1.3 million common SNPs. Although the potential of using the same methodology for meta-analysis in Africa is not clear, Band et al. [30] tried different methodologies for associations. Notably, two regions (*HBB* on chromosome 11 and *ABO* on chromosome 9), that are among the known authentic loci for malaria susceptibility and protection had numerious SNPs with association (P-value) value less than $< 5 \times 10^{-8}$. Although other regions showed significant associations with P-value $< 1 \times 10^{-6}$, their associations were not conclusive as they have never been replicated [30]. This study in particular highlighted the potential of imputation based meta-analysis in African population. That despite the challenges in Africans genetic make up, convincing results can be obtained when larger samples sizes and correct methodologies are applied [30]. In **Table** 1.2 below, we list the variants that have been identified and replicated through GWAS to be associated with malaria.

Table 1.2: *List of genes that have been reported to be associated with malaria through GWAS and have been replicated in some studies.*

Near Gene	CHR	SNP	Model	MAF	OR	P-Value	POPULATION	Reference
ATP2B4	1q32.1	rs10900585 (T/G)	Recessive	0.38(0.43)	0.60 [0.47-0.76]	2.0×10^{-5}	Ghana GWAS	[46]
				0.38(0.44)	0.69 [0.57-0.84]	1.9×10^{-4}	Ghanian Replication	
					0.65 [0.56-0.75]	6.1×10^{-9}	Ghanian GWAS Combined	
				0.33(0.37)	0.58 [0.40-0.85]	$5.3 * 10^{-3}$	Gambia Population	
					0.61	1.9×10^{-10}	Meta-analysis (Ghana+Gambia)	
		rs2365860 (A/C)	Recessive	0.33(0.39)	0.58 [0.45-0.74]	2.5×10^{-5}	Ghana GWAS	
				0.33(0.39)	0.68 [0.55-0.84]	3.9×10^{-4}	Ghanian Replication	
					0.63 [0.55-0.74]	1.5×10^{-8}	Ghanian GWAS Combined	
				0.27(0.30)	0.54 [0.36-0.81]	2.7×10^{-3}	Gambia Population	
					0.61	1.9×10^{-10}	Meta-analysis (Ghana+Gambia)	
		rs10900589 (T/A)	Recessive	0.33(0.39)	0.57 [0.44-0.74]	2.4×10^{-5}	Ghana GWAS	
				0.33(0.39)	0.69 [0.55-0.85]	5.1×10^{-4}	Ghanian Replication	
					0.63 [0.54-0.74]	2.1×10^{-8}	Ghanian GWAS Combined	
				0.27(0.29)	0.54 [0.36-0.81]	2.8×10^{-3}	Gambia Population	
					0.62	2.8×10^{-10}	Meta-analysis (Ghana+Gambia)	
		rs2365858 (C/G)	Recessive	0.33(0.39)	0.56 [0.43-0.73]	1.1×10^{-5}	Ghana GWAS	
				0.33(0.38)	0.68 [0.55-0.85]	5.9×10^{-4}	Ghanian Replication	
					0.63 [0.54-0.74]	5.1×10^{-8}	Ghanian GWAS Combined	
				0.27(0.29)	0.56 [0.37-0.83]	$4.4 * 10^{-3}$	Gambia Population	
					0.62	9.5×10^{-10}	Meta-analysis (Ghana+Gambia)	
ABO	9q34.2	rs8176719 (delG/G)	Dominant	0.36(0.29)	1.62 [1.35âĂŞ1.96]	1.2×10^{-7}	Ghana GWAS	[46]

Continued on next page

Table 1.2 – continued from previous page

Near Gene	CHR	SNP	Model	MAF	OR	P-Value	POPULATION	Reference
				0.37(0.28)	1.70 [1.48-1.96]	2.9×10^{-13}	Ghanian Replication	
					1.67 [1.50-1.86]	1.1×10^{-20}	Ghanian GWAS Combined	
				NA(0.16)	1.26 [1.11-1.44]	5×10^{-4}	Gambia Population	
					1.48	4.3×10^{-21}	Meta-analysis (Ghana+Gambia)	
		rs8176703 (C/A)	Dominant	0.093(0.057)	1.84 [1.42-2.39]	$= 5.1 \times 10^{-6}$	Ghana GWAS	
				0.091(0.056)	1.72 [1.41-2.10]	5.1×10^{-6}	Ghanian Replication	
					1.73 [1.48âĂŞ2.02]	4.0×10^{-12}	Ghanian GWAS Combined	
				NA(0.16)	1.26 [1.11-1.44]	5×10^{-4}	Gambia Population	
					1.48	4.3×10^{-21}	Meta-analysis (Ghana+Gambia)	
HBB	11p15.5	rs334 (A/T)	Heterozygous	0.053(0.072)	0.06 [0.038âĂŞ0.12]	2.5×10^{-21}	Ghana GWAS	[46]
				0.010(0.059)	0.15 [0.10-0.23]	1.6×10^{-18}	Ghanian Replication	
					0.011 [0.079-0.015]	1.6×10^{-18}	Ghanian GWAS Combined	
				NA	NA	1.3×10^{-28}	Gambia Population	
		rs372091 (C/T)	Heterozygous	0.029(0.068)	0.38 [0.028-0.52]	1.4×10^{-9}	Ghana GWAS	[46]
				0.028(0.058)	0.45 [0.34-0.59]	$3.6 * 10^{-8}$	Ghanian Replication	
					0.44 [0.36-0.54]	1.1×10^{-14}	Ghanian GWAS Combined	
				0.012(0.018)	0.63 [0.38-1.07]	0.085	Gambia Population	
					0.46	5.6×10^{-14}	Meta-analysis (Ghana+Gambia)	
MARVELD3	16q22.2	rs2334880 (T/C)	Additive	0.47(0.40)	1.31 [1.16-1.49]	$2.3 * 10^{-5}$	Ghana GWAS	[46]
				0.45(0.41)	1.20 [1.08-1.32]	$4.3 * 10^{-4}$	Ghanian Replication	
					1.24 [1.15-1.34]	3.9×10^{-8}	Ghanian GWAS Combined	
				0.40(0.38)	0.96 [0.81-1.13]	0.60	Gambia Population	

Continued on next page

Table 1.2 – continued from previous page

Near Gene	CHR	SNP	Model	MAF	OR	P-Value	POPULATION	Reference
					1.19	1.9×10^{-6}	Meta-analysis (Ghana+Gambia)	
SPATA3	2q37.1	rs6750230 (T/C)	Additive	0.37(0.43)	0.65	$1.6 * 10^{-5}$	Gambia GWAS	[10]
				0.38(0.40)	0.83	1.0×10^{-1}	Gambian Replication	
					0.74 [0.64-0.85]	3.2×10^{-5}	Ghanian GWAS Combined	
DDC	7p12.2	rs1451375 (A/C)	Dominant	0.18(0.21)	0.69	$3.0 * 10^{-4}$	Gambia GWAS	[10]
				0.19(0.22))	0.81	1.4×10^{-2}	Gambian Replication	
					0.75 [0.66-0.85]	6.1×10^{-6}	Ghanian GWAS Combined	
HBB	11p15.4	rs11036238 (C/G)	Trend	0.09(0.04)	0.65	$3.9 * 10^{-7}$	Gambia GWAS	[10]
				0.10(0.14)	0.81	6.8×10^{-6}	Gambian Replication	
					0.63 [0.66-0.72]	3.7×10^{-11}	Ghanian GWAS Combined	
SCOI	17p13.1	rs6503319 (T/C)	Trend	0.51(0.45)	1.28	$6.6 * 10^{-5}$	Gambia GWAS	[10]
				0.49(0.45)	1.14	2.1×10^{-2}	Gambian Replication	
					1.21 [1.12-1.31]	7.2×10^{-7}	Ghanian GWAS Combined	

1.5 Genotypes Imputation

1.5.1 Overview

Genotypes imputation is the process of estimating genotypes that are not directly measured in a sample of individuals under study. Through this, the evidence of the association of genetic markers that are not genotyped directly can be examined [48]. Imputation applies the knowledge of linkage of disequilibrium (LD) structure in the reference panel and the GWAS datasets to combine the data from the reference panel with the data from the study dataset thus predicting the untyped or missing genotypes in the target dataset [48, 49]. Genotypes imputation has two practical applications: In Genome Wide Association Studies (GWAS) and in fine mapping studies. In context of GWAS, genotype imputation is carried out across the whole genomes to predict the SNPs that are not typed into the study [20]. In fine mapping on the other hand, genotypes imputation is carried out in a more focused region of the genome to find the true causal variant. In either way, the baseline is to estimate variants that have not been typed into the study [48].

1.5.2 Application and Challenges in Genotypes Imputation

Genotypes imputation is not a new concept and its potential applications can be traced as early as the era of haplotypes phase inference methods [50]. During phasing, sporadic missing genotypes are imputed hence genotypes imputation is as old as phasing. However, genotypes imputation gained popularity in the era of GWAS , that saw emergence of larger and denser samples than before, thus allowing accurate imputation of missing genotypes [51]. Genotypes imputation presents a lot of benefits in genetics. Despite several application and success, challenges are inherent in genotypes imputation. Below, we give some of the popular applications of genotypes imputation, and challenges in genotypes imputation.

Application of Genotypes Imputation

Application in GWAS: Genotypes Imputation improves power in Genome Wide Association Studies by including into the study SNPs that were not typed. Usually, a high proportion of individuals have missing data at one or more markers, and thus removing these individuals in downstream analysis can substantially reduce the sample size of the data [51]. Again, many markers may fail to pass quality control hence are supposed to be removed in downstream analysis. Genotypes imputation presents a powerful technique for recovering SNPs that have failed to pass quality control, and also can be be used to estimates the SNPs that are not typed

into the study. This therefore maintains or improves the power of association studies [51]. A popular example of the power of imputation on GWAS is highlighted by the study by Zeggini et al. [52], where two novel associations (*PPARG* and *CDC123-CAMK1D*) were identified from the imputed SNPs. These associations, moreover, were confirmed via replication and genotyping [52].

Application in Fine Mapping: Fine mapping studies are studies that assess the genetic variations at a known GWAS risk locations with the aim of identifying the variants that have direct effect on the trait [53]. Genotypes imputation can be carried out in a specific region of the Genome to provide high resolution view of the associated region hence increasing the likelihood of identifying the true causal SNP. One of the best illustration of potential of imputation on fine mapping is from a study by Jallow et al. [10]. Using *HbS*, which is known as the best polymorphism for evaluating GWAS, Jallow et al. [10] sequenced 111kb region at the center of the GWAS signal on chromosome 11p15 using 62 individuals as a reference panel. They then applied IMPUTE program to impute approximately 2,500 samples from the GWAS and applied association test at each of the imputed SNPs. From their result, three imputed SNPs had stronger association than any of the SNPs genotyped on the initial GWAS. Moreover, the known polymorphism rs334 showed the strongest association with p value of $4.5 \times 10_{-14}$, which was far much stronger than the signal of association from the genotyped SNPs alone, which had a P-value of 3.9×10^{-7} [10]. Although this association signal was weaker than the association signal of genotyping the region directly (P-value$= 1.3 \times 10^{-28}$), one of the possible explanation for this is due LD patterns of African population, which strongly affects imputation.

Application in Meta-Analysis: As most GWAS susceptibility loci have small to moderate effects, sufficient statistical power can only be obtained when large sample sizes or densely typed markers are for associations [30]. Such statistical power is beyond the capacity of a single GWAS study, and perhaps, can be obtained from combining multiple GWAS studies [51]. Meta-analysis is the statistical procedure of integrating or combining data from several independent studies or studies from different genotyping platforms [54]. Through imputation, a homogeneous datasets from different study cohorts can be created [55], and a statistical test is conducted at each SNP, and across the collection using different meta-analysis models [30] . First application of imputation based meta-analysis was done in 2008, which identified novel associations that had been missed by previous GWAS based on single studies [56]. Since then, meta-analysis has been on the rise with many risk loci identified [54].

Current Challenges in Genotypes Imputation

Imputation accuracy decreases with minor alleles frequency. Consequently, imputation of rare and low frequency variants remains a very big challenge. Africa for example, which has the greatest diversity worldwide and the lowest LD patterns, has the lowest imputation accuracy worldwide [20]. To improve imputation accuracy for the low frequency variants, studies have suggested the use of diverse reference panels that contains more samples from diverse populations. This has seen reference panels like 1000 genomes being extended to contain more samples from Africa. Although slight improvement has been reported with this kind approach, the imputation accuracy for diverse population still remain at modest. Recently, there is so much hype about the development of a reference panel for African population, in which are near completion, and are expected to improve the power of association studies and imputation in particular. However, the benefits of these developments are yet to be seen. In one of the studies, Liu et al. [57] used whole genome data of 90 individuals from European populations to assess and compare the performance of three popular imputation tools(BEAGLE, IMPUTE2 [58] , and minimac [59]) using two reference panels: a European population specific reference panel and 1000 Genomes reference panel. Interestingly, the 1000G reference panel outperformed the populations European specific reference panel [57]. The question then is whether the African specific reference panel will improve the imputation performance in African population, and in diverse populations like the African Americans, or will it give similar experiences that have been reported by using specific reference panel in the European populations?

By the fact that imputation accuracy increases with increase in the size of the reference panels, there have been efforts of increasing the sizes of reference panels. As a result, reference panels are steadily increasing in sizes with the current largest reference panel having tens of thousands of samples. Moreover, this is set to increase to include hundreds of thousands of samples [60]. However, even though this would definitely improve the accuracy of imputed variants, the computation cost is also expected to increase. Increase in computation cost is a big challenge that current imputation tools must be able to deal with. As a consequence, imputation tools have devised several techniques for reducing the computational complexities like the use of pre-phasing which succeeded in reducing quadratic complexity, use of identity by descent, haplotype clustering and linear interpolation, use of specialized formats (like m3vcf, bref2, bref3) of reference panels to reduce file size and memory requirement and many more [60]. Nonetheless, total computation cost is still higher for most of the tools. Recently, Browning et al. [60] implemented a new imputation algorithm that has been shown to reduce the imputation

computation costs to less than a penny, and has been shown to be 100 times faster than the current imputation tools.

In malaria, as in most complex traits, imputation has been applied to improve the power of the association studies. To date, a number of studies have implemented various imputation tools and a lot of success have been reported [10, 30]. By the fact that malaria is the strongest force of evolutionary selection that has shaped the human genome, applying the best imputation tool in malaria studies can elucidate the potential of imputation and can be generalized in other traits.

1.6 Motivation of the Thesis

Malaria continues to be one of the leading causes of death globally; affecting more than 91 countries worldwide [61]. The global tally of malaria in 2015 was 212 million cases and 429,000 deaths [61]. It is estimated that over 90% of malaria related mortalities are reported in sub-Saharan Africa with the hardest hit group being pregnant women and children below the age of five [62]. Moreover on global scale, under-five years child death in sub-Saharan Africa is over 10% compared to developed countries where it is less than 1% [63]. This is primarily due to malaria and other infectious diseases [63]. Recently, a report of World Malaria Day 2017 approximated that over 70% of infants death is as a result malaria, killing a child every two minutes [61].

One of the interesting finding is that *P. falciparum* mainly kills children before reproductive age because it has the capacity to select emerging polymorphisms that offers protection against the severe form of the disease and death [64]. Thus a comprehensive understanding of the genetic basis of resistance/susceptibility to severe malaria can shed more lights into molecular mechanisms of drug resistance and protection, and can be applied in the development of effective treatments options and vaccines [30, 65]. An example of how genetic discoveries may be translated into the development of vaccines and treatments is about *Plasmodium vivax* [13]. The realization that Africans lack the Duffy blood group, caused by chemokine receptor gene popularly known as *DARC*, led to the molecular characterization of the crucial parasite protein that binds to the Duffy erythrocyte receptor [3]. Consequently, this led to the identification of a vaccine against *P. vivax* [66]. By the fact that malaria is known to be the strongest selection force that has shaped the human genomes, the genetic discovery on *P. falciparum* can not only revolutionize malaria research but also genetics research as a whole [13].

To date, the known genetic factors explains only a small proportion of host genetic resistance to malaria, an indication that most genetic factors are yet to be identified. GWA studies have been proven to be powerful tools for investigating the genetic architecture of human diseases including

Malaria [67]. However, the methods have been hugely suffering from different challenges including 1) the lack of translation of associated loci into suitable biological hypotheses [68], 2) the well-known problem of missing heritability [68], 3) the lack of understanding of how multiple modestly associated loci within genes interact to influence a phenotype [69], accuracy of genotypes imputation tools, and many more.

One way of improving the power of Genome Wide Association Studies is by increasing the sample size and by using a dense set of SNPs in the study [20]. Conrad et al. estimates that typically, 1.5 million SNPs in African populations have equivalent power of 0.6 million SNPs of HapMap phase 1 dataset of European descents [70]. This implies that African studies require more markers for GWAS to achieve adequate power than the European samples. Imputation has been shown to be the most cost effective way of including variants that are not typed into the study [20]. In malaria, imputation is not a new concept and has been applied in many studies with the first in Africa being by Jallow et al. [10]. This studies, Jallow et al. applied imputation with IMPUTE [71] program on the *HbS* locus using 62 individuals from Gambian population [10]. Their results showed that the imputation has the potential of boosting the GWAS signals. In particular, their GWAS signal improved from a signal of association value of 4×10^{-7} to 4×10^{-14} [10]. However, despite the successful results that have been obtained through imputation in African populations, with different genotypes imputation tools used in different studies, little is known as to which imputation tool is most appropriate when dealing with data from African ancestry. This therefore suggests that there is an urgent need of the assessment of the imputation tools in African populations to guide future researches. Given a rich source of information on both malaria GWAS and reference panels, it is now an intriguing time to evaluate the potential of imputation tools in imputing genetic data from African populations. **Table** 1.3 highlights some of the Malaria GWAS that have implemented imputation.

Table 1.3: *Example of malaria GWAS in Africa that have used imputation*

Authors	Journal	Publication Date	Tool	Title	Reference
Jallow et al.	Nature Genetics	2007	IMPUTE1	Genome-wide and fine-resolution association analysis of malaria in West Africa	[10]
Band et al.	PloS Genetics	2013	IMPUTE2 [58]	Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations	[30]
Ravenhall et al.	PloS Genetics	2018	IMPUTE2 [58]	Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania	[47]

Further, results from GWAS findings shows that much of the genetic basis for most of the complex traits is indeed a combination of hundreds or thousands of genetic variants with small effects [17]. This realization points to the fact that Single SNP analyses alone do not address the overall genomic architecture for most of the complex diseases [72]. Therefore, other methods that treats the entire genome as a unit or explores the overall effects of SNPs that fails to reach genome wide significance should be applied when analyzing complex traits. Moreover, it has been illustrated that much of the heritability of complex traits can be explained by analyzing the common variants that fail to reach genome wide significance threshold. Yang et al. for example used linear mixed model software (GCTA [73]) and illustrated that much of heritability of a trait can be explained by evaluating the effect of all SNPs [15]. Similarly, in malaria the estimated heritability is approximately 25% [11], with the replicated associations accounting for only a fraction of this estimate [74]. This implies that, just as in height, more insights can be gained by analyzing the variants that do not reach the GWAS significance threshold. We therefore propose to exploit this by using polygenic risk score analysis, which sums the risk alleles corresponding to a given trait or disease for an individual, weighted by the estimates of the effect sizes from GWAS published result or meta-analysis [17].

1.7 Objectives of the Thesis

We hypothesize that, malaria resistance is a complex trait that is under polygenic control as suggested by Mackinnon et al. [11]. We further suggest that in as much as genotypes imputation tools have been applied in studies of African ancestries, and in malaria studies in particular, better results can obtained if appropriate tool is used. This is only possible when the tools are properly evaluated using African genetic data and a recommendation given thereof. We therefore propose to asses the power of the genotypes imputation tools using whole genome data of African ancestry versus European ancestry individuals and admixed population. We apply two separate reference panels, 1000 genomes reference panel and an extended 1000 genomes reference panel that has more African samples, and finally give a recommendation of the best tool that has the highest imputation accuracy when dealing with genetic data from African populations. Also, by the fact that missing heritability is still large in most of diseases, and GWAS has not been able to account for this large “dark side of the genome", it is time to explore other approaches. So far, PRS has presented promising results in explaining the heritability that has never been accounted for by standard GWAS approaches. Nonetheless, this powerful technique has never been applied in any infectious disease. We thus explore the possible application of PRS in malaria and give

recommendations for future studies.

The following are the objectives of our study.

I) We propose to review the current tools used in imputation.

II) Compare the performance of genotypes imputation tools using the simulated datasets that mimic African and European genetics structure.

III) Access all malaria genome-wide associations studies (GWAS) data in African populations from MalariaGEN.

IV) Compare GWAS from genotypes imputation using the best tool identified in (ii) verses summary statistics imputed by ImpG.

V) Perform a meta-analysis of results in IV.

VI) Perform pathway enrichment from the results in V).

VII) Review PRS methods and compare the performance of different PRS methods using simulated datasets that mimic Africa populations.

VII) Apply the best PRS method from VII in genome-wide association study(GWAS) summary statistics of malaria to predict Malaria risk/resistance in order to understand Malaria-specific genetic architecture.

1.8 Outline of the Thesis

In **Chapter 2**, we discuss different mathematical approaches of genotypes imputation and some of the concepts that have revolutionized genotypes imputation. We also review some of the popular imputation tools and discuss different imputation quality metrics, and their comparison. In **Chapter 3**, we evaluate some of the popular imputation tools discussed in **Chapter 2**. We begin by first reviewing literature of previous studies that have compared different imputation approaches. We then describe our datasets and the simulation framework we applied in obtaining the datasets and finally present imputation results and the discussion. In **Chapter 4**, we apply the best tool identified from the evaluation in **Chapter 3** in imputing Malaria GWAS data from three African populations obtained from MalariaGEN. Also impute the summary statistics from the same populations with ImpG. We then compare the association result from imputed summary statistics data by ImpG and the summary statistics of GWAS from raw genotypes data imputed by the best imputation tool identified in **Chapter 3**. We conclude the chapter by performing a meta-analysis and pathway enrichment of the meta-analysis results. In **Chapter 5**, we review different PRS methods, classification and challenges in calculation of PRS. We review previous studies that have

compared different PRS methods, and finally compare different PRS methods using a simulated dataset that mimics Africa population. We then recommend the best PRS method based on the result. We finally apply the best performing PRS method based on the evaluation in predicting genetic liability/risk of severe malaria. In **Chapter 6**, we present conclusion and future work,

Chapter 2

Mathematical Review of Genotypes Imputation Approaches

2.1 Overview

A number of statistical methods have been proposed for genotypes imputation of GWAS data. However, the ultimate classification of these algorithms is difficult due to the fact that the implemented algorithms are often modified and combined in the newer versions of the imputation programs or implemented in completely new programs. Nonetheless, genotypes imputation tools can be broadly classified into four categories: the first category are tools that are based on standard statistical method like Multinomial models, linear regression with variable selection, Regression tree, Singular Value Decomposition, k-Nearest Neighbor (k-NN) method and many others. These tools, despite the fact that they can estimate the missing genotypes, they do not model the key features of the genetic data like recombination and linkage of disequilibrium (LD) [75]. As result, they have lower imputation accuracy as compared to the tools that incorporate the key characteristics of the genetic data. Schaid et al. [76] for example compared the performance of seven of these standard statistical methods (linear regression with variable selection, Regression tree, Singular Value Decomposition, k-Nearest Neighbor (k-NN)) with fastPHASE [77], which is based on population genetics principle. Their result showed that fastPHASE [77] had the lowest error rate, hence highest accuracy and outperformed all the other seven tools. The second category of tools that employs SNPs tagging approach in performing imputation. These tools includes PLINK [78], SNPMStat [79], TUNA [80], and UNPHASED [81]. Although these tools are very computationally efficient in carrying out imputation, they are less accurate since they do not utilize information across the entire chromosome in performing imputation. The third class of

genotypes imputation tools are tools based on haplotype matching. Here, a popular example is PWBT, which was developed recently by Durbin et al. [82]. Again, these tools, despite being computationally efficient, do not integrate across all possible mosaic haplotypes configurations hence giving less accurate estimates [83]. The fourth, and the last category of tools are those that are based on population genetics principle. Tools under this category includes BEAGLE, IMPUTE, MaCH, minimac [59] and many more. These tools apply hidden Markov Models (HMM) in imputing the missing genotypes and have been applied as the gold standard for performing genotypes imputation [84].

Here in this thesis, we will focus on the computational tools that are developed explicitly for genotypes imputation in GWAS, and employs Hidden Markov Model using Li and Stephen framework.

2.1.1 Li and Stephens HMM framework

Li and Stephens model offers a spectacular approach for analyzing LD patterns and modeling recombination hotspots [26]. This model, popularly known as Li and Stephens framework, has several advantages including ability to analyze LD patterns across multiple loci, can handle long chromosome stretches using the fact that haplotypes share contiguous stretches with each other despite individuals haplotypes being unique [26]. Moreover, it models the historical events, like mutations and recombination rates, hence offers an efficient application in the imputation of genotypes, and forms the underpinning of genotypes imputations [26].

In Li and Stephens framework, a subset of haplotypes are selected and used as reference set. This has since then been replaced with a reference panel like the 1000 Genome Project and the HapMap. Each marker in the reference panel represents a hidden state of the Hidden Markov Model (HMM). Observed genotypes are assumed to be imperfect mosaic of the reference haplotypes [26]. The state space of the HMM by Li and Stephens can be visualized as a matrix, where the columns are the markers in the reference panel and the rows are the individual haplotypes in the reference panel [54]. Each allele in the reference panel represents a state and the study haplotype, or genotypes are assumed to trace unobserved path through the matrix from the first reference marker to the last reference marker [54]. The study haplotype can thus be viewed as a combination of segments in the mosaic template, in which switching from one segment to another segment follows a Markov Chain [51]. This Markov chain incorporates the recombination event and therefore, a new haplotype can copy from different haplotypes from two consecutive loci. This simplifies the switching probabilities from one marker to the next, and the transition

probability does not depend on the haplotype being copied from. Alleles of the new haplotype can be different or close to the haplotypes they are copied from and this reflects the mutation or the genotyping error [26, 51]. Recombination and mutation parameters are determined as part of model fitting process [51].

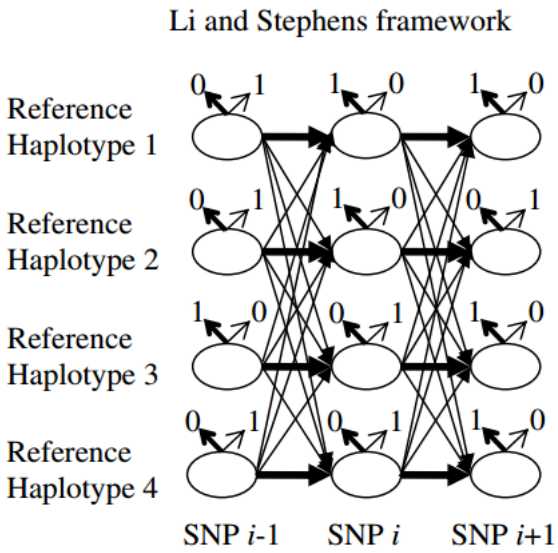


Figure 2.1: *Li and Stephens Hidden Markov Model framework for performing geneotypes imputation. The ovals represents the hidden state, defined by the reference markers. Bold arrows represent paths with the highest transition probabilities. Thin arrows allows for recombination between markers [51].*

To illustrate how Li and Stephens model is applied in imputation, Let H to be the set of reference haplotypes typed at M markers and h to denote the set of the target haplotypes. Denote the hidden states of the HMM by S and assume that each individual haplotype is independent of other individuals' haplotypes. The four elements of HMM includes: state space, initial state probabilities, transition probabilities and emission probabilities. Most imputation tools have similar initial, transition and emission probabilities. However, the only difference is how each tool defines the state space of the HMM. The general form of a hidden Markov model can be given as follows:

$$P(s|h,H,\theta,\rho) = P(S_1)\prod_{m=2}^{M} P(S_m|S_{m-1},\rho)\prod_{m=1}^{M} P(h_m|S_M,H,\theta) \tag{2.1}$$

Where $P(S_1)$ in Equation 2.1 denotes the initial probabilities, $P(S_m|S_{m-1},\rho)$ are the transition probabilities and $P(h_m|S_m,H,\theta)$ are the emission probabilities, ρ and θ represent recombination and error/mutation parameters. The initial probability for most tools is equaprobable in all configuration and is given by the following:

$$P(S_1) = \frac{1}{R}, \tag{2.2}$$

Where R in **Equation 2.2** is the number of haplotypes in the reference panel.

Transition probabilities on the other hand are defined by the following:

$$P(S_m = i|S_{m-1} = j) = \begin{cases} (1-\theta_{m-1}) + \frac{\theta_{m-1}}{R} & \text{if } i = j \\ \frac{\theta_{m-1}}{R}, & \text{Otherwise} \end{cases} \tag{2.3}$$

In Li and Stephens framework, the transition probabilities are defined based on the number of haplotypes in the reference panel, effective population size and the genetic distance between markers. Thus θ_{m-1} in *Equation 2.3* is defined as follows according to Li and Stephens framework:

$$\theta_{m-1} = 1 - e^{\frac{-4N_e d_{m-1}}{R}}, \tag{2.4}$$

where N_e in **Equation 2.4** is the effective population size, d_{m-1} is the genetic distance between the markers m and $m-1$, and R is the number of reference haplotypes in the reference panel. Therefore, any given state $(m-1, h)$ can be on the reference panel h with probability $\theta_{m-1} = 1 - e^{\frac{-4N_e d_{m-1}}{R}}$ or it can be on the same reference haplotype h due to recombination with probability $\frac{\theta_{m-1}}{R}$. Also, a state $(m-1, h)$ on reference haplotype h can transition to $(m, \hat{h})$ -where $h \neq \hat{h}$- with probability $\frac{\theta_{m-1}}{R}$

Emission probabilities on the other hand are given by the following:

$$P(h_m|S_m) = \begin{cases} 1-\epsilon_m, & \text{if } h_m = h_{im} \\ \epsilon_m, & \text{Otherwise} \end{cases} \tag{2.5}$$

Emission probabilities generally take into account the genotyping errors and mutations, modelled by the mutation parameter ϵ_m. **Equation 2.5** therefore implies that a state (m, h) emits allele carried by haplotype h at marker m with probability $1-\epsilon_m$ otherwise, it emits a different allele.

IMPUTE1 [85], IMPUTE2 [71] and IMPUTE4 [86] provides estimates of the fine recombination map, which can be downloaded from the program webpage. However, the user must specify effective population parameter which by default is 20,000 [71]. Emission parameter on the other hand is fixed internally by the program [71]. For BEAGLE tools, emission probabilities are defined in terms of error rate. By default, the tool uses $\epsilon_m = 0.0001$. MaCH, minimac [59], minimac 2 [87], minimac 3 [83] and minimac 4 [54] on the other hand estimate the parameters and are updated in each iteration. While this approach is flexible in that it can adopt to each data being analyzed, the parameters may not be estimated well which may consequently reduce the imputation accuracy [20].

2.2 Concepts Revolutionizing Genotypes Imputation

2.2.1 Pre-phasing

Early genotypes imputation methods like MaCH [88], IMPUTE [85], BEAGLE [89] and fasPHASE [77] were based on unphased geneotypes. These tools were very computationally intensive due to quadratic complexity. Typically, phasing of the target genotypes and subsequently imputation were done simultaneously on the same run of the algorithm. The realization that alleles can be imputed directly into phased target haplotypes motivated the separation of phasing step from the imputation step. This then has motivated the development of tools that can specifically require the target genotypes to be phased prior to the imputation, and has greatly reduced the imputation time and has motivated the implementation of additional computational optimizations.

2.2.2 Specialized Reference Panel Format and Clustering

Reading and storing large reference panel is a great challenge in imputation, and in genetics studies in general. The Variant Call Format (VCF) for example requires 4 byte for storage for every genotypes and can require several terabytes for millions of samples in the reference panel [90]. Although compressing the reference panel is an option that can greatly reduce the size of the reference panel, it does not address the issue of storage in memory during analysis. Moreover, time for decompressing large reference panel can sometimes exceed even the time required to perform imputation. Among the early reference formats that have been developed in imputation to reduce the size of reference panel are the Minimac3 VCF (M3VCF) [83], bref [60, 91]. M3VCF are created to adopt with the imputation algorithm implemented in minimac 3 [83]. Here, only unique haplotypes in each segment in the reference panel are stores as opposed to storing all the haplotypes for each segment. Bref (Binary reference format) on the other hand stores genotypes data in terms of blocks where each block contains a marker and genotype information for a set of subsequent markers. A marker can either be index coded (stores indices of haplotypes with non major alleles and specifically suited for low frequency markers) or sequence coded (stores the sequence of distinct alleles present in a given data block).

In short regions, identical haplotypes can be clustered together in which each cluster has a unique marker. HMM forward backward algorithm can then be applied to a unique cluster rather than all the markers of the reference panel. Some of the popular tools implementing clustering includes minimac 3 [83] [83], minimac4 [54], beagle 3 [89], beagle 4 [91] and beagle 5.0 [60]. The clustering in minimac is exclusively based on the reference panel and in precomputed prior to

imputation [83]. Clustering in BEAGLE on the other hand is based local clustering, which is applied on the genotyped markers of the target sample, and is implemented during imputation run [92].

2.2.3 Imputation via Linear Interpolation

Linear interpolation applies a two stage approach when calculating the HMM state probabilities. In the first stage, HMM forward backward algorithm is restricted on the genotyped markers [54, 91]. In the second stage, the HMM state probabilities for the missing markers are estimated via linear interpolation based on the genetic distance. Linear interpolation assumes that the genetic map position of markers are real numbers and applies the fact that over short distances between typed markers, HMM state probabilities changes smoothly from one marker to next bounding marker, and that HMM state probability of the missing marker can be estimated by a straight line [91]. More specifically, let x to be an untyped marker between two genotyped markers m and $m+1$. Suppose marker m carries allele a and marker $m+1$ carried allele b, then by linear interpolation, the probability that the imputed marker, x, carries allele a is given by the following:

$$P_a = \sum_{h \in H} (\lambda_{m,x} P(s_m = (m,h)) + (1 - \lambda_{m,x}) P(s_{m+1} = (m+1,h))) \tag{2.6}$$

Where $\lambda_{m,x} = \frac{g(m)-g(x)}{g(m+1)-g(m)}$, and $g(.)$ denotes the genetic marker position [91].

