Examining the relationship between genetic variation at G6PD and severe malaria

Shivang Shah
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2011

This thesis is dedicated to

my fiancé, family and friends,

for always reminding me of what is good and noble in life.

Acknowledgements

I would like to thank Dominic P. Kwiatkowski, Kirk A. Rockett, and Thomas E. Wellems for their encouragement and guidance as my supervisors, and Thomas N. Williams and Rick A. Fairhurst for guiding fruitful collaborative work, and providing continual advice and mentorship. Although too numerous to name individually, I would also like to thank my many labmates and colleagues across three continents for their tutelage and friendship. I additionally wish to acknowledge the National Institutes of Health Medical Scientist Training Program and Oxford-Cambridge Scholars Program for their continued support. Most importantly, I would like to thank my family, friends, and Kimberly for their love and support throughout my life.

Abstract

Examining the relationship between genetic variation at G6PD and severe malaria

Shivang S. Shah

Wolfson College

Submitted for the degree of Doctor of Philosophy

Trinity 2011

Glucose-6-phosphate dehydrogenase (G6PD) deficiency, a common heritable trait whose prevalence mirrors geographic patterns of historic malaria endemicity, is thought to confer a selective advantage owing to partial protection conferred against malaria. Direct evidence supporting this malaria protection hypothesis in the form of clinical association studies remains controversial, however, as conflicting results have been reported with respect to the strength and specificity of a protective effect, if any, conferred to carriers of G6PD deficiency-associated alleles.

This thesis examines genetic diversity at the *G6PD* locus, and then considers how such variation impacts both immediate molecular phenotypes and multifactorial clinical phenotypes. First, Chapter 3 presents a survey of variation at *G6PD* in several malaria-endemic areas, while Chapter 4 describes a novel technique for polymorphism discovery using pooled massively parallel sequencing. Next, in Chapters 5 and 6, I evaluate the link between genetic variation at the locus and G6PD enzyme activity, identifying major and minor determinants of G6PD deficiency state in an association study conducted in Kenya, and demonstrating a new technique for assaying G6PD deficiency at the level of an individual erythrocyte in a pilot project in Mali. Finally, Chapter 7 addresses the malaria protection hypothesis directly by conducting a fine-mapping case-control association study of severe malaria in the Gambia, where I found that G6PD deficiency alleles exhibited differential direction of association with respect to two important clinical syndromes– trending towards risk conferred to severe malarial anemia, and protection with respect to cerebral malaria. Overall, these findings suggest that future clinical association studies should consider heterogeneity at the genetic level, as well as at the level of molecular and clinical phenotypes in order to achieve a better mechanistic understanding of the relationship between G6PD deficiency and severe malaria.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Malaria

Malaria is a tropical infectious disease caused by parasites of the genus *Plasmodium*. More than half of the world's population is at risk, and each year hundreds of millions will develop a symptomatic infection, of which approximately 1 million will die of severe disease, most of whom are African children [1]. In endemic areas, the toll is felt not only in terms of morbidity and mortality, but also economically, creating a vicious cycle where the poorest are those most susceptible to disease, and those individuals and countries disproportionately infected are less economically productive [2].

1.1.1 Plasmodium infection and life cycle

The protozoan responsible for malaria was first identified in human blood by Laveran in the 1880s in Algeria and Italy [3]. Its mode of transmission via mosquito was demonstrated more than a decade later in India in birds by Ross [4], and in Italy in humans by Grassi, Bignami, and Bastianelli [5]. Later work showed that the mosquitoes that transmit human malaria are those of the *Anopheles* genus, and that the host range for *Plasmodium*, apart from humans, includes birds, reptiles, and mammals such as rodents and primates [6]. Species of *Plasmodium* known to infect humans include *P. vivax*, *P. ovale*, *P. malariae*, *P. knowlesi*, and *P. falciparum*, the virulent parasite most associated with mortality, which is thought to have originated from a related gorilla strain, though the time of emergence is still unknown [7].

The life cycle of *Plasmodium* traverses a diverse range of environments in both host and vector (Figure 1.1). Following gametocyte fertilization and ookinete development in the mosquito midgut, the sporozoites produced migrate to the salivary gland. During the blood meal of a female mosquito, tens to hundreds of infective sporozoites are injected into host subcutaneous tissue, after which the sporozoites migrate to the liver where they invade and develop inside hepatocytes.

Parasite surface receptors that mediate hepatocyte binding and invasion include the circumsporozoite protein (CS) and thrombospondin-related adhesive protein (TRAP). The CS protein has also been found to exert an immunosuppressive role in downregulating the appropriate cellular immune response elicited

in hepatocytes, doing so indirectly via competitive inhibition of nuclear import of canonical transcription factor NF-κB, thus leading to reduced class I major histocompatibility complex (class I MHC) expression on the cell surface [8]. Sporozoites in hepatocytes multiply and differentiate during a symptomless pre-erythrocytic stage which lasts 1-2 weeks, forming thousands of merozoites which then burst into the hepatic sinuses and are competent to invade red blood cells (RBC). With *P. vivax* and *P. ovale*, a dormant, hypnozoite form can persist in the liver and produce recrudescences many months after an initial infection [9].

The erythrocytic stage is characterized by regular cycles of invasion and exponential growth, which produce the paroxysmal fevers associated with malaria.

Invasion of RBC by the parasite is mediated by several parasite surface receptors found in specialized organelles called rhoptries and micronemes at the apical surface of the merozoite. These receptors possess Duffy-binding like (DBL) and reticulocyte-binding like (RBL) domains that mediate their interaction with host RBC receptors such as sialic acid and members of the glycophorin family [10]. The diversity of the repertoire of such invasion receptors appears to determine the invasion efficiency and specificity, and by extension, the virulence of certain parasite strains by virtue of the constraint on parasitemia. For example, *P. vivax* is strongly dependent on host RBC being at the reticulocyte stage and also expressing Duffy antigen [11], and as a consequence rarely leads to high parasitemias and generally causes far milder disease as a result. By contrast, one particularly virulent strain of rodent parasite *P. yoelli* is able to not only invade reticulocytes, but other RBC as well, thus allowing it to generate high parasitemia and cause more severe disease. Similarly, *P. falciparum* has a wide array of DBL- and RBL-containing receptors such that it is competent to invade a variety of RBC types and thus achieve higher parasitemias and cause more severe disease [10].

Invasion is followed by maturation from an immature ring form to a trophozoite. During this process, nutrient intake is crucial, and parasites utilize a plasmodial surface anion channel (PSAC) to assist in the delivery of sugars, amino acids, purines, and vitamins to the maturing trophozoite [12]. Amino acids are also obtained via uptake and catabolism of host hemoglobin, performed in a specialized acidic compartment called the digestive vacuole (food vacuole). In the process of digesting hemoglobin, the liberated heme must be polymerized into hemozoin (malaria pigment) in order to limit the production of reactive oxygen species [13], and several drugs, most notably chloroquine, act to inhibit this process of heme polymerization as a major aspect of their mechanism of action [14]. Lastly, a putative endosymbiont plastid organelle called the apicoplast is critical in fatty acid and heme biosynthesis, and also represents a novel target for drug development [15].

After several rounds of mitotic division during a schizont stage, six to twenty new merozoites are produced which then burst the host RBC and find new RBC to invade. In the course of natural infection (untreated), an individual will harbor both asexual (merozoites, ring-form and mature trophozoites) and sexual forms (gametocytes) of the parasite in the blood, the latter of which can complete the life cycle when taken up during a mosquito blood meal [10]. The commitment to gametocytogenesis is multifactorial and an area of active research, but an inhospitable host milieu (i.e. decreased new host RBC for invasion or insufficient nutrients) is certainly one precipitating factor, as *P. vivax*, which is limited in the total number of RBC it can infect owing to its reticulocyte selectivity, produces gametocytes very soon after emergence from the liver [10], while with *P. falciparum* gametocytogenesis may not occur for several days or weeks.

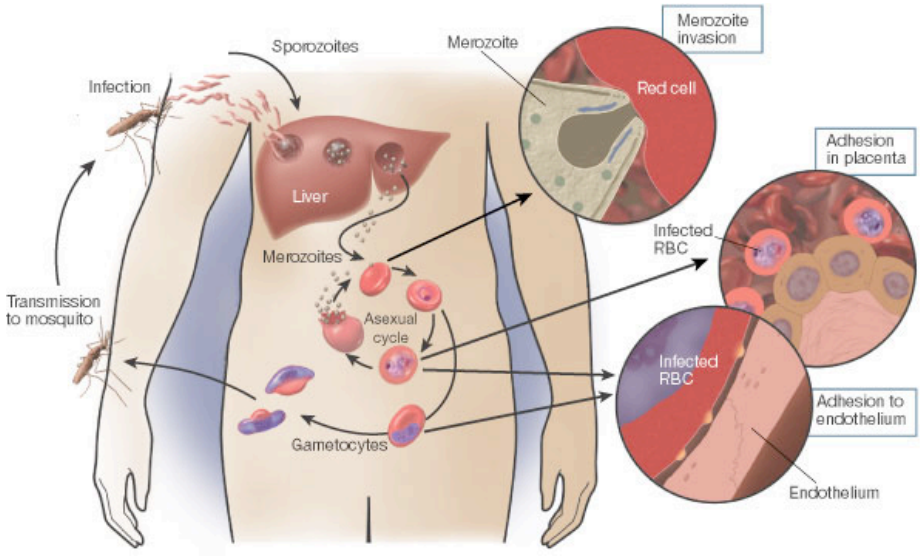


Figure 1.1: Schematic of life cycle of *P. falciparum*. Reproduced from Miller et al. [10].

1.1.2 Morbidity and mortality

While the etiology of malaria can be traced to *Plasmodium*, the pathophysiology is more complex, and much remains to be understood about the determinants of particular disease manifestations.

1.1.2.1 Infection

The first step in the development of malaria is infection. In a given population, a certain proportion of individuals will harbor parasites, and this parasite prevalence rate (Figure 1.2) is a function of exposure to infectious bites, which in turn depends on mosquito ecology, and thus is strongly influenced by temperature and humidity, and epidemiological control measures, such as bednet usage. Regions can be classified based on parasite prevalence rate [16] as follows: hypoendemic - less than 10%, mesoendemic -

between 10% and 50%, hyper-/holoendemic - 50% or higher. For example, in southern Mali, transmission is seasonal but high (hyperendemic), whereas in the Gambia [17], transmission is seasonal but slightly lower (mesoendemic).