Minimac4 [54] forms the real line interval by using the midpoints of the non overlapping marker intervals and applies the posterior probabilities of the genotyped marker within a given flunking region to impute the missing markers. Linear interpolation is currently implemented in BEAGELE4.1 [91], BEAGLE 5.0 [60] and minimac4 [54].

2.2.4 Imputation Via the Web Service

Imputation servers allows users to upload phased or unphased GWAS data and receive a downloadable output of phased and imputed data. Imputation servers offers a number of advantages to researchers including: enabling researcher to spend much time focusing on their research rather than learning how each tool of the imputation works, offers a platform for researcher to share and consolidate their research and allows access to reference panel that are very restrictive and cannot be made publicly available by storing the reference panel data behind the firewall. Currently, there are two imputation servers: the University of Michigan Imputation

server which uses minimac3 [83] for imputation and Wellcome Sanger Institute imputation server that uses PBWT for imputation.

2.3 Review of Population Genetics Imputation Tools

Imputation tools based on population genetics principle can be classified into two classes: Tools before the pre-phasing era and tools after the pre-phasing era. We discuss each category below.

2.3.1 Imputation Tools Before the Pre-phasing Era

Impute v1 [85] uses Li and Stephens HMM framework in which each individuals HMM is conditioned on the set of reference haplotypes and a set of parameters. Let i be an individual's genotype in the target GWAS. Then the HMM used by IMPUTE v1 [85] is given by the following:

$$P(G_i|H,\theta,\rho) = \sum_S P(G_i|S,\theta)P(S|H,\rho) \tag{2.7}$$

Where $S = \{S_1,\ldots,S_M\}$ with each $S_m = (S_{m1}, S_{m2})$ and $S_{mk} = \{1,\ldots,R\}$. $P(G_i|S,\theta)$ is the emission probability and depends on the mutation parameter θ. On the other hand, $P(S|H,\rho)$ is the transition probability and is dependent on the recombination parameter ρ. MaCH uses a similar HMM framework to that of IMPUTE [85] except that in MaCH [88], the HMM of an individual is conditioned on the current haplotype estimate of other individuals. Transmission and emission probabilities are similar to that of IMPUTE v1 [85].

FastPHASE [77] is the earliest tool that implemented the population genetics approach based on Li and Stephens model, and can carry out both phasing and imputation. Here, the HMM model is based on the assumption that over tightly linked regions, haplotypes tends to cluster into groups of similar patterns. It employs a similar transition probabilities as IMPUTE v1 [85]. However, emission probabilities are derived based on the allele frequencies of each cluster rather the mutation rate as in IMPUTE v1 [85]. Moreover, fastPHASE [77] uses Expectation Maximization (EM) algorithm approach in estimating the parameters rather than Markov Chain Mornte Carlo (MCMC) approach employed in IMPUTE v1 [85].

BEAGLE [89] uses a similar clustering algorithm as that of fastPHASE [77]. However, BEAGLE does not use Li and Stephens model but instead uses a BEAGLE [93] algorithm that is similar to Li and Stephens model. HMM model under BEAGLE [93] algorithm is a directed cyclic graph with variable number of hidden states at each marker. Hidden states are represented as clusters

at each node of the graph [89]. BEAGLE algorithm has fewer states at each marker hence speeds up computation [93]. To achieve fewer number of states, BEAGLE implement a two step running procedure: the bifurcation step and the pruning step [93]. In the bifurcation step, a bifurcation tree is constructed that describes how the set of haplotypes are constructed across the entire set of haplotypes. A weighting is then implemented on each edge of the tree by the number of haplotypes that goes through the edge. In the pruning step, the tree is pruned whereby at each level of the tree pairs of nodes are compared in terms of their downstream haplotype frequencies. Nodes that are similar are merged hence resulting to a more parsimonious characterization of the dataset [93].

IMPUTE2 [58] is an improved version of IMPUTEv1 [85]. It uses both phased and unphased study data as input hence referred as a flexible approach. The algorithm first partitions the SNPs into two sets: a set T that is typed both in the study sample and the reference panel, and a set U that is missing in the study sample but typed in the reference panel. Imputation is then performed in two steps. In the first step, genotypes in the set T are phased using equation 2.7 of IMPUTE v1 [85]. It then applies equation 2.1 in the second step to impute alleles in the set U. This way, much computational effort is applied on the first step and the second step is typically very first since its haploid [58]. The algorithm alternates between phasing and imputation using a MCMC framework [58]. The difference between IMPUTE2 [71] and IMPUTE1 [85] is that IMPUTE1 [85] is analytical and it integrates over all possible possible haplotype configuration for each study individual and can only be implemented when the study individuals are treated independently. This therefore sacrifices LD information in the target data and the computational burden increases with the increase in the size of the reference panel. IMPUTE2 [71] implements haplotype sampling strategy in a MCMC framework which scales down with the size of the reference panel.

2.3.2 Imputation Tools After the Pre-phasing Era

Prephasing was first introduced by Howie et al. [59] in 2012 after an observation that imputation tools spend more computational effort to account for the phases of unknown GWAS genotypes. Both analytical and sampling strategies employed by the imputation programs are very computationally intensive, and the computational cost for each approach increases with increase in the size of the reference panel.

Minimac [59] is the first stand alone imputation tool to work with only phased genotypes data. It was developed as part of MaCH algorithm, and relies on the phased output from

MaCH, although it could also accept phased data from other tools as well. Minimac is based on the assumption that the haplotypes underlying GWAS datasets have been estimated correctly. Nonetheless, the accuracy of imputation depends on how well the haplotypes are estimated. It then applies a standard HMM to estimate the marginal probabilities of the missing alleles in each GWAS haplotype using the reference haplotypes as the template. The advantages that come with prephasing is that the haplotypes are estimated once and can be re-used many times during the program runs as opposed to standard imputation tools where the most likely haplotype for each individual is estimated each time imputation is done with an updated reference panel or new reference panel.

Minimac2 [87] is the second tool to be released after pre-phasing era. It's an improvement of minimac [59] to reduce imputation computation time and memory requirement through software engineering approaches. It implements a manual loop vectorization as opposed to the automatic loop vectorization which is implemented in modern compilers by default. This is motivated from the observation that the default automatic vectorization may fail in some instances, especially when the optimization is complex. Minimac2 [87] also implements parallel processing through OpenMP and improves on data locality access.

Minimac3 [83] applies the state space reduction, which is similar to that implemented in BEAGLE3 [89]. Here, the genome is first subdivided into consecutive haplotypes blocks, such that each block contains similar haplotypes. While BEAGLE 3 [89] uses a fixed number of markers per segment/block, minimac 3 [83] applies dynamic programming algorithm to select an optimal number of markers in each block. HMM iteration is then performed only on the unique haplotypes in each block. To achieve the same state space as that of IMPUTE2 [58] and minimac [59], it applies a reversible function that maps the reduced space to the original space.

BEAGLE4 [91] is an improvement of BEAGLE 3 [89] and is the first BEAGLE algorithm to implement Li and Stephens HMM framework. It applies a state space reduction in which markers are aggragated using a fixed 0.005 cM into a single aggregate marker. The allele frequency of the aggragate marker is given in terms of the observed allele frequencies of the constituent markers and the HMM state space is defined in terms of the state space of the aggregate marker. This distinguishes BEAGLE4 [91] from other methods, since the state space are in terms of aggregate markers rather than in terms of all the markers in the reference panel [91]. The forward backward algorithm of the HMM is then performed on the aggregate markers and the missing genotypes are imputed via linear interpolation. BEAGLE 4 [91] is the first imputation algorithm to implement linear interpolation strategy. This is motivated by the fact that over short distances, changes in state probability of a given allele within the short distance changes smoothly and can be

approximated by a straight line.

Minimac4 [90] builds from minimac 3 [83] and applies further a state space reduction approach to reduce the computational complexity experienced from minimac 3 [83] and to increase speed. Developers refers to this as an aggressive state space reduction [90]. Here, state space reduction is applied only on the genotyped markers as opposed to all markers in a genomic segment as was the case in minimac 3 [83], hence referred to as an aggressive state space reduction. It then applies a reversible function that maps reduced space to an aggressive reduced space. To impute the missing markers, a non overlapping flanking region of each genotyped marker is defined based on the midpoint of the genotyped markers. For each flanking region, ungenotyped markers are imputed using the posterior probabilities of the genotyped markers. To reduce computation time, the algorithm only transforms haplotype templates with high posterior probabilities in each flanking region other than using all the haplotypes in the flunking region when imputing the missing markers.

BEAGLE 5.0 [60] presents a slight improvement of BEAGLE 4 [91]. It however presents three computation improvement on BEAGLE 4 [91]. First, instead of using full reference panel like in BEAGLE 4 [91] to impute the missing alleles, BEAGLE 5.0 [60] conditions the HMM on a composite reference panel whose composite reference haplotypes comprise of regions that share Identity by State (IBS) with both the reference panel and the target sample. The composite reference haplotype is generated in a preprocessing step in which the full reference panel is reduced to a small composite reference panel, which is mosaic of the reference haplotypes. While previous versions of BEAGLE were based on short fixed window, based on the hamming distance which had to be specified (e.g ≤ 0.5), this approach had issues for example the window size could too small hence imputation accuracy will be reduced as the segments are truncated by the window boundaries. Composite reference panel allows for the imputation of both long and short Identity by Descent (IBD)segments, and moreover the composite haplotype segments are mosaic of the haplotypes in the reference panel, with all reference panels haplotypes that are associated with the IBS segments are included in the composite reference panel. Second, the posterior probabilities are obtained via linear interpolation just like in BEAGLE4.1 [89], with the only difference being that the posterior probabilities are calculated when the VCF file is being created not like in BEAGLE 4[89] where the probabilities of each individual marker is calculated immediately after every state iteration. Third, BEAGLE 5.0 [60] process the reference panel in Bref3 (Binary Reference 3) format which breaks the chromosome into consecutive non overlapping intervals, and alleles are stored in form of distinct sequences with a pointer for each haplotype that carries a given marker for major alleles and for the non major allele, the index of

the haplotype that carries the allele is stored. This format reduces memory requirement since the number of distinct allele sequence is less than the number of haplotypes [60]. **Table 2.1** shows the summary of population genotyoes imputation tools.

Table 2.1: *Summary of popular genotypes imputation tools, their underlying methods (Hidden Markov Models (HMM) parameters), publication year and citations.*

Tool	HMM state space	HMM Parameter Functions	Author	Year	Title of publication	Number of citations as at May 2019
fastPHASE	Composed of all genotype configurations from a fixed number of localized haplotype clusters	Depends on recombination and mutation parameters. The parameters estimated using EM algorithm	Scheet et al. [77]	2006	A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase	1682
IMPUTE v1	Uses all genotype configurations from all the reference haplotypes	Uses a fine scale recombination map that is fixed and provided internally by the program	Marchini et al [85]	2007	A new multipoint method for genome-wide association studies by imputation of genotypes	2247

Continued on next page

Table 2.1 – continued from previous page

Tool	HMM state space	HMM Parameter Functions	Author	Year	Title of publication	Number of citations as at May 2019
IMPUTE v2	Uses all possible reference haplotypes	Depends on fine scale recombination map that is fixed and provided internally by the program	Howie et al [71]	2009	A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies	2801
BEAGLE 3 [89]	All genotype configurations from a variable number of localized haplotye clusters	Empirical model with no explicit parameter functions	Browning et al. [89]	2009	A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals	1299

Continued on next page

Table 2.1 – continued from previous page

Tool	HMM state space	HMM Parameter Functions	Author	Year	Title of publication	Number of citations as at May 2019
MaCH	Uses all genotype configurations from all the reference haplotypes	Depends on recombination rate, mutation rate and genotyping error. Parameters are fit using MCMC or EM algorithm	Li et al. [48]	2010	MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes	1717
minimac v1	Use all possible reference haplotypes	Similar to MaCH	Howie et al.[59]	2012	Fast and accurate genotype imputation in genome-wide association studies through pre-phasing	1156
minimac 2	Uses all possible reference haplotypes	Similar to MaCH	Fuchsberger et al. [87]	2014	minimac2: faster genotype imputation	216
minimac 3 [83]	Uses all unique allele sequences observed in reference data in a small genomic segment	Similar to MaCH. However,the parameter estimates are pre-calculated and fixed	Das et al.[83]	2016	Next-generation genotype imputation service and methods	385

Continued on next page

Table 2.1 – continued from previous page

Tool	HMM state space	HMM Parameter Functions	Author	Year	Title of publication	Number of citations as at May 2019
BEAGLE 4 [91]	Uses all unique allele sequences observed in reference and target data at each of aggregate genotyped marker	Depends on recombination rates and error rates that are fixed and pre-calculated	Brian L. Browning and Sharon R. Browning [91]	2015	Genotype Imputation with Millions of Reference Samples	282
minimac 4 [90]	Collapsed allele sequences from reference data that match at genotyped positions in small genomic segments	Similar to minimac3	Das et al [90]	2017	Minimac4 - Faster Imputation through Aggressive State Space Reduction of Hidden Markov Models	
BEAGLE 5 [60]	Similar to BEAGLE 4 [91]	Similar to BEAGLE4.1	Browning et al. [60]	2018	A One-Penny Imputed Genome from Next-Generation Reference Panels	11

2.4 Measure of Imputation Quality and Accuracy

Imputation quality metrics allows the evaluation and assessment of the imputation outputs. Weakness in genotypes imputation has a potential bias in meta-analysis and Genome Wide Association Studies (GWAS) [55], thus must be properly evaluated. Each software implements their own quality metric for evaluating the quality of imputation outputs. Below, we present a discussion of the quality metrics from some of the popular imputation tools.

2.4.1 Beagle Allelic R^2

BEAGLE allelic R^2 is an estimate of the squared correlation between the imputed allele dosage and the true allele dosage [89]. Although the true allele dosage is not known, it can be estimated from posterior genotypes probabilities. Let X represents the true genotype, Y imputed posterior probabilities, and Z be the genotype with the highest probability. Also, suppose X and Z are inform of genotype dosages ($x \in X, z \in Z$ can take values in $\{0,1,2\}$). Then, R^2 is given as follows:

$$\begin{aligned} R^2 &= \frac{COV(X,Z)^2}{Var(X)Var(Z)} \\ &= \frac{\left(\frac{1}{N}\sum_j \left(z_i E[X|y_j]\right) - \frac{1}{N^2}\sum_j \left(E[X|y_j]\right)\sum_j z_j\right)^2}{\left[\frac{1}{N}\sum_j z_j^2 - \frac{1}{N^2}\left(\sum_j z_j\right)^2\right]\left[\frac{1}{N}\sum_j E[X^2|y_j] - \frac{1}{N^2}\left(\sum_j E[X|y_j]\right)^2\right]} \end{aligned} \tag{2.8}$$

To estimate $E[X|y_j]$ and $E[X^2|y_j]$, BEAGLE makes an assumption that the posterior probabilities are correctly calibrated and $P(X = k|Y = y_j) = y_j(k)$. Hence $E[X|y_j] = y_j(1) + 2y_j(2)$ and $E[X^2|y_j] = y_j(1) + 4y_j(2)$

$$R^2 = \frac{\left[\sum_{j=1}^{N} z_j e_j - \frac{1}{N}\sum_{j=1}^{N} z_j \sum_{j=1}^{N} e_j\right]^2}{\left[\sum_{j=1}^{N} f_j - \frac{1}{N}\left(\sum_{j=1}^{N} e_j\right)^2\right]\left[\sum_{j=1}^{N} z_j - \frac{1}{N}\left(\sum_{j=1}^{N} z_j\right)^2\right]} \tag{2.9}$$

Where $e_j = p_j + 2p_j$, i.e the allele dose of the j^{th} SNP, $f_j = p_j + 4p_j$

2.4.2 Impute Info

Impute info is based on measuring the relative statistical information of the population allele frequency. By assuming that all the genotypes data are observed, then the full likelihood is given as follows:

$$\begin{aligned} L(\theta_j) &= \prod_{i=1}^{N} \theta_j^{G_{ij}} (1-\theta_j)^{(2-G_{ij})} \\ U(\theta) &= \frac{d \log L(\theta)}{d\theta} \\ &= \frac{\left(\sum_{i=1}^{N} G_{ij}\right) - 2N\theta_j}{\theta_j \left(1-\theta_j\right)} \\ I(\theta) &= \frac{dU(\theta)}{d\theta} \\ &= \frac{\left(\sum_{i=1}^{N} G_{ij}\right)}{\theta_j^2} + \frac{2N - \left(\sum_{i=1}^{N} G_{ij}\right)}{\left(1-\theta_j\right)^2} \end{aligned} \tag{2.10}$$

Info measure is thus given by the following [71]:

$$\begin{aligned} I_A &= \frac{E[I(\hat{\theta})] - Var(U(\hat{\theta}))}{E[I(\hat{\theta})]} \\ &= \begin{cases} 1 - \frac{f_{ij} - e_{ij}^2}{2N\hat{\theta}(1-\hat{\theta})} & \text{if } \hat{\theta} \in (0,1) \\ 1, & \text{if } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases} \end{aligned} \tag{2.11}$$

Where N is the total number of individuals in the study, $\hat{\theta}$ is the sample allele frequency, e_{ij} is the expected allele frequency and $f_{ij} = P_{n,1} + 4P_{n,2}$, where P is the imputed genotypes probability [71].

2.4.3 minimac $\hat{r}^2$

Refers to an estimate of the correlation between imputed and the true/unoberved genotype [83]. Let $\hat{p}$ to denote the alternate allele frequency and D_i to denote the imputed allele probability for the i^{th} haplotype and n to denote the number of haplotypes. Then,

$$\hat{r}^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1-\hat{p})} \tag{2.12}$$

Minimac $\hat{r}^2$ is based on the assumption that poorly imputed genotypes counts will tends towards the expectation of population allele frequencies. Particularly, if p is the frequency of the imputed

allele, then the genotype counts is estimated as $2p$. The genotype dosage on the other hand is estimated by summing the posterior probability of the alternate ellele. For example, suppose the posterior probability of allele at each haplotype is 0.94 and 0.97. Then the genotype dose will be just $0.94 + 0.97 = 1.91$.

2.4.4 Comparing the Quality Metric for Different Imputation Tools

To determine whether the imputation quality metric for different imputation are similar, Marchini et al. [20] used a simulated a case-control data with 2000 samples on chromosome 22 based on HapMap 2 haplotypes. Their results shows that the information measure are highly correlated and thus can be compared directly without the need of further conversion [20].

Besides the built in accuracy statistics, other popular measures for assessing the imputation accuracy includes Imputation Quality Score (IQS), corncordance rate and the squared correlation (Rsq). While concordance rate measure the proportion of chance agreement between the imputed and true genotypes, IQS on the other hand adjusts on the chance agreement and is generally based on Cohen's Kappa statatistics of chance agreement [94]. Both the measures compares the imputed and genotyped data via masking approach. In contrast, concordance rate is less preffered since it regards most variants as well imputed even if that is not the case [94]. Ramnarine et al. further observed that for the common variants, Imputation Quality Score (IQS) and the BEAGLE R^2 provided a similar assessment accuracy, and that they can differ considerably for the rare variants. Nevertheless, the choice of accuracy statistics matters most for the rare variants more than the common variants, and that for the common variants IMPUTE2 info, BEAGLE Rsq and squared correlation (Rsq) between the true and imputed genotypes produce similar assessment for the imputation accuracy [94]. This is similar to an evaluation by Marchini et al. [20], who showed that the internal imputation metric from the tools are highly correlated. However, for low frequency variants, neither BEAGLE Rsq, nor squared correlation nor IMPUTE info is recommended for evaluation but instead, IQS should be considered [94].

Chapter 3

Evaluation of Current Genotypes Imputation Tools Through Data Simulation

3.1 Introduction

Genotypes imputation has led to the discovery of thousands of genetic associations which would have been missed from the initial studies [60, 95]. Africa is believed to be the ancestral home for human population and is characterized with the greatest recombination events, leading to high diversity. Indeed studies have demonstrated that Africa has the highest genetic diversity worldwide [96]. Consequently, high diversity reduces the accuracy with which the missing genotypes can be imputed thus making GWA studies in African populations difficult [96]. Further, Africa is known to be leading in the burden of disease, ranging from communicable diseases to non communicable diseases worldwide [97]. To address this burden, proper assessment of imputation tools need to be done in African populations to give guideline for the choice of tool that is more appropriate when dealing with genotypes data from African descent. This will definitely improve the power for association studies in African populations and populations with high genetic diversity.

We assessed the performance of IMPUTE2 [58] and minimac3 [83], minimac4 [90], and BEAGLE 4 [91] in imputing simulated data both from African and European population. IMPUTE4 [86] , BEAGLE 5 [60] and minimac4 [90] are among the most current tools that were developed recently. Although they have been shown to produce superior results in most studies [60], no study has ever evaluated their performances, including the previous versions, using African population study data as the point of reference.

In this chapter, we use data from African and European populations to assess the performance

of the popular and current tools used in imputation to estimate the missing genotypes in simulated dataset both from African and European populations. We then present a comparison of these tools, between and within the population. Finally, we give a recommendation of the tool that is appropriate when dealing with data, especially from African origin.

3.2 Review of Literature

Advancement of high throughput genotyping technologies and a comprehensive catalogue of human genetic variants have been very instrumental to researchers in finding the genetic variants that are associated with complex traits [79]. However, missing data is inevitable in any studies and in genetics in particular. Even with well designed high quality genotyping platforms, some SNPs will be missing at some sites either due to failures by the assays or missing by design [79]. If the missing data are not accounted for or are even excluded from association studies, the genetic studies will suffer from loss of power. Genotypes imputation offers a cost effective way of recovering variants that have lower call rate or are missing completely in the study sample through the use of imputation methods [20]. The imputation methods use known information like patterns of LD between the missing SNPs and their typed flunking SNPs in making inference or estimating the untyped or missing genetic variants [75]. Imputation therefore increases the number of genetic variants that can be tested for association hence improving the power of association studies.

Genotypes imputation can either be carried out across the whole genome or in a particular region of the genome [20]. When carried out across the whole genome, imputation can facilitate the identification of novel susceptibility loci in association studies by making more variants available for testing for the association. Additionally, GWAS data from different genotyping platforms can be merged in a meta-analysis hence increasing the sample sizes, which consequently improves the power of association studies [20, 98]. When imputation is carried within a specific region of the genome, the goal is to fine-map a known susceptibility loci. Thus, the imputation can find all the variants that are in LD with a given loci therefore making it possible to find the causal variant.

Comparing the performances of different imputation tools can either be based on a specific reference panel or based on different sets of reference panels. Usually, when evaluating the performance of different genotypes imputation tools using a specific reference panel, the objective is explicitly compare different imputation tools and to find the best tool that gives the highest imputation accuracy for a given population. However, under this settings the choice of a given

specific reference panel must be known in advance, for examples can be based on some previous findings. On the other hand, when using more than one reference panel, then objective is to compare different reference panels and to find the best reference panel that maximizes the imputation quality for a given population. Moreover, different imputation tools can be used for each of reference panel under evaluation and thus can help in the identification of the best tool, from a given set of tools that gives the highest imputation accuracy.

Several studies have evaluated the performance of different genotypes imputation methods [58, 59, 60, 98, 99]. However, most of these studies have been focusing on the European populations. In one of the studies for example, Nothnagel et al. compared the performances of four imputation tools (BEAGLE, IMPUTE, MACH, and PLINK) using German population as the target sample [100]. This study recommended using either MaCH or BEAGLE for practical use when imputing GWAS data from German descent [100]. Nonetheless, no such studies has been carried out in African populations. A notable exception, however, is a study by Howie et al. [58], that used MalariaGEN datasets from Africa to compare the performance of two imputation tools: IMPUTE2 and BEAGLE. Using two reference panels, Gambian reference panel (GMB), which was obtained by randomly extracting 100 individuals from the MalariaGEN and Gambian+Ghanians +HM3.afr (GMB+GHN+HM3.afr) reference panel that had 100 Ghanian from MalariaGEN, 100 Gambians from MalariaGEN and HM3.afr (phase 3 HapMap sample from Africa) that had 822 haplotypes from ASW, LWK, MKK, and YRI. This study showed that IMPUTE2 [58] is more recommended for studies of African descent [58]. However, despite the fact this conclusion has been applied in most studies, as has been witnessed from the fact that majority of studies from African origin that require imputation aspect, have actually used IMPUTE2 [58] as illustrated in **Table 1.3**, this research only evaluated the performance of two tools. More tools have been developed since then, and even the tools used in this analysis have undergone several modifications [60]. The question then is, does the new tools, or the previous tools which have been modified or improved to their present versions customized enough to handle a diverse populations, like the African population? Additionally, the reference panels used by this study [58] were very small in size and contained very few haplotypes as compared to the present reference panels. Notably, the fact that the reference panel (GMB+GHN+HM3.afr) performed better than the population specific reference panel (GMB) showed that a reference panel, like 1000 genomes project, that contains more samples from across different populations should be prioritized over a population specific reference panel [58].

In other studies, Howie et al. evaluated the effect of pre-phasing on genotype imputation using three imputation tools, MaCH, IMPUTE2 [58] and minimac [59]. By using GWAS dataset of

2,490 individuals from Wellcome Trust Case Control Consortium 2 (WTCCC2) for the 1958 British Birth Cohort, Howie et al. [59] showed that pre-phasing of genotypes data improves imputation performances. On accuracy, IMPUTE2 [58] and minimac [59] recorded better and similar results across populations. However, low imputation accuracy were observed from the African American [59]. Particularly, using genotypes data from Women's Health Initiative (WHI) study that had 8,421 African Americans with 829,370 SNPs that remained after quality control, Howei et al. [59] recorded an average R^2 of 0.690 using 60 CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) + 59 YRI (Yoruba in Ibadan, Nigeria) of the 1000 Genomes project reference panel. This, however, improved slightly to 0.693 when pre-phasing was done with MaCH and later imputed with minimac [59]. Similar results were obtained when using 1000 genomes reference panel that had 283 EUR (European)+172 Africans. Additionally, highest average R^2 result of 0.73 was obtained from 1000 genomes project reference panel that had 381 EUR+174 Africans [59]. This again clearly shows that better imputation accuracy results can be obtained when a larger and a more diverse reference panel is used in performing imputation. Again, from the fact lower imputation accuracy were recorded from African Americans also highlighted how population diversity affects the performance of imputation tools.

Similarly, in other studies, to address the question of the choice of a reference panel, Liu et al. [98] used deep Whole Genome Sequencing (WGS) data of 90 European population (EUR) to assess the performances of three imputations tools; BEAGLE, IMPUTE2 and minimac 2 using two reference panels: Single population and Multi-population reference panels (referred to ALL in the study) that had 1092 individuals from 14 populations. Liu and his colleagues demonstrated that IMPUTE2 and minimac had higher imputation accuracy than BEAGLE, and that using a multi-ethnic reference panel like 1000 Genomes Project is more beneficial than using a single population reference reference panel, even when handling a specific population data like the European dataset [98]. Additionally, in recent studies Vergara et al. [99] compared the performance of three reference panel: Consortium on Asthma among African Ancestry Populations in the Americas (CAAPA), Haplotype Reference Consortium (HRC) and 1,000 Genomes project (1000G), using 3,747 African American populations recruited from two cohorts: HCV and COPDGene cohorts. Again, 1000G reference panel, which is a multi-ethnic reference panel, recorded the best performance both in coverage and in accuracy [99].

Here in this study, we evaluate and compare the potential of popular genotypes imputation method in imputing GWAS data from African, European and admixed populations.

3.3 Materials and Methods

3.3.1 Study Data and Reference Panels

We simulated three datasets for a five way admixed individuals using fractalSIM [101]. Informations about each of the datasets is summarized in **Table 3.1** and simulation details can be found in the next section. We further simulated homogeneous datasets that mimics African and European ancestries. The information of simulation parameters for the homogeneous datasets had been previously described in [101]. Each simulated population had three datasets as described in the **Table 3.1** below.

Table 3.1: *Simulated Data of African, European and Admixed Population*

Population	No of Individuals	Number of SNPs
AFRICAN	1,000	9,139,969
	3,000	9,139,969
	5,000	9,139,969
EUROPEAN	1,000	9,139,969
	3,000	9,139,969
	5,000	9,139,969
ADMIXED	1,000	623,330
	3,000	623,330
	5,000	623,330

3.3.2 Simulation approaches for the admixture datasets

For admixture simulations, we simulated case-control admixture datasets with five way admixture process at single point using HapMap3 datasets, under null and causal disease model. We extracted each of the parental population from HapMap3 using PLINK. The parental populations comprised of samples from EUR (European Ancestry), MAFR (Mixed African Ancestry), SAS (South Asian Ancestry), WAFR (West African Ancestry) and EAS (East Asian Ancestry), with each contributing genetic population of 0.15, 0.35, 0.10, 0.30 and 0.10 respectively. A description for each of the parental population is given in **Table 3.2**. Simulation approaches for homogeneous African population and European populations can be obtained from [101].

The simulation parameters were chosen in such a way that the simulated dataset would mimic the real dataset as much as possible. Indeed Mugo et al. [101] evaluated the resultant dataset

and demonstrated that the simulated dataset were similar to the parental real dataset in terms of minor allele frequency distribution and LD structure [101]. Consequently, the imputation performance on the simulated dataset would reflect that of the real population.

Table 3.2: ***Ancestral populations for the admixture simulations. ASW are African ancestry in SW USA, GBR are the British from England and Scotland, IBS represents Iberian populations in Spain, FIN are the Finnish in Finland, LWK are the Luhya in Webuye, Kenya, ESN are the Esan in Nigeria, ACB are the African Caribbean in Barbados, YRI represents Yoruba in Ibadan, Nigeria, GWD are the Gambian in Western Division âĂŞ Mandinka, MSL are the Mende in Sierra Leone, PJL are the Punjabi in Lahore, Pakistan, BEB are the Bengali in Bangladesh, ITU are the Indian Telugu in the U.K, STU are the Sri Lankan Tamil in the UK, CHB are the Han Chinese in Beijing, China, CDX are the Chinese Dai in Xishuangbanna, China, CHX are the Han Chinese South, China, KHV represents Kinh in Ho Chi Minh City, Vietnam and JPT represent Japanese in Tokyo, Japan***

Ancestry	DESCRIPTION	HapMAP3 Samples	Proportion
EUR	European Ancestry	CEU, GBR, IBS, FIN	0.15
MAFR	Mixed African Ancestry	LWK, ACB and ASW	0.35
WAFR	West African Ancestry	YRI, ESN, GWD, MSL	0.30
SAS	South Asian Ancestry	PJL, BEB, ITU, STU	0.10
EAS	East Asian Ancestry	CHB, CDX, CHS, KHV, JPT	0.10

For the null disease model admixture simulations, we simulated a null disease model for the parental populations in chromosomes 1, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 16, 18, 19, 21 and 22 in all the pre-admixture simulation. For the disease causal model, we simulated a causal model in chromosome 2, 6, 11, 15 and 20. **Table 3.3** summarizes the SNPs and their relative risk of the disease in each of the parental population.

Table 3.3: *Relative Risk for various disease SNPs. HET represents the heterozygous risk effect and HOMO represents the homozygous risk. BP represent the Base Pair Position while SNP is the SNP ID. SAME implies the risk effect and position is the same in all the ancestral populations*

ANCESTRY	CHROMOSOME	SNP	BP	HET/HOMO RISK
SAME	2	rs17042838	113843337	1.001, 1.002
		rs13410964	113843283	1.001, 1.002
SAME	6	rs2232238	29942857	1.001, 1.003
MAFR & WAFR	11	rs7106136	64748278	0.981, 0.980
		rs10897540	64757496	0.980, 0.983
SAS		rs7106136	64748278	0 .995, 0.996
		rs10897540	64757496	0.997, 0.996
EAS		rs7106136	64748278	0.991, 0.990
		rs10897540	64757496	0.990, 0.993
EUR		rs7106136	64748278	1.010, 1.015
		rs10897540	64757496	1.005, 1.001
MAFR, WAFR & EUR	15	rs365314	62606413	0.9998, 0.9995
SAS & EAS		rs365314	62606413	1.0002, 1.0001
MAFR, WAFR SAS, & EAS	20	rs6107104	25922993	1.00001, 1.00003
EUR	20	rs6115375	25871801	1.00001, 1.00003

3.3.3 Phasing and Imputation Using Different Tools

We extracted chromosome 1 to 22 from each of the datasets and phased each chromosome with EAGLEv2.3. For each chromosome, we obtained a phased vcf output format file. Imputation was carried out using five popular bioinformatics tools: IMPUTE2, IMPUTE4, MINIMAC3, MINIMAC4 and BEAGLE4. For each imputation run, we used 1000 Genome Project reference panel (1KG) as the reference panel. We performed imputation by splitting each chromosome into 5Mb overlapping chunks and thereafter specified the lower and upper bounds for each chunk in each imputation tool for each imputation run. IMPUTE2 and IMPUTE4 allow this option by specifying the *int* command preceded by indicating the lower and upper bound of each chunk. Similarly BEAGLE4 and BEAGLE 5 has an option for specifying the chromosome and the region for performing the imputation, in which case can be a chunk of interest. Mininac3 and minimac 4 on the other hand have an option of specifying a given region of the chromosome using *–start* and *–end* flags. Each imputation chunk was submitted in parallel clusters.

We assessed the imputation accuracy using the internal quality metric obtained from each

imputation program and the imputation concordance. First, we modeled the distribution of imputation accuracy for each tool verses the minor allele frequencies. For the imputation accuracy calculation, we only considered the imputed variants that had minor allele frequencies greater than 0.05. Further we categorized the imputation accuracy into three classes: The first class represented imputed variants with $maf > 0.05$ and the imputation accuracy less than 0.4. We regarded this class of imputed variants as *poorly imputed*. The second class were variants with $maf > 0.05$ and the imputation accuracy $>= 0.4$ but less than 0.7. We adopted the name *moderately imputed* for this class. The third class were imputed variants that had $maf > 0.05$ and the imputation accuracy $>= 0.7$ and the class was considered as *well imputed* variants. Finally, we summarized the results in form of tables and figures.

3.4 Results

We compared 5 imputation approaches in imputing simulated datasets that mimic African populations, European populations and an admixed populations. We used 1,000 genome project reference panel for imputing the missing variants. Details of this reference panel can be found from their publication [102]. We used several criteria to evaluate the imputation performance. First, we modeled the relationship between minor allele frequencies and the imputation accuracy (using the internal imputation quality metric for each tool). **Figure 3.1 (a)** shows the relationship between the imputation accuracy and the minor allele frequencies in Africa populations for the 1,000 samples dataset, while **Figure 3.1 (b)** and **Figure 3.1 (c)** for European and Admixed datasets respectively. As expected, our result shows that imputation accuracy increases with increase in minor alleles frequencies for all the populations. This is consistent with the previous studies that modeled the relationship between minor allele frequencies and the imputation accuracy [58]. Interestingly, the effects were more pronounced in the African populations, across the simulated datasets with different sample sizes than all the other populations as shown in **Figure 3.1** below. Similar trends were displayed by 3,000 samples and 5,000 samples datasets as illustrated in **Figure 6.1** and **Figure 6.2** in the **Appendix** section.

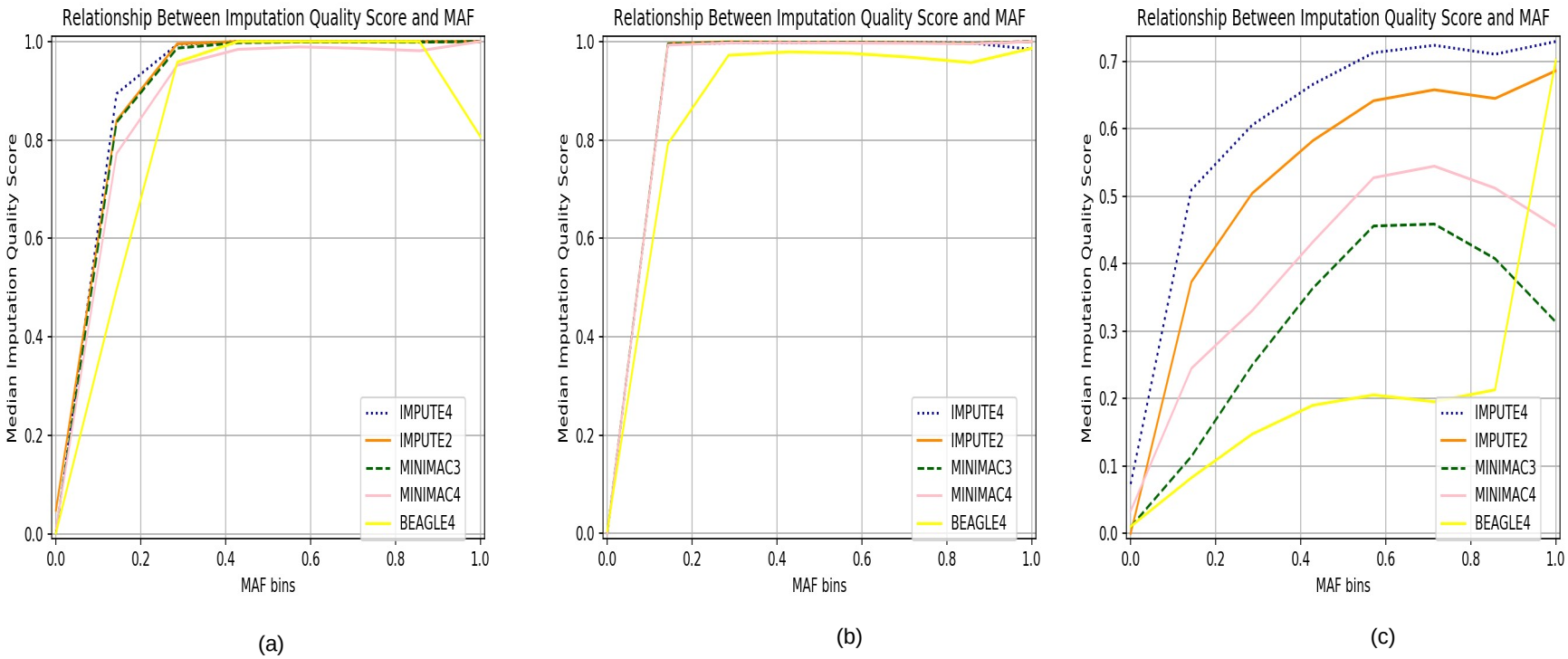


Figure 3.1: *Relationship between minor allele frequency and the imputation accuracy at different minor allele frequency bins for 1,000 samples from African population (a), European population (b) and admixed population (c).*

We then evaluated the imputation accuracy per simulated population dataset to determine the best imputation approach for each population. We applied two criteria in evaluating the imputation accuracy. First, we calculated the percentage number of variants that were considered as well imputed, moderately imputed and poorly imputed for each imputation method in each simulated population dataset as described in the method section. We then compared the imputation accuracy of the tools using different minor allele frequency bins: $maf < 0.01$ (rare variants), $0.01 \leq maf \leq 0.05$ (less common variants) and $maf > 0.05$ (common variants).

Table 3.4 represents the percentages for the imputed variants with $maf > 0.05$ at various thresholds for the three populations. In Africa populations, for the 1,000 samples dataset, IMPUTE 2 recorded the highest percentage (74.697%) of well imputed variants, followed by IMPUTE4 (66.78%), with BEAGLE4.1 recording the least percentage (35.43%) for the well imputed variants (**Table 3.4**). Similar trends were recorded for the 5,000 samples and 3,000 samples data with each case BEAGLE4.1 performing worst (**Table 3.4**). However, for the moderately and poorly imputed variants, BEAGLE4.1 recorded the highest percentages across all the African datasets, with exception in the 5,000 samples where it recorded 22.88%, which was slightly lower in percentage than minimac 4 (23.59%) (**Table 3.4**). The accuracy obtained from the admixed datasets recorded similar trend to that of African datasets. However, a lower percentage of well imputed variants were recorded across all the admixed samples, with IMPUTE2 recording the highest percentage of well imputed across all the admixed datasets. Additionally, with exception of IMPUTE2, more than half of the imputed variants from the admixed samples were falling within moderately and poorly imputed categories for all the tools. In BEAGLE4.1 for example, more than 74% of the variants were poorly imputed (**Table 3.4**).

For the European datasets, more than 85% of the imputed variants were falling within the threshold of well imputed for all the tools, across all the datasets. Minimac3 recorded the highest percentage (94.53%) for the well imputed variants, followed by IMPUTE4 (94.07%) across the 1,000 samples European dataset (**Table 3.4**). However, for the 3,000 samples and 5000 samples datasets, IMPUTE4 recorded the highest percentage of well imputed variants (97.07% for the 3,000 samples and 98.68 % for the 5,000 samples). Overall, BEAGLE 4.1 recorded the lowest percentage for the well imputed variants across all samples. Both the tools recorded less than 5% of the variants as moderately imputed and less than 11% of the variants as poorly imputed across all the European samples (**Table 3.4**).

Table 3.4: *Percentage Number of Imputed variants at different imputation thresholds. Imputed variants that were having* > 0.7 *imputation accuracy were classified as well imputed (Well-Imp), while those with imputation accuracy between 0.4 and 0.7 were considered as moderately imputed (Mod-Imp) otherwise poorly imputed (Poorly-Imp)*

	Well-Imp	Mod Imp	Poorly Imp	Well-Imp	Mod Imp	Poorly Imp	Well-Imp	Mod Imp	Poorly Imp
Tools		**Africa 1000**			**European 1000**			**Admix 1000**	
IMPUTE2	74.7	17.69	7.61	86.6	2.16	3.78	55.38	36.31	8.47
IMPUTE4	66.78	18.69	14.53	94.07	3.16	10.23	44.4	36.46	19.13
MINIMAC3	64.14	17.76	18.1	94.53	1.75	3.72	14.9	27.26	57.84
MINIMAC4	61.98	23.67	14.35	93.98	2.66	3.35	19.99	38.98	41.02
BEAGLE4	35.43	24.75	39.82	85.21	4.79	9.99	7.44	16.9	76.68
		Africa 1000			**European 1000**			**Admix 1000**	
IMPUTE4	67.13	18.66	14.21	97.07	0.96	1.97	44.11	36.5	19.39
IMPUTE2	74.64	17.75	7.61	78.83	16.45	4.72	55.39	36.08	8.47
MINIMAC3	64.19	17.72	18.09	94.62	1.66	3.72	5.77	12.43	81.79
MINIMAC4	62.21	23.66	14.13	94.19	2.42	3.39	19.96	38.95	41.08
BEAGLE4	34.5	24.61	40.89	85.75	4.83	9.42	7.21	16.11	76.68
		Africa 1000			**European 1000**			**Admix 1000**	
IMPUTE4	66.46	17.68	7.66	98.67	22.46	33.87	44.32	36.38	19.29
IMPUTE2	74.66	18.8	14.75	78.31	13.13	8.55	54.93	36.31	8.76
MINIMAC3	64.28	17.7	18.02	94.58	1.71	3.7	14.89	27.11	57.99

Continued on next page

Table 3.4 – continued from previous page

	Well-Imp	Mod Imp	Poorly Imp	Well-Imp	Mod Imp	Poorly Imp	Well-Imp	Mod Imp	Poorly Imp
MINIMAC4	62.24	23.59	14.17	94.12	2.52	3.37	19.92	38.83	41.23
BEAGLE4	33.63	22.88	43.49	83.77	5.37	10.86	7.98	17.99	74.03

Table 3.5 shows the comparison of imputation accuracy at different bins for different simulated datasets. For the simulated African datasets, across all the sample sizes, IMPUTE2 recorded the highest imputation accuracy for the common variants (80.21 for the 1,000 samples data, 80.23% for the 3000 samples data and 80.20 % for the 5,000 samples data) and less common variants (72.11% for the 1000 samples, 72.31 % for the 3,000 samples and 72.24 %) (i.e common variants are variants with $maf > 0.05$ and the less common variants are the variants with $0.01 \leq maf \leq 0.05$). However, for rare variants (varaints with $maf < 0.01$) for the same simulated datasets, IMPUTE4 recorded the highest imputation accuracy followed by minimac 4, then minimac 3. However, BEAGLE4 recorded the lowest imputation accuracy at all minor allele frequency levels (**Table 3.5**).

For the simulated European datasets, minimac 3 recorded the highest imputation accuracy followed by minimac 4, IMPUTE4, IMPUTE2 and BEAGLE4 respectively for the common variants. However, for the less common variants, IMPUTE4 recorded the highest imputation accuracy followed by BEAGLE4 recording the least across all samples as shown in **Table 3.5**.

For the admixed individuals, IMPUTE2 recorded the highest imputation accuracy across all the simulated datasets for both the common variants and the less common variants. IMPUTE4 recorded the highest imputation accuracy across all the simulated datasets for the rare variants regardless of the population. Imputation accuracy for the simulated admixed datasets recorded similar trend to that of the African samples across all the tools with the only difference being on the percentage accuracies (**Table 3.5**).

Table 3.5: *Percentage Imputation Accuracy at different minor alleles frequency (MAF) bins.*

	Africa 1000			European 1000			Admix 1000		
	MAF>0.05	**0.01<=MAF<0.05**	**MAF<0.01**	**MAF>0.05**	**0.01<=MAF<0.05**	**MAF<0.01**	**MAF>0.05**	**0.01<=MAF<0.05**	**MAF<0.01**
IMPUTE2	80.21	72.11	7.020	92.41	86.85	12.33	69.46	51.02	7.300
IMPUTE4	75.64	59.84	43.42	94.10	90.15	76.13	62.79	37.42	29.28
MINIMAC3	73.94	57.52	12.11	94.4	88.18	17.08	37.94	11.42	0.970
MINIMAC4	73.58	59.50	12.86	94.31	88.84	16.98	46.33	24.47	3.33
BEAGLE4	57.14	40.92	5.23	85.77	63.09	10.08	20.51	8.310	1.050
	Africa 3000			**European 3000**			**Admix 3000**		
	MAF>0.05	**0.01<=MAF<0.05**	**MAF<0.01**	**MAF>0.05**	**0.01<=MAF<0.05**	**MAF<0.01**	**MAF>0.05**	**0.01<=MAF<0.05**	**MAF<0.01**
IMPUTE2	80.23	72.13	7.270	81.35	75.48	14.46	69.52	51.12	8.26
IMPUTE4	75.94	60.20	42.07	94.02	89.79	75.29	62.61	37.47	26.09
MINIMAC3	74.01	57.76	13.31	94.54	87.95	19.53	22.80	11.96	0.990
MINIMAC4	73.77	59.76	13.38	94.45	88.66	19.41	46.31	24.57	3.67
BEAGLE4	56.64	40.50	6.170	86.66	63.39	12.04	20.16	8.130	1.080
	Africa 5000			**European 5000**			**Admix 5000**		
	MAF>0.05	**0.01<=MAF<0.05**	**MAF<0.01**	**MAF>0.05**	**0.01<=MAF<0.05**	**MAF<0.01**	**MAF>0.05**	**0.01<=MAF<0.05**	**MAF<0.01**
IMPUTE2	80.2	72.24	7.270	82.21	74.27	11.21	69.28	51.11	8.660
IMPUTE4	75.38	59.64	41.30	93.68	88.86	75.24	62.70	37.83	24.68
MINIMAC3	74.06	58.08	13.42	94.55	88.04	20.23	37.85	11.51	1.010
MINIMAC4	73.75	59.97	13.49	94.45	88.72	20.10	46.24	24.61	3.720

Continued on next page

Table 3.5 – continued from previous page

	MAF>0.05	0.01<=MAF<	MAF<0.01	MAF>0.05	0.01<=MAF<	MAF<0.01	MAF>0.05	0.01<=MAF<	MAF<0.01
BEAGLE4	55.45	40.19	5.960	84.98	58.01	12.02	21.23	8.480	1.260

Finally, we compared the imputation concordance the common variants. We restricted the comparison to common variants since it has been previously shown that direct comparison of concordance rate for low frequency or rare variants generally leads to incorrect assessment of imputation accuracy [94]. **Table 3.6** shows the comparison for imputation concordance for different datasets. Just like in imputation accuracy, IMPUTE2 recorded the highest imputation concordance with at least 99.2% concordance across all the simulated African datasets. IMPUTE4 equally recorded higher concordance with at least 97.8% across all the African datasets whereas minimac3, minimac4 and BEAGLE 4 recorded slightly lower imputation concordance percentage with BEAGLE 4 recording the lowest. Similar trend was observed for the admixed datasets, with IMPUTE2 recording the best concordance percentage and BEAGLE4 recording the least. However, for simulated European datasets, IMPUTE4 recorded the highest concordance percentage that all the other tools. Interestingly, we obtained higher concordance percentages across all the tool with each tool attaining at least 99.90% concordance as shown in **Table 3.6**.

Table 3.6: *Percentage imputation concordance for variants with maf>0.05.*

	Africa 1,000	**European 1,000**	**Admix 1,000**
IMPUTE2	99.21	99.90	90.92
IMPUTE4	97.97	99.99	87.09
MINIMAC4	89.02	99.93	88.86
MINIMAC 3	89.64	99.97	88.93
BEAGLE4	87.14	99.96	87.57
	Africa 3,000	**European 3,000**	**Admix 3,000**
IMPUTE2	99.27	99.96	91.12
IMPUTE4	97.88	99.99	87.51
MINIMAC4	89.30	99.93	87.87
MINIMAC 3	90.71	99.97	89.19
BEAGLE4	88.53	99.93	87.57
	Africa 5,000	**European 5,000**	**Admix 5,000**
IMPUTE2	99.21	99.96	91.31
IMPUTE4	97.97	99.99	87.50
MINIMAC4	89.02	99.94	87.79
MINIMAC 3	89.64	99.97	89.14
BEAGLE4	87.14	99.93	87.57

3.5 Discussion and Recommendation

Imputation has been recognized as one of the gold standard methods in Genome Wide Association Studies (GWAS) owing to its potential of increasing the number of markers that can be tested for association, hence improving the power of association studies [60] through fine-mapping, meta-analysis and functional GWAS. Several tools for performing genotypes imputation have been proposed and as such, their evaluations are very important as it can inform studies of the best tool that suits a given population. By using simulated genotypes that mimics African, European and admixed populations, we performed imputation with five popular, current and widely used imputation methods: IMPUTE4, minimac 4, BEAGLE4, minimac 3 and IMPUTE2 using 1000 Genomes Project reference panel. All the imputation programs requires the genotypes data to be pre-phased with exception of IMPUTE2 and BEAGLE 4 that perform both phasing and imputation. However, for better imputation experience, the developers recommends pre-phasing the study data. IMPUTE4 is the latest version of IMPUTE tools and represents a modification of IMPUTE2 to accept only a pre-phased study data. Minimac4 on the other hand is the latest version of minimac tools, and represents recent software advancements that have been developed to efficiently handle large reference panel at a lower computation cost and high accuracy [54]. We used 1000 genome project reference panel as the reference panel for imputing the missing variants. Details for the reference panel can be found from the developers website [102]. Our results, however, mainly focuses on the simulated dataset of African populations, which has had little attention as far as the evaluation of imputation performances for various methods in different populations are concerned. Additionally, we compared the imputation performances for the simulated African populations with that of the European populations and a multi-way admixed populations (as summarized in **Table 3.3**), which actually is the largest admixture individuals ever evaluated. Schurz et al. [103], for example, recently evaluated imputation performace on the South African colored (SAC) population, which is considered as a 5 way multi-admixed individuals, which as at that time was considered the largest admixed individuals ever evaluated. Our study therefore raises very serious considerations, since international migration expands at an exponential rate, we will soon have populations that are more than 5 way admixed individuals and as such, current tools should be advancing for the challenge that comes with admixed populations.

To put the results into context, we also modeled the relationship between minor allele frequencies and the imputation performance for each tool. Previous studies have demonstrated that the imputation accuracy increases with the increase in the frequency of the minor alleles.

As expected, and as shown in **Figure 3.1**, and **Figure 6.1** and **Figure 6.2** in the **Appendix** section, our results suggested similar trends for all the imputation tools across all simulated African and European populations. However, the simulated admixture population that recorded mixed results where the imputation accuracy increases and declines at some point with increase in minor alleles frequency. We nevertheless address a number of questions including: what is the best imputation approach for the African populations? How do current imputation approaches in Africa population compare with European populations and the admixed populations? Does the recent version of the imputation tools customized enough to handle African populations? Does the size of the study sample affects the imputation performance?

Based on the overall imputation performance, our results suggest that IMPUTE2 and IMPUTE4 should be the preferred choice when imputing data from homogeneous African decent, especially from West and Central Africa, and can make better imputation prediction than minimac 3, minimac 4 and BEAGLE4.1. However, BEAGLE 4 consistently gave lower imputation performance across all the three datasets of the simulated African origin. For the European populations descent, most of the softwares gave comparatively higher imputation quality. Indeed, all the softwares recorded a mean of more than 85% imputation quality for the common variants and a mean of over 60% for the less common variants. Based on the common variants alone, minimac 3 recorded the highest imputation quality, followed by minimac 4 and IMPUTE4. However, for the rare and less common variants IMPUTE4 recorded the highest imputation accuracy. In fact, IMPUTE4 recorded a mean of over 75 % for the rare variants across all the European samples, which was higher than what was attained by best tool (IMPUTE2) in the simulated African datasets based on common variants. For the admixed individuals, we obtained lower imputation qualities across all the samples. Nevertheless, IMPUTE2 gave the highest imputation accuracy and BEAGLE4.2 recorded the least imputation quality. Our results highlights that although Africa populations could be diverse, admixed individuals could be more diverse than the African populations and hence can present serious challenges for the imputation programs as evidenced by the results. Moreover, majority of African populations are highly admixed and similar results may be obtained in such populations.

Our results also suggests that imputation performance does not depend on the size of the study data but rather on the size of the reference panel, supporting previous finding [91]. Particularly, our study recorded very low or no effect of the imputation quality as a function of the size of the study sample for all the populations, across all the simulated datasets. Previously, Huang et al. [104] for example observed that when the sample size is small (about 10 individuals) and no reference panel is used, the boost in accuracy is generally very high as more individuals are

added into the study data. However, this reaches a plateau at some point, and the accuracy does not improve with increase in the number of individuals. In contrast, when a reference panel is used in performing imputation, then the imputation quality generally increases with increase in the size of the reference panel and does not generally depend with the size of the study sample [104]. Similarly, Zheng et al. [105] showed that using the largest and a diverse reference panel like the 1000 genome project reference panel, that we used in this study, results to better imputation quality. This is because, larger and a diverse reference panel can contain more information on the parental diversity than a population specific reference panel, hence improving the chances of imputing the rare variants [58].

Further, we observed that even though the imputation tools have been advancing to the newer versions, or newer tools being developed altogether, our results suggest that these advancements are so far not customized enough to capture the low linkage of disequilibrium and higher diversity within the African populations and the admixed populations. Surprisingly, all the newer version of the imputation programs recorded lower imputation accuracy than their older versions for both the African and Admixed populations. It this therefore very clear that the newer version or the new imputation programs are being developed to reduce the running time in processing large reference panels, and not to account for the genetic diversity and patterns of LD like those found in African and admixed populations. Nevertheless, for the European populations, there was a substantial improvement in imputation quality. Although not assessed by the current study, the computation speed for all the newer versions improved considerably and the experience were far much better as compared with the older versions.

Previous studies that compared the performance of genotypes imputation in different populations have shown that European population have higher imputation accuracy than the African and the admixed populations. Huang et al for example evaluated imputation accuracy on 29 populations [104]. Based on their results, the highest accuracy was observed from the European populations, then East Asian population, followed by East Asia, Central Asia, South Asia, America, Oceania and Middle East, and Africa recorded the least. However, our study extends this further and applies the recent and a more diverse reference panel, 1000G reference panel, and the most recent version of the imputation programs, as opposed to Huang et al. [104] that only evaluated the performance based on only one imputation program (MACH), which has been overtaken by the newer version of the imputation programs. In other studies, Hancock et al. also evaluated imputation performance in African American individuals using four imputation programs: IMPUTE2, MaCH, BEAGLE, MaCH-Admixed. Interestingly, IMPUTE2 recorded the overall highest imputation quality of 0.68 when using the reference panel that comprises of

YRI+CEU+ASW, which dropped to 0.55 when the whole 1000 Genomes reference panel was used, similar to our findings for the admixed populations for the common variants. Remarkably, this findings illustrated that the comparison of imputation metric can be assessed directly by the internal metric from the imputation programs. Specifically, the evaluation based on internal quality metric and that of masked analysis generated the same conclusion as far as the best performing software is concerned [106]. Elsewhere, Liu et al. compared the performance of IMPUTE2, minimac and BEAGLE 3 in European populations [98]. IMPUTE2 and minimac performed better than BEAGLE, and that the speed of both the programs, although not evaluated here, were superior to that of BEAGLE. Similarly Nothnagel evaluated the performances of imputation tools using German Descent populations[100] and recommended either BEAGLE or MACH for performing imputation the German descents.

The inability of the imputation tools to capture this diversity highlights that indeed imputation tools have still a long way to go and should be prepared for such challenges. Our study highlights the need of newer and robust imputation tools that can handle the diversity and low LD patterns of the admixed and African populations.

In conclusion our study is the first to evaluate potential of genotyoes imputation tools in African populations using the most current software tools. African and Admixed population recorded very low quality imputation for both low frequency SNPs (SNPs with MAF between 0.01 and 0.05) and rare SNPs (with MAF<0.01). Irregardless of the software used, or the best software from the results, our study highlights that future tools should exploit models that can capture the diversity and patterns of LD found in African and admixed populations. Moving forward, IMPUTE2 should still be the preferred method when imputing data from African or admixed populations regardless of its computation cost.

Chapter 4

Raw Genotypes Verses Summary Statistics Imputation on Malaria GWAS from MalariaGen

4.1 Introduction

Hidden Markov model based imputation methods have been applied as the gold standard for performing genotypes imputations [107]. However, these methods can only be applied to individual level genotypes [84], which in most cases may not be readily available due to the logical constraints or confidentiality [107] and the process for data sharing can be very time consuming given data sharing agreements that have to be met [108]. Nevertheless, it has been shown that comparative power in association studies can be achieved when the imputation is implemented at summary statistics level, which in most cases are readily available from the published findings [84]. Indeed studies have demonstrated that imputation from summary statistics has a potential of recovering almost the same signal as that of individual level genotypes imputation with minimal or no increase in false positive association rate [84]. Moreover, summary statistics based imputation is known to be very fast and computationally more efficient than the individual level genotypes imputation methods [84, 107]. Additionally, it has been shown to improve the power of enrichment in most loci [84] thus making it very essential for enrichment analysis.

There exists many summary statistics based imputation tools. Some of the most commonly used summary statistics based imputation methods include ImpG [84], DIST [107], DISTMIX [109] (which is a modification of DIST to handle both homogeneous and admixed population), DISSCO [108] and many more. To date, although the comparison of different summary statistics

based imputation is still lacking, ImpG [84] (which applies Multivariate Gaussian method to estimate associations at various missing SNPs) is considered the most popular method. In fact, in one of the studies that evaluated the performance of this tool demonstrated that it can recover as much signal as that obtained via raw genotypes based imputation [84]. Particularly, this study showed that summary statistics based imputation with this method (ImpG) can recover up-to approximately 87% of the effective sample size, which is almost the same as individual level genotypes imputation, which can recover up-to 89% of the effective sample size [84].

While several studies have compared summary statistics based imputation [84, 107, 109], evaluation of the performance of any summary statistics based imputation in African populations is still lacking. In this chapter, we compare the performance of two imputation approaches: Imputation via summary statistics using IMPG [84] and imputation on the raw genotypes data using IMPUTE2 [58] , which we have identified as the best tool for imputing GWAS data from African populations. We then perform a meta-analysis of the imputed datasets. Finally, we perform a pathway analysis using the meta-analysis results.

4.2 Comparison of GWAS from Raw Genotypes Data Imputed with IMPUTE2 [58] and Summary statistics from ImpG

4.2.1 Study Data, Imputation and Quality Control

We obtained both raw genotypes and summary statistics data of children diagnosed with severe malaria for the Kenyan population from MalariaGen website, which is available for public access. The data was initially applied in GWAS of severe malaria [30]. All the quality control steps, DNA extraction and sequencing are well described in the publication paper ([30]).

The study datasets contained case/control subjects that had severe malaria from the African populations: Kenya, Gambia and Malawi. We did quality control with PLINK 2.0 [110] on 3,142 individuals (1,505 cases and 1,474 controls and 163 missing phenotypes) from the Kenyan population, 4,179 from The Gambia and 4,473 from Malawi. We excluded the imputed SNPs that had less than 0.7 imputation $info$ and then applied missingness test (geno>0.05), minor allele frequency test (maf<1%) and Hardy Weignberg Equilibrium (–hwe include-nonctrl <0.0001). Additionally, we applied a less stringent HWE on both cases and controls (hwe<0.0000001) and conducted a heterozygosity check removing individuals that deviated more than 3 times the mean heterozygosity rate. Further, we performed relatedness check ($\hat{\pi} > 0.1875$ (median of 2nd degree

relative and 3rd degree relatives), removing individuals with the lowest call rate in each of the closely related individuals pairs. We finally retained curated datatsets with a total of 14,663,086 SNPs (for the 1,505 are cases and 1,474 are controls), 7,558,176 SNPs (for the 2429 cases, 2,491 controls) and 8,193,112 SNPs (for the 1,193 cases and 1,321 controls) for Kenya, Gambia and Malawi populations respectively.

To evaluate the possibility of population substructure and self reported ethnicity in the datasets, we characterized the presence of population substructure in the data which can bias the GWAS findings by performing Principal Component Analysis (PCA). We applied PLINK 2.0 in calculating the first top ten principal components (PCs) for each of the curated datasets. We then plotted all the individuals on the first two PCs as shown in **Figure 4.1** for Kenya **(A)**, for Malawi **(B)** and **(C)** for The Gambia.

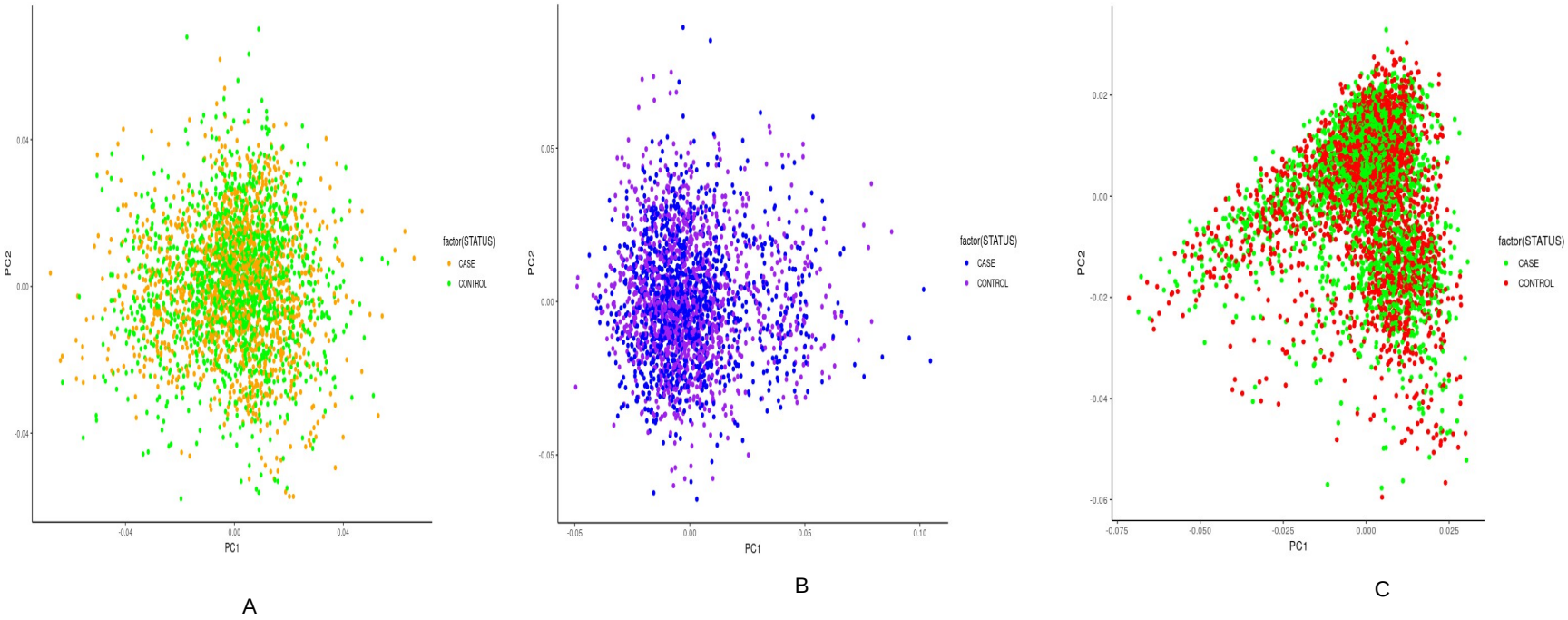


Figure 4.1: *Principal Component Analysis plots for Kenya* ***(A)****, Malawi* ***(B)*** *and Gambia* ***(C)*** *using the first two top Principal Components (PCs).*

As expected, we observed substantial population substructure across all the datasets. Some individuals across all the populations could be assigned to a distinct ethnic group whereas others seem to have complex ancestries, and could not be assigned to any specific ethnic group. Importantly cases and controls were not separated using the plot of the forst two PCs as shown in the **Figure 4.1**. Nevertheless, we included the first ten PCs as covariates in the association analysis on the curated datasets to account for the self reported ethnicity and population substructure. Additionally, we applied a mixed model approach implemented in EMMAX in calculating the kinship matrix, which we thereafter included in the logistic regression. All the analysis were based on the additive model of association.

For the summary statistics data, we retained only SNPs that had $r^2pred > 0.6$, which is an internal accuracy measure by ImpG, for the downstream analysis. There were 7,835,854 for the Kenyan sample, 6,986,174 SNPs for the Gambian sample and 7,958,269 SNPs from Malawi.

4.2.2 Results

Association Result for Raw Genotypes Based Imputation

Figure 4.2 (a), 4.3 (a), and **Figure 4.4 (a)** shows the Q-Q plot for Kenya, Gambia and Malawi respectively. The genomic control (λ_{GC}) for Kenya was 1.054428, 1.050381 for The Gambia and 1.052303 for Malawi. All these genomic control values are acceptable and suggest very little deviation from the null expectation. On the other hand, **Figure 4.2 (b), 4.3 (b)**, and **Figure 4.4 (b)** shows the respective Manhattan plots for Kenya, The Gambia and Malawi. The genome wide significance threshold was set at 5×10^{-8} and is represented by the red line in each Manhattan plot. No SNP surpassed this threshold for Malawi and Gambia associations. However, there were some SNPs on chromosome 11 from the Kenya dataset that surpassed the threshold value.

For the summary statistics, the genomic control values were $0.9957279, 0.988425, 0.9983602$ for Kenya, Gambia and Malawi respectively. All these values suggest a little deflation of the P-values, and little departure from the null expectation.