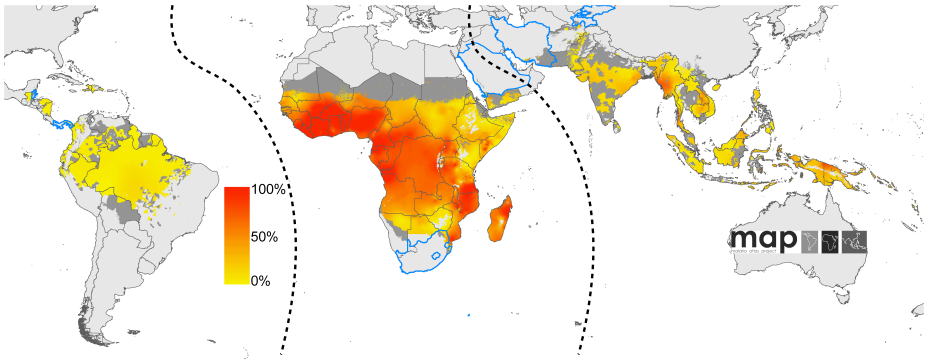


Figure 1.2: Global distribution of *P. falciparum* in 2007. Color-graded according to parasite rate in children between 2 and 10 years of age, with low risk zones colored in grayscale. Diagram generated by Malaria Atlas Project [1].

1.1.2.2 Mild disease

In most individuals, the presence of parasitemia at any given time is not associated with symptoms. As parasites begin to multiply in number, symptoms may appear, though the correlation between parasite density and morbidity is poor [18, 19], and while lower parasite densities can be useful in excluding malaria as a diagnosis in high transmission areas [1], very high parasite densities do not necessarily result in severe or even overt disease. When mild malaria symptoms do develop, they include fever (by definition), chills, headache, and malaise, and occasionally gastrointestinal symptoms such as vomiting or diarrhea [20]. The development of splenomegaly and anemia may follow, but the disease generally resolves with or without medication in the majority of individuals.

1.1.2.3 Severe illness

In a minority of symptomatic cases ($< 1\%$), mild malaria can progress to more severe disease. Life-threatening forms of malaria are associated with certain key clinical signs and biomarkers [21]. In children, who contribute overwhelmingly to the global burden of malaria mortality, severe malaria may be understood as three overlapping clinical entities (Figure 1.3). It is important to note that two or even all three of these syndromes may be present in a given individual, and this presentation correlates with poorer prognosis, as does the presence of hypoglycemia, jaundice, or acute renal failure [20, 21].

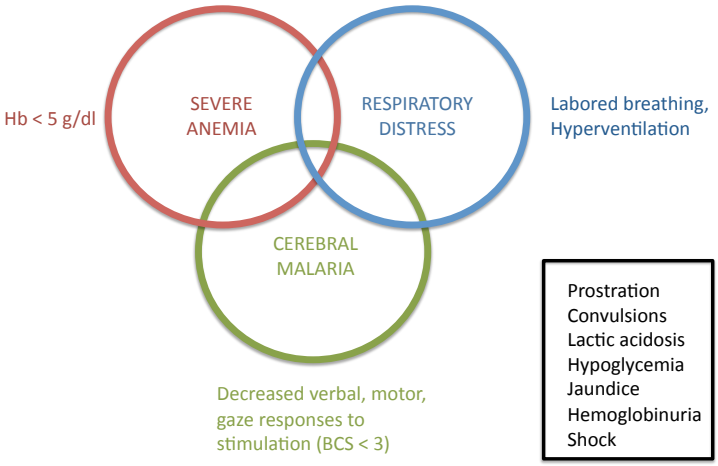


Figure 1.3: Schematic of severe malaria syndromes. While other useful clinical conditions indicative of severe malaria have been identified (black box), three especially common (and overlapping) syndromes include severe anemia, cerebral malaria, and respiratory distress.

Severe malarial anemia Severe malarial anemia (SMA) is defined by the World Health Organization (WHO) as the presence of hemoglobin level of less than 5 g/dL [22], and is independently associated with mortality of less than 5% [21], though this may be complicated by secondary jaundice or renal failure. SMA is thought to be multifactorial in nature, with infected RBC lysis being an obvious, but minor contributor, as parasitemia is rarely high enough to explain the drop in hemoglobin levels.

One major factor contributing to SMA is the destruction of uninfected RBC, which appears to occur principally via increased extravascular clearance by splenic macrophages, whose activity can be enhanced by increased production of gamma interferon (IFN-γ) by CD4+ T cells [23]. Indeed, the hallmarks of oxidatively-stressed, senescent RBC are found in individuals with SMA, including exposed phosphotidylserine groups and increased complement and autologous IgG deposition at the RBC membrane [24]. Additionally, in untreated experimental infections in human neurosyphilis patients, the time course of SMA development was found to be closely linked to splenic enlargement [25].

Apart from intravascular hemolysis and extravascular splenic clearance, the impaired production of new RBC acts to exacerbate and prolong SMA. Impaired erythropoesis is evident in the decreased reticulocyte count found in human SMA patients [26], and also in the impaired development of erythroid precursors in mice [27]. It is thought that the suppression of hematopoesis may be mediated by IFN-γ and/or tumor necrosis factor alpha (TNF-α), both of which are upregulated during the course of a malaria infection [24].

Finally, apart from malaria, other superimposed conditions may serve to increase risk to severe anemia, including the presence of hookworm infection or iron deficiency [28]. And while case mortality is

low for SMA, this mortality represents that observed at a hospital or health center, where, in contrast to CM and RD, a safe, effective, and affordable treatment option exists– transfusion. In reality, then, the case fatality rate for SMA in the absence of treatment is probably much higher.

Cerebral malaria Cerebral malaria (CM) is defined qualitatively by the presence of impaired consciousness as measured by a coma score, and is associated with mortality approaching 10% [21]. In Africa, generally the Blantyre coma score is used, which assesses ocular, motor, and verbal responsiveness on a scale of 0 (unconscious) to 5 (normal).

Pathology studies strongly implicate infected RBC (iRBC) sequestration in the cerebral microvasculature in the pathogenesis of CM (Figure 1.4). The first such observation was made in 1894 by Marchiafava and Bignami, who observed iRBC sequestration in brain capillaries of Italians who died of malaria with cerebral complications [29]. A later pathology study in Thai adults has demonstrated signs of cerebral endothelial activation including upregulation of ICAM-1 and E-selectin, with iRBC binding co-localized to expression patterns of those endothelial receptors [30]. Even more recently, pathology studies of Malawian children found that iRBC sequestration, along with thrombosis and ring hemorrhages, was associated with blood brain barrier disruption, as well as axon and myelin damage [31], and that the presence of extraerythrocytic malaria pigment was strongly associated with microvascular pathology [32]. Additionally, unique to CM as compared to nonmalarial causes of coma is retinal vascular pathology, including hemorrhages and white spots [32].

Less well understood is how iRBC mediate the pathogenesis of CM, and the evidence is still best characterized as associative and not causal. For example, sites of iRBC sequestration often contain numerous other blood cells including monocytes/macrophages, neutrophils, and platelets. Experiments in rodent malaria suggest that CM is an immunopathological disorder, mediated by certain subsets of T cells active in innate immunity and especially by IFN-γ, as mice deficient in either IFN-γ or its cognate receptor fail to develop CM [33]. It is thus thought that CM involves some combination of local impairment of vascular reactivity leading to ischemia as well as systemic processes such as endothelial activation, coagulation pathway activation, and generalized immune activation mediated by cytokines such as IFN-γ and TNF-α [34]. The role of nitric oxide is especially interesting as a readily diffusible compound with diverse biological activities including some that would seem to counteract the development of CM, e.g. vasodilation, downregulation of endothelial activation, inhibition of platelet aggregation, but also those that might exacerbate it, e.g. neuronal excitotoxicity and oxidative damage produced as part of the effector immune response by macrophages and neutrophils via inducible nitric oxide synthase [35]. Finally, it is important to note that non-immune factors such as hypoglycemia may also play a role in exacerbating the symptoms and pathology of CM.

Respiratory distress Respiratory distress (RD) is characterized by labored breathing and/or hyperventilation, and is associated with mortality of greater than 20% [21]. It is thought to be a respiratory compensation for metabolic acidosis caused by lactate generated by anaerobic glycolysis, but may also be exacerbated by pulmonary vascular congestion caused by heart failure secondary to SMA. While no clear molecular mechanism associated with RD has yet been identified, it seems logical that factors that impair oxygen delivery (e.g. severe anemia) or blood flow (e.g. systemic microvascular pathology caused by iRBC sequestration) would serve to increase blood lactate levels, and thus act to increase the risk for RD. More recently, evidence implicating a hyperinflammatory state has been observed in RD patients in Ghana, characterized by increased levels of IFN-γ, TNF-α, and neopterin, a marker of macrophage activation [36].

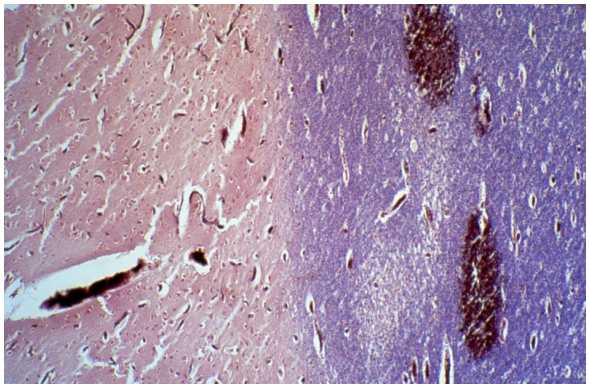
1.1.2.4 Modifiers of clinical burden

Several factors in mosquito, host, and parasite can influence the development of mild disease and the progression to severe disease.