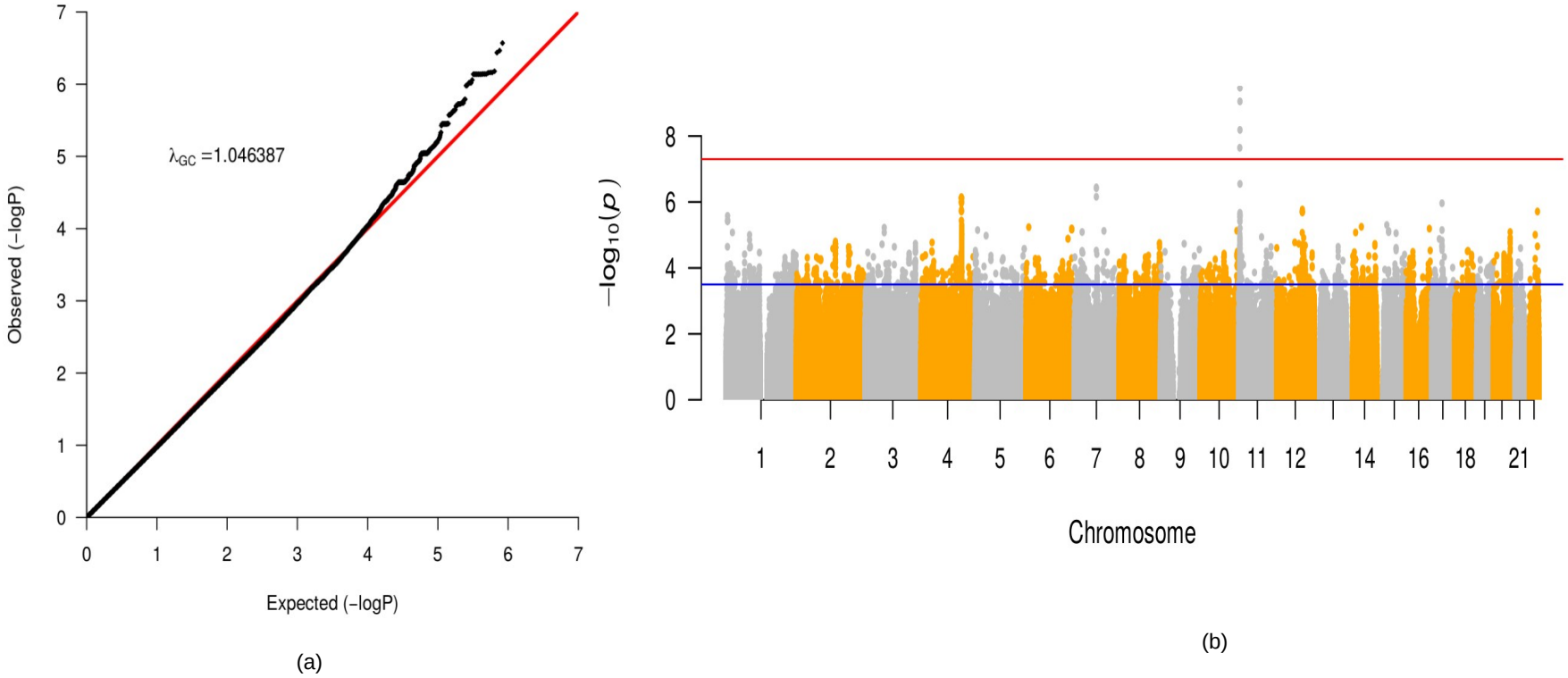


Figure 4.2: *QQ-Plot and Manhattan plot for GWAS of raw genotypes from Kenya. The Genomic control value after adjusting for the covariates was 1.054428 and suggest very little departure from the null expectation. Some SNPs, on chromosome 11, surpassed the genome wide threshold.*

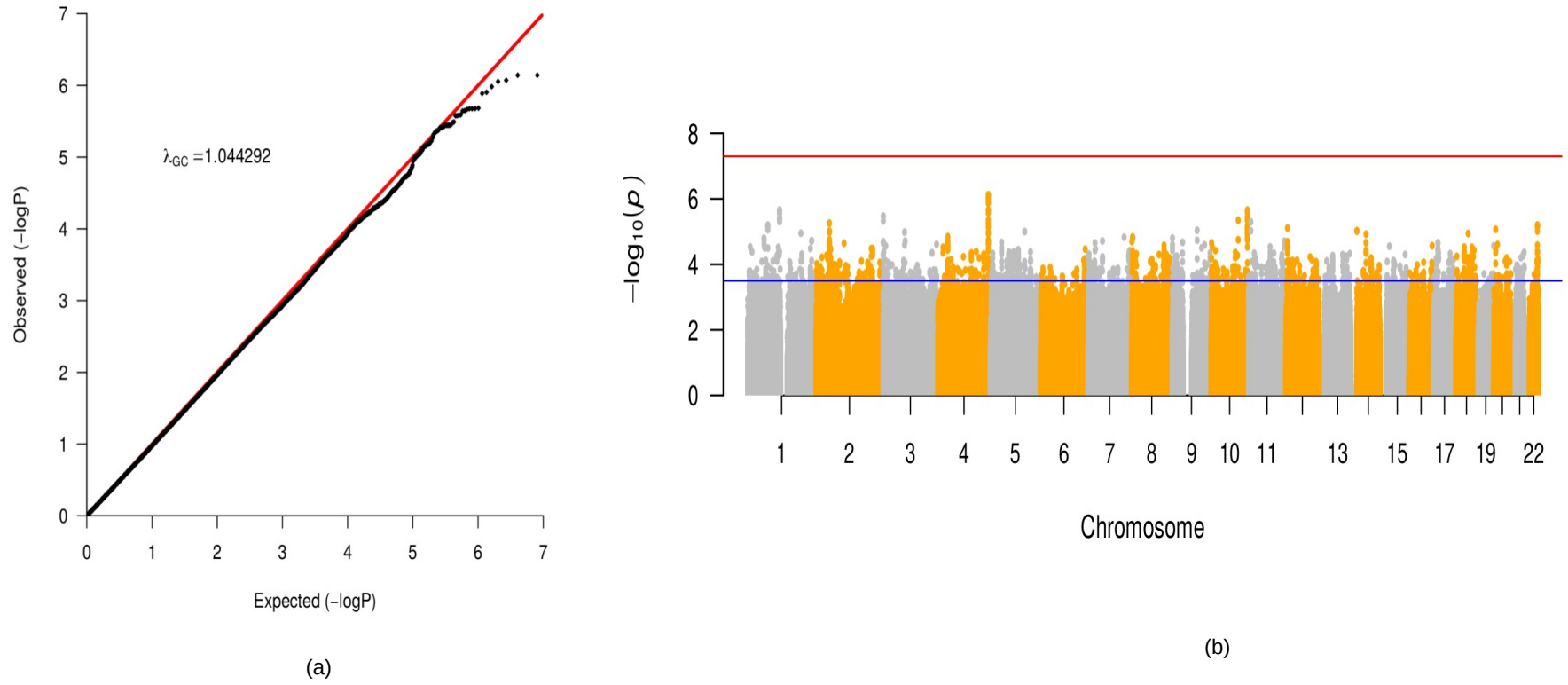


Figure 4.3: *QQ-Plot and Manhattan plot for GWAS of raw genotypes from The Gambia. The Genomic control value after adjusting for the covariates was 1.044292 and suggest very little departure from the null expectation. No SNP attained the genome wide threshold.*

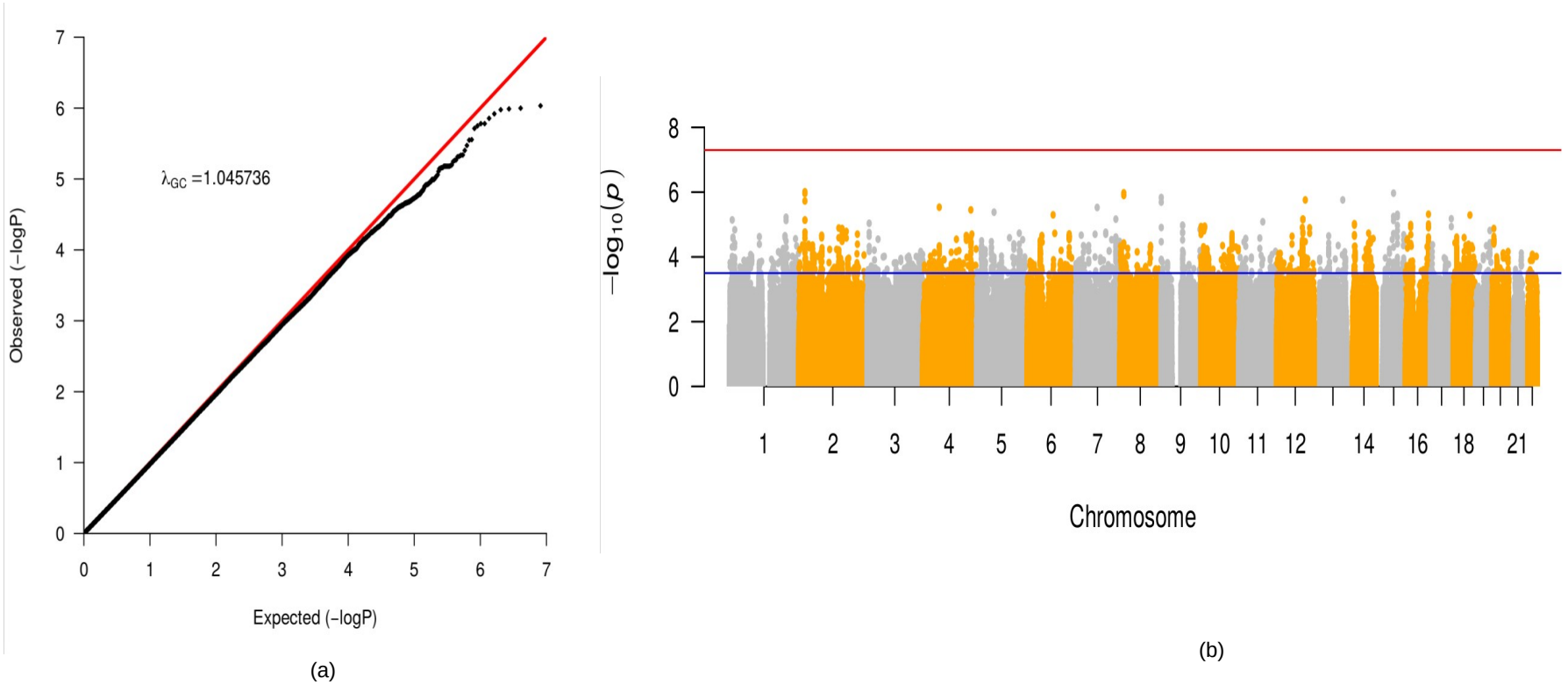


Figure 4.4: *QQ-Plot and Manhattan plot for GWAS of raw genotypes from Malawi. The Genomic control value after adjusting for the covariates was 1.045736 and suggest very little departure from the null expectation. No SNP attained the genome wide threshold.*

Association Result for Raw ImpG Imputed Summary statistics

Figure 4.5 (a), 4.6 (a), and **Figure 4.7 (a)** shows the Q-Q plot for Kenya, The Gambia and Malawi respectively. The genomic control (λ_{GC}) were $0.9957279, 0.988425, 0.9983602$ for Kenya, Gambia and Malawi respectively. All these values suggest a little deflation of the P-values, and little departure from the null expectation. Similarly, **Figure 4.5 (b), 4.6 (b)**, and **Figure 4.7 (b)** shows the respective Manhattan plots for Kenya, The Gambia and Malawi. Just like the raw genotypes imputed by IMPUTE2 association, we set the genome wide significance threshold at 5×10^{-8} and is represented by the red line in each Manhattan plot. However, no SNP surpassed the significance threshold for all the datasets.

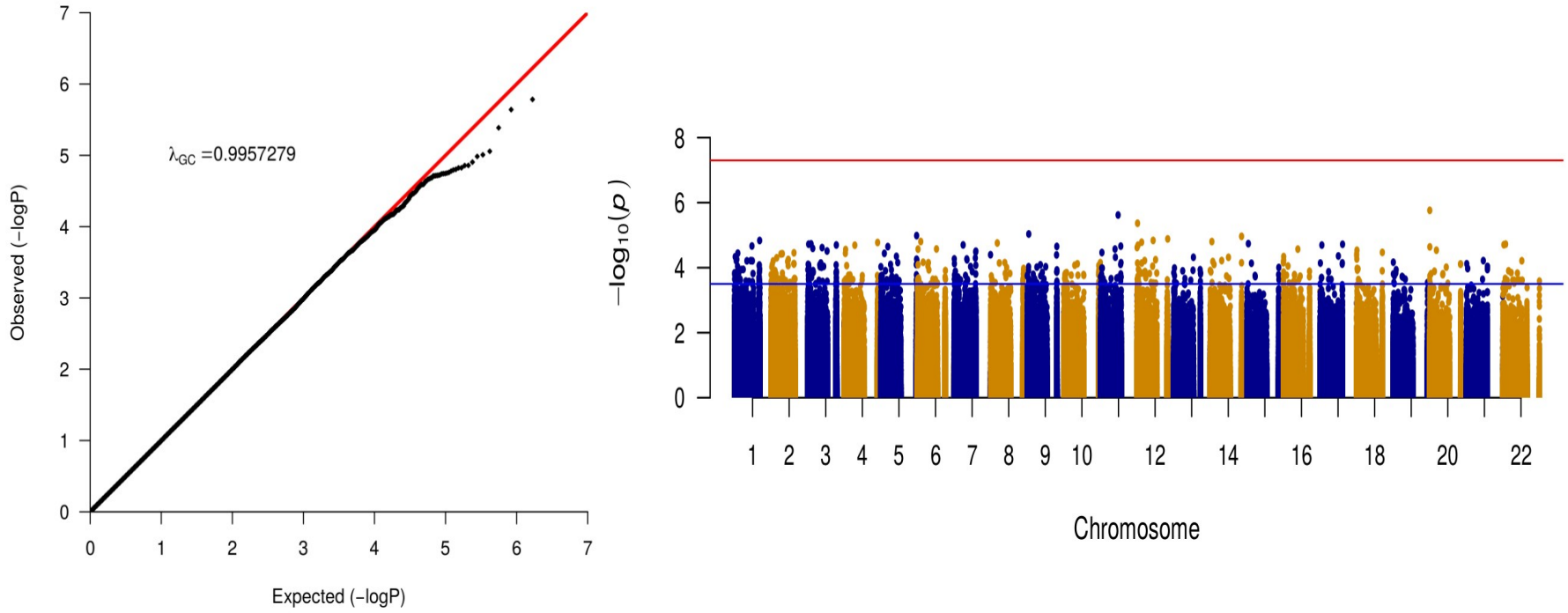


Figure 4.5: *QQ-Plot and Manhattan plot for ImpG imputed summary statistics from Kenya. The Genomic control value after excluding the strand ambiguous SNPs and SNPs with $r^2 pred < 0.6$ was 0.9957279. No SNP attained the genome wide threshold.*

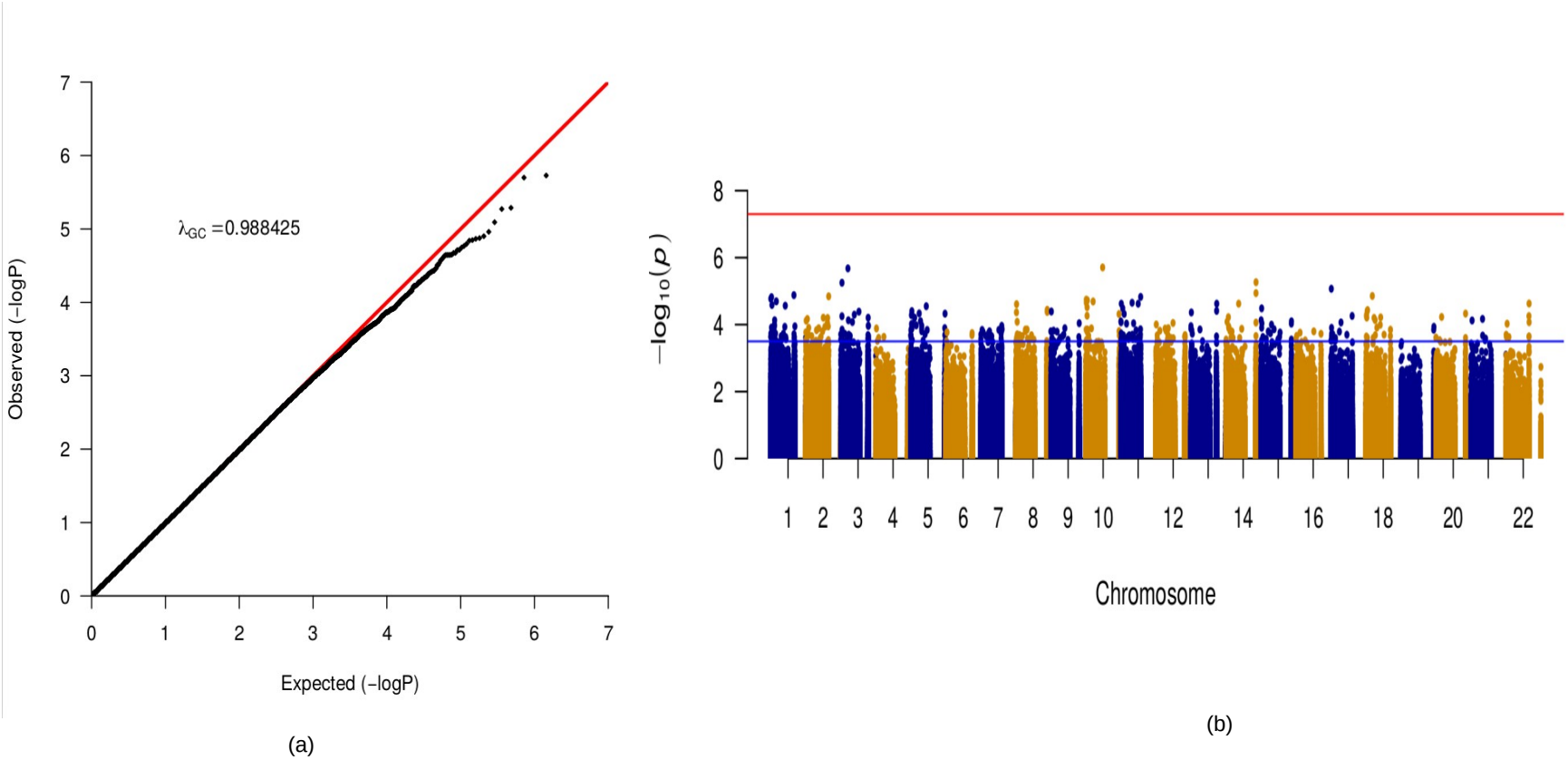


Figure 4.6: *QQ-Plot and Manhattan plot for ImpG imputed summary statistics from Kenya. The Genomic control value after excluding the strand ambiguous SNPs and SNPs with* $r^2 pred < 0.6$ *was 0.988425. No SNP attained the genome wide threshold.*

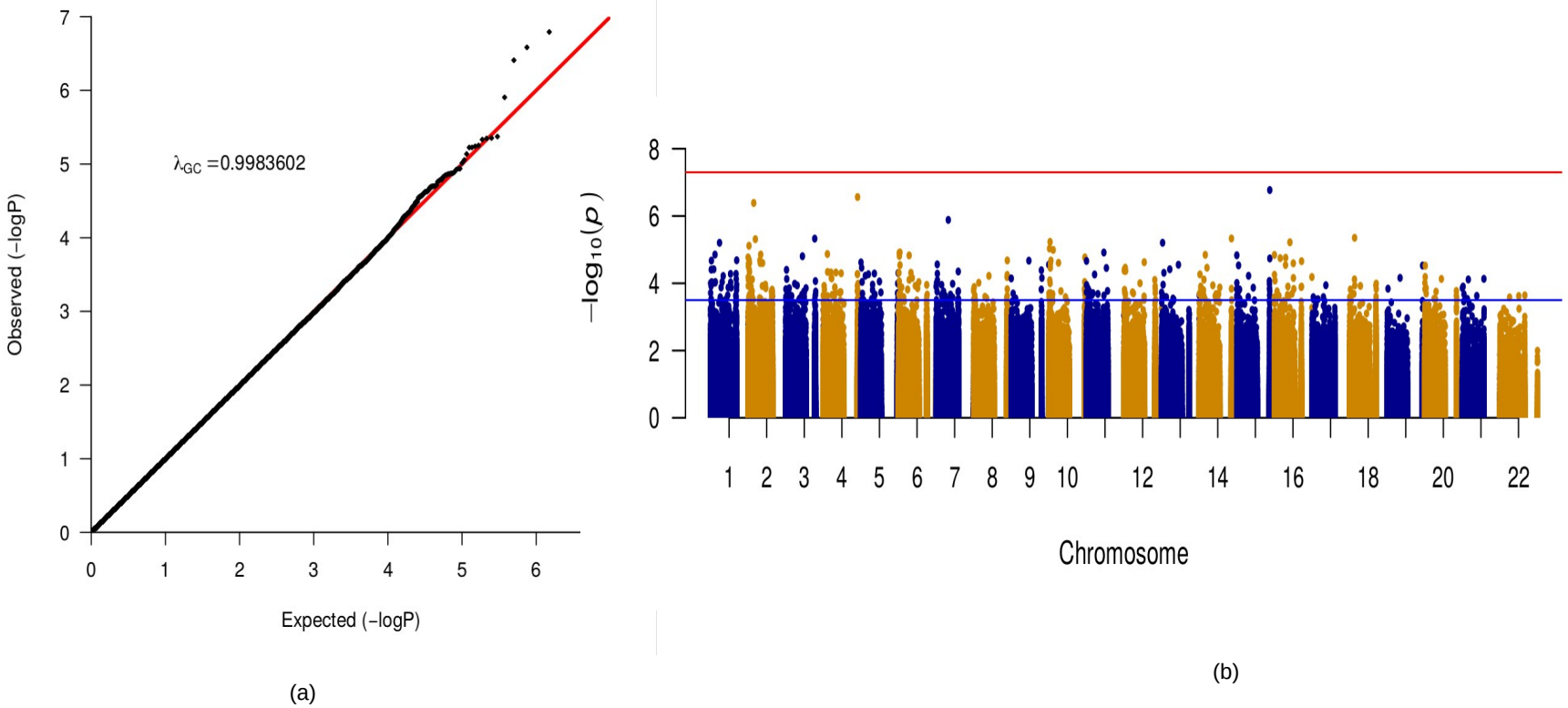


Figure 4.7: *QQ-Plot and Manhattan plot for ImpG imputed summary statistics from Kenya. The Genomic control value after excluding the strand ambiguous SNPs and SNPs with $r^2 pred < 0.6$ was 0.9983602. No SNP attained the genome wide threshold.*

Comparison of GWAS from imputed raw genotypes and imputed summary statistics

To break down the results further, we lowered the genome wide threshold at various P-values and compared the number of SNPs at each P-value thresholds for association results from the raw genotypes data and the association result for the ImpG summary statistics. **Table 4.1** shows the number of SNPs at various P-value thresholds. At the stringent genome wide threshold of P-value $<= 5 \times 10^{-8}$, there were 9 significant SNPs from the Kenyan GWAS from the raw genotypes. However, there were no significant SNPs for both GWAS of raw genotypes from Malawi and The Gambia at this stringent GWAS threshold. Similarly there were no significant SNPs for all the population from the ImpG summary statistics at this threshold as shown in **Table 4.1**. At P-value of $< 5 \times 10^{-7}$, we obtained 12 SNPs from Kenyan GWAS of raw genotypes imputed by IMPUTE2 and 3 SNPs from Malawi GWAS from ImpG. Finally, at P-value $< 5 \times 10^{-6}$, there were no significant SNPs for the raw genotypes data from Malawi GWAS whereas Kenya and Gambia had 76 and 39 respectively. On the other hand, for ImpG imputed summary statistics, Malawi had the highest number of SNPs that attained P-value of $< 5 \times 10^{-6}$, followed by Kenya with 3 SNPs and The Gambia with 2 SNPs.

Table 4.1: *Signifiant SNPs at various thresholds from the GWAS of raw genotypes from Kenya, Malawi and Gambia imputed with IMPUTE2 [58] , and the GWAS of imputed summary statistics using ImpG from the same populations respectively.*

P-value	Population	Number of SNPs (IMPUTE2 [58])	Number of SNPs(ImpG)
P-value $<= 5.0 \times 10^{-8}$	Malawi	0	0
	Kenya	9	0
	Gambia	0	0
P-value $<= 5.0 \times 10^{-7}$	Malawi	0	3
	Kenya	12	0
	Gambia	0	0
P-value $<= 5.0 \times 10^{-6}$	Malawi	0	8
	Kenya	76	3
	Gambia	39	2

Finally, **Table 4.2** list the SNPs that were identified to be associated with severe malaria with P-Value $< 5.0 \times 10^{-8}$. All the 9 SNPs were located on chromosome 11 and were identified from the association studies of the Kenyan raw genotypes data after imputation with IMPUTE2. Of these, 3 SNPs were located on the *HBG2* gene. The remaining SNPs were not mapped to any gene from the dbSNP database. The other SNPs that were identified at various P-value thresholds

are listed in the **Appendix 6**.

Table 4.2: *Table showing the SNPs that were identified to be associated with severe malaria with P-value* 5.0×10^{-8}. * *represents the variants that were found in the dbSNP but was not mapped to any gene and # represent SNPs that were not found in the dbSNP but existed the association file.*

CHR	SNP	BP	Gene	A1	A2	MAF	BETA	SE	P
11	rs143022364	4617306	*	A	G	0.072437	-0.141144	0.025209	$2.15648e^{-08}$
11	rs12295158	5252794	*	G	A	0.0955631	-0.130534	0.0224625	$6.20201e^{-09}$
11	rs112075505	5276402	*HBG2*	T	TTTAAAG	0.072314	-0.160681	0.0255521	$3.20877e^{-10}$
11	rs112035597	5277116	*HBG2*	A	G	0.072314	-0.160681	0.0255521	$3.20877e^{-10}$
11	rs113981422	5277117	*HBG2*	A	C	0.072314	-0.160681	0.0255521	$3.20877e^{-10}$
11	rs12292063	5302406	*	A	G	0.0744936	-0.155516	0.025345	$8.46414e^{-10}$
11	kgp12988299	5304648	#	G	A	0.0609265	-0.207343	0.0274388	$4.13903e^{-14}$
11	rs111978456	5311492	*	GC	G	0.0683305	-0.185135	0.0262375	$1.7121e^{-12}$
11	rs145843585	5321510	*	C	CAGG	0.0593421	-0.216745	0.0279629	$9.10445e^{-15}$

4.2.3 GWAS-based Imputation: Discussion

Our study focused on the comparison of two imputation approaches: imputation via summary statistics and imputation via raw genotypes data from Kenya, Malawi and Gambian populations. We have shown that imputation via raw genotypes has more power in identifying SNPs that are associated with a phenotype. For example, we obtained a total of 11 significant associations from the GWAS of dataset that was imputed with IMPUTE as opposed to no significant association from the summary statistics that was imputed with ImpG. Moreover, our results further suggests that despite the fact that imputation via summary statistics is possible, the P-values generated from such studies are in most cases deflated and could not be corrected even upon application of further QC. As such, it almost impossible to assess the bias from GWAS findings due to population stratification or hidden relatedness from such associations. Previously, in the initial application of ImpG [84], Pasaniuc et al. obtained a $\lambda_{GC} = 0.94$ and they further observed that this could be improved up-to $\lambda_{GC} = 1.00$ upon inclusion of the pairwise correlation among the SNPs from the GWAS data. However, this information may not be available from most summary statistics and thus we did not consider that possibility in our analysis.

In other studies, Lee et al. evaluated the performance of DISTMIX [109], which is also a summary statistics based imputation program, and IMPUTE2 [58] in imputing data from Psychiatric Genetic Consortium Schizophrenia Phase 2 [109]. Their results show that GWAS conducted from data imputed by IMPUTE2 [58] has more potential in identifying more SNPs associated with any given phenotype. Nonetheless, the authors showed that some associations were identified by DISTMIX that were never identified by IMPUTE2 based association just like in our findings. Additionally, although they showed that DISTMIX can reduce the false positive errors that can bias the GWAS findings, this study did not calculate the genomic control value which has a potential of quantifying whether the GWAS results are inflated or deflated.

Out of the 9 significant SNPs, all from chromosomes 11, three were located on the protein coding gene, *HBG2*. The *HBG2* gene is known to be expressed in the fetal liver, spleen and bone marrow [111]. Previously, this gene, has been tested with association with Malaria [10] from the Gambian samples. However, there was only one SNP (from Jallow et al. studies) within 50 kb region of *HBG2* and *HBG1* hence was insufficient to warrant any association [10]. *HBG2* gene on the other hand has been previously shown to be associated with sickle cell disease in Tanzania [112]. Interestingly, all the three SNPS had a P-Value 3.20877^{-10} with BETA equal to -0.160681. However, the five other significant SNPs had only been validated via frequency and clustering and were not mapped to their respective genes and one SNP could not be identified from the

dbSNP database.

In the next **Section 4.3**, we apply a meta-analysis of these studies and try if we can replicate and improve the association signal of these findings in a meta-analysis sample.

4.3 Meta-Analysis

4.3.1 Methodology

We then sought to compare the meta-analysis using the summary statistics from ImpG and the meta-analysis using summary statistics from the raw genotypes imputed by IMPUTE2 [58]. Only SNPs that were common in all the datasets (Kenya, Gambia and Malawi) were retained for meta-analysis. We used METASOFT [113] tool, which has the potential of carrying out meta-analysis using both fixed, random effect, binary effect and the Han and Eskin random-effects model [113, 114]. We used an in-house python script to obtain the set of SNPs that were common among the datasets in each case. To control any possible confounding from population stratification, we first run all the SNPs without $-lambda_{mean}$ and $-lambda_{hetero}$ parameters in the METASOFT. We then obtained the values as $-lambda_{mean} = 0.974629$ and $-lambda_{hetero} = 0.599252$ from the log files of the meta-analysis run for the raw genotypes GWAS data imputed by IMPUTE2 [58] and $-lambda_{mean} = 4.859256$ and $-lambda_{hetero} = 1.005534$ for the ImpG summary statistics. We then repeated the meta-analysis by supplying these values to the METASOFT as recommended [114]. Variants with meta-analysis p-values less than $< 5 \times 10^{-8}$ were considered significant.

4.3.2 Results and Discussion

We obtained a total of 7,805,875 common SNPs across the three studies for IMPUTE2 based meta-analysis and a total of 840,249 from the ImpG based meta-analysis. **Table 4.3** summarizes the number of SNPs at various P-value thresholds.

Table 4.3: *Number of SNPs at different P-value threshold for different models from the meta-analysis of GWAS that was generated by datasets that were imputed with IMPUTE2 and meta-analysis of summary statistics that was imputed by ImpG.*

P-Value	Fixed Effect		Random Effect		Binary Effect	
	IMPUTE2	ImpG	IMPUTE2	ImpG	IMPUTE2	ImpG
5×10^{-8}	1	10	1	7	2	5
5×10^{-7}	2	49	2	42	3	14
5×10^{-6}	19	150	6	119	17	73

We identified a total of 27 SNPs above the P-value threshold of $< 5.0 \times 10^{-6}$ for the IMPUTE2 based meta-analysis. **Table 6.1** in **Appendix** lists all these SNPs, their base pair position (BP), the genes where there are located on and their P-values under various models. Of these SNPs, only two SNPs had a P-value of $< 5.0 \times 10^{-7}$ across all the models (random effect, fixed effect and binary effect). Of the two SNPs, one SNP was identified to be significantly associated with severe malaria at P-value $< 5.0 \times 10^{-8}$ across all the models. This SNP was identified as rs12295158 located in the *HBB* gene, which causes sickle cell disease for the homozygotes and has a protective effect for the heterozygotes [30]. **Table 4.4** further list the SNPs that had a P-value of $<= 5.0^{-7}$ across all the meta-analysis model for IMPUTE2 based meta-analysis.

Table 4.4: *Top SNPs that had a P-Value of less that* 5.0^{-8} *across all the models of meta-analysis of the GWAS from datasets imputed by IMPUTE2. P-FE and B-FE are the P-value and beta under fixed model; P-RE and B-RE are the P-value and beta under random effect; P-BE is the P-value under binary effect.*

SNP	BP	GENE	P- FE	BETA-FE	P-RE	BETA-RE	P-BE
rs3837432	205757	*B3GNTL1,SCGB1C1*	7.62690×10^{-12}	-0.0373725	7.62690×10^{-8}	-0.0373725	5.04733×10^{-7}
rs12295158	5252794	*HBB*	2.87576×10^{-12}	-0.125517	2.87576×10^{-12}	-0.125517	9.63597×10^{-12}

On the other hand, for ImpG based meta-analysis, 7 SNPs attained a P-value threshold of $< 5.0 \times 10^{-7}$ across all the meta-analysis models (random effect, fixed effect and binary effect). Table 4.5 lists all these SNPs, their base pair position, genes they are located on among others. Of these, only two SNPs had P-value of $< 5.0 \times 10^{-8}$ across all the meta-analysis models. These SNPs were rs183731078, which is located in *RFX3* gene, and rs8096513, which is located in *DLGAP1* gene. *RFX3* is a protein coding gene that encodes transcription factors, and is a member of the gene family of the regulatory factors X [115]. *DLGAP1* on the other hand is also a protein coding gene and has been shown to be associated with diseases like schizophrenia and Obsessive-Compulsive Disorder among others [116]. Several other SNPs showed strong but not conclusive associations using different models. A total of 159 SNPs attained a P-value of $< 5.0 \times 10^{-6}$. **Table 6.2** in the **Appendix** list all these other SNPs.

Table 4.5: ***SNPs that had a P-Value of*** 5×10^{-7}

SNP	BP	GENE	P- FE	BETA-FE	P-RE	BETA-RE	P-BE	P-Kenya	P-Malawi	P-Gambia
rs60577152	5079600	*ITPR1*	1.78471×10^{-8}	0.948862	1.78471×10^{-8}	0.948862	1.21524×10^{-7}	0.000166278	0.00652561	0.00100309
rs74857376	1731779	*GMDS*	2.67359×10^{-8}	0.929216	2.67359×10^{-8}	0.929216	1.93380×10^{-7}	0.00132840	0.00199223	0.000851119
rs114760297	6534810	*LY86-AS1*	3.24272×10^{-8}	0.980051	3.24272×10^{-8}	0.980051	2.23611×10^{-7}	0.000407766	0.00628715	0.000915931
rs183731078	3780522	*RFX3*	8.39562×10^{-9}	0.899794	8.39562×10^{-9}	0.899794	4.46842×10^{-8}	0.000157072	0.0115578	0.000163379
rs111685758	114426898	*RBM19*	1.46762×10^{-8}	0.890990	1.46762×10^{-8}	0.890990	1.07231×10^{-7}	0.00109744	0.00130871	0.000827569
rs112736328	128188145	*LOC440117*	1.27092×10^{-8}	1.00007	1.27092×10^{-8}	1.00007	9.10594×10^{-8}	0.00480794	0.000325555	0.000595629
rs8096513	4455491	*DLGAP1*	1.42909×10^{-9}	1.03503	1.42909×10^{-9}	1.03503	1.00692×10^{-8}	0.00455867	1.35397×10^{-5}	0.00106408

4.3.3 Identification of significant SNPs using m-value cut-off

To ensure that the SNPs identified were indeed significantly associated with severe malaria. We calculated meta-analysis P-value and restricted the calculation to only SNPs that were identified to have an effect by applying an m-value approach implemented in METASOFT, and then computing the combined P-value. We retained only SNPs that had an m-value of > 0.9 in at least two studies and a combined meta-analysis P-value of $< 5.0 \times 10^{-8}$. **Figure 4.8, 4.9** and **4.10** shows the forest plots of these SNPs together with their respective P-M plots. For this, rs12295158 on the *HBB* gene that was obtained from the meta-analyis of datasets that was imputed with IMPUTE2 met this criteria and obtained a meta-analysis P-value of 1.06^{-14} . On the other hand, for ImpG summary statistics meta-analysis, we again retained all the two SNPs, rs183731078 which belongs to **RFX3** gene and rs8096513, which is located in *DLGAP1* gene, with combined meta-analysis P-Value of 7.69×10^{-9}, 1.30×10^{-9} respectively. **Figure 4.8, 4.9** and **4.10** illustrates the forest plots of these SNPs.

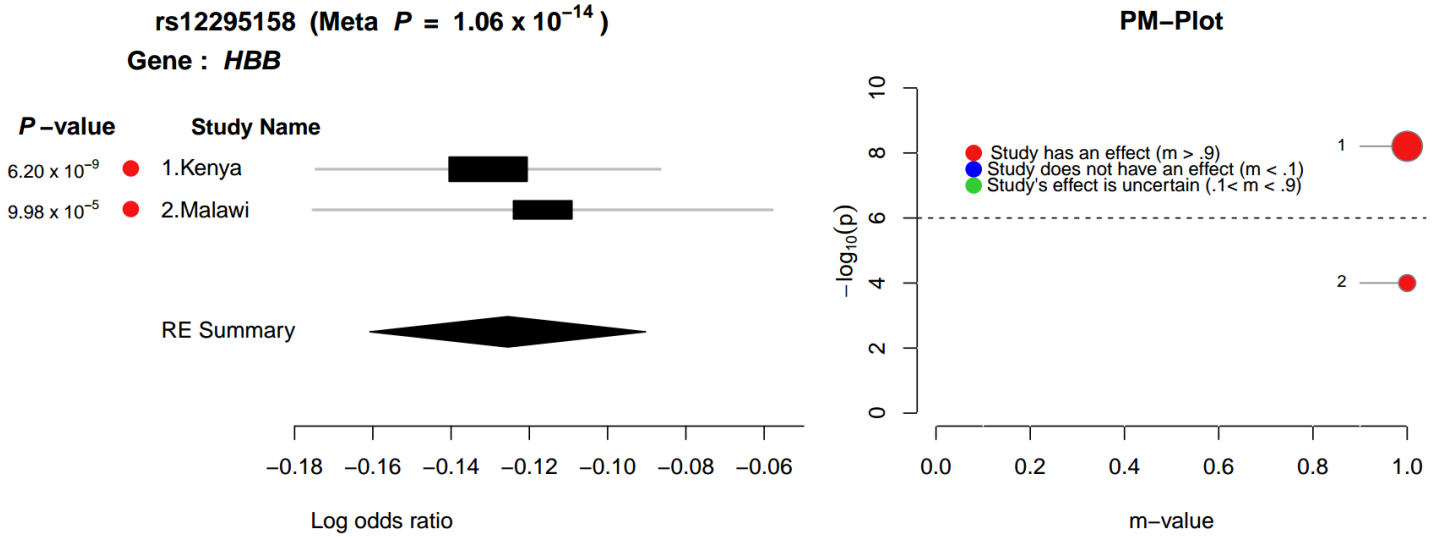


Figure 4.8: *Forest plot of the rs12295158 of the HBB gene from the meta-analysis of the case-control datasets of severe malaria from Kenya, Malawi and Gambia. The data was first phased and imputed with IMPUTE2 and finally a meta-analysis was performed on the summary statistics of the imputed datasets. The combined P-value for the meta-analysis was* 1.06×10^{-14} *with logs of odds ratio of* ≈ -0.13 *implying a protective effect*

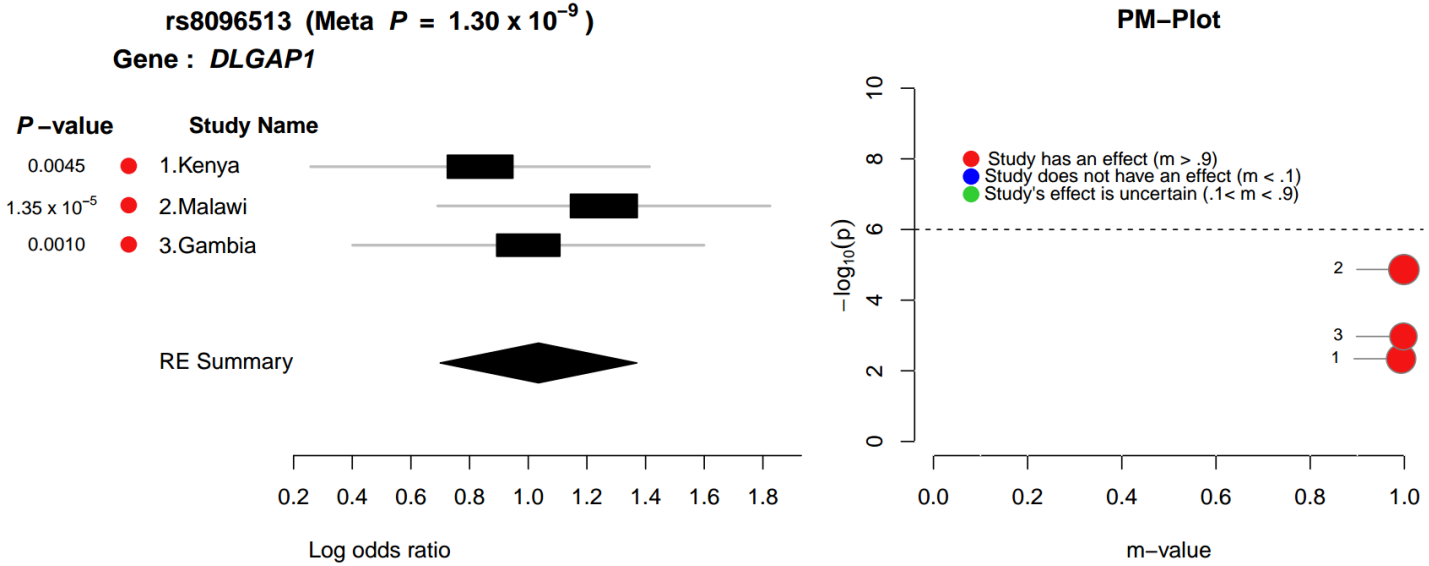


Figure 4.9: *Forest plot of rs8096513 of the DLGAP1 gene from the meta-analysis of the case-control dataset of severe malaria from Kenya, Malawi and Gambia from the summary statistics that was imputed by ImpG. The combined meta-analysis P-value was* 1.30×10^{-9} *with logs of odds ratio* ≈ 1.0*, which implies the effect is neither protective nor increases the risk.*

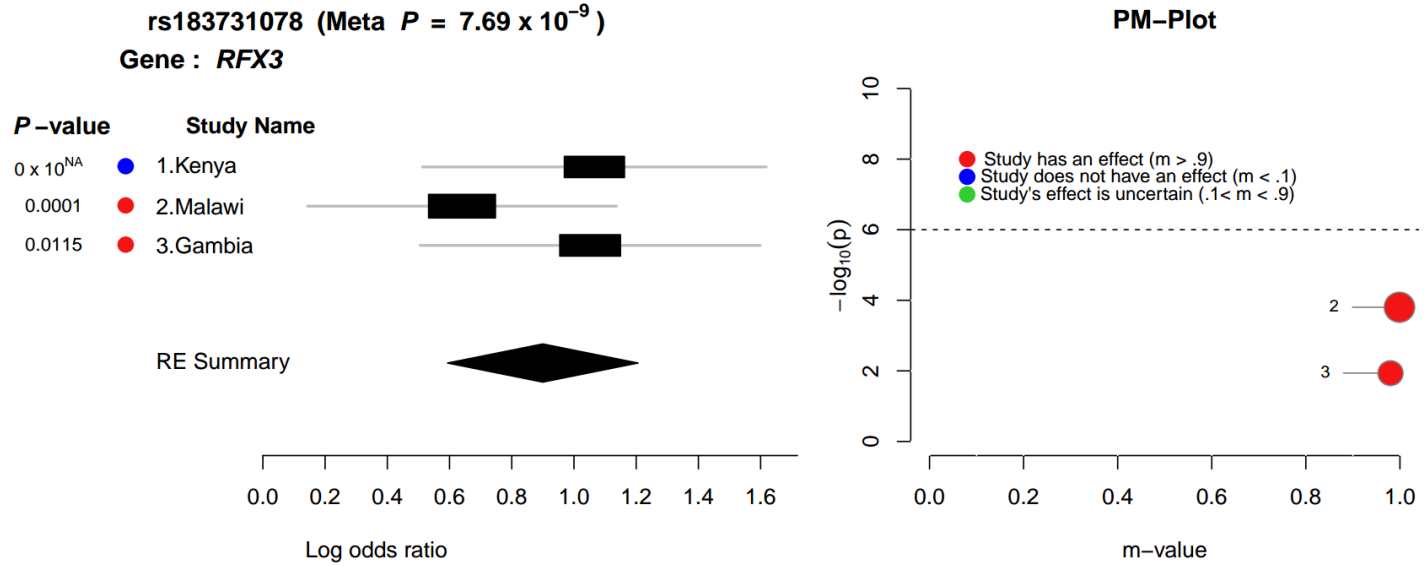
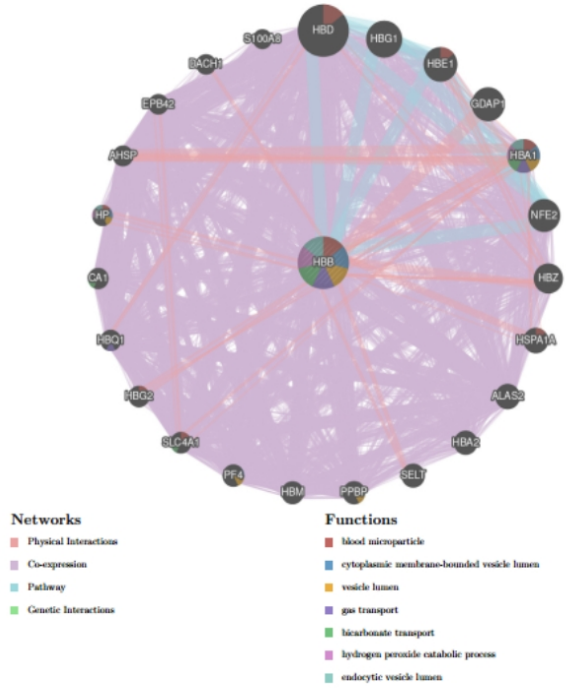


Figure 4.10: *Forest plot of rs183731078 of the RFX3 gene from the meta-analysis of the case-control dataset of severe malaria from Kenya, Malawi and Gambia from the summary statistics that was imputed by ImpG. The combined P-value for the meta-analysis was* 7.69×10^{-9} *with logs of odds ratio of* ≈ 0.9 *implying a protective effect.*

4.3.4 Pathway analysis and identification

To put the results into biological context, we then sought to understand the biological significance of genes identified from ImpG meta-analysis and IMPUTE2 based meta-analysis, and their relationship with the phenotype of interest (severe malaria). For this analysis, we used the genes obtained from the meta-analysis as the query genes and applied Genemania in predicting the genes they are related to in terms of co-expression, physical interaction, genetic interaction and shared pathways. Genemania is one the most popular and powerful web based application tool that has been applied in predicting pathway gene set of the genes identified from a given study [117]. Genemania, by default, predict twenty other additional genes that are functionally similar to the query list of genes. However, for this analysis, we adjusted the number of genes until we obtained the prediction network with the best FDR (False Discovery Rate) values. We performed three types of prediction: First, we used genes identified from IMPUTE2 as query list of genes; then genes identified via ImpG as the query list of genes, and finally we used both genes identified from ImpG and IMPUTE2 meta-analysis as the query list of genes.

For IMPUTE2 meta-analysis, we obtained a network of 23 related genes with 5,304 interactions. **Figure 4.11** shows the resulting network that was generated by the best FDR values and genes together with their scores. *HBD* was identified as gene that had the strongest association with a score of 0.005. The top biological functions which were associated with this network were blood microparticle at a P-value of $8.813 \times 10-9$, cytoplasmic membrane-bounded vesicle lumen, vesicle lumen, and bicarbonate transport both at a P-value of 1.977×10^{-4} and gas transport at P-value of 7.561×10^{-4}.

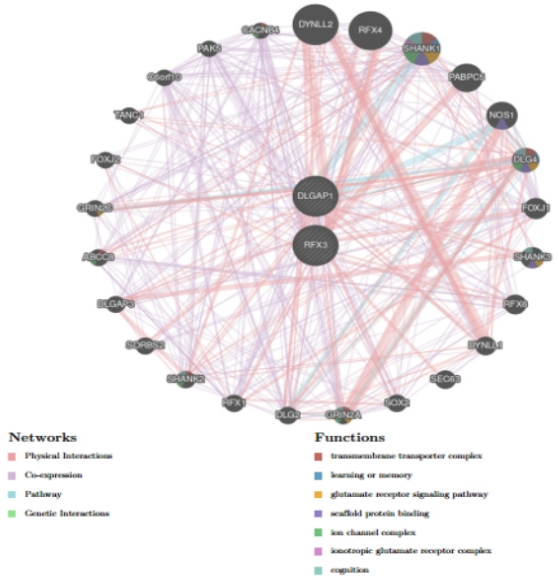


Biological pathways that are enriched by the Network of genes identified from IMPUTE2 based meta-analysis.

Gene	Score
HBB	0.8195460437543921
HBD	0.005592007900498 1305
HBG1	0.003122783792412 0982
HBE1	0.002749178917167 9715
GDAP1	0.002632789644391 7864
HBA1	0.002600638957405 7542
NFE2	0.002456232347866 9046
HBZ	0.002002262230113 6828
HSPA1A	0.001309121458450 0799
ALAS2	0.001296771727658 5189
HBA2	0.001070924104396 7976
SELT	0.001059664490173 9172
PPBP	0.000942828409871 6988
HBM	0.000912788104175 7768
PF4	0.000806715678127 9006
SLC4A1	0.000672979149161 8675
HBG2	0.000660377715671 6228
HBQ1	0.000633491718579 0712
CA1	0.000632942867114 9051
HP	0.000597407345212 4085
AHSP	0.000565780056075 2713
EPB42	0.000513288092714 1703
DACH1	0.000492956114226 2839
S100A8	0.000418701796613 24114

Figure 4.11: *The resulting network from the interaction of the query genes and other genes using biological databases in Genemania*

For ImpG meta-analysis, we obtained a network of 25 related genes with 483 interactions. **Figure** 4.12 shows the resulting network that was generated by the best FDR values and genes together with their scores. *DYNLL2* was identified as gene that had the strongest association with a score of 0.002. The top biological functions which were associated with this network were transmembrane transporter complex, learning or memory, glutamate receptor signaling pathway, scaffold protein binding, ion channel complex and ionotropic glutamate receptor complex both at a P-value of 7.88×10^{-5} .

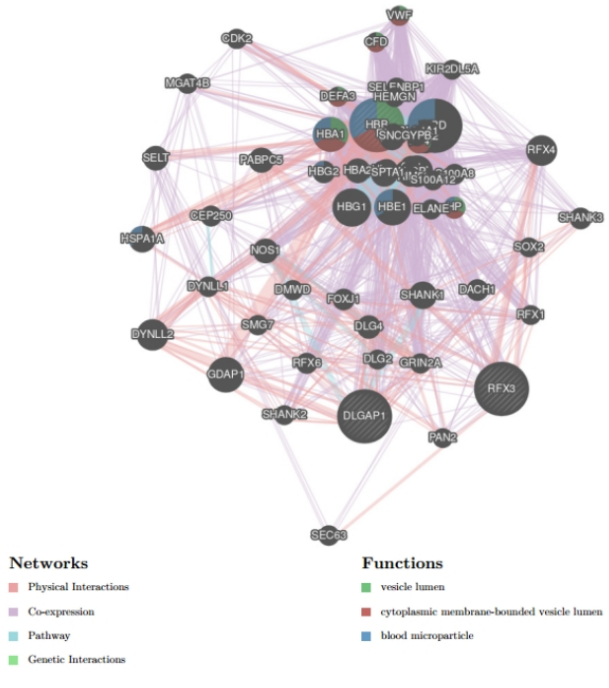


Biological pathways that are enriched by the Network of genes identified from ImpG based meta-analysis.

Gene	Score
DLGAP1	0.8874625592434371
RFX3	0.8630983378429147
DYNLL2	0.0020527209609225716
RFX4	0.0019045157365560894
SHANK1	0.0014733696796449536
PABPC5	0.0011164397023533668
NOS1	0.0010669023048037185
DLG4	0.0008276253726496874
FOXJ1	0.0005762271346020 92
SHANK3	0.0005727774793805 04
RFX6	0.0004701486547664113
DYNLL1	0.0003812042665500 8366
SEC63	0.0003804148904644 2
SOX2	0.0002908606293010596
GRIN2A	0.0002895397318622983
DLG2	0.0002837225317700587
RFX1	0.0002805079550479 1504
SHANK2	0.0002781984146917327
SORBS2	0.0002660115838986066
DLGAP3	0.0002507643845776619
ABCC8	0.0002503727475920 4253
GRIN2C	0.0002366960835217 2578
FOXJ2	0.0002269250500097 0753
TANC1	0.0002182064810945 5165
C6orf10	0.0002075343030192 3004
PAK5	0.0002042195347127 2403
CACNB4	0.0001896157116574 8703

Figure 4.12: *The resulting network from the interaction of the query genes and other genes using biological databases in Genemania*

Finally, using both the genes identified from ImpG and IMPUTE2 meta-analysis as the query genes. We obtained a network of 60 genes with 10,391 interactions. **Figure 4.13** shows the resulting network that was generated by the best FDR values and genes together with their scores. Again, *HBD* was identified as gene that had the strongest association with a score of 0.005. The top biological functions which were associated with this network were cytoplasmic membrane-bounded vesicle lumen at a P-value of $1.15 \times 10-7$, blood microparticle at a P-value of $1.15 \times 10-7$ and blood microparticle at a P-value of 0.0000509.



Biological pathways that are enriched by the Network of genes identified from both ImpG and IMPUTE2 based meta-analysis.

Gene	Score	Gene	Score
DLGAP1	0.884269596269422	*AHSP*	0.0005872008633979187
RFX3	0.8602316244785984	*HP*	0.0005760972640222461
HBB	0.8175013476953622	*FOXJ1*	0.0005619151324336991
HBD	0.05592363538394424 8	*RFX2*	0.0005560232522456299
HBG1	0.0035496189472748463	*SHANK3*	0.0005532351763919974
HBE1	0.0032283845958120394	*EPB42*	0.0005109877202182367
HBA1	0.002851976857052796	*DACH1*	0.0005053625030487785
GDAP1	0.0026135630560678025	*CEP250*	0.0004640874320915866 5
NFE2	0.002438961471842971	*S100A8*	0.0004138656082212089
HBZ	0.0023078931722088808	*MB*	0.0004062507057979725
RFX4	0.0021110224661603505	*DYNLL1*	0.0003772471898326746
DYNLL2	0.0019892023641364664	*CYGB*	0.0003713519458927727 6
SHANK1	0.0014373891170473 83	*GYPA*	0.0003679354082580244
DLGAP3	0.0013766979487909 17	*SMG7*	0.0003599842745741766
HBA2	0.0013725341746637643	*DMWD*	0.0003582962324253924 4
HSPA1A	0.0012959571265615533	*ELANE*	0.0003577830525067771
DLGAP2	0.0012910993219270028	*SNCA*	0.0003535452615257561 6
ALAS2	0.0012805911193074504	*NGB*	0.0003506896730467446
HBM	0.0012236204948908869	*GRIN2A*	0.0003425390560234942
DLGAP4	0.0011309061295977 64	*CDK2*	0.0003415036324481701 7
HBG2	0.0010861243529288744	*SEC63*	0.0003379366922872107
NOS1	0.0010427680810028428	*TRIM58*	0.0003277108723540589 4
SELT	0.0010276384844395 97	*S100A12*	0.0003210720061035399
DLGAP5	0.0010254991325437 39	*RFX8*	0.0003100594403421808
PABPC5	0.0009900720496885 95	*RFX7*	0.0003085977254855820 6
HBQ1	0.0009448528574407256	*MGAT4B*	0.0003075046828663885 5
PPBP	0.0009250422035498862	*DEFA3*	0.0003021734838826728 4
PF4	0.0008152982736326919	*VWF*	0.0003001312710755877
DLG4	0.0008079522435325281	*CFD*	0.0002872993699313819 4
RFX1	0.0007607191983137751		
RFX6	0.0007335894007701738		
SLC4A1	0.0006781926800818461		
CA1	0.0006406350851264619		

Figure 4.13: *The resulting network from the interaction of the query genes and other genes using biological databases in Genemania*

4.3.5 Pathway Enrichment Analysis

Finally to gain the mechanistic insights of the genes that were identified to be functionally related to the genes identified from the meta-analysis, we performed an enrichment analysis of both the networks (from ImpG interaction network, from IMPUTE2 meta-analysis network and both) to ascertain pathway relevance to the disease for each network. To achieve this, we used enrichR to perform pathway enrichment analysis for both the network generated by IMPUTE2 meta-analysis genes, ImpG meta-analysis genes and finally enrichment of network generated by both IMPUTE2 and ImpG meta-analysis genes. Further, we also explored disease/drugs that are associated with each of the network using enrichR.

For IMPUTE2 based meta-analysis network enrichment, African trypanosomiasis was the most enriched pathway with a combined score of 768.55 and a P-value of 1.148×10^{-5}, followed by malaria with a combined score of 536.77 and a P-value of 2.697×10^{-5} using KEGG 2019 Human reference database. Other enriched pathways are listed in **Table 6.3** in the **Appendix**. Moreover, according to OMIM disease database, this network was associated with anemia with a P-value of 7.948×10^{-7}, malaria with a P-value of 0.01667. Other diseases associated with this

network is listed in **Table 4.6**.

Table 4.6: ***Diseases that were identified from OMIM disease database to be associated with the pathway network from IMPUTE2 based meta-analysis.***

Name	P-value	Adjusted P-value	Odds Ratio	Combined score
anemia	7.948×10^{-7}	0.00007154	54.64	767.49
Malaria	0.01667	0.7504	59.52	243.68
charcot-marie-tooth disease	0.03192	0.9576	30.86	106.31
blood	0.04234	0.9527	23.15	73.19

On the other hand, For ImpG based meta-analysis, according to KEGG 2019 human database reference network, Glutamatergic synapse was identified as the top enriched pathway with a P-value of 1.311×10^{-10} and a combined score of 1035.00, followed by Cocaine addiction at P-value of 3.878^{-5} with a combined score of 460.66. Other additional top pathways are listed in **Table 6.4** in the **Appendix** section. The top diseases that were identified to be associated with this network were microphthalmia at a P-value of 0.02535, autism at a P-value of 0.02535. Other diseases associated with this network is listed in **Table 4.7**

Table 4.7: ***Diseases that were identified from OMIM disease database to be associated with the pathway network from ImpG based meta-analysis.***

Name	P-value	Adjusted P-value	Odds Ratio	Combined score
microphthalmia	0.02535	1.000	38.99	143.27
autism	0.02535	1.000	38.99	143.27
diabetes mellitus,type 2	0.04234	1.000	23.15	73.19
epilepsy	0.07547	1.000	12.77	33.00
diabetes mellitus	0.07547	1.000	12.77	33.00
ataxia	0.07797	1.000	12.35	31.50
diabetes	0.09652	1.000	9.88	23.09

Finally, for the enrichment of the network generated by both the genes from IMPUTE2 and ImpG meta-analysis, malaria was identified as the most enriched pathway with a combined score of 284.57 and a P-value of 0.00001703. African trypanosomiasis was the second most enriched

pathway with a score of 217.44 and a p-value of 0.00002144. **Table 6.5** lists all other additional pathways that were enriched. The top diseases that were identified to be associated with this network were malaria at a P-value of 0.000867, anemia at a P-value of 0.00004077. Other diseases associated with this network is listed in **Table 4.8**.

Table 4.8: ***Diseases that were identified from OMIM disease database to be associated with the pathway network from ImpG based meta-analysis.***

Name	P-value	Adjusted P-value	Odds Ratio	Combined score
malaria	0.0008672	0.03902	45.35	319.74
anemia	0.00004077	0.003669	20.82	210.41
blood	0.005742	0.1723	17.64	91.00
dementia	0.03716	0.8361	26.46	87.10
autism	0.05821	1.000	16.71	47.51
parkinson disease	0.06709	1.000	14.43	38.99
charcot marie tooth disease	0.08171	1.000	11.76	29.45

Our study suggests that while it is possible to perform imputation, both at summary statistics level or at raw genotypes level, much can be gained in a given study if imputation is performed in two steps especially for meta-analysis: First at raw genotypes level using the best performing genotypes imputation algorithm, and then at the summary statistics level after performing associations. Our studies show that this increases the chance of identifying association that would have been missed. Interestingly, based on the adjusted P-values, only one disease (anemia with an adjusted P-value of 0.00007154), was identified to be significantly associated with the network generated from IMPUTE2 meta-analysis enrichment and for ImpG based meta-analysis, no disease attained the significance level of 0.05 using the adjusted P-value. However, for the combined network, two diseases were identified to be significantly associated with the network: anemia with adjusted P-value of 0.003669 and malaria with adjusted P-value of 0.03902.

Previously, it has been shown that there is almost 99% correlation between the effect size and P-value of association of the masked SNP and the imputed SNPs with ImpG [84, 109]. This implies that the association values from the imputed summary statistics are as good as the association values that can be obtained when the SNPs are directly genotyped. Thus going by this claim, and as proven from the enrichment analysis, we can confidently report that both imputation via summary statistics and raw genotypes can improve the chance of identifying the

associations that were not typed in the original study. However, it is universally accepted that imputing the raw genotypes via the Hidden Markov Model (HMM) tools is the gold standard [84] for performing imputation. Thus the previous studies have instead recommended the use of summary statistics based imputation to avoid the restrictions that are encountered on accessing the individual level genotypes data [84, 118]. Similarly, we cannot confirm which approach is superior than the other especially when both the summary statistics and raw genotypes data can be accessed. Instead, we recommend implementing both of the approaches if possible in any given study: impute the raw genotypes and then impute the summary statistics after performing the associations.

Chapter 5

Evaluation of Polygenic Risk Scores Methods

5.1 Overview

The objective of epidemiology is to identify and characterize the genetic risk factors that are associated with complex diseases or trait [72]. Knowledge of the risk can then be applied by clinicians in prevention, diagnosis, prognosis and treatment of a particular disease. Genome Wide Association Studies (GWAS) have been successful in the identification of single nucleotide polymorphism (SNPs) and genes that are associated with a wide range of complex diseases [36]. Nevertheless, majority of the variants that have been identified through GWAS to be significantly associated with complex diseases/traits, and have been replicated across studies, explains just a small proportion of variability in the trait of interest, hence limiting their predictive power [17]. Moreover, results from GWAS findings have suggested that much of the genetic basis for most of the complex traits is a combination of small effects of hundreds or thousands of genetic variants [17]. This realization indicates that Single SNP-based analyses alone do not address the overall genomic or polygenic architecture for most of the complex diseases [72]. Thus for complex traits, if the polygenicity is not accounted for, then individuals genome wide prediction or the genetic profile is compromised [17].

Polygenic Risk Score (PRS), or Polygenic Score, is the sum of risk alleles corresponding to a given trait or disease for each individual in the study weighted by the estimates of the effects size from GWAS published result or met-analysis [17]. PRS therefore summarizes the genome wide data into a single variable which measures an individual liability or tendencies to a trait or phenotype of interest [36]. Although there are other methods (like GCTA [73] and LDSCORE

[119])that have been proposed and can summarize the risk of common SNPs that do not achieve stringent GWAS threshold, it is only PRS that summarizes these effects at the individual level [120].

In this chapter, we present a brief overview of Polygenic Risk Score, a review of PRS methods and a brief literature review. We then simulate GWAS datasets and compare the performances of several PRS methods and compare the performances of different PRS methods using a simulated data that mimics African population. We finally apply the best PRS method from the simulation study in GWAS dataset of malaria.

5.2 Polygenic Risk Scores (PRS) Applications and Challenges

PRS for each individual in a given study is calculated by computing the sum of the risk alleles corresponding to a phenotype of interest, and the scores are weighted by the effects estimates from GWAS summary statistics. PRS has a number of applications: It may be used to detect shared genetic etiology among traits [17]. This is achieved by training the score on one trait and testing the score against another trait and if there are associations, then there is shared genetic basis. When there is shared genetic etiology, then this shows that there exist a common molecular etiology between the traits and this may be applied in developing new treatments or can be applied in identifying individuals at risk. Moreover, this knowledge of shared etiology can also be used to pinpoint problems with the current diagnosis or nosology

PRS may also be used in risk prediction and for stratifying individuals into different risk groups. For example, PRS was applied in schizophrenia and demonstrated that individuals whose PRS values were falling within the upper 10 percentiles were shown to be having 10 times chance of developing schizophrenia than the individuals whose PRS were falling with the lower 10 percentile. Similarly, in a recent study of breast cancer, individuals whose PRS were falling within the top percentile had the highest overall lifetime risk of 32.6% of breast cancer [38]. PRS may therefore be applied as a guide in providing intervention or treatment options according to the risk groups.

5.2.1 Challenges in the Calculation of PRS

Besides very many applications, there exists a number of challenges in the calculation of PRS. First, the total number of SNPs that should be included in the calculation of PRS is not known [121]. Usually, a p-value threshold has been applied in the selection of SNPs for inclusion. However,

the optimal P-value for inclusion is generally unknown. Consequently, as in most studies, PRS is calculated at a range of P-values and the p-value that gives the best prediction accuracy (or the p-value that gives the highest correlation or association with the phenotype in the validation dataset) is adapted [17, 122]. Nevertheless, this kind of approach is less useful if the phenotype is not available in the target dataset. Mak et el. suggested a down-weighting of the weights with the SNPs' local discovery rate, in which an individual shrinkage factor is estimated using a data driven approach [121]. This has been shown to give comparative predictive performance with the best predictive P-value. However, this kind of approach is not implemented in most tools.

Another issue in the calculation of PRS is linkage of Disequilibrium (LD). Some PRS methods do not take into account that some SNPs are in LD with each other. This lead to biased estimates of the scores. It is however recommended to prune the SNPs before inclusion in the PRS, as was done with Purcell et al. [123]. However, this kind of approach has a potential of discarding the most predictive SNPs and as a result reducing the prediction power. Wray et al. [124] has since then suggested clumping, which removes SNPs that are in LD based on their p-value of associations. Other methods have recommended inclusion of LD matrix from the reference panel to control LD.

5.3 Review of PRS

5.3.1 Overview

Polygenic risk scores analysis requires two inputs; (1) the GWAS base dataset which is essentially the GWAS summary statistics data of an association between genotypes and a given phenotype and (2) the target dataset, which is just the GWAS data of the target/study sample [17].

Let G denote the genotype matrix for the GWAS dataset. Let the number of markers be M and let the number of individuals in the target population be N. Then the general form for a standard polygenic risk score of an individual i is defined follows;

$$\text{PRS}_{P_T,i} = \sum_{m=1}^{M} G_{im}\hat{\beta}_m \tag{5.1}$$

Where G_{im} is the genotype dosage for individual i at marker/SNP m and β_m is the estimate of the effect size of the m^{th} variant from the discovery/base GWAS summary statistics. Usually, standard PRS methods begins by removing SNPs in LD using a LD prunning procedure and then applies a P-value threshold in selection of SNPs to be included for the calculation of PRS. This

kind of approach, however, can remove more predictive SNPs during the pruning process hence reducing the predictive values. Nonetheless, this was the first approach used in the first study that implemented PRS.

To improve prediction accuracy, several methods have been proposed including, LD clumping as opposed to pruning, Bayesian methods, Penalized regression methods, and including functional annotations. We present a review of these methods in the next section.

5.3.2 Classification of PRS Methods

Methodologies for the construction of the polygenic risk scores differ primarily across two dimensions: 1) how weights used in PRS calculations are generated, and 2) how to determine which SNPs to include in the PRS calculation. Particularly, one of the fundamental problems at the early stage of the development of PRS methods was the fact that there is no inherent information of LD from summary statistics. Thus if SNPs in a given loci are in high LD with one another, and all are included in the score, then definitely the score will be inflated [122]. Consequently, several methods have been suggested to control LD including LD pruning, LD clamping, using Bayesian methods, penalized regression, and including functional annotation. As a consequence, PRS methods may be classified based on how they model the LD. Thus, we can classify PRS methods into two general classes: Bayesian PRS and non Bayesian PRS methods.

Non Bayesian PRS Methods

PRSice

PRSice is the first dedicated PRS software. It was developed in 2014 by Euesden et al. [17] and represent one of the simplest current PRS methods. PRSice uses a similar model as that of **Equation 5.1**, except that it applies a LD clumping and a P-value threshold in the calculation of PRS. In this kind of settings, pairs of SNPs in LD are clumped based on their P-values [17]. For example, for two linked SNPs, one pair of the linked SNPs with less significant value is excluded from the calculation of PRS.

PRSice has the capability of calculating the PRS at various P-value thresholds thus obtaining the best predictive threshold. However, over-fitting is common in case of high resolution PRS and that the P-value from the PRS could be inflated. Thus, the authors suggests that a permutation test should be performed to minimize the over-fitting [17].