Transmission intensity With respect to the mosquito, absolute numbers, population sex ratio, and sporozoite prevalence rate all influence the population dynamics of malaria risk in areas of different endemicity [37]. In high transmission areas, the incidence of clinical disease will approach 100% during the first five years of life, but drop off after that, whereas in moderate transmission areas, the incidence will be lower, but more constant over time. Work done in the Gambia and Kenya has also shown an inverse relationship between transmission intensity and severe malaria risk, where the highest transmission areas experienced proportionally fewer severe malaria clinical episodes as compared to low and moderate transmission zones [38]. Interestingly, this trend held for all forms of severe malaria, including SMA and CM, though the effect was less pronounced for SMA, which might explain the different proportions of the two syndromes seen in these two areas of Africa. Whereas CM predominates by a 2:1 ratio in the Gambia, a 1:1 ratio of CM:SMA is seen in Kenya [39].

Host immune response The most important and well-established aspect of the host response to infection is the development of adaptive immunity [34]. This acquired immunity is both humoral and cellular, age- and exposure-dependent, and is non-sterile (i.e. it controls, but does not prevent future infection). In hyperendemic regions, this immunity is generally negligible before the age of five, but well-developed by the age of ten. It is also thought that immunity to severe disease may be acquired relatively rapidly, possibly even after a single infection [40]. Immune mechanisms serve to prevent and control the course of infection at various stages in the life cycle, can be modulated by the parasite, and may even serve to exacerbate the course of disease pathogenesis.

Marchiafava and Bignami, *On summer-autumn malaria fevers.*



Wellcome Library

Figure 1.4: Pathology of cerebral malaria. The top diagram shows the original drawings of Marchiafava and Bignami that first identified the presence of infected RBC in the cerebral microvasculature [29]. The bottom photomicrograph (courtesy: Wellcome Library) shows a pale area of demyelination and two ring hemorrhages in the white matter (on right, with myelin stained in purple), and widespread sequestration of infected RBC in the microvasculature of both white and grey matter.

Pre-erythrocytic stage immunity Even before sporozoite migration to the liver, circulating antibodies can serve to neutralize and opsonize the sporozoite inoculum delivered by the mosquito. Importantly, however, antibodies directed against sporozoite invasion receptors such as CS protein confer neither sterile immunity nor clinical protection [41].

Should the sporozoite succeed in invading the hepatocyte, several mechanisms exist to control the development of infection, notably the cellular adaptive immune response characterized by IFN-γ production and effector CD8+ T cells. In infected hepatocytes, IFN-γ serves to enhance antigen presentation by class I MHC and expression of co-stimulatory molecules, both of which then prime antigen-specific CD8+ effector T cells for cytolysis via secretion of perforins and granzymes.

The pre- and intra-hepatic stages of infection represent attractive targets for vaccines as the parasite load during this time period is limited and thus tractable. However, while irradiated intrahepatic sporozoites have been shown to confer sterile immunity to infectious challenge in both rodents and humans [42], vaccines such as RTS,S which utilize the immunodominant epitope of CS protein have only been able to confer limited, nonsterile protection in recent clinical trials [43]. Approaches to enhance the immunogenicity of the vaccine include coupling its administration with stronger adjuvants as well as expressing the antigen using viral vectors such as modified vaccinia Ankara (MVA) which are known to elicit strong cellular immune responses [44].

Erythrocytic stage immunity During the erythrocytic stage, both humoral and cellular immunity act to control and eliminate infection. The importance of the humoral response is highlighted by the fact that B cell depleted mice are unable to fully clear parasitemia when infected with *P. chabaudi chabaudi* [45]. Likewise, the critical role of cellular immunity is strengthened by the observation that CD4+ T cells and elaboration of IFN-γ are required elements for clearance of parasitemia in mouse models, with CD4+ T cell depleted and IFN-γ deficient mice both exhibiting persistent parasitemia, the latter also associated with increased mortality [46, 45].

Blood stage immunity acts to both limit parasite density and virulence. Antibodies directed against certain merozoite surface receptors can serve to limit invasion efficiency and compromise parasite fitness. Additionally, CD4+ T cells play an important role, through direct and indirect means. Through direct interaction with antigen-specific B cells, they induce activation of antibody class switching for production of IgG which opsonize iRBC and free merozoites for immune clearance by macrophages, which can also be directly activated by CD4+ T cells. Macrophage activation is also mediated indirectly via IFN-γ produced by CD4+ T cells.

Erythrocytic stage immunity also serves to limit parasite virulence. Examples include antibodies directed

against parasite-derived immunogenic molecules such as GPI, or ones that impede parasite adhesive phenotypes mediated by variant surface antigens such as PfEMP1 [34]. The role of this important parasite surface antigen in immunity and pathogenesis is further explored below.

Finally, antibodies that opsonize the gametocyte stage parasite serve to limit transmission back to the mosquito vector. As such, a transmission blocking vaccine that targets this stage might be an effective intervention in limiting parasite prevalence rate in *Anopheles* and thus serve to reduce the number of infectious bites experienced by those living in malaria-endemic areas [10].

Innate immunity Innate immune cells such as dendritic cells (DCs) and natural killer (NK) cells are also important in controlling infection [34]. DCs function primarily as 'professional' antigen-presenting cells to CD4+ T cells which then directly and indirectly (via production of IFN-γ) act to stimulate macrophages, and also promote antibody class-switching from IgM to IgG. NK cells serve as cytotoxic effector cells that directly eliminate infected RBC, a function which can help to control higher levels of parasitemia, but which alone is inadequate for fully eliminating an infection [47].

Immunomodulatory role of the parasite The parasite is also capable of modulating the immune response. The glycosylphosphotidylinositol (GPI) anchors of certain *Plasmodium* membrane antigens are strongly immunogenic pathogen-associated molecular patterns (PAMPs) and can induce the production of large amounts of interleukin-4 from NK T cells (a variant of T cell with nonspecifically avid T cell receptors) and TNF-α from macrophages via binding to certain Toll-like receptors [48, 49]. This elaboration of cytokines can precipitate a systemic inflammatory reaction that can serve to exacerbate the fever, headaches and nausea that accompany malarial illness [34].

Another way in which parasites can alter the host immune response is via inhibition of DC maturation, thus impairing antigen presentation, CD4+ T cell activation and IFN-γ production. This effect, mediated by the binding of PfEMP1 on the infected RBC surface to CD36 on the DC membrane [50], thus acts to prevent effective elaboration of the erythrocytic stage adaptive immune response.

Parasite virulence factors Finally, the parasite itself impacts disease presentation in species- and strain-specific ways. The most important distinction lies between *P. falciparum* and all other species, where the former is considerably more virulent and accounts for the vast majority of global malaria mortality. Apart from the immunopathogenic effects of the GPI anchor mentioned earlier, specific strains of *P. falciparum* are also associated with increased virulence by virtue of their effects on the adhesion properties of infected RBC [51]. Such properties include rosetting, or the ability of infected RBC to bind uninfected RBC, and endothelial cytoadherence, which refers to the ability of infected RBC to bind to endothelial cells. These phenomena are mediated by parasite-encoded variant surface antigens of the *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) family.

The presence of parasite-derived surface antigens on the RBC surface was first suggested by early experiments performed in primate systems. In these studies, it was found that schizont-infected RBC could be agglutinated using sera from the infected animal, but not from an uninfected animal [52]. If an infection was allowed to proceed for long enough, antigenically-distinct recrudescences could be produced where sera from earlier waves of parasitemia could not agglutinate infected RBC from later waves [53]. The antigen mediating this phenomenon was parasite-encoded, stage specific (ring stage iRBC could not agglutinate), and clone-specific, as sera generated from an infection with one clone could not agglutinate infected RBC from another clone. Immune recognition of this antigen seemed to be important with respect to clinical protection, too, as adults generally had sera that was more broadly cross-reactive to different parasite isolates than did children [54].

Later work would demonstrate that while the parasite exports several proteins to the RBC, the key mediator in the agglutination phenotype, as well as in cytoadherence and rosetting was PfEMP1 [55]. This high molecular weight protein is encoded by any one of nearly sixty *var* genes [56] in a mutually exclusive fashion, where only one variant is expressed at a time, with variant switching thought to occur at a variable but low rate [51]. The tremendous genetic diversity at the *var* loci, coupled with the agglutination experiments strongly suggests that PfEMP1 is a major parasite variant antigen, and that the phenomenon of *var* switching represents a critical avenue for immune evasion.

Cytoadherence is another mechanism for immune evasion as PfEMP1 is thought to mediate the iRBC sequestration seen in CM. Infected RBC attached to endothelial receptors such as CD36 or ICAM-1 are no longer in the peripheral circulation and can thus no longer be cleared by splenic macrophages. Importantly, cytoadherence is not restricted to endothelial cells; iRBC can also bind platelets [57] and white blood cells [58]. Binding of iRBC to these various cell types in the cerebral microvasculature may mediate a pro-inflammatory and pro-coagulant state thought to characterize the immunopathogenesis of CM (see earlier discussion of CM).

Apart from the pathology data implicating sequestration and cytoadherence in cerebral vascular and neuronal damage, *in vitro* work examining the cytoadherence of field isolates from patients with severe malaria demonstrate stronger cytoadherence to ICAM-1 in CM patients [59, 60], which comports with the increased ICAM-1 expression seen in pathological specimens [30]. Rosette formation, mediated by PfEMP1 binding to complement receptor 1, is similarly associated with disease severity in similar studies, including one in the Gambia looking at CM [61] and another in Kenya examining severe malaria [62]. Additionally, there is one PfEMP1 variant which binds chondroitin sulfate A and is strongly associated with placental malaria, a severe complication of pregnancy in primigravid women associated with intrauterine growth retardation and increased maternal and fetal mortality [63].