PRS uses observed genotypes or imputed posterior probabilities that have been converted

to dosages. However, it relies on PLINK 1.07 for the conversion. Recently, Chen et al. [125] implemented a new PRS software called **PRSoS** that accommodates both genotypes and imputed posterior probabilities. Moreover, PRS on the Spark (PRSoS) does not discard strand ambiguous SNPs as was in the case in PRSice [17]. Nonetheless, both PRSos and PRSice uses a similar model for calculation of PRS. PRSoS uses the following model to calculate PRS from imputed genotypes posterior probabilities.

$$\text{PRS}_{P_T,i} = \sum_{m=1}^{M} \hat{\beta}_m \{2P(AA)_m + P(AB)_m\} \tag{5.2}$$

Where $P(AA)_m$ and $P(AB)_m$ in **Equation 5.2** is the probability of homozygous genotype with two copies of effect allele (AA) at the SNP m and the probability of heterozygous genotype with one copy of the effect allele at SNP m [125].

POlygenic Ld-Adjusted RIsk Score (POLARIS)

POLARIS [126] was published in 2017 and applies a similar approach like that of PRSice in the calculation of PRS. However as opposed to PRSice which accounts for LD via clumping, POLARIS [126] accounts for LD between SNPs via spectral decomposition of the SNP correlation matrix. Vector of the target genotypes are replaced by a vector of adjusted dosages. The PRS of POLARIS is given by the following:

$$\begin{aligned} PRS_i &= \sum_{j=1}^{M} \beta_j \left(\sum_{k=1}^{M} \frac{1}{\lambda_k} x_k x_k^T g \right) \\ &= \sum_{i=1}^{M} \beta_j \left(\sum_{k=1}^{M} \frac{1}{\sqrt{\lambda_k}} x_k(j) \sum_{j=1}^{M} x_k(j) g_j \right) \end{aligned} \tag{5.3}$$

Where x_k are the eigen vector of an $M \times M$ matrix of SNPs and λ_k is the eigen value corresponding to the eigen vectors. To avoid instability, a ridge parameter $\lambda_0 = \frac{1}{N}$ is introduced to **Equation 5.3** to obtain the following.

$$\begin{aligned} PRS_i &= \sum_{j=1}^{M} \beta_j \left(\sum_{k=1}^{M} \sqrt{\frac{1+\lambda_0}{\lambda_k + \lambda_0}} x_k(j) \sum_{j=1}^{M} x_k(j) g_j \right) \\ &= \sum_{j=1}^{M} \beta_j \hat{g}_j \end{aligned} \tag{5.4}$$

P+T funct LASSO

P+T funct LASSO was introduced by shi et al. [127] in 2016. This method was motivated by the fact selection of SNPs using a particular threshold for the inclusion of PRS calculation may affect

the density of the effect size estimate and thus may result into upward bias estimate, an effect known as Winner's curse. To correct for this effect, Shi et al proposed two shrinkage methods: shrinkage via Maximum Likelihood Estimation (MLE) and shrinkage via lasso regularization. Additionally, Shi et al.[127] also introduced a 2D PRS, in which SNPs are partitioned into two groups: S1, which includes SNPs with higher priors, and S2 includes the SNPs with lower priors. The high prior SNPs and the low prior SNPs are obtained from an external annotation datasets. A differential treatment is then applied to both the two groups (S1 and S2) in the calculation of PRS. Thus, PRS under this settings PRS is given as follows:

$$PRS(p_{T_1}, p_{T_2})_i = \sum_{m \in S_1} \hat{\beta}_m I(p_m < p_{T_1}) g_{im} + \sum_{m \in S_2} \hat{\beta}_m I(p_m < p_{T_2}) g_{im} \tag{5.5}$$

Where SNPs in S_1 are selected at less rigorous threshold than SNPs in S_2. This approach has been shown to perform better than the popular 1D PRS in **Equation 5.1** and g_{im} is the genotype dosage for individual i at SNP m.

MLE shrinkage is implemented by assuming that the estimate of the effect sizes of SNP $m \int \{1, \ldots, M\}$ follows a normal distribution $\hat{\beta_m} \sim \mathcal{N}(\beta_m, \hat{sigma}_m)$. A shrinkage estimator that maximizes the likelihood of $P(\hat{\beta_m} | P_m < \alpha)$ is then used as the new estimate of the effect size, $\hat{\beta}_m^{mle}$. However, this approach is highly intensive for large SNPs and at a grid of p-values [127]. Thus, Shi et el. proposes a less intensive method that applies a lasso regularization.

Consider an objective function below,

$$f(\beta) = (y - X\beta)^T (y - X\beta) + 2\lambda \sum_j \beta_j \tag{5.6}$$

Where X is the standardized genotypes and y are standardized phenotype. Then by assuming that SNPs are independent, and let $\hat{\beta}_m = \sum_{i=1}^{N} (y_j - \bar{y}) x_{im}$ (where N are the number of individuals), then

$$\hat{\beta}_m^{lasso} = sign(\hat{\beta}_m) \left| |\hat{\beta_m}| - \lambda \right| I(\hat{\beta_m} > \lambda) \tag{5.7}$$

$$= sign(\hat{\beta}_m) \left| |\hat{\beta_m}| - \lambda \right| I(P_m > \alpha) \tag{5.8}$$

Since $P_m < \alpha$ is equivalent to $\hat{\beta_m} > \lambda$. And PRS calculation follows by substituting $\hat{\beta}_m^{lasso}$ in the equation.

lassosam

Introduced in 2017 by Mak et al. [121]. Lassosam uses a similar lasso regularization as that used by P+T funct lasso [127]. Here however, the main motivation was to not to correct the Winners' curse as was in the case of P+T funct lasso [127] but rather to account for LD using an external datasets. Thus, despite having similar objective function in **Equation 5.6**, here X^TX can be viewed as LD between the SNPs, which can be estimated from the reference panel. On the other hand, XTy can be viewed as SNP-wise association, which can be obtained from the summary statistics. Thus, making it no-longer an regularization problem on the penalty term since the genotypes used in the estimation of LD are different from the genotypes used in the estimation of summary statistics. Mak et al. [121] applies another regularization on **Equation 5.6** to obtain an elastic net problem given by the following.

$$f(\beta) = y^Ty + (1-s)\beta^T X_r^T X_r \beta - 2\beta^T r + 2\lambda(||\beta||_1^1) \tag{5.9}$$

Where $r = X^Ty$, to note that the genotypes used in the construction of $X_r^T X_r$ and X^Ty are not the same. A coordinate descent algorithm is applied on **Equation 5.9** to obtain an estimate of each β_m.

$$\beta^{(t)} = \begin{cases} \frac{sign(v_j^{(t)})\left|v_j^{(t)} - \lambda\right|}{\tilde{X}_j^T \tilde{X}_j + s} & \text{if } |v_j^{(t)}| - \lambda > 0 \\ 0, & \text{Otherwise} \end{cases} \tag{5.10}$$

Where $v_j^{(t)} = r_i - \tilde{X}_i^T \left(\tilde{X}\beta_i^{(t-1)} - \tilde{X}\beta_i^{(t-1)}\right)$

Bayesian PRS Methods

Just like in non Bayesian approaches, the main idea for the use of Bayesian methods was to overcome lower prediction power as a result of not accounting for LD. Bayesian methods applies a prior on the genetic architecture and LD from the reference panel. Among the popular tools under Bayesian method include LDpred, LDpred-inf, AnnoPred, and many more.

LDpred and LDpred-inf

LDpred-inf and LDpred were the first PRS methods using Bayesian methods to be developed in 2015 by vilhjalmsson et al. [122]. LDpred-inf and LDpred applies a Bayesian approach in the calculation of PRS by assuming priors from the genetic architecture and LD information

from the reference panel [122]. The prior has a point normal mixture distribution and has two parameters: (a) Heritability explained by genotypes, which is estimated from GWAS summary statistics, and (b) a fraction of causal markers with non zero effects. This prior allow also for the non infinitesimal genetic architectures.

The posterior mean effect is calculated by conditioning on the genetic architecture and LD, which is approximated from the reference panel or from an independent validation data. For the unlinked markers, an assumption is made that distant markers are not linked and thus the posterior mean effect sizes under a small region pf LD can be estimated as follows:

$$E[\beta_m^l | \hat{\beta}_m^l, D] = \left(\frac{M}{N h_g^2} I + D_m \right)^{-1} \hat{\beta}_m^l \tag{5.11}$$

Where h_g^2 denotes heritability explained by SNPs, M is the number of SNPs in the region l, N is the number of samples in the target sample, D_l is the LD matrix within the region l and D is the LD matrix estimated from the reference panel [122]. This posterior effect estimate can be calculated analytically.

When $D_l = I$, then **Equation 5.11** generalizes to a posterior with infinitesimal priors. Thus this generalization can be used to calculate the posteriors for the SNPs that are linked [122]. Equation one is computed analytically by LDpred-inf [122] and PRS for each individual is obtained by substituting $\hat{\beta_m}$ in **Equation 5.1** by posterior mean effect estimated from **Equation 5.11**.

LDpred is the modification of LDpred-inf [122]. Here, a non infinitesimal Gaussian mixed prior is used to estimate the posterior mean of the causal effect sizes. However, the posterior mean effect size under this settings cannot be derived analytically. Thus, LDpred applies MCMC Gibbs sampling to numerically approximate these estimates. To ensure convergence, a shrink probability (given by $f = \min(1, \frac{\hat{h}_g^2}{(\bar{h_g}^2)_i})$ of being causal is applied at each iteration. Additionally, Rao-Blackwellization is further applied to speed convergence [122].

Annopred [128]

Annopred was developed by Hu et al. [128] in 2017 and is the first Bayesian PRS method to include functional annotations [128].

AnnoPred implements a three stage framework: First, GWAS signals are enriched using external annotation data and are assigned in different categories based on the enrichment. Second, an empirical prior on the SNPs effect size is then estimated based on annotation assignment and

signal enrichment. Finally, the prior together with the marginal summary statistics and the LD matrix are used to infer the posterior effect size for each SNP in a joint modeling framework [128]. Posterior effect size if given by:

$$E_A[\beta_j|\hat{\beta}, \hat{D}], \tag{5.12}$$

where $\hat{D}$ in **Equation 5.12** is the LD matrix estimated from the reference panel or in the target datasets.

Annopred applies LD score regression to partition trait heritability by annotation. Suppose there are K+1 partition labeled as $S_0 \ldots, S_K$, such that the estimated SNP heritability per partition is given as follows $Var(\beta_i) = \sum_{j:i \in S_j} \hat{\tau}_j$, where τ_j is the i^{th} SNP on partition j of K.

AnnoPred implements two priors to ensure flexibility against different architectures. First prior assumes spike and a slab distribution of the effect size given by $\beta \sim p_0 N(0, \frac{\hat{\sigma}_i^2}{p_0}) + (1 - p_0)\delta_o$. This prior assumes that each annotation category has the same number of causal SNPs although the effect sizes are different across the annotation categories. Second prior on the other hand assumes same effect sizes across annotation categories but different proportions of SNPs.

Let T_i denote SNPs with the similar annotation assignment with SNP i, and that M_{T_i} is the total number of such collections. Also define total heritability is given as follows $H_0^2 = p_0 M_0 V$. Then heritability of SNPs within a given collection T_i is given as $H_{T_i}^2 = p_{T_i} M_{T_i} V$, where p_{T_i} is the proportion of causal SNPs in T_i, σ_0 is a Dirac function, V is the variance of the causal $V = Var(\beta_{causal})$, M_0 is the total number of SNPs, and p_0 is overall proportion of SNPs. The prior is thus given by the following;

$$B_i \sim p_{T_i} N(0, V) + (1 - p_T)\delta_o \tag{5.13}$$

Where $V = \frac{H_0}{p_0 N_o}$ and $P_{T_i} = p_0 \frac{M_o H_{T_i}^2}{M_{T_i} H_0^2}$. And the posterior effect size for the two priors is given by,

$$f(\beta_b|\hat{\beta}_b, \hat{D}_b) \equiv N(E[\hat{\beta}_b|\beta_b, \hat{D}_b], var(\hat{\beta}_b|\beta_b, \hat{D}_b) \prod_{i \in b} f(\beta_i) \tag{5.14}$$

$$\begin{cases} N(E[\hat{\beta}_b|\beta_b, \hat{D}_b], var(\hat{\beta}_b|\beta_b, \hat{D}_b) \prod_{i \in b} \left[p_0 N(0, \frac{\sigma_i^2}{p_0}) + (1 - p_0)\delta_0 \right] \\ N(E[\hat{\beta}_b|\beta_b, \hat{D}_b], var(\hat{\beta}_b|\beta_b, \hat{D}_b) \prod_{i \in b} \left[p_{T_i} N(0, \frac{H_0}{p_0 N_0}) + (1 - p_0)\delta_0 \right] \end{cases} \tag{5.15}$$

Deriving $E[\beta_b|\hat{\beta}_b, \hat{D}_b]$ from the join distribution of β is difficult hence drawn from a Gibbs sampling of $f(\beta|\hat{\beta}_b \hat{D}_b)$, using the sampling mean as an approximation of $E[\beta_b|\hat{\beta}_b, \hat{D}_b]$. The posteror effect sizes are then used as the new weight for the calculation of an individual score using the standard PRS in **Equation 5.1**.

LDpred -inf -funct and LDpred -funct.

Both LDpred -funct inf and LDpred-funct are modifiations of LDpred-inf to include the functional information in the calculation of PRS, just like in the case of PRS with winner's curse correction via Maximum Likelihood estimation (mle) or lasso. However, the difference between these methods is that, LD-pred funct inf and LDpred- funct applies a Bayesian approach in the calculation of PRS while PRS with winner's curse correction via mle or lasso applies the frequentist approach. Thus, the functional information are incorporated in the priors under these methods [129].

LDpred inf funct, uses a similar priors to LDpred inf by [122] except that here, the priors are assumed to have the following distribution $\beta \sim \mathcal{N}(0, c * \sigma_m^2)$, where σ_m is the expected per SNP heritability under the baseline LD model that incorporates the functional annotations and c is the normalizing constant and is given by $c = \frac{1}{h_g^2} \sum_{m=1}^{M} I(\sigma_m > 0)$. Thus, the posterior mean effect size under this model is just the same as LDpred-inf except that it is additionally conditioned on the per SNP heritability. Particularly, posterior mean under this model is obtained by solving the following system equation.

$$\left(N \times D + \frac{1}{c} \begin{bmatrix} \frac{1}{\sigma_1^2} & \dots & 0 \\ & & \\ 0 & & \frac{1}{\sigma_1^2} \end{bmatrix} \right) E\left[\beta | \hat{\beta}_i, \sigma_1^2, \dots \sigma_{m+}^2\right] = N\hat{\beta} \tag{5.16}$$

Where m_+ in **Equation 5.16** denotes the number of SNPs whose expected per SNP heritability is greater than zero. Thus, under this model the PRS of an individual in the study is obtained by substituting the $\hat{\beta}_m$ in **Equation 5.1** by the posterior effect size estimates from **Equation 5.16**.

LDpred-funct uses exactly the same model as that of LDpred-funct-inf in **Equation 5.16** except that the posterior mean effect sizes are further regularized via a cross validation. The posterior mean effect size estimates are first partitioned into K bins and a PRS is calculated for each bin. Specific weight of the bins are optimized via cross validation using the validation dataset. The overall PRS under this method is given by the following:

$$PRS_{LDpred_{funct}} = \sum_{k=1}^{K} \alpha_k PRS(k) \tag{5.17}$$

Where α_k is the specific weight for each bin and is optimized via a cross validation [122] and K is the number of bins.

Table 5.1 summarizes some of the popular PRS methods, their models, publication year among others.

Table 5.1: *Summary of some of the popular PRS tools*

SOFTWARE	Model	Accounting for LD	YEAR	COMMENT	Effect Size Data format	Target Format	REF
SOFTWARE	Model	Accounting for LD	YEAR	COMMENT	Effect Size Data format	Target Format	REF
PLINK	Standard PRS	LD Clumping	2009	First tool to be applied in the calculation of PRS	Summary statistics	PLINK FORMAT	[78]
PRSice	Standard PRS	LD Clumping	2014	First Standalone PRS software	Summary statistics	PLINK (bed/bim/fam)	[17]
LDpred & LDpred inf	Bayesian PRS	Calculate LD from the reference Panel	2015	First Bayesian PRS software	Summary Statistics	PLINK (bed/bim/fam)	[122]
Annopred	Bayesian PRS	Calculate LD from the reference Panel	2017	Bayesian PRS	Summary Statistics	PLINK (bed/bim/fam)	[128]
Lassosum	Bayesian PRS	Calculate LD from the reference Panel	2017	Penalized Regression PRS	Summary Statistics	PLINK (bed/bim/fam)	[121]
PRSoS	Standard PRS	Spectral Decomposition	2018	First PRS to implement Spectral Decomposition	Summary statistics	Oxford (.gen/ .sample),Variant Call Format (vcf)	[125]

POLARIS	Standard PRS	Applies Principal Component Analysis (PCA)	2018	First tool to apply PCA to adjust for LD	Summary statistics	PLINK (bed/bim/fam)	[126]
LDpred -inf -funct & LDpred-funct	Bayesian PRS	Calculate LD from the reference Panel	2018	Bayesian PRS	Summary statistics	PLINK (bed/bim/fam)	[129]
Multi-ethnic PRS	Bayesian PRS	Calculate LD from the reference Panel and Target data	2018	Bayesianregression PRS	Summary statistics	PLINK (bed/bim/fam)	[130]
PRS-CS	Bayesian PRS	Calculate LD from the reference Panel	2019	Bayesian regression PRS	Summary statistics	PLINK (bed/bim/fam)	[131]
EBPRS	Bayesian PRS	No LD calculation and No training data	2020	Bayesian PRS	Summary statistics		[132]

5.4 Literature Review: Comparison of PRS Methods

Polygenic risk score methods have been evaluated by a number of studies. In one of the studies Mek et al. [121] compared the performance of their method (lassom, which uses penalized regression approach to estimate the weights used in the PRS) with LDpred [122] (which applies a Bayesian approach in the estimation of the posterior mean of the effect sizes) using a simulated data of seven diseases from Wellcome Truct Case Control Consortium (WTCCC) that had 358,179 SNPs and 12,744 cases and 2,859 controls after quality control . Their results suggest that lassom [121] is faster and more accurate than LDpred. Importantly, this study also compared the effect of using lassosam when P-value threshold is applied (by activating the *p-thres* flag in the tool). Interestingly, lassosam without p-value thresholding and clumping still performed better than LDpred. Thus this approach seems to offer a solution to reduction in prediction power that arises from exclusion of some SNPs based on their P-value [121].

Hu et al. [128] compared the performance of AnnoPred with other four best prediction methods: PRS method based on P-value threshold (PRS_{sig}), PRS based on P-value theshold and Prunning (PRS_{P+T}) and LDpred and PRS_{all} (PRS based on all SNPs). Using both real and simulated data, Hu et al. [128] showed AnnoPred performs better than all the other PRS methods in all settings using both real and simulated data. Interestingly, PRS method based on all markers demonstrated the lowest performance nearly in all the settings thus highlighting the importance for selecting a subset of SNPs in the calculation of PRS. However, although not assessed by this study, lassosam has been shown to give better prediction accuracy by using all the SNPs rather than a just a subset of the SNPs. Additionally, Hu et al. [128] demonstrated that methods estimating the PRS weights based on Bayesian framework (LDpred and AnnoPred) had almost similar performances and outperformed the non Bayesian methods in all settings as illustrated in table 5.2. This study moreover highlighted that including functional annotations in the calculation of PRS improves the performances of PRS method [128].

Table 5.2: *Correlation between PRS and different traits in the real data set [128]. PRS(all) represent PRS computed with all the P-value, PRS(sig) for the SNPs that were identified to be associated with a given trait, PRS(P+T) implies SNPs were first clumped that a P-value threshold was applied in selecting the SNPs. Annopred outperformed all the approaches.*

Disease/Trait	**PRS(sig)**	**PRS(all)**	**PRS(P+T)**	**LDpred**	**AnnoPred**
Crohn's Disease	0.27	0.229	0.32	0.325	**0.343**
Breast Cancer	0.084	0.055	0.12	0.122	**0.137**
Rheumatoid Arthritis	0.204	0.114	0.248	0.282	**0.287**
Type-II Diabetes	0.165	0.156	0.204	0.202	**0.22**
Celiac Disease	0.11	0.136	0.18	0.197	**0.213**

Recently, Marquez et al. [129] compared the performances of five PRS methods P+T, LDpred-inf [122], P+T-funct-LASSO, LDpred-funct-inf [129], and LDpred-funct [129] methods using a simulated dataset from the UK biobank interim release. On average, method tha incorporate functional annotations showed better results than methods that do not include the functional annotation. In particular, incorporating functional annotations led to over 17% improvement in the method and further, regularization led to over 27% improvement compared the most accurate method that does not include functional annotation [129]. In overall, LDpred-funct had the best result and P+T produced the least result. These findings are consistent with the previous findings by [128], that illustrated that the potential of functional enrichment in the calculation of PRS.

Interestingly, comparison and evaluation of the Polygenic Risk Scores methods have mainly been done by their authors, who in most cases compare their new method with the existing methods. Surprisingly, nearly all these new methods, from the reported results seems to be superior than other existing methods. Moreover, tools like lassosam, which seem to perform very well, has never been compared with the new methods like LDpred-funct [129] and Annopred [128]. Here, we will compare the performance of Polygenic Risk methods which have been shown to be superior in most of the studies. So far, previous studies have done the comparison using European populations and no study has assessed the performance of different PRS methods using a diverse population, like African. Thus, here we use a simulated both from Africa to compare the performances of different PRS tools.

5.5 Evaluation of PRS methods using Simulated Data

5.5.1 Study Data and Methodology

We used the simulated dataset of 5,000 samples that mimics African populations. How the dataset was obtained is described in **Section 3.3**. Additionally, we obtained the summary statistics of the case/control datasets datasets from Mugo et al. [101]. How the summary statistics was generated is well described in the publication paper [101].

PRS calculations for all the tools are not similar. Some of the programs are stand alone programs and do not rely on other programs to compute the score whereas others like PRScS can only weight the SNPs and thereafter relies on PLINK in generating the individual risk scores. Unless mentioned, we applied the default program settings in generating the scores.

PRSice can generate the risk scores at a range of P-values in a single run and has a potential of determining the P-value that gives the best predictive risk scores. PRSice does this in a simple way by allowing users to specify the lower interval using the flag $--lower$ and the upper interval using the flag $--upper$ and an interval for each iteration using the flag $--interval$. We retained only the best predictive PRS, which was generated at P-value of 5.0×10^{-5}. Moreover, as recommended by the developers, all the SNPs that were in LD were clumped prior to performing the PRS. For fair and uniform comparison of the tools, we restricted the calculation of PRS for each of the tools using only the best predictive P-value that was obtained from PRSice.

PRS calculation with LDpred involves three steps. First, the program synchronizes the study genotypes with the summary statistics in which an HDF5 file is generated. This HDF5 file contains the synchronized genotypes and reduces the processing time for the subsequent runs. The program then re-weights the SNPs effects sizes using either a Gibbs sampling approach or can also apply P-value by first clumping SNPs depending on what the user has activated. Our study, however, mainly focused on the comparison of PRS methods that implement clumping and P-value threshold hence we did not consider gibbs sampling algorithm of the tool. Finally, the program generates risk scores of the individuals at various P-value thresholds. We restricted the calculation of the scores using one standard P-value of $5.0e^{-5}$ that was predicted by PRSice as the most predictive threshold value.

For PLINK, we followed a similar approach as that of PRSice. However, provide an option for specifying a range file which guides the program on what P-value range to generate the scores. This setting generates scores for an individual using all the P-values within the range. To circumnavigate this, we developed an in-house python script that selects all the SNPs from

the summary statistics that satisfy only the P-value of 5.0×10^{-5} as estimated from PRSice. We then applied this new summary statistics file in computing the scores. We applied a similar approach for the PRSoS and PRScS. However, PRScS requires input from a reference panel which can be downloaded from the programs website. Moreover, PRScS cannot generate PRS on it owns but rather relies on PLINK in generating the scores. We thus applied PRScS in weighting the summary statistics and thereafter applied PLINK in generating the scores.

To compare the performances of the tools, we evaluated the potential of each tool in predicting the phenotype of each individual in the study. To do this, we implement a logistics regression model in R and regress phenotypes of individuals against their risk scores. We compare how much variance in the phenotypes is explained by the PRS generated from each tool. Finally, we fit Receiver Operator Curve (ROC), which has a potential of assessing the specificity and sensitivity of each tool in prediction. Such an assessment can help in determining how the model is able to distinguish between cases and controls, given the PRS scores. Moreover, we use Area Under Curve (AUC) to quantify how well the PRS scores can be used in predicting an individual phenotypes.

5.5.2 Results and Discussion

Table 5.3 shows the predicted r^2 for different tools when the scores are applied in predicting the phenotypes for each individual in the study. PRSoS gave a slightly better prediction performance followed by PRSice. However, the other tools gave very low prediction with PLINK giving the lowest prediction. To measure the accuracy of the scores in predicting the phenotypes, we calculated AUC (Area Under Curve) for different tools as represented in column three of **Table 5.3**. The highest AUC was obtained from PRSoS, followed by PRSice. However, in overall, the AUC values were very low and depicted only $\approx$ 50% to 52% that the PRS scores can distinguish cases from controls. Again, LDpred(P+T) and PRScs gave the most uncertain predictive accuracies.

Table 5.3: ***Percentage variation of the phenotypes explained by the PRS across differnet PRS tools.***

Tool	percentage Rsq	AUC
LDpred(P+T)	6.710072^{-4}	0.5007
PRSoS	0.01736	0.511
PRScS	1.27118×10^{3}	0.5002
PLINK	8.1990×10^{-7}	0.5018
PRSice	0.01398	0.5091

The plots displayed by ROC in **Figure 5.1** further shows how the model has very low predictive accuracy in predicting the phenotype.

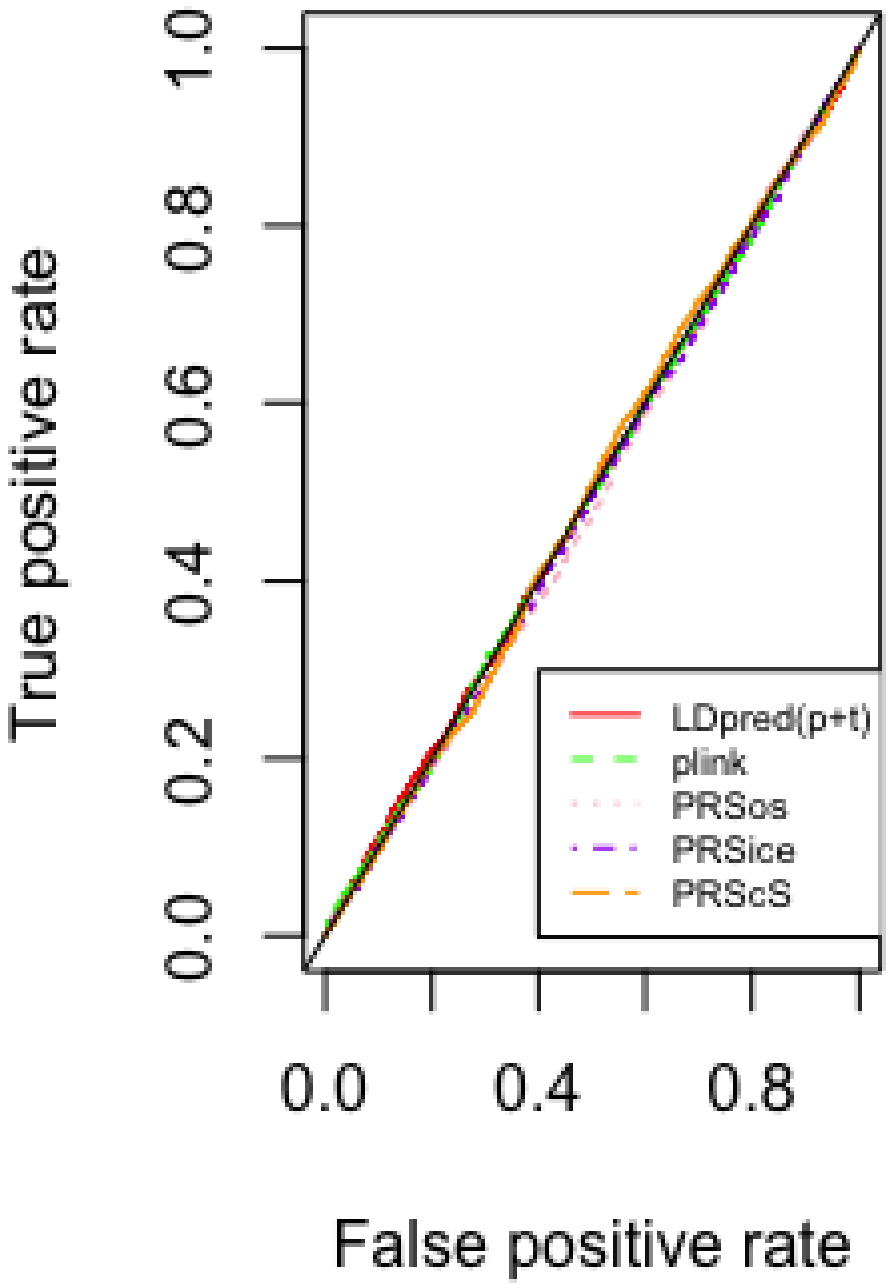


Figure 5.1: *Plot of sensitivity and specificity of different PRS tools. LDpred(p+t) represents the result from the LDpred that implements clumping and P-value threshold when calculating the scores.*

To put our result into global context, Marquez et al. [129] previously compared the performance of five (i.e P+T, P+T funct lasso, LDpred-funct-inf and LDpred-pred-funct) PRS methods using both simulated and real datasets. Their results shows that the methods that incorporates functional annotation into the PRS calculation performs better than the standard PRS methods. In other similar studies, Hu et al also demonstrated that method that incorporates functional annotations in the calculation of PRS outperforms the general P+T methods [128]. However, these studies only evaluated basic P+T methods and the result may not reflect the overall performance of all the P+T methods. By the fact that these studies obtained higher prediction power than our current study, one reason could be due to the fact that their study settings and our study settings are completely different. European populations are well characterized with high level of LD than African populations, and as such, obtaining a higher prediction accuracy is not surprising. Also, their study had larger sample size than our current study which automatically boosts the performance of the PRS methods [16].

In other studies, Ge et al. [131] compared the performance of their newly developed method (PRS-CS) with LDpred, P+T, LDpred-inf, PRS unadjusted and PRS-CS auto. PRS-CS had the best overall performance in all the comparison criteria. This again highlights the fact that, although we did not apply the functional PRS methods in our comparison, our conclusion may not change since the other methods obtained almost similar performance with the PRS-CS, which has been shown to outperform the functional PRS methods.

Our study shows that the PRS methods have almost similar prediction accuracy when similar approaches are applied when generating the scores. However, the prediction accuracy of PRS in Africa populations is still very low and thus necessitates the development of newer algorithms that may adapt well to all populations, including the Africa populations which bears the greatest burden of most diseases.

5.6 Application of PRS in Malaria

One of the aim of genetics studies is to understand how the genetic variation contributes to the phenotypic variations. GWAS has been a tool of choice for more than a decade, and a number of variants that are associated with complex traits have been identified and replicated. However, the puzzle that remains for geneticists is that even the strongest genetic associations explain only a small fraction of the genetic variance and that the genome-wide significant hits explains only a small variation of the phenotypic variance. In malaria for example, less than half of the genetic variation is explained by the erythrocyte associated polymorphism, that for a long a long time

have been hypothesized as the major disorders for malaria susceptibility [47].

Recently, it has been shown that many causal loci whose effect sizes do not reach GWAS significance threshold explains much of the heritability [15]. In Schizophrenia for example, it was shown that common variants explain much of the expected heritability than the GWAS significant hits [78]. The same has been reported in height and subsequent studies, that have together highlighted the importance of analyzing the small effect SNPs that fail to attain the genome-wide significant threshold across a wide variety of traits [15, 78]. These findings suggests that complex traits are driven by the accumulation of weak effects that together drive the disease susceptibility/protection. In this section, we explore how the non genome-wide significant variants can predict malaria susceptibility. We use some of the malaria GWAS published data from African population to estimate the polygenic risk scores using the best predictive polygenic risk score identified from the previous chapter.

5.6.1 Materials and Methods

We apply the study data and the summary statistics data described in **Section 4.2.1** in calculating the PRS for each of the individual in the study sample from Kenyan population. We chose Kenya because we obtained the highest number of SNPs from the imputation of summary statistics by ImpG. The summary statistics had a total of 19,973,364, from which we obtained a total of 7,666,662 common SNPs across the target data and the base data. We then proceeded with the calculation of PRS by first clumping all the SNPs that were in LD to avoid obtaining an inflated scores. We applied a clumping P-value of 1.000 and $r^2 = 0.1$ thereby only retaining 381,364 SNPs.

Finally, we calculated PRS at various P-value thresholds using PRSoS, which we identified in the previous **Section 5.2** as the best performing PRS method under P+T PRS methods. Since PRSoS cannot directly determine the best predicting threshold, we applied PRSice in estimating the best predictive threshold, which we subsequently applied in generating the score for all the individuals using PRSoS. The scores were then regressed in a logistic regression with the phenotypes to determine the relationship between the scores and the phenotypes, and how much variation on the phenotypes can be explained by the scores.

5.6.2 Results and Discussion

We identified P-value of 0.0246714 as the best predictive threshold value for the generation of PRS from the Kenyan sample datasets. There were only 1,043 SNPs at this threshold, which we used in generating the scores.