1.1.3 Therapeutic and control measures

1.1.3.1 Antimalarial drugs

Malaria is curable with chemotherapeutic agents, including those that interfere with the metabolism of folate (e.g. sulfadoxine, pyrimethamine) or hemoglobin (e.g. chloroquine), and several others whose mechanism remain a mystery. The latter includes two mainstays of severe malaria therapy, quinine and artemisinin, both of which are derived from natural products (cinchona bark and sweet wormwood, respectively) and have been used as traditional therapies for hundreds of years [20]. Despite the existence of curative therapies, drug resistance has been a longstanding problem that began with resistance to chloroquine in the 1960s in southeast Asia, and threatens each new generation of antimalarials, most recently the first-line drug artemisinin [64]. An additional problem that specifically affects treatment and public health management of *P. vivax* and *P. ovale* is that primaquine, the only drug available to provide radical cure of hypnozoite-stage parasites, is contraindicated in patients with glucose-6-phosphate dehydrogenase deficiency [20], a relatively common trait in malaria-endemic areas (discussed below). Finally, apart from antimalarial medications, supportive therapies may include acetaminophen to treat fever, phenobarbital to treat seizures, and transfusion therapy to treat severe anemia and concomitant hyperparasitemia [20].

1.1.3.2 Population-level control of malaria

At a population level, control of malaria can, in theory, be achieved in a variety of ways, with the ultimate aim being the attainment of transmission rates of zero in a particular region (elimination) or from the world as a whole (eradication) [65].

The most important, evidence-based method to date remains vector control, which includes the use of insecticide-treated nets (ITN) and indoor residual spraying of homes with insecticide, along with elimination of mosquito breeding sites (e.g. dredging of standing water in a swamp). For example, ITN usage is associated with a roughly 50% reduced incidence of mild malaria, and widespread spraying of the insecticide DDT in the 1950s allowed many Mediterranean countries to eliminate malaria entirely [66].

Another critical aspect to malaria control involves the parasite itself. Continuous transmission occurs when many hosts are infected, each serving as a reservoir for mosquito-transmissible gametocytes. Current approaches to rational vaccine and drug design take all aspects of the transmission cycle into account in efforts to reduce transmission rate, akin to generating 'herd immunity'. Vaccines currently in development or in clinical trials seek to curb the initial development of infection (pre-erythocytic stage), the exponential growth of parasite density (erythrocytic stage), or the transmission of parasite to mosquito (gametocyte stage)[67]. Similar efforts are being undertaken in drug discovery, where the ideal drug candidate is one that targets not only the asexual trophozoite, but also the sexual-stage gametocyte [68].

Finally, there are the more general issues relating to nutrition and health care infrastructure, both of which have significant impacts on the morbidity and mortality associated with malaria. For example, malnutrition-related growth stunting is strongly associated with both increased prevalence of mild and severe disease [69]. Likewise, it has been shown that ease of access to medical treatment improves disease outcomes [70]. But while addressing these structural inequalities is difficult, it has considerable bearing on the efficacy of any vector- or parasite-focused approaches, as was demonstrated after the last malaria eradication campaign in the 1950s, when many countries were able to transiently eliminate transmission, only to have it resurface years later owing to poor surveillance and lack of rapid removal of reservoirs of infection [71].

1.2 G6PD and the host susceptibility genetics of malaria

1.2.1 Glucose-6-phosphate dehydrogenase

1.2.1.1 Physiological role

Glucose-6-phosphate dehydrogenase (G6PD) catalyses the rate-limiting first step in the pentose phosphate pathway (PPP), generating NADPH during the oxidation of glucose-6-phosphate (G6P) to 6-phosphogluconolactone (Figure 1.6). Nicotinamide adenine dinucleotide phosphate (NADPH) is a cofactor critical to a variety of fundamental metabolic processes, including:

a. fatty acid and nucleic acid synthesis

b. antioxidant defense – regeneration of reduced glutathione (via glutathione reductase)

c. respiratory burst in immune effector cells (NADPH oxidase)

d. endothelial function – generation of nitric oxide (NO synthase)

e. hemoglobin catabolism - conversion of heme to biliverdin (heme oxygenase)

G6PD is a housekeeping enzyme constitutively expressed in the cytosol of all cell types, and is found in all living species, including prokaryotes. This includes the *Plasmodia* species, which express a unique variant of G6PD that is conjugated to gluconolactolase (the second enzyme in the PPP), and thus acts as a dual-function enzyme [72]. Human G6PD in its active form is a homodimer or homotetramer of a 56 kDa (515 amino acid [a.a.]) monomer composed of three principal domains [73, 74]:

a. N-terminal domain (a.a. 27-197) – contains the catalytic NADP binding site

b. alpha-helical linker (a.a. 198-206) – location of the conserved substrate (G6P) binding site

c. C-terminal domain (a.a. 207-515) – contains the quaternary interaction domain and the structural NADP binding site (both of which maintain thermodynamic stability)

1.2.1.2 Regulation of G6PD activity

Direct regulation Key allosteric inhibitors of G6PD activity include sex steroid precursor dihydroepiandrosterone (DHEA) [75], and cell cycle master regulator p53, which inhibits G6PD activity by preventing dimerization [76]. Indirect regulation of G6PD activity is provided in the form of processes that alter substrate, product, and/or co-factor concentration. For example, if cellular energy stores are depleted, G6P will be shunted into glycolysis, decreasing G6PD activity, and if NADPH consumption is high, G6PD activity will increase owing to reduced product and NADPH concentration. This latter scenario is especially relevant in the red blood cell (RBC), where basal G6PD activity is at 1-2% of its theoretical maximum [77], but must increase rapidly in response to oxidant stressors, as NADPH is required to regenerate antioxidant glutathione, and G6PD is the only source of NADPH in RBC.

Genetic regulation Regulatory sequence at *G6PD* is best characterized for the promoter region (Figure 1.5). A 1243 bp CpG island, located at chrX:153774894-153776136 on hg19/GRCh37, contains 138 CpG motifs and is a bidirectional site of transcriptional repression via DNA methylation for both *G6PD* and *IKBKG*. Additionally, the promoter contains binding sites for several transcription factors, notably the specificity factors 1 and 3 (SP1 and SP3), as well as cyclic AMP response binding and modulatory proteins (CREB and CREM).

SP1 and SP3 serve to enhance and repress transcription, respectively, with their binding to GC-box motifs (5'-KGGGCGGRRY-3' and its complement) demonstrated via gel-shift assay using core *G6PD* promoter sequence, and their functional role demonstrated by examining chloramphenicol acetyltransferase (CAT) expression driven by the *G6PD* promoter in HeLa cells co-transfected with either SP1- or SP3-expressing plasmids [78, 79].

CREB and CREM compete for binding to the same sequence (5'-CAGAAACAGTATGACGATAGGCAGAT-3'), which is located at chrX:153776348-153776373 on hg19/GRCh37, telomeric to the aforementioned CpG island. Experiments done *in vitro* and *in vivo* using a mouse model suggest that CREM acts as a transcriptional repressor, and may contribute to impaired vascular reactivity via a mechanism that involves aldosterone-mediated upregulation of CREM, which leads to impaired nitric oxide bioavailability [80].

Finally, it is important to note that that bioinformatic predictions and preliminary data from ENCODE (available at UCSC Genome Browser) suggest that *G6PD* may harbor additional binding sites for a variety of important transcription factors including heat shock factor and steroid hormone response binding protein.

With respect to epigenetics, the most important feature is that, as an X chromosome gene, one copy

is subject to transcriptional repression in females via random X chromosome inactivation (XCI). While many genes on chromosome X are able to escape XCI, *G6PD* is not [81], and in each terminally differentiated somatic cell in the female body, only one copy is expressed, leading to a mosaic expression of any G6PD functional phenotypes (discussed further below and in Chapters 5 and 6). While XCI itself is thought to be random, there is some evidence of a skewing of XCI with increasing age [82], and furthermore there exists the possibility of *in situ* natural selection of cells expressing a particular G6PD functional variant.

G6PD in the erythrocyte As the erythrocyte is uniquely dependent on G6PD for NADPH production, the absence of G6PD leads to the inability to regenerate reduced glutathione. Glutathione, a tripeptide which acts as an electron donor in its reduced form, is important in the neutralization of reactive oxygen species, and serves to maintain an appropriately reductive cellular environment. In doing so, peroxidative damage to proteins and lipids is prevented. Impaired protein structure and function can lead to enzyme inactivation or loss of proper cytoskeletal architecture. Lipid peroxidation can lead to alterations in membrane fluidity dynamics and pore formation which can ultimately result in susceptibility to osmotic lysis. Glutathione also plays in important role in maintaining the structural integrity of hemoglobin by maintaining reduced sulfhydryls which when oxidized can lead to the production of hemoglobin precipitates known as Heinz bodies. Thus, it indirectly plays a critical role in the principal RBC function of oxygen delivery.

G6PD outside the erythrocyte G6PD also plays an important role in cells that are not solely dependent on it for NADPH, and this may be because of dose-dependent effects that cannot wholly be overcome by compensatory mechanisms or redundant pathways for generating NADPH.

Vascular reactivity For example, in endothelial cells, NADPH is a necessary cofactor for nitric oxide synthase (NOS), which generates nitric oxide (NO), a membrane-permeable free radical which exerts key local vascular effects important in small vessel homeostasis. It functions in preventing ischemia, as NO induces vascular smooth muscle relaxation ensuring adequate local blood flow. For example, in experiments done *in vivo* in mice, aldosterone, a natural inhibitor of G6PD was found to inhibit vascular reactivity in a dose-dependent fashion [80]. NO also serves in an immunomodulatory capacity, by controlling the expression of receptors such as ICAM-1 that are associated with leukocyte adhesion. Finally, it acts as a local check on coagulation, by inhibiting platelet aggregation. *In vitro* experiments using bovine endothelial cells show that G6PD-derived NADPH may be important in vessel remodeling as well, as deficient G6PD activity in this system limited the ability of endothelial cells to initiate angiogenesis in response to vascular endothelial growth factor [83].