To determine the relationship between the scores and the phenotypes, we performed a logistic regression of the phenotypes and the scores in R. **Figure 5.2** shows the results of the logistic regression.

```
Call:
glm(formula = PHENO2 ~ SCORE, family = binomial(link = "logit"), data = malaria)

Deviance Residuals:
Min      1Q  Median      3Q     Max
-1.357  -1.224   1.069   1.129   1.223
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.869      1.176   2.441   0.0147 *
SCORE         -5.143      2.198  -2.340   0.0193 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4303.0  on 3111  degrees of freedom
Residual deviance: 4297.5  on 3110  degrees of freedom
AIC: 4301.5

Number of Fisher Scoring iterations: 3
```

Figure 5.2: *Logistic regression of severe malaria with the individuals Polygenic Risk Scores.*

At significant level of $\alpha = 0.05$, severe malaria was found be to significantly associated with the scores. The logs of having severe malaria malaria was 2.869, with a unit change in score was associated with -5.143 change in the odds of having severe malaria. We further calculated the proportion of variance in the phenotypes that can explained by the scores. For this, we used r^2, which can be calculated by **Equation 5.18** from the logistic regression output.

$$\hat{R^2} = 1 - \frac{\text{Residual Deviance}}{\text{Null deviance}} \tag{5.18}$$

The scores could only explained 1.28% of the phenotype as determined by the equation.

We then compared the phenotypic variance explained by the GWAS SNPs and that explained by the PRS scores. For this, we used GCTA to estimate the SNP heritability using the Restricted Maximum Likelihood (*reml*) approach. We obtained $h^2_{snp} = 24.65$ of which the GWAS significant SNPs could explain up-to 10.03% of this variation. By the fact the variation explained by the GWAS significant SNPs was ten folds higher than the variation explained by the PRS, our study therefore confirms the low predictive power of PRS methods in non European populations as has been previously observed [133].

Previously, PRS has been shown to perform best for highly heritable traits like height [15]. For

example, in Attention Deficit/Hyperactivity Disorder (ADHD) for example, which is known to be highly heritable ranging from 70-80% (estimated from twin studies), PRS from the largest GWAS sample (55,374 samples) attained Nagelkerke's R^2 of 5.5% [134]. Similarly, in Autism Spectrum Disorder (ASD) which has an estimated heritability of 50%, PRS attained a Nagelkerke R^2 of 2.5% [135]. Thus, for malaria, in which the host genes are estimated to accounts for approximately 25% of the outcome [13] further justifies why we obtained lower PRS scores.

In conclusion, we emphasize the need for better PRS methods that can be generalizable to all populations and to all traits.

Chapter 6

Conclusion and Future Work

In conclusion, we evaluated the performance of five imputation methods (BEAGLE4, IMPUTE2, IMPUTE4, minimac3 and minimac4) in imputing simulated genotypes datasets that mimics three populations: African, European and admixed populations. We have shown that based on imputation accuracy, IMPUTE2 should be given more preference when imputing genetic data from African and admixed populations. However, for the European samples, the imputation tools gave comparative performance and generally recorded higher accuracy than the African and admixed samples across all the datasets. Interestingly, we observed that newer versions of the imputation program recorded higher imputation accuracy that their older versions for the European datasets. However, for the African and the admixed samples the tools recorded lower imputation accuracy than their older versions. Our study thus confirms that the current improvement in imputation programs are no longer focusing on modeling the high genetic diversity and lower LD of populations like the African but are rather focusing on reducing the computational complexity and handling large reference panels and samples [60].

We have also shown that more power in association studies can be achieved when imputation is performed both at raw genotypes level and summary statistics level. As a consequence, we recommend that if the individual level genotypes data are available, we suggest performing imputation on the raw genotypes and subsequently on the summary statistics level to maximize the gains and improve the chance of detecting the associations. Importantly, we have demonstrated that both imputation via summary statistics and imputation via raw genotypes can be applied to any population data and have different potentials.

Finally, we evaluated the performance of five PRS (LDpred(p+t), PLINK, PRSoS, PRSice, PRScS) methods on African populations. We have shown that of the five P+T methods, PRSoS gives the best prediction accuracy than the other five methods. We have also demonstrated that

PRS can be applied in predicting the risk of an infectious disease like severe malaria. However, the prediction rate is very low and may fail distinguish the cases from the controls.

Our future work will be building on the current imputation methods and exploit ways on how the low LD level and high diversity can be modeled to improve imputation performance. We will also package all the scripts used in this research in form of a software that can help researchers select any given imputation tool of their choice and perform imputation on their studies with just a click.

———————

References

[1] Charles CJ Carpenter, Greg W Pearson, Violaine S Mitchell, Stanley C Oaks Jr, et al. *Malaria: obstacles and opportunities*. National Academies Press, 1991.

[2] Malaria. CDC, Centers for Disease Control and Prevention , https://www.cdc.gov/malaria/about/biology/index.html, Accessed July 2018.

[3] F Verra, VD Mangano, and D Modiano. Genetics of susceptibility to plasmodium falciparum: from classical malaria resistance genes towards genome-wide association studies. *Parasite immunology*, 31(5):234–253, 2009.

[4] Karen Hayton and Xin-zhuan Su. Drug resistance and genetic mapping in plasmodium falciparum. *Current genetics*, 54(5):223–239, 2008.

[5] Philip J Rosenthal. *Antimalarial chemotherapy: mechanisms of action, resistance, and new directions in drug discovery*. Springer Science & Business Media, 2001.

[6] Zenglei Wang, Mynthia Cabrera, Jingyun Yang, Lili Yuan, Bhavna Gupta, Xiaoying Liang, Karen Kemirembe, Sony Shrestha, Awtum Brashear, Xiaolian Li, et al. Genome-wide association analysis identifies genetic loci associated with resistance to multiple antimalarials in plasmodium falciparum from china-myanmar border. *Scientific reports*, 6, 2016.

[7] Umberto DâĂŹAlessandro. Malaria elimination: Challenges and opportunities. In *Towards Malaria Elimination-A Leap Forward*. IntechOpen, 2018.

[8] Balbir Singh, Lee Kim Sung, Asmad Matusop, Anand Radhakrishnan, Sunita SG Shamsul, Janet Cox-Singh, Alan Thomas, and David J Conway. A large focus of naturally acquired plasmodium knowlesi infections in human beings. *The Lancet*, 363(9414):1017–1024, 2004.

[9] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[10] Muminatou Jallow, Yik Ying Teo, Kerrin S Small, Kirk A Rockett, Panos Deloukas, Taane G Clark, Katja Kivinen, Kalifa A Bojang, David J Conway, Margaret Pinder, et al. Genome-wide and fine-resolution association analysis of malaria in west africa. *Nature genetics*, 41(6):657, 2009.

[11] Margaret J Mackinnon, Tabitha W Mwangi, Robert W Snow, Kevin Marsh, and Thomas N Williams. Heritability of malaria in africa. *PLoS medicine*, 2(12):e340, 2005.

[12] Alphaxard Manjurano, Nuno Sepúlveda, Behzad Nadjm, George Mtove, Hannah Wangai, Caroline Maxwell, Raimos Olomi, Hugh Reyburn, Christopher J Drakeley, Eleanor M Riley, et al. Usp38, frem3, sdc1, ddc, and loc727982 gene polymorphisms and differential susceptibility to severe malaria in tanzania. *The Journal of infectious diseases*, 212(7):1129–1139, 2015.

[13] Dominic P Kwiatkowski. How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics*, 77(2):171–192, 2005.

[14] Kirk A Rockett, Geraldine M Clarke, Kathryn Fitzpatrick, Christina Hubbart, Anna E Jeffreys, Kate Rowlands, Rachel Craik, Muminatou Jallow, David J Conway, Kalifa A Bojang, et al. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature genetics*, 46(11):1197–1204, 2014.

[15] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[16] Frank Dudbridge. Polygenic epidemiology. *Genetic epidemiology*, 40(4):268–272, 2016.

[17] Jack Euesden, Cathryn M Lewis, and Paul F OâĂŹreilly. Prsice: polygenic risk score software. *Bioinformatics*, 31(9):1466–1468, 2014.

[18] Timothy Mak, Robert Milan Porsch, Shing Wan Choi, and Pak Chung Sham. Polygenic scores without external summary statistics. *bioRxiv*, page 252270, 2018.

[19] Stephan Ripke, Alan R Sanders, Kenneth S Kendler, Douglas F Levinson, Pamela Sklar, Peter A Holmans, Dan-Yu Lin, Jubao Duan, Roel A Ophoff, Ole A Andreassen, et al.

Genome-wide association study identifies five new schizophrenia loci. *Nature genetics*, 43(10):969, 2011.

[20] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499, 2010.

[21] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[22] Daniel L Hartl, Andrew G Clark, and Andrew G Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[23] Janis E Wigginton, David J Cutler, and Gonçalo R Abecasis. A note on exact tests of hardy-weinberg equilibrium. *The American Journal of Human Genetics*, 76(5):887–893, 2005.

[24] Ping Zeng, Yang Zhao, Cheng Qian, Liwei Zhang, Ruyang Zhang, Jianwei Gou, Jin Liu, Liya Liu, and Feng Chen. Statistical analysis for genome-wide association study. *Journal of biomedical research*, 29(4):285, 2015.

[25] David J Balding. A tutorial on statistical methods for population association studies. *Nature reviews genetics*, 7(10):781, 2006.

[26] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[27] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[28] Leading Edge. Are genome-wide association studies of infection any value?

[29] Chris CA Spencer, Zhan Su, Peter Donnelly, and Jonathan Marchini. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5):e1000477, 2009.

[30] Gavin Band, Quang Si Le, Luke Jostins, Matti Pirinen, Katja Kivinen, Muminatou Jallow, Fatoumatta Sisay-Joof, Kalifa Bojang, Margaret Pinder, Giorgio Sirugo, et al. Imputation-based meta-analysis of severe malaria in three african populations. *PLoS genetics*, 9(5):e1003509, 2013.

[31] Yan Du, Jiaxin Xie, Wenjun Chang, Yifang Han, and Guangwen Cao. Genome-wide association studies: inherent limitations and future challenges. *Frontiers of medicine*, 6(4):444–450, 2012.

[32] Yi Wang, Yi Li, Chunhong Qiao, Xiaoyu Liu, Meng Hao, Yin Yao Shugart, Momiao Xiong, and Li Jin. Nuclear norm clustering: a promising alternative method for clustering tasks. *Scientific reports*, 8(1):10873, 2018.

[33] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446, 2010.

[34] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, 2009.

[35] Johanna Jakobsdottir, Michael B Gorin, Yvette P Conley, Robert E Ferrell, and Daniel E Weeks. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS genetics*, 5(2):e1000337, 2009.

[36] Cathryn M Lewis and Evangelos Vassos. Prospects for using risk scores in polygenic medicine. *Genome medicine*, 9(1):96, 2017.

[37] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, 381(9875):1371–1379, 2013.

[38] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K Bolla, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34, 2019.

[39] Hellen Gelband, Claire B Panosian, Kenneth J Arrow, et al. *Saving lives, buying time: economics of malaria drugs in an age of resistance*. National Academies Press, 2004.

[40] Robert W Snow, Carlos A Guerra, Abdisalan M Noor, Hla Y Myint, and Simon I Hay. The global distribution of clinical episodes of plasmodium falciparum malaria. *Nature*, 434(7030):214, 2005.

[41] Franklin A Neva. Looking back for a view of the future: observations on immunity to induced malaria. *The American journal of tropical medicine and hygiene*, 26(6_Part_2):211–215, 1977.

[42] Gabriel Otieno, Joseph K Koske, and John M Mutiso. Transmission dynamics and optimal control of malaria in kenya. *Discrete Dynamics in Nature and Society*, 2016, 2016.

[43] Laurence Florens, Michael P Washburn, J Dale Raine, Robert M Anthony, Munira Grainger, J David Haynes, J Kathleen Moch, Nemone Muster, John B Sacci, David L Tabb, et al. A proteomic view of the plasmodium falciparum life cycle. *Nature*, 419(6906):520, 2002.

[44] DJ Weatherall. Thalassaemia and malaria, revisited. *Annals of Tropical Medicine & Parasitology*, 91(7):885–890, 1997.

[45] Melanie J Newport and Chris Finan. Genome-wide association studies and susceptibility to infectious diseases. *Briefings in functional genomics*, 10(2):98–107, 2011.

[46] Christian Timmann, Thorsten Thye, Maren Vens, Jennifer Evans, Jürgen May, Christa Ehmen, Jürgen Sievertsen, Birgit Muntau, Gerd Ruge, Wibke Loag, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, 489(7416):443–446, 2012.

[47] Matt Ravenhall, Susana Campino, Nuno Sepúlveda, Alphaxard Manjurano, Behzad Nadjm, George Mtove, Hannah Wangai, Caroline Maxwell, Raimos Olomi, Hugh Reyburn, et al. Novel genetic polymorphisms associated with severe malaria and under selective pressure in north-eastern tanzania. *PLoS genetics*, 14(1):e1007172, 2018.

[48] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387–406, 2009.

[49] Joshua T Burdick, Wei-Min Chen, Gonçalo R Abecasis, and Vivian G Cheung. In silico method for inferring genotypes in pedigrees. *Nature genetics*, 38(9):1002–1004, 2006.

[50] ME Hawley and KK Kidd. Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86(5):409–411, 1995.

[51] Sharon R Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, 124(5):439–450, 2008.

[52] Eleftheria Zeggini, Laura J Scott, Richa Saxena, Benjamin F Voight, Jonathan L Marchini, Tianle Hu, Paul IW de Bakker, Gonçalo R Abecasis, Peter Almgren, Gitte Andersen,

et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638, 2008.

[53] Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc, and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 31(12):i206–i213, 2015.

[54] Sayantan Das, Gonçalo R Abecasis, and Brian L Browning. Genotype imputation from large reference panels. *Annual review of genomics and human genetics*, 19:73–96, 2018.

[55] Elisabeth M Van Leeuwen, Alexandros Kanterakis, Patrick Deelen, Mathijs V Kattenberg, P Eline Slagboom, Paul IW De Bakker, Cisca Wijmenga, Morris A Swertz, Dorret I Boomsma, Cornelia M Van Duijn, et al. Population-specific genotype imputations using minimac or impute2. *Nature protocols*, 10(9):1285–1296, 2015.

[56] Cristen J Willer, Serena Sanna, Anne U Jackson, Angelo Scuteri, Lori L Bonnycastle, Robert Clarke, Simon C Heath, Nicholas J Timpson, Samer S Najjar, Heather M Stringham, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics*, 40(2):161, 2008.

[57] Li Li, Yun Li, Sharon R Browning, Brian L Browning, Andrew J Slater, Xiangyang Kong, Jennifer L Aponte, Vincent E Mooser, Stephanie L Chissoe, John C Whittaker, et al. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PloS one*, 6(9):e24945, 2011.

[58] Bryan Howie, Jonathan Marchini, and Matthew Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6):457–470, 2011.

[59] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–959, 2012.

[60] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[61] World Health Organization et al. World malaria day 2017: key messages. 2017.

[62] Peter B Bloland, World Health Organization, et al. Drug resistance in malaria. 2001.

[63] Robert E Black, Saul S Morris, and Jennifer Bryce. Where and why are 10 million children dying every year? *The lancet*, 361(9376):2226–2234, 2003.

[64] Christine M Cserti and Walter H Dzik. The abo blood group system and plasmodium falciparum malaria. *Blood*, 110(7):2250–2258, 2007.

[65] Yik-Ying Teo, Kerrin S Small, and Dominic P Kwiatkowski. Methodological challenges of genome-wide association analysis in africa. *Nature Reviews Genetics*, 11(2):149–160, 2010.

[66] Elinor K Karlsson, Dominic P Kwiatkowski, and Pardis C Sabeti. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, 15(6):379, 2014.

[67] Philip W Hedrick. Population genetics of malaria resistance in humans. *Heredity*, 107(4):283–304, 2011.

[68] George W Comstock. Tuberculosis in twins: A re-analysis of the prophit survey 1–3. *American Review of Respiratory Disease*, 117(4):621–624, 1978.

[69] Emile R Chimusa, Mamana Mbiyavanga, Gaston K Mazandu, and Nicola J Mulder. ancgwas: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. *Bioinformatics*, 32(4):549–556, 2015.

[70] Donald F Conrad, Mattias Jakobsson, Graham Coop, Xiaoquan Wen, Jeffrey D Wall, Noah A Rosenberg, and Jonathan K Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics*, 38(11):1251, 2006.

[71] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

[72] Frank Dudbridge, Nora Pashayan, and Jian Yang. Predictive accuracy of combined genetic and environmental risk scores. *Genetic epidemiology*, 42(1):4–19, 2018.

[73] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[74] Gavin Band, Quang Si Le, Geraldine M Clarke, Katja Kivinen, Christina Hubbart, Anna E Jeffreys, Kate Rowlands, Ellen M Leffler, Muminatou Jallow, David J Conway, et al. New

insights into malaria susceptibility from the genomes of 17,000 individuals from africa, asia, and oceania. *BioRxiv*, page 535898, 2019.

[75] Yu-Fang Pei, Jian Li, Lei Zhang, Christopher J Papasian, and Hong-Wen Deng. Analyses and comparison of accuracy of different genotype imputation methods. *PloS one*, 3(10):e3551, 2008.

[76] Zhaoxia Yu and Daniel J Schaid. Methods to impute missing genotypes for population data. *Human genetics*, 122(5):495–504, 2007.

[77] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[78] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[79] DY Lin, Y Hu, and BE Huang. Simple and efficient analysis of disease association with missing genotype data. *The American Journal of Human Genetics*, 82(2):444–452, 2008.

[80] Dale R Nyholt, Chang-En Yu, and Peter M Visscher. On jim watson's apoe status: genetic information is hard to hide. *European Journal of Human Genetics*, 17(2):147, 2009.

[81] Frank Dudbridge. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Human heredity*, 66(2):87–98, 2008.

[82] Richard Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

[83] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284, 2016.

[84] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

[85] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[86] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. Genome-wide genetic data on˜ 500,000 uk biobank participants. *BioRxiv*, page 166298, 2017.

[87] Christian Fuchsberger, Gonçalo R Abecasis, and David A Hinds. minimac2: faster genotype imputation. *Bioinformatics*, 31(5):782–784, 2015.

[88] Eric Yi Liu, Mingyao Li, Wei Wang, and Yun Li. Mach-admix: genotype imputation for admixed populations. *Genetic epidemiology*, 37(1):25–37, 2013.

[89] Brian L Browning and Sharon R Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223, 2009.

[90] Sayantan Das. Next generation of genotype imputation methods. 2017.

[91] Brian L Browning and Sharon R Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[92] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[93] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[94] Shelina Ramnarine, Juan Zhang, Li-Shiun Chen, Robert Culverhouse, Weimin Duan, Dana B Hancock, Sarah M Hartz, Eric O Johnson, Emily Olfson, Tae-Hwi Schwantes-An, et al. When does choice of accuracy measure alter imputation accuracy assessments? *PloS one*, 10(10):e0137601, 2015.

[95] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2016.

[96] Lucy Huang, Mattias Jakobsson, Trevor J Pemberton, Muntaser Ibrahim, Thomas Nyambo, Sabah Omar, Jonathan K Pritchard, Sarah A Tishkoff, and Noah A Rosenberg. Haplotype variation and genotype imputation in african populations. *Genetic epidemiology*, 35(8):766–780, 2011.

[97] Shona Dalal, Juan Jose Beunza, Jimmy Volmink, Clement Adebamowo, Francis Bajunirwe, Marina Njelekela, Dariush Mozaffarian, Wafaie Fawzi, Walter Willett, Hans-Olov Adami, et al. Non-communicable diseases in sub-saharan africa: what we know now. *International journal of epidemiology*, 40(4):885–901, 2011.

[98] Qian Liu, Elizabeth T Cirulli, Yujun Han, Song Yao, Song Liu, and Qianqian Zhu. Systematic assessment of imputation performance using the 1000 genomes reference panels. *Briefings in bioinformatics*, 16(4):549–562, 2015.

[99] Candelaria Vergara, Margaret M Parker, Liliana Franco, Michael H Cho, Ana V Valencia-Duarte, Terri H Beaty, and Priya Duggal. Genotype imputation performance of three reference panels using african ancestry individuals. *Human genetics*, 137(4):281–292, 2018.

[100] Michael Nothnagel, David Ellinghaus, Stefan Schreiber, Michael Krawczak, and Andre Franke. A comprehensive evaluation of snp genotype imputation. *Human genetics*, 125(2):163–171, 2009.

[101] Jacquiline W Mugo, Ephifania Geza, Joel Defo, Samar SM Elsheikh, Gaston K Mazandu, Nicola J Mulder, and Emile R Chimusa. A multi-scenario genome-wide medical population genetics simulation framework. *Bioinformatics*, 33(19):2995–3002, 2017.

[102] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[103] Haiko Schurz, Stephanie Julia Pitts, Paul David Van Helden, Gerard Tromp, Eileen Garner Hoal, Craig Kinnear, and Marlo Möller. Evaluating the accuracy of imputation methods in a five-way admixed population. *Frontiers in Genetics*, 10:34, 2019.

[104] Lucy Huang, Yun Li, Andrew B Singleton, John A Hardy, Gonçalo Abecasis, Noah A Rosenberg, and Paul Scheet. Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, 84(2):235–250, 2009.

[105] Hou-Feng Zheng, Jing-Jing Rong, Ming Liu, Fang Han, Xing-Wei Zhang, J Brent Richards,

and Li Wang. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One*, 10(1):e0116487, 2015.

[106] Dana B Hancock, Joshua L Levy, Nathan C Gaddis, Laura J Bierut, Nancy L Saccone, Grier P Page, and Eric O Johnson. Assessment of genotype imputation performance using 1000 genomes in african american studies. *PLoS One*, 7(11):e50610, 2012.

[107] Donghyung Lee, T Bernard Bigdeli, Brien P Riley, Ayman H Fanous, and Silviu-Alin Bacanu. Dist: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927, 2013.

[108] Zheng Xu, Qing Duan, Song Yan, Wei Chen, Mingyao Li, Ethan Lange, and Yun Li. Dissco: direct imputation of summary statistics allowing covariates. *Bioinformatics*, 31(15):2434–2442, 2015.

[109] Donghyung Lee, T Bernard Bigdeli, Vernell S Williamson, Vladimir I Vladimirov, Brien P Riley, Ayman H Fanous, and Silviu-Alin Bacanu. Distmix: direct imputation of summary statistics for unmeasured snps from mixed ethnicity cohorts. *Bioinformatics*, 31(19):3099–3104, 2015.

[110] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):7, 2015.

[111] Vidhya Gopalakrishnan, Parthiban Purushothaman, and Anusha Bhaskar. Proteomic analysis of plasma proteins in diabetic retinopathy patients by two dimensional electrophoresis and maldi-tof-ms. *Journal of Diabetes and its Complications*, 29(7):928–936, 2015.

[112] Julie Makani, Stephan Menzel, Siana Nkya, Sharon E Cox, Emma Drasar, Deogratius Soka, Albert N Komba, Josephine Mgaya, Helen Rooks, Nisha Vasavda, et al. Genetics of fetal hemoglobin in tanzanian and british patients with sickle cell anemia. *Blood*, 117(4):1390–1392, 2011.

[113] Buhm Han and Eleazar Eskin. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, 88(5):586–598, 2011.

[114] Buhm Han and Eleazar Eskin. Interpreting meta-analyses of genome-wide association studies. *PLoS genetics*, 8(3):e1002555, 2012.

[115] Patrick Emery, Bénédicte Durand, Bernard Mach, and Walter Reith. Rfx proteins, a novel family of dna binding proteins conserved in the eukaryotic kingdom. *Nucleic acids research*, 24(5):803–807, 1996.

[116] Jun-Ming Li, Chao-Lin Lu, Min-Chih Cheng, Sy-Ueng Luu, Shih-Hsin Hsu, and Chia-Hsiang Chen. Genetic analysis of the dlgap1 gene as a candidate gene for schizophrenia. *Psychiatry research*, 205(1-2):13–17, 2013.

[117] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220, 2010.

[118] Geraldine M Clarke, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121, 2011.

[119] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[120] Shing Wan Choi, Timothy Shin Heng Mak, and Paul O'reilly. A guide to performing polygenic risk score analyses. *BioRxiv*, page 416545, 2018.

[121] Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480, 2017.

[122] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592, 2015.

[123] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'donovan, Patrick F Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

[124] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507, 2013.

[125] Lawrence M Chen, Nelson Yao, Elika Garg, Yuecai Zhu, Thao TT Nguyen, Irina Pokhvisneva, Shantala A Hari Dass, Eva Unternaehrer, Hélène Gaudreau, Marie Forest, et al. Prs-on-spark (prsos): a novel, efficient and flexible approach for generating polygenic risk scores. *BMC bioinformatics*, 19(1):295, 2018.

[126] Emily Baker, Karl Michael Schmidt, Rebecca Sims, Michael C O'Donovan, Julie Williams, Peter Holmans, Valentina Escott-Price, and with the GERAD Consortium. Polaris: Polygenic ld-adjusted risk score approach for set-based analysis of gwas data. *Genetic epidemiology*, 42(4):366–377, 2018.

[127] Jianxin Shi, Ju-Hyun Park, Jubao Duan, Sonja T Berndt, Winton Moy, Kai Yu, Lei Song, William Wheeler, Xing Hua, Debra Silverman, et al. Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS genetics*, 12(12):e1006493, 2016.

[128] Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, and Hongyu Zhao. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS computational biology*, 13(6):e1005589, 2017.

[129] Carla Marquez-Luna, Steven Gazal, Po-Ru Loh, Nicholas Furlotte, Adam Auton, Alkes L Price, 23andMe Research Team, et al. Modeling functional enrichment improves polygenic prediction accuracy in uk biobank and 23andme data sets. *bioRxiv*, page 375337, 2018.

[130] Carla Márquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Alkes L Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, 41(8):811–823, 2017.

[131] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1776, 2019.

[132] Shuang Song, Wei Jiang, Lin Hou, and Hongyu Zhao. Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLOS Computational Biology*, 16(2):e1007565, 2020.

[133] L Duncan, H Shen, B Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and B Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications*, 10(1):1–9, 2019.

[134] Ditte Demontis, Raymond K Walters, Joanna Martin, Manuel Mattheisen, Thomas D Als, Esben Agerbo, Gísli Baldursson, Rich Belliveau, Jonas Bybjerg-Grauholm, Marie Bækvad-Hansen, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, 51(1):63, 2019.

[135] Daniel J Weiner, Emilie M Wigdor, Stephan Ripke, Raymond K Walters, Jack A Kosmicki, Jakob Grove, Kaitlin E Samocha, Jacqueline I Goldstein, Aysu Okbay, Jonas Bybjerg-Grauholm, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature genetics*, 49(7):978, 2017.

[136] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2):117, 2017.

[137] Ellen M Leffler, Gavin Band, George BJ Busby, Katja Kivinen, Quang Si Le, Geraldine M Clarke, Kalifa A Bojang, David J Conway, Muminatou Jallow, Fatoumatta Sisay-Joof, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science*, 356(6343):eaam6393, 2017.

[138] Jonathan K Pritchard, Joseph K Pickrell, and Graham Coop. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology*, 20(4):R208–R215, 2010.

[139] Noah A Rosenberg, Michael D Edge, Jonathan K Pritchard, and Marcus W Feldman. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, medicine, and public health*, 2019(1):26–34, 2018.

[140] NR Wray, SH Lee, D Mehta, AA Vinkhuyzen, F Dudbridge, and CM Middeldorp. Polygenic methods and their application to psychiatric disorders and related traits. *J Child Psychol Psychiatry*, 55:1068–1087, 2014.

[141] The Malaria Genomic Epidemiology Network. A global network for investigating the genomic epidemiology of malaria. *Nature*, 456(7223):732, 2008.

[142] Deepti Gurdasani, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, Louise Iles, Martin O Pollard, Ananyo Choudhury, et al. The african genome variation project shapes medical genetics in africa. *Nature*, 517(7534):327, 2015.

[143] Rasika Ann Mathias, Margaret A Taub, Christopher R Gignoux, Wenqing Fu, Shaila Musharoff, Timothy D O'Connor, Candelaria Vergara, Dara G Torgerson, Maria Pino-Yanes, Suyash S Shringarpure, et al. A continuum of admixture in the western hemisphere revealed by the african diaspora genome. *Nature communications*, 7:12522, 2016.

[144] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75, 2015.

[145] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348, 2013.

[146] Jennifer L Asimit and Eleftheria Zeggini. Imputation of rare variants in next-generation association studies. *Human heredity*, 74(3-4):196–204, 2012.

[147] Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature*, 526(7572):253, 2015.

[148] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual review of genomics and human genetics.*, 10:387–406, 2009.

[149] Lynn B Jorde et al. Genetic variation and human evolution, 2003.

[150] William Amos and John Harwood. Factors affecting levels of genetic diversity in natural populations. *PHILOSOPHICAL TRANSACTIONS-ROYAL SOCIETY OF LONDON SERIES B BIOLOGICAL SCIENCES*, 353:177–186, 1998.

[151] Sharon R Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, 124(5):439–450, 2008.

[152] Jennifer L Asimit and Eleftheria Zeggini. Imputation of rare variants in next-generation association studies. *Human heredity*, 74(3-4):196–204, 2013.

[153] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 2009.

[154] Donghyung Lee, T Bernard Bigdeli, Brien P Riley, Ayman H Fanous, and Silviu-Alin Bacanu. Dist: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927, 2013.

[155] Tdentical by descent. wikiibd, https://isogg.org/wiki/Identical_by_descent, Accessed October 2018.

[156] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[157] Daniel Shriner, Adebowale Adeyemo, Guanjie Chen, and Charles N Rotimi. Practical considerations for imputation of untyped markers in admixed populations. *Genetic epidemiology*, 34(3):258–265, 2010.

[158] Jared O'Connell, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu, Sheila Ulivi, Massimiliano Cocca, Michela Traglia, Jie Huang, Jennifer E Huffman, Igor Rudan, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, 10(4):e1004234, 2014.

[159] Brian Desany, Xinghua Shi, Xiaodong Fang, Philip Awadalla, Sarah Lindsay, Simon Gravel, Lisa D Brooks, Stephen T Sherry, Chris Hartl, Phil Lacroute, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 2010.

[160] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443, 2016.

[161] Alan Adolphson, Steven Sperber, and Marvin Tretkoff, editors. *p-adic Methods in Number Theory and Algebraic Geometry*. Number 133 in Contemporary Mathematics. American Mathematical Society, Providence, RI, 1992.

[162] Mario Mitt, Mart Kals, Kalle Pärn, Stacey B Gabriel, Eric S Lander, Aarno Palotie, Samuli Ripatti, Andrew P Morris, Andres Metspalu, Tõnu Esko, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage wgs-based imputation reference panel. *European Journal of Human Genetics*, 25(7):869, 2017.

[163] Peng Lin, Sarah M Hartz, Zhehao Zhang, Scott F Saccone, Jia Wang, Jay A Tischfield,

Howard J Edenberg, John R Kramer, Alison M Goate, Laura J Bierut, et al. A new statistic to evaluate imputation reliability. *PloS one*, 5(3):e9697, 2010.

[164] Julian B Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna MM Howson, Adam Auton, Simon Myers, Andrew Morris, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294, 2012.

[165] Sarah L Spain and Jeffrey C Barrett. Strategies for fine-mapping complex traits. *Human molecular genetics*, 24(R1):R111–R119, 2015.

[166] Jason P Wendler, John Okombo, Roberto Amato, Olivo Miotto, Steven M Kiara, Leah Mwai, Lewa Pole, John O'Brien, Magnus Manske, Dan Alcock, et al. A genome wide association study of plasmodium falciparum susceptibility to 22 antimalarial drugs in kenya. *PloS one*, 9(5):e96486, 2014.

[167] Wanqing Wen, Xiao-ou Shu, Xingyi Guo, Qiuyin Cai, Jirong Long, Manjeet K Bolla, Kyriaki Michailidou, Joe Dennis, Qin Wang, Yu-Tang Gao, et al. Prediction of breast cancer risk based on common genetic variants in women of east asian ancestry. *Breast Cancer Research*, 18(1):124, 2016.

[168] Theresa Wimberley, Christiane Gasse, Sandra Melanie Meier, Esben Agerbo, James H MacCabe, and Henriette Thisted Horsdal. Polygenic risk score for schizophrenia and treatment-resistant schizophrenia. *Schizophrenia bulletin*, 43(5):1064–1069, 2017.

[169] International Schizophrenia Consortium et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748, 2009.

[170] Lisa C Ranford-Cartwright and Jonathan M Mwangi. Analysis of malaria parasite phenotypes using experimental genetic crosses of plasmodium falciparum. *International journal for parasitology*, 42(6):529–534, 2012.

[171] Hiasindh Ashmi Antony and Subhash Chandra Parija. Antimalarial drug resistance: An overview. *Tropical parasitology*, 6(1):30, 2016.

[172] Jacques Le Bras and Rémy Durand. The mechanisms of resistance to antimalarial drugs in plasmodium falciparum. *Fundamental & clinical pharmacology*, 17(2):147–153, 2003.

[173] Christian Timmann, Jennifer A Evans, Inke R König, André Kleensang, Franz Rüschendorf, Julia Lenzen, Jürgen Sievertsen, Christian Becker, Yeetey Enuameh, Kingsley Osei Kwakye,

et al. Genome-wide linkage analysis of malaria infection intensity and mild disease. *PLoS genetics*, 3(3):e48, 2007.

[174] Ronald L Nagel and EF Jr Roth. Malaria and red cell genetic defects. *Blood*, 74(4):1213–1221, 1989.

[175] Frank B Livingstone. Malaria and human polymorphisms. *Annual review of genetics*, 5(1):33–64, 1971.

[176] Sunil Parikh and Philip J Rosenthal. Human genetics and malaria: relevance for the design of clinical trials, 2008.

[177] P Sabeti, S Usen, S Farhadian, M Jallow, T Doherty, M Newport, M Pinder, R Ward, and D Kwiatkowski. Cd40l association with protection from severe malaria. *Genes and immunity*, 3(5):286–291, 2002.

[178] Eric Akum Achidi, Tsiri Agbenyega, Stephen Allen, Olukemi Amodu, Kalifa Bojang, David Conway, Patrick Corran, Panos Deloukas, Abdoulaye Djimde, Amagana Dolo, et al. A global network for investigating the genomic epidemiology of malaria. *Nature*, 456(7223):732–737, 2008.

[179] Louis H Miller, Steven J Mason, David F Clyde, and Mary H McGinniss. The resistance factor to plasmodium vivax in blacks: the duffy-blood-group genotype, fyfy. *New England Journal of Medicine*, 295(6):302–304, 1976.