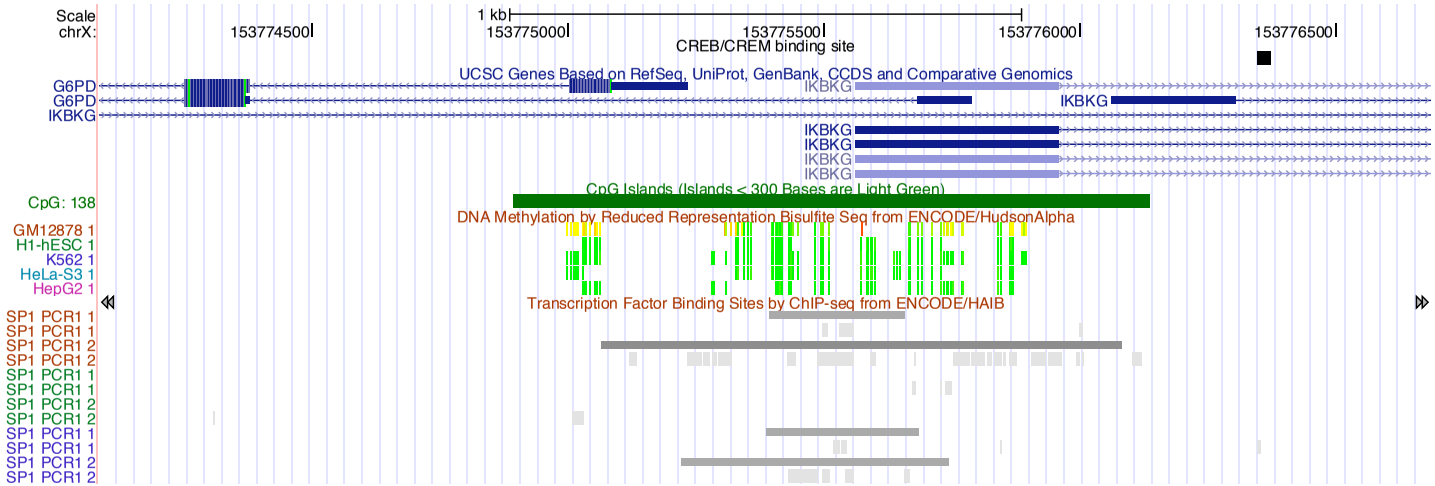


Figure 1.5: Promoter region of *G6PD*. This UCSC Genome Browser graphic shows the CREB/CREM transcription factor binding site (top track, black), which is shared by the *G6PD* and *IKBKG* genes (upper tracks, blue) and located telomeric to a large CpG island (middle track, dark green). Also shown are DNA methylation sites (middle tracks, light green) and SP1 binding sites as determined by ChIP-seq (lower tracks, gray) in different cell lines.

Immune function during infection G6PD also functions in the respiratory burst generated by NADPH oxidase in immune cells such as macrophages and neutrophils. NADPH oxidase produces superoxide, which by itself and in concert with NO produced by inducible NOS found in these cells (the reaction of superoxide with NO produces the peroxynitrite radical), generates oxidant species that generate peroxidative damage of proteins, lipids and nucleic acids of pathogen species both intra- and extracellular. In fact, deficiency of G6PD can lead to an X-linked form of chronic granulomatous disease [84], a disorder characterized by impaired macrophage function leading to increased susceptibility to recurrent infection. However, it is also possible that there may be beneficial consequences to an impaired respiratory burst, as G6PD deficiency has been associated with fewer atherosclerotic lesions in a mouse model of dyslipoproteinemia [85].

There is also limited evidence that G6PD deficiency is associated with decreased production of IL-10 and IFN-γ [86]. While the latter represents a hindrance to launching an effective cellular adaptive immune reponse, the decreased production of anti-inflammatory IL-10 might serve to enhance T cell priming by infected nucelated cells and more generally suppress the induction of a pathogenic systemic inflammatory response. As IL-10 levels are elevated in severe malaria, this may be one aspect of how G6PD deficiency might exert a protective effect.

Breakdown of hemoglobin Finally, G6PD also plays a role in hemoglobin catabolism by providing NAPDH to heme oxygenase (HO), which converts heme to biliverdin, liberating iron and carbon monoxide in the process. Evidence from rodent models of malaria suggest that upregulation of HO is important in reducing the amount of pro-oxidant free heme present in the plasma, which can act as a cytotoxic agent, and thus increased HO activity might therefore be protective against experimental cerebral malaria. [87]. As G6PD deficient individuals are at risk for oxidant-induced hemolysis generally, the impairment of HO activity might serve to exacerbate the effects of an acute hemolytic episode or chronic subclinical episodes.

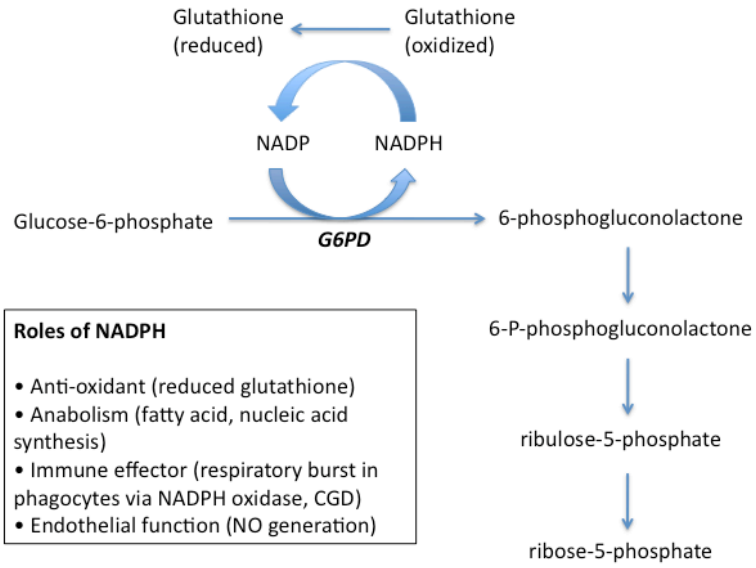


Figure 1.6: The physiologic role of G6PD. Glucose-6-phosphate dehydrogenase catalyses the first step of the pentose phosphate pathway, generating NADPH, a cofactor important to a variety of metabolic processes.

1.2.1.3 Deficiency of G6PD

History Deficiency of G6PD is the most common enzymopathy in man, affecting hundreds of millions worldwide, and is found in regions as diverse as southern Europe, the Middle East, south and southeast Asia, and Africa[88]. Its clinical features have been recognized since antiquity, but it was not until the middle of the twentieth century that it was recognized as a heritable trait. The observation of acute hemolysis in African-American soldiers upon taking the antimalarial drug primaquine catalyzed research on the phenomenon that ultimately led to the identification of G6PD deficiency as the underlying cause [89, 90]. That primaquine-sensitivity was heritable and intrinsic to the RBC itself was demonstrated by transfusion of radiolabeled RBC from primaquine-sensitive individuals into normal individuals, with hemolysis observed only in radiolabeled cells upon primaquine administration [91]. The defect was then localized to G6PD when it was noted that primaquine-sensitive individuals were unable to efficiently reduce oxidized glutathione as a consequence of decreased NADPH production by G6PD [92].

Functional genetics Pedigree studies would reveal that G6PD deficiency was an X-linked disorder, and interestingly, kinetic studies of G6PD activity in the RBC of female heterozygotes would provide the first functional evidence of X chromosome inactivation in humans [93]. Over decades of research, hundreds of mutations in the G6PD gene have been described throughout the world, and several of these have been found to be relatively common with population frequencies of greater than 1% (Figure 1.7). Although initial protein electrophoresis and kinetic studies revealed many spurious variants owing to inter-laboratory variation in methodology, more recent work using DNA sequencing has confirmed that there are well over one hundred distinct genetic variants [94]. Allelic heterogeneity at *G6PD* is characterized by highly variable biochemical and clinical penetrance, graded according to severity by WHO on

a scale from 1 to 5, with severe variants (class I) that cause chronic nonspherocytic hemolytic anemia (CNSHA) being quite rare. The majority of functional polymorphisms described to date are nonsynonymous variants identified by cDNA sequencing that localize to different regions of the peptide sequence.

While many severe (class I) variants localize to peptide sequence corresponding to the dimer interface (a.a. 380-430), for most common variants there is no precise understanding of the molecular cause of the deficiency state, though none appear to be related to catalytic function, and several lines of evidence suggest there may be decreased thermal stability leading to enzyme breakdown, especially in older RBC [95, 96]. One African variant in particular that has been well studied is the 202/A- (rs1050828; C>T; Val68Met) mutant ('202' refers to the cDNA coordinate).

One key early observation was that the G6PD activity of nucleated cells was higher than that of RBC in G6PD deficient individuals, with nucleated cell activity approaching near-normal levels in 202/A- individuals [97], suggesting that there was no primary defect in production or catalytic function of the enzyme. Following on from this, G6PD activity was examined in young versus old RBC, using density gradient fractionation to separate RBC based on *in vivo* age [98, 95]. Here it was found that 202/A- individuals exhibited near normal G6PD activity in their youngest RBC (reticulocytes), but exhibited rates of RBC age-dependent decreases in activity that were much greater than that for wild-type G6PD, suggesting a dramatic degradation or inactivation of enzyme over the course of the RBC lifespan. As the ability of anti-G6PD sera to inhibit the specific activity of G6PD did not differ between wild type and 202/A- enzyme, it was suggested that degradation was the more likely explanation [98]. Further biochemical work has suggested that an increased sensitivity to sulfhydryl oxidation in the 202/A- enzyme might lead to enzyme inactivation by both decreasing affinity for glucose-6-phosphate, but also rendering it more sensitive to thermal inactivation [96].

Interestingly, there is not always a clear relationship between *in vitro* G6PD activity and clinical episodes of acute hemolysis. Some low activity variants like 1360/Union (G>A; Arg454Cys) cause little hemolysis (class 2), while some moderate activity variants like 1180/Alhambra (C>G; Val394Leu) lead to a chronic hemolytic anemia (class 1). One possible explanation for the discrepancy lies in the reaction conditions used in the *in vitro* G6PD activity assay itself, where non-physiological NADPH concentrations are generally employed. As NADPH competes with NADP+ for access to the same binding site, it can thus act to inhibit dimerization and optimal enzyme function. Indeed, this was found to be the case in the aforementioned example, where NADPH was able to preferentially bind 1180/Alhambra over 1360/Union at physiological NADPH concentration, thus inhibiting formation of the active dimer [99].

Apart from the 202/A- variant, there are three other deficiency variants found in Africa and relevant to this thesis, though biochemical studies of these are few. These include the 542/Santamaria (T>A;

Asp181Val; class 2), 680/Mexico City (C>A; Arg227Leu; class 3), and 968/Betica (A>G; Leu323Pro; class 3) variants. Though limited, there is some evidence suggesting that thermal instability may play a role in enzymatic deficiency for the 542/Santamaria and 680/Mexico City mutants [100]. Thermodynamic and kinetic data could not be found for 968/Betica.

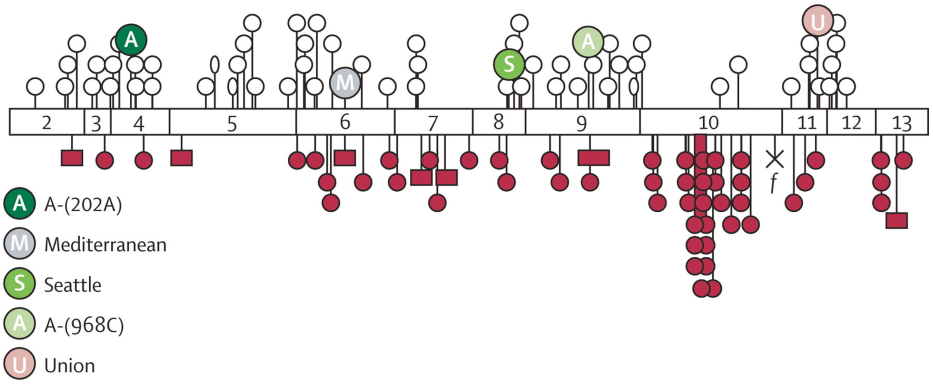


Figure 1.7: Mutations associated with G6PD deficiency. Schematic illustrating the location of nonsynonymous substitutions associated with G6PD deficiency along the coding sequence of *G6PD*, with milder variants (class II, III, and IV) shown atop the exon blocks (white), and the most severe variants (class I) shown below (red). Especially common mutations are indicated by name. Reproduced from Cappellini and Piomelli [73]

Clinical implications While the prevalence of G6PD deficiency trait is high, the clinical penetrance is generally mild, and the vast majority of affected individuals are unaware of their condition [90]. Nonetheless, serious complications can occur, usually as a result of oxidant stress generated by certain foods, medications, or infections. For example, ingestion of the fava bean can rapidly (24 h) lead to hemolysis (favism), as can taking a variety of drugs, including primaquine, as well as certain quinolones and sulfonamides, though drug-induced hemolysis generally develops less rapidly (8-10 days). G6PD deficiency is also associated with blackwater fever, a poorly understood, infrequent complication of malaria treated with quinine (itself not an oxidant) characterized by massive hemolysis and hemoglobinuria culminating in acute renal failure and death [101]. The most common precipitant of hemolysis in G6PD deficiency is infection or febrile illness [73], with common causes being hepatitis, pneumonia, and cytomegalovirus infection. Acute or subacute hemolysis triggered by these environmental insults can result in a profound anemia characterized Heinz bodies (denatured hemoglobin precipitates in the RBC), unconjugated bilirubinemia, and hemoglobinuria, with the resultant ischemia and acute tubular necrosis (caused by hemoglobin casts) leading to acute renal failure [73]. Treatment includes eliminating the offending insult and occasionally transfusion, but the hemolysis is usually self-limiting and only rarely life-threatening.

Neonatal jaundice, characterized by unconjugated bilirubinemia in the absence of anemia or reticulocytosis (i.e. jaundice not secondary to hemolysis), is thought to result from impaired bilirubin conju-

gation in the liver, with risk augmented by prematurity, exposure to oxidant stressors, and Gilbert's syndrome, a common conjugation defect in glucuronosyltransferase activity caused by promoter mutations in *UGT1A1* [102]. In contrast to hemolytic anemia, neonatal jaundice is a life-threatening condition that can lead to permanent global neurologic deficits (kernicterus), and must be treated promptly with phototherapy and exchange transfusion [73].

In certain individuals with chronic nonspherocytic hemolytic anemia (CNSHA), neonatal jaundice may be especially severe and the threat of hemolytic anemia constant. In these patients, who have sporadic mutations of the class I variety, hemolysis is extravascular and chronic, leading to splenomegaly and gallstones, with transfusion often required as maintenance therapy throughout life [103].

G6PD deficiency has also been found to be associated with immune-associated chronic diseases, including diabetes [104, 105, 106] and cardiovascular disease [85, 107, 108]. It has additionally been linked to X-linked chronic granulomatous disease [84] and necrotizing enterocolitis [109]. In animal models, G6PD has been found to play an important role in endothelial function (via NO synthase) with deficiency resulting in impairment of both vascular reactivity [80] and angiogenesis [83].

1.2.2 Malaria protection hypothesis

1.2.2.1 Geographical correlation between malaria endemicity and erythrocyte polymorphisms

In 1948, Haldane first postulated that the high frequency of thalassemia in some populations was due to an increased fitness among heterozygotes conferred by resistance to malaria [110]. This so-called 'malaria protection hypothesis' posited that the global frequency distribution of erythrocyte genetic variants such as those associated with thalassemia, hemoglobin S, and G6PD deficiency was reflective of their geographical coincidence with the historical distribution of malaria (compare Figures 1.2 and 1.8). Given that susceptibility to infectious disease in general is highly heritable [111], and that malaria exerts outsize influence on early childhood mortality in endemic areas, it was thought to be a powerful agent of natural selection in humans [112].

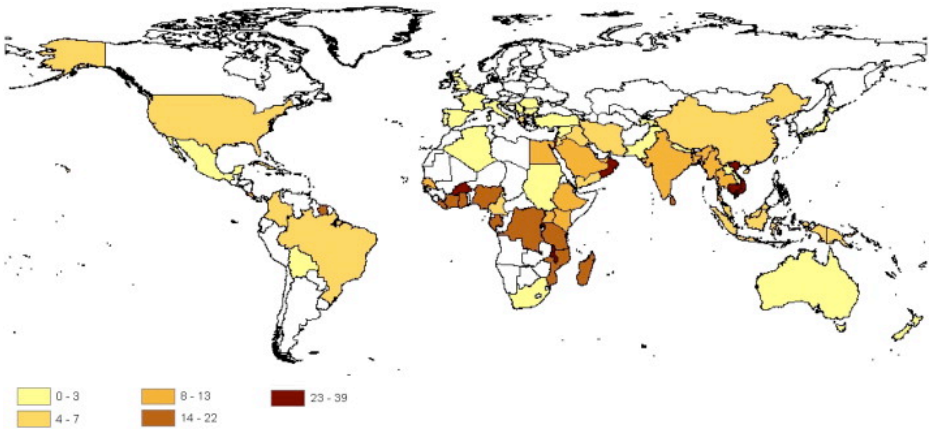


Figure 1.8: Global distribution of G6PD deficiency. As estimated in the meta-analysis by Nkhoma et al. [88], with percent prevalence color-graded according to the scale shown.

Much of the work seeking to prove this hypothesis has focused on geographic evidence in the form of microepidemiological surveys. For α-thalassemia, the evidence has been quite strong, with independent observations in Nepal [113] and Melanesia [114] of strong population differentiation for α-thalassemia allele frequency that follow patterns of historical endemicity for malaria. Furthermore, in one Nepalese ethnic group with allele frequency approaching 80%, seven-fold reduced malaria incidence compared to other groups was observed [115].

By contrast, geographical evidence for G6PD deficiency has been mixed. In East Africa, Allison [116] found relatively low frequencies of G6PD deficiency ($< 3\%$) among highland ethnic groups (Kikuyu and Masai), as compared to seven lowland groups (range: 15% to 27%), as did Flatz and Sringam [117] in Thailand (highland: 0-3%, lowland: 11-18%). However, in Papua New Guinea, Kidson and Gorman [118] found a more complicated pattern– while there seemed to be a general inverse correlation between altitude and frequency, at a finer scale, they found individual ethnic group frequencies that belied this trend, with some groups in holoendemic areas exhibiting frequencies less than 1%, and one highland group with a frequency of 5%. In a similar vein, low frequencies have been reported in malaria-endemic areas of Indonesia (1-3%, [119]), and conversely relatively high frequencies (6-9%) have been reported on the non-malarious islands of Yap and Palau in Micronesia [120].

1.2.2.2 Natural selection at *G6PD*

Population genetics More recent work has focused on using information on the genetic architecture surrounding known G6PD functional alleles to infer population genetic history, i.e. the age of functional variants or the magnitude of associated selection coefficients. For example, Tishkoff et al. found that microsatellite haplotype diversity was reduced on chromosomes bearing Mediterranean and African functional alleles [121], the latter result also corroborated by sequencing results that demonstrated reduced

nucleotide variability on the functional allelic background [122]. Other work has extended these observations, literally, by noting the extreme uniformity of functional allele haplotypes, even when considering polymorphisms hundreds of kilobases away from the variant [123, 124]. This phenomenon can be formalized using long-range haplotype tests which look for high-frequency alleles with unusually high extended haplotype homozygosity (EHH), the idea being that under neutrality, EHH surrounding an allele drifting to high-frequency would fall off with distance due to the effects of recombination and mutation. These tests have yielded convincing results for specific consideration of one African [123] and one Asian [125] variant, but have yet to be detected in unbiased, genome-wide surveys [126, 127].

Balancing selection As discussed earlier, the adverse clinical implications of G6PD deficiency may serve to counterbalance any malaria-protective effects.

Chief among these is neonatal jaundice in the newborn, which recent studies suggest has a case fatality rate between 20 and 40% [128, 129]. Additionally, those infants who manage to survive suffer long-term neurologic consequences; a small study in Kenya found that greater than 40% of such infants were unable to stand or sit up independently [130]. Taken together, the profound morbidity and mortality associated with neonatal jaundice suggests that in early infancy, a net negative selective pressure may operate on G6PD deficiency alleles, as the incidence of malaria in newborns is negligible due to protection conferred by maternal antibodies during the first six months of life.

In early childhood (6 months to 5 years of age), however, when the risk for severe malaria is greatest, a strong protective effect conferred against such life-threatening illness would likely result in a net positive selective pressure on G6PD deficiency alleles.