[180] A Moreno, I Caro-Aguilar, SS Yazdani, AR Shakri, S Lapp, E Strobert, H McClure, CE Chitnis, and MR Galinski. Preclinical assessment of the receptor-binding domain of plasmodium vivax duffy-binding protein as a vaccine candidate in rhesus macaques. *Vaccine*, 26(34):4338–4344, 2008.

[181] Robert W Snow, Carlos A Guerra, Juliette J Mutheu, and Simon I Hay. International funding for malaria control in relation to populations at risk of stable plasmodium falciparum transmission. *PLoS medicine*, 5(7):e142, 2008.

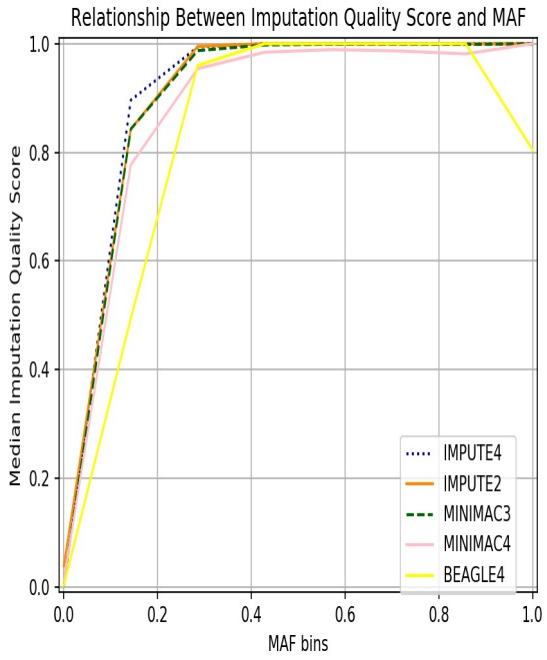
[182] Jane Achan, Ambrose O Talisuna, Annette Erhart, Adoke Yeka, James K Tibenderana, Frederick N Baliraine, Philip J Rosenthal, and Umberto D'Alessandro. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malaria journal*, 10(1):144, 2011.

[183] Najia K Ghanchi, Andreas Mårtensson, Johan Ursing, Sana Jafri, Sándor Bereczky, Rabia Hussain, and Mohammad A Beg. Genetic diversity among plasmodium falciparum field isolates in pakistan measured with pcr genotyping of the merozoite surface protein 1 and 2. *Malaria journal*, 9(1):1, 2010.

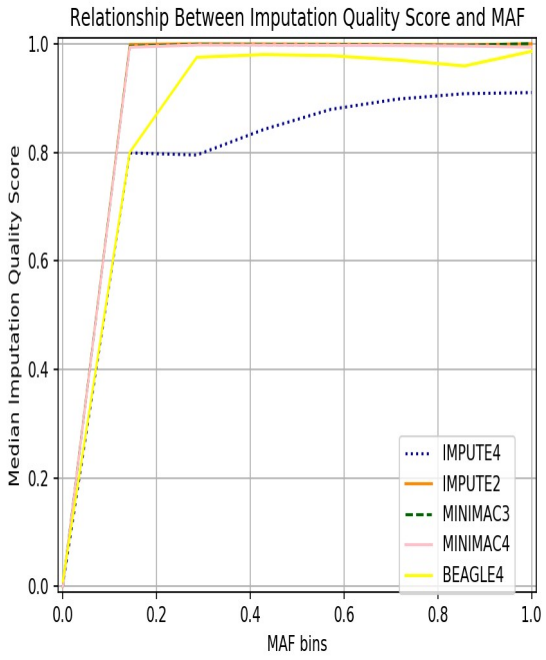
[184] Malaria Genomic Epidemiology Network, Kirk A Rockett, Geraldine M Clarke, Kathryn Fitzpatrick, Christina Hubbart, Anna E Jeffreys, Kate Rowlands, Rachel Craik, Muminatou Jallow, David J Conway, et al. Reappraisal of known malaria resistance loci in a large multicenter study. *Nature genetics*, 46(11):1197, 2014.

Appendix

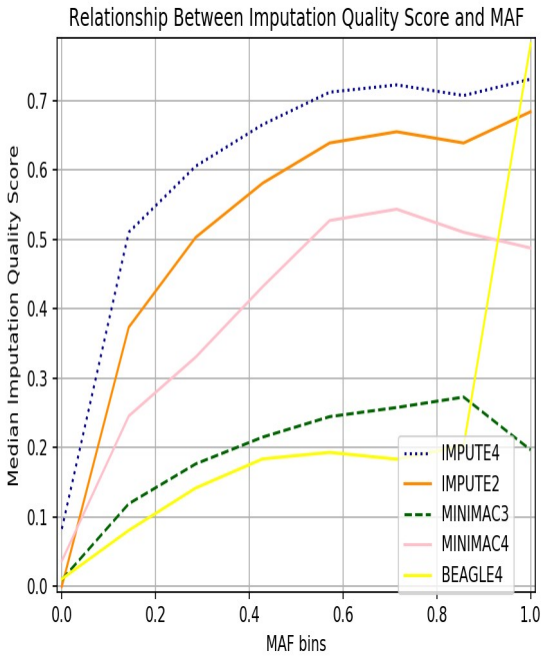
Relationship Between Minor Allele Frequency and Imputation Accuracy



(a)



(b)



(c)

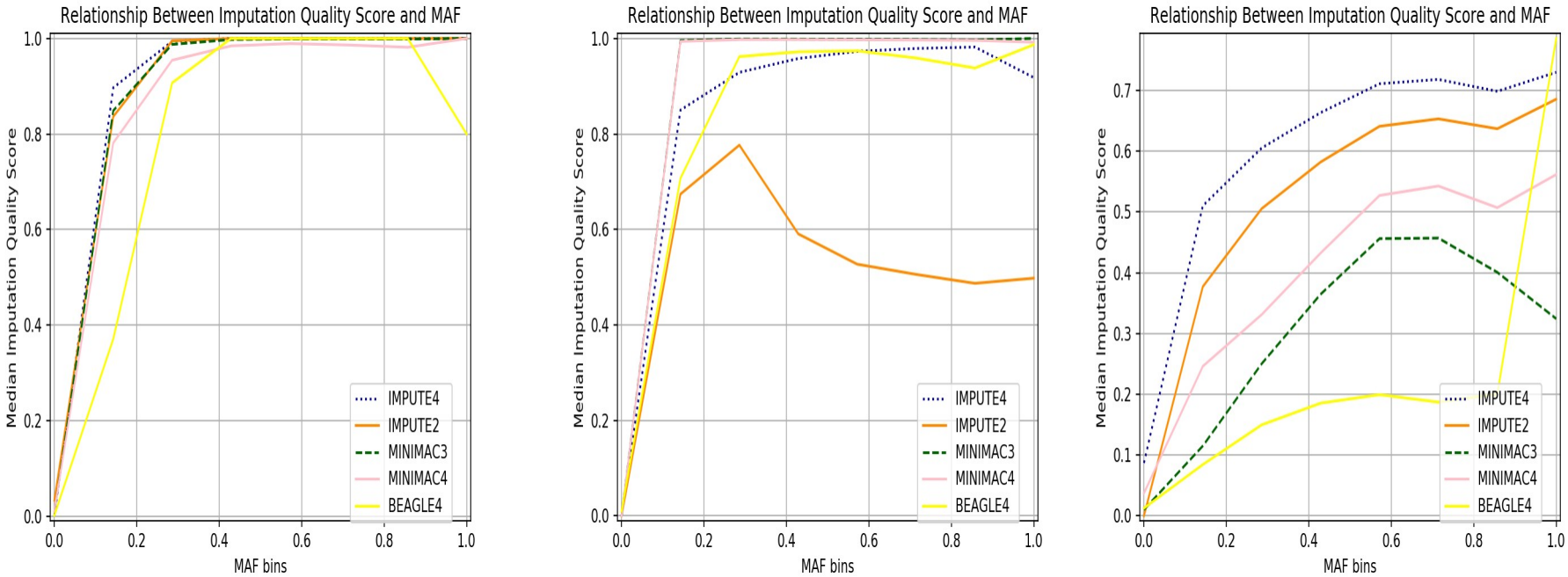


Figure 6.2: *Relationship between minor allele frequency and the imputation accuracy at different minor allele frequency bins for 5,000 samples from African population (a), European population (b) and admixed population (c).*

Meta-analysis

Table 6.1: *SNPs with P-value less than* 5.0^{-6} *from the meta-analysis of the GWAS data that was imputed with IMPUTE2 prior to performing association. We considered the SNPs whose either has a Fixed effect P-value (P-FE) or random effect P-Value (P-RE) or binary effect P-Value (B-PE) or both less than* 5.0^{-6}

SNP	BP	GENE	P-FE	P-RE	P-BE
rs11809587	175128444	*TNN*	$4.68588\times^{06}$	$4.68588\times^{06}$	$1.60938\times^{05}$
rs1116396	138518857	*THSD7B*	0.000250746	0.0619837	1.62011×10^{-06}
rs36023051	182180385	*LINC00290*	0.000845379	0.515529	4.53619×10^{-06}
rs13179872	128149876	*FBN2*	0.000394406	0.124655	1.85935×10^{-06}
rs9505102	7272006	*RREB1*	$4.96513\times^{06}$	$5.32487\times^{06}$	$2.42561\times^{05}$
rs9505103	7272069	*RREB1*	$4.72385\times^{06}$	$6.57653\times^{06}$	$2.32246\times^{05}$
rs17439102	78217826	*MAGI2*	0.0273898	0.467282	3.22411×10^{-06}
rs112908647	22411637	*NEBL*	0.000487464	0.154957	1.57679×10^{-06}
rs61744500	131306493	*MKI67*	2.61506×10^{-06}	0.00812510	2.21348×10^{-06}
rs61859848	131308685	*MKI67*	3.02920×10^{-06}	0.0106009	2.25074×10^{-06}
rs7098519	131309256	*MKI67*	3.03723×10^{-06}	0.0106344	2.22169×10^{-06}
rs76175170	131310661	*MKI67*	2.27526×10^{-06}	0.00720084	2.06792×10^{-06}
rs3837432	205757	*SCGB1C1*	$7.62690\times^{08}$	$7.62690\times^{08}$	$5.04733\times^{07}$
rs143022364	4617306	*OR52K2*	0.00217830	0.457040	7.76036×10^{-08}
rs16906303	4830219	*OR52R1*	$2.63212\times^{06}$	$2.63212\times^{06}$	$9.11006\times^{06}$
rs141935051	4849164	*OR52R1*	0.00752670	0.169865	6.45535×10^{-07}

Continued on next page

Table 6.1 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs4290259	5241282	*OR51V1*	5.37584×10^{-05}	0.0871173	3.20555×10^{-08}
rs12295158	5252794	*HBB*	$2.87576\times^{12}$	$2.87576\times^{12}$	$9.63597\times^{12}$
rs146674260	5271354	*HBG1*	0.000130571	0.273425	9.54751×10^{-07}
rs1188152	56423445	*KTN1*	$4.60589\times^{06}$	$6.19290\times^{06}$	$1.55828\times^{05}$
rs9652671	83365188	*CDH13*	4.54358×10^{-06}	0.00189990	5.25585×10^{-06}
rs4426340	83365327	*CDH13*	4.54358×10^{-06}	0.00189990	5.25429×10^{-06}
rs4782771	83366384	*CDH13*	4.99713×10^{-06}	0.00198876	9.14488×10^{-06}
rs1364298	83367122	*CDH13*	2.68637×10^{-06}	0.000313878	8.74813×10^{-06}
rs9953918	56005645	*NEDD4L*	1.10599×10^{-05}	0.0141790	3.94872×10^{-06}
rs4891690	66156630	*LOC643542*	4.73859×10^{-06}	0.000418924	1.32482×10^{-05}
rs11467740	59815685	*LOC284757*	$2.59144\times^{06}$	$2.59144\times^{06}$	$1.58367\times^{05}$

Table 6.2: *SNPs with P-value less than* 5.0^{-6} *from the meta-analysis of the summary statistics that was imputed by ImpG. We considered the SNPs whose either the fixed effect P-value (P-FE) or the random effect P-Value (P-RE) or the binary effect P-Value (B-PE) or both was less than* 5.0^{-6}

SNP	BP	GENE	P-FE	P-RE	P-BE
rs7524250	7395704	*CAMTA1*	0.000317038	0.0525284	5.93874×10^{-07}
rs61106622	41967195	*SCMH1*	1.99278×10^{-06}	1.99278×10^{-06}	1.21338×10^{-05}
rs12750340	75469124	*C1orf173*	3.37512×10^{-07}	3.37512×10^{-07}	1.51650×10^{-06}
rs74630680	77549114	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.92387×10^{-06}
rs113847753	77553124	*ST6GALNAC5*	1.15613×10^{-06}	1.15613×10^{-06}	6.38819×10^{-06}
rs77124392	77558424	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.90840×10^{-06}
rs112995621	77559806	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.94666×10^{-06}
rs74741768	77570705	*ST6GALNAC5*	3.00578×10^{-06}	3.00578×10^{-06}	1.96573×10^{-05}
rs77213925	77579836	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.93160×10^{-06}
rs6661634	77583455	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.96447×10^{-06}
rs111682848	77588467	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.90401×10^{-06}
rs75214944	77589671	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.92406×10^{-06}
rs6686876	77631258	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.93024×10^{-06}
rs145316166	77643355	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.92971×10^{-06}
rs115743261	77645114	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.92288×10^{-06}
rs149115101	77648752	*ST6GALNAC5*	3.82529×10^{-07}	3.82529×10^{-07}	1.93955×10^{-06}
rs112279760	77669142	*ST6GALNAC5*	1.21620×10^{-06}	1.21620×10^{-06}	4.56487×10^{-06}

Continued on next page

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs6695084	77670921	*ST6GALNAC5*	1.15613×10^{-06}	1.15613×10^{-06}	6.31856×10^{-06}
rs77195732	77676413	*PIGK*	3.82529×10^{-07}	3.82529×10^{-07}	1.96801×10^{-06}
rs75874098	77682109	*PIGK*	2.45495×10^{-06}	2.45495×10^{-06}	1.25046×10^{-05}
rs76763532	77692380	*PIGK*	8.78525×10^{-07}	8.78525×10^{-07}	5.09951×10^{-06}
rs116840202	107904682	*NTNG1*	4.90251×10^{-07}	4.90251×10^{-07}	2.83702×10^{-06}
rs79377552	107909266	*NTNG1*	2.16055×10^{-06}	2.16055×10^{-06}	1.09458×10^{-05}
rs116436654	228895380	*RHOU*	2.34258×10^{-06}	2.30439×10^{-05}	6.50083×10^{-06}
rs6675755	230643364	*PGBD5*	4.66609×10^{-06}	4.66609×10^{-06}	2.45171×10^{-05}
rs74142924	230652839	*PGBD5*	4.33178×10^{-06}	0.00629534	3.06147×10^{-06}
rs10175893	34253742	*LTBP1*	0.00540118	0.0845111	1.02977×10^{-07}
rs13392908	34258290	*LTBP1*	0.581972	0.157430	2.80662×10^{-08}
rs76864510	107735018	*ST6GAL2*	4.64868×10^{-06}	4.64868×10^{-06}	2.81567×10^{-05}
rs75002532	107735966	*ST6GAL2*	4.64868×10^{-06}	4.64868×10^{-06}	2.81798×10^{-05}
rs13392296	115626894	*DPP10*	1.65407×10^{-07}	1.65407×10^{-07}	1.15118×10^{-06}
rs10190446	115631727	*DPP10*	2.87663×10^{-07}	2.87663×10^{-07}	1.92089×10^{-06}
rs76429767	197041698	*DNAH7*	2.81824×10^{-06}	0.00405477	2.30113×10^{-06}
rs73084588	215679061	*BARD1*	1.90077×10^{-07}	7.88713×10^{-06}	5.92873×10^{-07}
rs60577152	5079600	*ITPR1*	1.78471×10^{-08}	1.78471×10^{-08}	1.21524×10^{-07}
rs73128543	7036405	*GRM7*	2.27040×10^{-06}	2.27040×10^{-06}	1.42071×10^{-05}
rs6801701	38868808	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	9.19238×10^{-06}
Continued on next page					

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs6790184	38869118	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	9.11948×10^{-06}
rs7644278	38873347	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	9.07330×10^{-06}
rs113867650	38876770	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	9.22085×10^{-06}
rs75748123	38881110	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	9.08303×10^{-06}
rs77023151	38885483	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	9.01370×10^{-06}
rs78731117	38886084	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	8.98716×10^{-06}
rs61752574	38888524	*SCN10A*	2.38864×10^{-06}	2.38864×10^{-06}	8.87832×10^{-06}
rs7629694	38889984	*SCN11A*	2.38864×10^{-06}	2.38864×10^{-06}	8.94973×10^{-06}
rs74732633	38905642	*SCN11A*	2.38864×10^{-06}	2.38864×10^{-06}	8.98191×10^{-06}
rs116432495	73943963	*PROK2*	1.29507×10^{-07}	1.29507×10^{-07}	8.42123×10^{-07}
rs4677367	73944476	*PROK2*	3.05474×10^{-06}	3.05474×10^{-06}	1.57590×10^{-05}
rs9836886	73944903	*PROK2*	1.29507×10^{-07}	1.29507×10^{-07}	8.46182×10^{-07}
rs73848574	96923828	*EPHA6*	4.85690×10^{-06}	5.63185×10^{-05}	1.93300×10^{-05}
rs116320429	152846312	*MBNL1*	1.00582×10^{-06}	0.000153015	2.04835×10^{-06}
rs76855250	161389107	*NMD3*	4.32569×10^{-06}	4.32569×10^{-06}	2.58286×10^{-05}
rs12108147	161389695	*NMD3*	4.32569×10^{-06}	4.32569×10^{-06}	2.59110×10^{-05}
rs16857277	183018012	*MCF2L2*	2.71795×10^{-06}	7.65299×10^{-05}	8.91227×10^{-06}
rs113612631	188752605	*LPP*	5.05151×10^{-08}	1.14659×10^{-05}	1.95541×10^{-07}
rs73048763	188764868	*LPP*	1.88569×10^{-06}	3.32246×10^{-05}	8.91315×10^{-06}
rs141591558	188766723	*LPP*	1.78999×10^{-06}	0.000206797	7.52236×10^{-06}

Continued on next page

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs59432893	6905143	*PPP2R2C*	1.08491×10^{-06}	1.55504×10^{-06}	4.44545×10^{-06}
rs181709025	120611523	*FAM170A*	5.45164×10^{-07}	5.45164×10^{-07}	2.84954×10^{-06}
rs111660480	125947116	*ALDH7A1*	2.30888×10^{-06}	2.30888×10^{-06}	1.47310×10^{-05}
rs72798658	133892880	*PHF15*	3.05338×10^{-06}	0.00175681	4.36424×10^{-06}
rs6859169	133898564	*PHF15*	3.05338×10^{-06}	0.00175681	4.57340×10^{-06}
rs73790073	137205147	*KLHL3*	2.86578×10^{-06}	2.86578×10^{-06}	1.83164×10^{-05}
rs79703803	159944078	*C1QTNF2*	1.80329×10^{-06}	0.000146461	2.80630×10^{-06}
rs80290517	159946431	*C1QTNF2*	2.81613×10^{-07}	2.81613×10^{-07}	1.40945×10^{-06}
rs78464555	159946738	*C1QTNF2*	1.80329×10^{-06}	0.000146461	2.78683×10^{-06}
rs111352897	159946930	*C1QTNF2*	1.80329×10^{-06}	0.000146461	2.84610×10^{-06}
rs11953703	159947961	*C1QTNF2*	5.85374×10^{-06}	0.00422000	2.01250×10^{-06}
rs74857376	1731779	*GMDS*	2.67359×10^{-08}	2.67359×10^{-08}	1.93380×10^{-07}
rs114760297	6534810	*LY86-AS1*	3.24272×10^{-08}	3.24272×10^{-08}	2.23611×10^{-07}
rs56187477	6784474	*LY86-AS1*	1.64348×10^{-07}	1.64348×10^{-07}	1.02150×10^{-06}
rs116296279	9899074	*OFCC1*	4.46455×10^{-06}	0.000209404	1.20248×10^{-05}
rs57691377	12087410	*HIVEP1*	3.68113×10^{-06}	3.68113×10^{-06}	2.14357×10^{-05}
rs77100571	16632468	*SOSTDC1*	0.000128744	0.0349951	2.56638×10^{-06}
rs143251670	16632910	*SOSTDC1*	0.000128744	0.0349951	2.66335×10^{-06}
rs75593119	16633320	*SOSTDC1*	0.000128744	0.0349951	2.69632×10^{-06}
rs149880498	22747569	*STEAP1B*	3.14753×10^{-06}	3.14753×10^{-06}	1.41378×10^{-05}

Continued on next page

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs112267225	24321385	*STK31*	6.92305×10^{-07}	6.07719×10^{-05}	2.87304×10^{-06}
rs59255437	36955957	*HERPUD2*	2.44041×10^{-06}	2.44041×10^{-06}	1.57665×10^{-05}
rs73455755	127653379	*SND1*	1.44683×10^{-06}	1.44683×10^{-06}	7.31437×10^{-06}
rs115752921	4116498	*CSMD1*	9.42619×10^{-08}	9.42619×10^{-08}	5.07114×10^{-07}
rs192443459	3780518	*RFX3*	1.45649×10^{-07}	1.45649×10^{-07}	9.57891×10^{-07}
rs183731078	3780522	*RFX3*	8.39562×10^{-09}	8.39562×10^{-09}	4.46842×10^{-08}
rs75676131	14034754	*MPDZ*	4.74487×10^{-06}	4.74487×10^{-06}	2.73662×10^{-05}
rs71513209	15097606	*FREM1*	3.59854×10^{-06}	3.59854×10^{-06}	1.48299×10^{-05}
rs7867725	71836731	*TJP2*	4.29884×10^{-06}	4.29884×10^{-06}	2.29749×10^{-05}
rs118182351	124556385	*GSN*	1.23932×10^{-06}	8.00207×10^{-05}	3.54347×10^{-06}
rs56101037	53375308	*PRKG1*	2.25421×10^{-06}	2.25421×10^{-06}	1.46884×10^{-05}
rs114302523	94038690	*CPEB3*	4.15755×10^{-06}	4.15755×10^{-06}	2.57750×10^{-05}
rs75642696	120233642	*FAM204A*	1.73933×10^{-06}	0.000379639	4.87215×10^{-06}
rs59104649	129150690	*LOC728065*	1.13557×10^{-06}	1.13557×10^{-06}	7.33642×10^{-06}
rs78402894	132740552	*GLRX3*	1.35824×10^{-06}	1.35824×10^{-06}	8.33035×10^{-06}
rs115573913	132758348	*GLRX3*	1.40427×10^{-07}	1.40427×10^{-07}	9.22566×10^{-07}
rs115084777	132758950	*GLRX3*	1.40427×10^{-07}	1.40427×10^{-07}	9.16619×10^{-07}
rs79272561	132776876	*GLRX3*	1.52671×10^{-07}	1.52671×10^{-07}	1.03125×10^{-06}
rs114651413	132777100	*GLRX3*	1.52671×10^{-07}	1.52671×10^{-07}	1.02525×10^{-06}
rs181397896	132777863	*GLRX3*	1.52671×10^{-07}	1.52671×10^{-07}	1.01880×10^{-06}

Continued on next page

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs76221219	132778067	*GLRX3*	1.52671×10^{-07}	1.52671×10^{-07}	1.00810×10^{-06}
rs76615755	4262041	*NUP98*	1.89052×10^{-06}	1.89052×10^{-06}	9.77961×10^{-06}
rs113564957	4329853	*NUP98*	8.53370×10^{-07}	2.31360×10^{-05}	2.65907×10^{-06}
rs151320773	18795103	*LDHA*	4.65557×10^{-06}	0.0110859	1.00997×10^{-05}
rs79413720	30805302	*FSHB*	2.21549×10^{-06}	2.66723×10^{-05}	7.01145×10^{-06}
rs76845508	131001813	*ADAMTS15*	3.00579×10^{-06}	5.14595×10^{-05}	1.03086×10^{-05}
rs111772662	12974808	*DDX47*	0.686188	0.159730	1.29569×10^{-06}
rs113092010	12985262	*DDX47*	0.686188	0.159730	1.35366×10^{-06}
rs35074374	102940479	*IGF1*	3.20926×10^{-06}	5.96108×10^{-05}	1.48835×10^{-05}
rs61935454	102945889	*IGF1*	3.20926×10^{-06}	5.96108×10^{-05}	1.46406×10^{-05}
rs73399125	106534519	*NUAK1*	4.62196×10^{-06}	4.62196×10^{-06}	2.62717×10^{-05}
rs73403013	114424029	*RBM19*	3.56843×10^{-06}	3.56843×10^{-06}	2.30344×10^{-05}
rs111685758	114426898	*RBM19*	1.46762×10^{-08}	1.46762×10^{-08}	1.07231×10^{-07}
rs116600250	114428274	*RBM19*	4.30673×10^{-07}	4.30673×10^{-07}	2.87893×10^{-06}
rs73406439	115255955	*TBX3*	1.25080×10^{-07}	5.92204×10^{-06}	5.17466×10^{-07}
rs114641714	116114648	*TBX3*	3.51323×10^{-06}	3.51323×10^{-06}	2.16666×10^{-05}
rs112442238	116123041	*TBX3*	1.20645×10^{-06}	1.20645×10^{-06}	7.94417×10^{-06}
rs112736328	128188145	*LOC440117*	1.27092×10^{-08}	1.27092×10^{-08}	9.10594×10^{-08}
rs78970694	128189633	*LOC440117*	3.09530×10^{-06}	3.09530×10^{-06}	2.01577×10^{-05}
rs143383128	78159614	*MYCBP2*	7.73814×10^{-08}	7.73814×10^{-08}	5.38455×10^{-07}
Continued on next page					

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs142696984	78161441	*MYCBP2*	7.73814×10^{-08}	7.73814×10^{-08}	5.48333×10^{-07}
rs143732367	78169793	*MYCBP2*	3.61018×10^{-06}	3.61018×10^{-06}	2.03343×10^{-05}
rs140567953	78170405	*MYCBP2*	3.61018×10^{-06}	3.61018×10^{-06}	2.07125×10^{-05}
rs11841309	78172830	*MYCBP2*	3.61018×10^{-06}	3.61018×10^{-06}	2.03531×10^{-05}
rs4001146	98413112	*RAP2A*	1.12537×10^{-06}	1.12537×10^{-06}	5.26342×10^{-06}
rs74109414	98416186	*RAP2A*	1.12537×10^{-06}	1.12537×10^{-06}	5.29779×10^{-06}
rs55666653	98422488	*RAP2A*	1.12537×10^{-06}	1.12537×10^{-06}	5.19218×10^{-06}
rs75401524	98430416	*RAP2A*	1.12537×10^{-06}	1.12537×10^{-06}	5.28575×10^{-06}
rs75798371	106625513	*LINC00343*	1.51209×10^{-06}	1.51209×10^{-06}	1.02336×10^{-05}
rs17115107	44948742	*LRFN5*	1.56948×10^{-07}	0.000359026	1.80256×10^{-07}
rs111362305	56306095	*KTN1*	2.09212×10^{-06}	2.09212×10^{-06}	1.25386×10^{-05}
rs147525481	56306157	*KTN1*	2.09212×10^{-06}	2.09212×10^{-06}	1.26139×10^{-05}
rs113215666	56306969	*KTN1*	2.09212×10^{-06}	2.09212×10^{-06}	1.25446×10^{-05}
rs112482790	56307284	*KTN1*	2.96203×10^{-06}	2.96203×10^{-06}	1.84075×10^{-05}
rs16953734	71670073	*THAP10*	2.94000×10^{-06}	2.94000×10^{-06}	1.94058×10^{-05}
rs148623690	71676470	*THAP10*	2.92790×10^{-06}	2.92790×10^{-06}	1.77925×10^{-05}
rs139400081	71676639	*THAP10*	2.94000×10^{-06}	2.94000×10^{-06}	1.93139×10^{-05}
rs115582725	72584085	*CELF6*	1.81779×10^{-06}	1.81779×10^{-06}	1.18490×10^{-05}
rs114254006	72595098	*CELF6*	4.59392×10^{-06}	4.59392×10^{-06}	2.94669×10^{-05}
rs60466172	72623423	*CELF6*	4.59392×10^{-06}	4.59392×10^{-06}	2.92612×10^{-05}
Continued on next page					

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs114369106	72625278	*CELF6*	4.54524×10^{-07}	4.54524×10^{-07}	3.10077×10^{-06}
rs144749334	94395197	*CHD2*	1.16841×10^{-06}	2.26880×10^{-05}	5.44624×10^{-06}
rs11631144	101339706	*ASB7*	1.82093×10^{-06}	1.82093×10^{-06}	1.14890×10^{-05}
rs113469389	26047975	*HS3ST4*	7.78709×10^{-09}	5.64390×10^{-07}	4.60911×10^{-08}
rs113909212	26049230	*HS3ST4*	3.94124×10^{-09}	1.29372×10^{-06}	2.52969×10^{-08}
rs113604782	26049477	*HS3ST4*	3.54940×10^{-08}	2.28640×10^{-06}	1.94107×10^{-07}
rs72483415	26060758	*HS3ST4*	1.38042×10^{-07}	1.38042×10^{-07}	9.77389×10^{-07}
rs41528147	79555485	*WWOX*	2.33765×10^{-06}	2.33765×10^{-06}	7.11190×10^{-06}
rs7193361	81010343	*MAF*	1.83405×10^{-06}	3.01221×10^{-06}	5.66695×10^{-06}
rs73264863	86286588	*KIAA0513*	4.93818×10^{-06}	4.93818×10^{-06}	2.25562×10^{-05}
rs73264899	86294373	*KIAA0513*	4.93818×10^{-06}	4.93818×10^{-06}	2.26502×10^{-05}
rs8096513	4455491	*DLGAP1*	1.42909×10^{-09}	1.42909×10^{-09}	1.00692×10^{-08}
rs11876683	58022895	*CCBE1*	4.69757×10^{-06}	1.94651×10^{-05}	1.38194×10^{-05}
rs75304175	74118555	*ZADH2*	1.64927×10^{-06}	1.64927×10^{-06}	8.87783×10^{-06}
rs12608785	32796314	*TSHZ3*	1.89773×10^{-06}	1.89773×10^{-06}	9.81669×10^{-06}
rs144751221	33492965	*SLC7A9*	4.47657×10^{-07}	4.47657×10^{-07}	2.84429×10^{-06}
rs113748720	33495080	*SLC7A9*	1.17117×10^{-07}	1.17117×10^{-07}	7.81696×10^{-07}
rs74758634	56755150	*ZSCAN5A*	2.53343×10^{-06}	2.53343×10^{-06}	1.38622×10^{-05}
rs61484690	5863637	*PROKR2*	3.71262×10^{-06}	3.71262×10^{-06}	2.47105×10^{-05}
rs77889812	25139678	*VSX1*	3.43642×10^{-06}	1.03800×10^{-05}	9.57629×10^{-06}
Continued on next page					

Table 6.2 – continued from previous page					
SNP	**BP**	**GENE**	**P-FE**	**P-RE**	**P-BE**
rs143241795	40094621	*ERG*	2.74099×10^{-06}	2.74099×10^{-06}	1.54720×10^{-05}
rs114055013	27526491	*CRYBA4*	4.37168×10^{-06}	6.36760×10^{-06}	1.55806×10^{-05}

6.1 Pathway Enrichment Analysis

Table 6.3: ***Biological pathways that are enriched by the Network of genes identified from IMPUTE2 based meta-analysis.***

Index Name	P-value	Adjusted P-value	Odds Ratio	Combined score
African trypanosomiasis	0.00001148	0.003537	67.57	768.55
Malaria	0.00002697	0.004154	51.02	536.77
Nitrogen metabolism	0.02021	1.000	49.02	191.25
Collecting duct acid secretion	0.03192	1.000	30.86	106.31
Prion diseases	0.04119	1.000	23.81	75.94
Glycine, serine and threonine metabolism	0.04694	1.000	20.83	63.73
Porphyrin and chlorophyll metabolism	0.04923	1.000	19.84	59.75
Legionellosis	0.06399	1.000	15.15	41.65
Chemokine signaling pathway	0.02160	1.000	8.77	33.64
Antigen processing and presentation	0.08847	1.000	10.82	26.25

Table 6.4: ***Biological pathways that are enriched by the Network of genes identified from ImpG based meta-analysis.***

Name	P-value	Adjusted P-value	Odds Ratio	Combined score
Glutamatergic synapse	1.311×10^{-10}	4.038×10^{-8}	45.48	1035.00
Cocaine addiction	0.00003878	0.005972	45.35	460.66
Amyotrophic lateral sclerosis (ALS)	0.00004376	0.004492	43.57	437.34
Nicotine addiction	0.001326	0.06808	37.04	245.38
Vasopressin-regulated water reabsorption	0.001603	0.06172	33.67	216.69
Circadian entrainment	0.0002973	0.02289	22.91	186.05
Arginine biosynthesis	0.02798	0.6630	35.27	126.14
Long-term potentiation	0.003676	0.1132	22.11	123.96
Amphetamine addiction	0.003784	0.1060	21.79	121.50
Protein export	0.03061	0.6734	32.21	112.28

Table 6.5: ***Biological pathways that are enriched by the Network of genes identified from both IMPUTE2 and ImpG meta-analysis.***

Index Name	P-value	Adjusted P-value	Odds Ratio	Combined score
Malaria	0.00001703	0.005244	25.92	284.57
African trypanosomiasis	0.0002144	0.02201	25.74	217.44
Glutamatergic synapse	0.00002975	0.004581	13.92	145.12
Vasopressin-regulated water reabsorption	0.008484	0.6533	14.43	68.82
Cocaine addiction	0.01044	0.6432	12.96	59.11
Amyotrophic lateral sclerosis (ALS)	0.01127	0.5788	12.45	55.84
Nitrogen metabolism	0.05224	1.000	18.67	55.12
Arginine biosynthesis	0.06414	1.000	15.12	41.52
Staphylococcus aureus infection	0.01947	0.7495	9.34	36.78
Protein export	0.07003	1.000	13.80	36.70