Finally, in late childhood (10 years of age and higher), when adaptive immunity has fully developed and thus the likelihood of developing severe malaria decreases precipitously, it is likely that no strong selective forces operate, though it is possible that the rates of upper respiratory illness or viral hepatitis in an area, or exposure to other factors known to precipitate hemolytic crises in G6PD deficient individuals, may play a modest role.

What the differential selection in infancy versus childhood suggests is that G6PD deficiency alleles may be subject to a balancing selection process, where the same alleles are first depleted and then enriched in a population, resulting in no strong net directional selection on allele frequency over a short generational timescale. Age-stratified allele frequency data for G6PD deficiency might help clarify the different selective forces at work, as it has for sickle cell trait (HbAS), where an early study in the Congo found that the frequency of HbAS was 0.103 in children under 3 y, while it rose to 0.226 in children between 3 and 10 y [131].

1.2.2.3 Association studies

Association studies offer direct, empirical evidence with which to evaluate the malaria protection hypothesis.

Mild clinical phenotypes Clinical phenotypes associated with asymptomatic infection or uncomplicated malaria include parasite prevalence rate, parasite density, and clinical episode frequency. For example, the earliest work establishing the canonical protective effect for sickle trait (HbAS) found both decreased parasite rate in natural settings and decreased parasite density in experimental infection [132]. Later work has generally corroborated this result, including two large recent studies conducted in Kenya [133] and Ghana [134], each finding lower parasite densities in HbAS individuals, and the former additionally observing a decreased frequency of symptomatic clinical episodes. [132, 133, 134]

Studies examining the association of mild malaria clinical phenotypes with G6PD deficiency, however, have provided little consensus. While Allison and Clyde found slight differences in both *P. falciparum* rate and density in Tanzania [135], most later work has not replicated this finding, including studies in the Congo [131], Nigeria [136], Mali [19], and Thailand [137, 125]. More vexingly, the studies from Thailand also examined association with *P. vivax* density and found diametrically opposite results, with one observing an increased parasite rate [137] and the other decreased parasite density [125] in G6PD deficient individuals.

Severe malaria phenotypes Mild malaria phenotypes are poorly correlated with disease burden, and are thus of questionable clinical relevance and also less pertinent to discussions of evolutionary fitness and the malaria protection hypothesis. More important to consider are those phenotypes associated with severe malaria, whether binary (e.g. case-control status) or continuous (e.g. serum lactate levels), and more specifically, those features associated with particular clinical subphenotypes. As severe malaria itself is quite rare (less than 1% of all cases), conducting an adequately-powered study is quite costly both in terms of time and resources, but such work offers the potential to reveal effects that might be subtle or even unobservable in studies of mild malaria.

Large case-control studies examining variant hemoglobins S and C have revealed protective effects of even greater magnitude than those found in studies of uncomplicated malaria phenotypes, and evidence of protection has been consistent across independent investigations conducted in several different regions, including Nigeria [138], Ghana [134], the Gambia [139], Mali [18], Burkina Faso [140], and Kenya [141]. The relationship between G6PD deficiency and severe malaria is less clear, however, for reasons both methodological and biological. Past studies have differed with respect to clinical definitions used and also in how they define and type the G6PD deficiency trait, with the latter also being of biological relevance in that *G6PD* is an X-linked gene (unlike the globins), which complicates interpretation of effect size

and statistical significance in gender-stratified analyses.

The earliest association study of severe malaria was performed in Nigeria by Gilles et al. [138], who specifically investigated cerebral malaria (200 controls, 100 cases), and found a strong protective effect in both males and females, estimating that G6PD deficient individuals overall had about 70% reduced risk of developing CM.

Two later studies also considered specific clinical subphenotypes (SMA and CM), including one conducted by Ruwende et al. [39] in the Gambia (421 controls, 556 cases [174 SMA, 382 CM]) and Kenya (292 controls, 250 cases [115 SMA, 135 CM]), and another by Guindo et al. [19] in Mali (2765 controls, 271 cases [223 CM, 48 SMA]). Both studies assessed trait carriage via genetic typing of a major African variant, G6PD202 ("A-"), and a pooled meta-analysis of both studies [19] presented strong evidence for protection against CM ($OR = 0.44$, $P = 0.002$), but not SMA ($OR = 0.70$, $P = 0.48$). Consistent with this finding, association with severe malaria as a whole met statistical significance ($P < 0.05$) only in those countries where CM comprised a majority of cases (The Gambia = 69% and Mali = 82%).

However, both studies should be viewed cautiously in light of particular methodological issues relating to genotyping. In the study done by Ruwende et al., the frequency estimate in the Gambia for the 202T allele ($DAF = 0.059$) may be inflated relative to two recent published reports suggesting the frequency is much lower in the Gambia and Senegal (between 1-3%), which may explain why their enzyme activity data showed only a modest reduction in G6PD activity for 202T hemizygotes of 50%, as opposed to 80-90% that has been observed by other investigators. And in the work presented by Guindo et al., where it is claimed that protection is conferred only to males, the observed female genotype distribution for controls at G6PD202 (TT: 40, CT: 148, CC: 1011) is grossly inconsistent with frequencies expected under an assumption of Hardy-Weinberg equilibrium ($P = 6.6 \times 10^{-22}$, χ^2 test), which might suggest cryptic population structure or systematic genotyping error. As the typical significance threshold in genome-wide association studies is $P < 10^{-7}$, their claims with respect to gender-specificity may not be accurate.

Guindo et al. are not the only investigators to claim a sex-specific protective effect. Indeed, while they espouse a male-specific protective effect, others have claimed the exact opposite. Following on from the suggestions of Haldane and Allison regarding heterozygote advantage for thalassemia and sickle cell trait, Luzzatto [142] and Bienzle [136] sought to explain balancing selection for an X-linked trait by positing that the heterozygote state, present only in females, was associated with decreased parasite rate and density. In addition to using response variables poorly correlated with disease burden, the specific conclusions drawn in the Bienzle et al. study are incongruous– in males, the nondeficient A allele is found to be protective, and in females, only one type of heterozygote (A-/B but not A-/A) is protected.

Given that several other studies have found no significant difference in the protective effect in both males and females [138, 39, 143], it is possible that sex-specific differences in association at *G6PD* have been overstated, as has been the case for many other genes [144].

1.2.2.4 Laboratory studies

Experimental work can offer not only further indirect corroboration of the malaria protection hypothesis, but also insights into the cellular and molecular mechanisms underlying protection, such as those affecting parasite fitness and virulence or those concerning host adaptation to infection.

Parasite fitness and virulence An obvious starting point from which to investigate mechanisms of protection conferred by the hemoglobinopathies and G6PD deficiency is to look at how the variant RBC itself impacts parasite biology. Perhaps the simplest metrics to consider in this regard are invasion efficiency and growth rate. Experiments using sickle trait RBC have been informative in this regard, as investigators have found impaired invasion and growth of parasites in AS erythrocytes [145, 146], and also increased propensity for infected AS cells to sickle [147], which would serve to enhance splenic clearance. Similar experiments have been conducted using G6PD deficient RBC, but have yielded mixed results, with some researchers finding evidence of invasion preferences [142] or aberrant *in vitro* growth phenotypes [148], and still others not [149, 150]. Some have even suggested that the parasite upregulates its own G6PD in response to deficiency in the host RBC [151], but this was later observed to be a methodological artifact by the same group [152]. In general, the ambiguity in the invasion/growth rate data mirrors that found in the mild malaria clinical data where neither *in vivo* parasite density nor the risk of asymptomatic malaria significantly differ between G6PD deficient and normal individuals. Overall, this suggests that unlike with sickle trait, protection may not be associated with parasite survival inside the variant RBC.

More important to consider are those virulence phenotypes that impact the syndromes of severe malaria, including adherence phenotypes associated with the PfEMP1 parasite variant antigen family [51]. Two such phenotypes include rosetting, or the ability of infected RBC to bind uninfected RBC, and endothelial cytoadherence, which refers to the ability of infected RBC to bind to endothelial cells. Both rosetting and cytoadherence are associated with clinical severity, and, more specifically, the pathogenesis of cerebral malaria via mechanisms thought to involve changes in local blood flow, hemostasis, and endothelial activation [62, 30]. Impaired rosette formation has been observed for infected blood group O RBC *in vitro* [153], as has impaired cytoadherence in HbAS and HbAC RBC [154, 155]. Similar experiments have not yet been reported for G6PD deficiency, but would be of great interest given their demonstrated utility in elucidating key mechanistic aspects of protection for these other erythrocyte polymorphisms.

Host adaptation The age-dependent acquisition of adaptive immunity significantly reduces the risk of severe malaria, and there is some evidence that suggests that variant RBC may modulate this process. In sickle trait, for example, it has been observed that protection from symptomatic infection increases with age until 10 y, which likely indicates that HbAS accelerates the development of acquired immunity [156]. This is quite plausible given that infected HbAS RBC sickle more easily and are thus more susceptible to splenic clearance and presentation to antigen presenting cells. As ring-infected G6PD deficient RBC are more susceptible to phagocytosis *in vitro* [150], it is possible that G6PD may also modulate host immunity in a similar fashion, though this has yet to be demonstrated *in vivo*. Additionally, non-immune mechanisms of host adaptation, such as tolerance to infection, as demonstrated recently for sickle trait with respect to heme catabolism [87], may prove to be important as well.

1.2.2.5 Refining the hypothesis

The picture that emerges from decades of research on G6PD and the malaria protection hypothesis is one rife with incongruity. A closer look at the underlying assumptions and evidentiary support reveals that biases and hyperbole have shaped the narrative over a half-century.

The malaria protection hypothesis overall suffers from experimental bias in its focus on the erythrocyte. When Haldane first presented his hypothesis in the middle part of the twentieth century, hematology research was at its zenith (Figure 1.9), as blood was an easily obtainable source of tissue for protein biochemistry, and screening for hemoglobin and G6PD variants was facile with the advent of new gel electrophoresis techniques [90]. Of course, it intuitively makes sense that the red cell would be a strong target for natural selection by malaria, and the near complete absence of Duffy antigen on RBC in sub-Saharan Africa [112] is almost certainly due to strong protection conferred against *P. vivax* (Duffy antigen is an important invasion receptor [11]), but our modern understanding of the pathogenesis of malaria suggests that the development of mild and severe disease involves more than simply parasite-specific effects in the RBC. And the fallacy of focusing on traits that are highly differentiated between malarious and nonmalarious regions is demonstrated by consideration of the ABO blood group. The protective effect of blood type O is fairly well-established both epidemiologically [157] and mechanistically [153], and yet it is not highly differentiated between populations based on malaria endemicity.

Figure 1.9: Experimental bias towards the erythrocyte. The ratio between citation records found in Google Scholar for RBC specific articles (search terms: 'RBC', 'erythrocyte,' 'red cell,' and 'red blood cell") versus all cell biology articles as a whole (search term: 'cell'), is shown for each decade beginning between 1920 and 2000.

Additionally, while the hypothesis arose out of a desire to explain the high frequency of certain RBC polymorphisms in historically malaria-endemic regions, it is worth considering explanations other than selection. It is first important to distinguish between the high prevalence of G6PD deficiency as a biochemical trait and the high frequency of a particular allele associated with G6PD deficiency. In an area of relative allelic homogeneity, these two are the same, but in many places, such as West Africa, southeast Asia, and Oceania, there may be multiple alleles that contribute to prevalence of deficiency trait, each individually of modest or low frequency [103]. It is easy to overstate or misinterpret the prevalence of G6PD deficiency, when the locus produces a phenotype that is easily scorable, and may simply have an elevated rate of mutation, and/or an ability to tolerate new mutation well (i.e. produce a molecular phenotype of limited clinical penetrance). Additionally, the impact of demographic change is often overlooked, especially recent change occurring within the last 10000 years. For example, if a high frequency allele is present in one ethnic group, and that group rapidly expands in size and geographic range through displacement of other groups, the effects of genetic drift are enhanced, and an allele can rapidly rise to high frequency over a large area by chance [158]. This possibility should be considered when considering that some of the highest frequencies of the 202/A- G6PD deficiency allele are found near the anthropological epicenters of the Bantu expansion in Africa, near Nigeria and the Congo [122]. An additional point that has bearing on discussion of G6PD allele frequency concerns balancing selection.

If the protective effect is indeed as strong as has been suggested (nearly a 50% reduced risk of severe malaria, see above), then a similarly strong deleterious effect should also exist to produce a condition of balancing selection, which does not comport with the relatively mild adverse clinical penetrance of G6PD deficiency trait, where most G6PD deficient individuals are unaware of their status (in contrast to sickle cell anemia).

Finally, it is worth examining whether the strength of evidence for protection is consistent with a process of natural selection. As mentioned earlier, the most useful studies to consider are case-control studies of a clinical phenotype with significant morbidity and mortality, i.e. a phenotype that might have some bearing on fitness. To date, the only evidence demonstrating protection in an association study of severe malaria comes from studies in only one part of the world– West Africa (see above). More evidence from diverse populations with multiple independent functional alleles is needed to confirm the hypothesis and definitively demonstrate a causal role for G6PD enzyme deficiency itself, as opposed to a merely associative role.

1.3 Disease and diversity – a genetics revolution

1.3.1 A map of human genetic diversity

1.3.1.1 Sequencing the genome

The Human Genome Project (HGP) began in 1987, and published a draft version of the *H. sapiens* genome in 2001 [159], providing researchers with an atlas for the blueprint of life and setting the stage for a decade of remarkable advances in genomic science and biomedicine. As monumental an achievement as it was, the HGP draft contained sequence data from only four individuals, and is more appropriately viewed as the necessary scaffold upon which a more plastic and complex picture of the human genome could be revealed. Indeed, the human genome is not a single genome– it is six billion genomes, found in individuals with unique and shared characteristics, some observable and some not, many of which contribute significantly to disease susceptibility.

1.3.1.2 Cataloguing genetic variation across the genome

The International HapMap Consortium began in 2002 as an effort to complement the HGP and build a map of worldwide human genetic diversity that could then be marshaled to identify the genetic components of common diseases, like cardiovascular disease and diabetes [160]. The 'common variant - common disease' hypothesis holds that susceptibility to common disease has a strong genetic component consisting of the additive effects of many common variants, and as such, susceptibility to common disease can be treated as a quantitative character following an approximately normal probability distribution as given by the central limit theorem. The HapMap project surveyed global genetic variation by random and targeted resequencing of 269 individuals from four populations of relative genetic homogeneity, revealing

more than 6 million novel single nucleotide polymorphisms (SNPs) in its first release [161]. This contribution represented a significant increase in both the depth and breadth of genetic diversity known, as the public database for SNP data (dbSNP) contained only 2 million SNPs at the time, half obtained from HGP, and the other half via partial genome sequencing of only 24 individuals performed by The SNP Consortium [162].

1.3.1.3 Next generation sequencing

With the development of high-throughput sequencing modalities, our knowledge of genome-wide sequence and structural variation has continued to grow at an ever increasing pace [163]. While sequencing of the 3 billion base pairs of the HGP genome took well over a decade and cost billions of dollars to complete, obtaining a full human genome sequence today is commonplace, fast (2-3 weeks), and cheap (less than $10000 USD) [164]. New consortia such as 1000 Genomes Project [165] have now generated sequence data from the genomes of thousands of individuals, using massively-parallel sequencing methods such as 454/Roche pyrosequencing, ABI SOLiD ligation-based sequencing, and Solexa/Illumina sequencing-by-synthesis. Today, it is estimated that there are nearly 40 million SNPs that have been identified in the human genome (roughly 99% of all predicted variants), and full sequence data and/or dense genome-wide genotype data for thousands of individuals is now publicly available [166].

1.3.2 Linking genetic variation to biological function

1.3.2.1 Genome wide association studies

Efforts by these consortia have also revealed how alleles on a chromsome are associated with one another in allelic groups called haplotypes, with the strength of association between SNPs quantifiable as linkage disequilibrium (LD). The observation that the genome comprised blocks of high LD separated by defined recombination hotspots [167], suggested that at most genomic loci, a few common haplotypes could explain most of the genetic variation. Genome-wide, then, a group of hundreds of thousands of SNPs that tagged these haplotypes (Figure 1.10) could efficiently describe most human genetic variation [168]. The development of SNP chips that could type these hundreds of thousands (and now, millions) of variants, along with the appropriate analytical tools, allowed for the promise of HGP and HapMap to be realized in the form of genome-wide association studies (GWAS) on common diseases such as coronary artery disease and bipolar disorder [169], and traits such as body mass index [170] and height [171]. Like the genomic projects, these GWAS studies have generally been developed by consortia involving multiple centers working together to combine large data sets, generating increased statistical power in the process. To date, well over 1000 loci associated with a variety of common diseases and traits have been described (in addition to nearly 3000 Mendelian disorder-related genes), and have illuminated new biological pathways involved in disease, which may provide new leads for drug development [166].

The implication of the massive influx of next-generation sequencing data for GWAS and association

studies in general is that the level of resolution to which an association signal can be resolved is increased, as is the power to detect the effects of rare variants. At present, this increased power is achieved through a statistical procedure called imputation [172], where dense sequence or genotype data from a subset of individuals genetically-related to a particular case-control study population allows researchers to guess at, or impute, genotypes at SNPs that weren't typed in the case-control group, using haplotype relationships between typed and untyped SNPs (as inferred from the imputation panel, in which all SNPs are typed). In the future, when it becomes cost-effective, imputation-based methods will most likely be supplanted by whole-genome sequencing of all individuals in an association study, as the full picture of both sequence variation and structural variation (insertions, deletions, inversions, etc.) is superior to data only from SNPs.

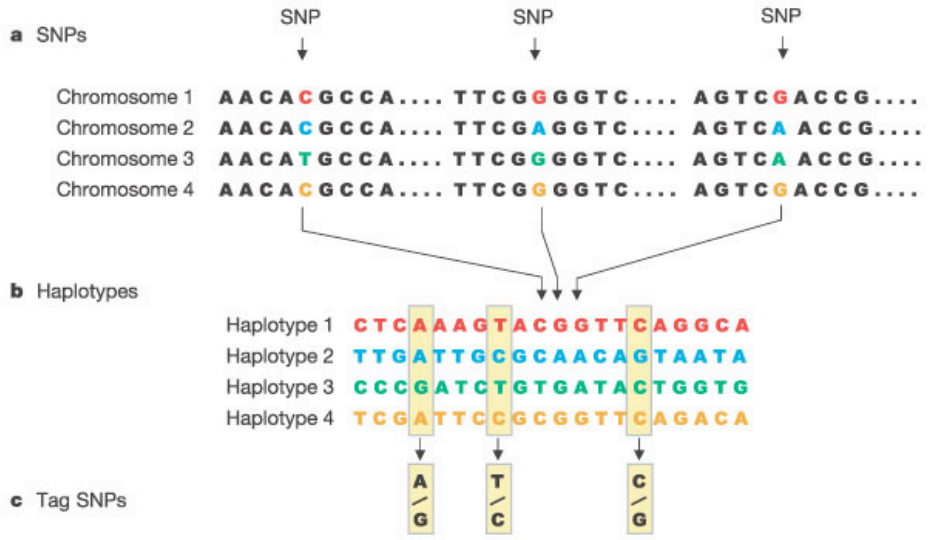


Figure 1.10: Principle of GWAS. If four haplotypes (b) represent the genetic variability space along the length of a chromosome (a), the information content present across the haplotypes can be efficiently represented by only three tag SNPs (c). Reproduced from the International HapMap Consortium [160].

1.3.2.2 Identifying functional genetic variants associated with disease

Fine-mapping association studies One challenge in the current GWAS era concerns how to follow-up on putative hits in an initial association scan. The process of precisely resolving an association signal at a locus is called fine-mapping, and involves surveying genetic variation at a locus via resequencing, and then association testing of as many variants discovered as possible in an attempt to confirm and resolve the signal. The underlying assumption is that if a single causal variant exists, there should be one or more SNPs typed that are in comparatively stronger LD with it (relative to other SNPs at the locus) and thus exhibit stronger signals of association.

The first step in fine-mapping, resequencing of the candidate locus, can be prohibitively expensive depending on the length of the region to be interrogated and the sensitivity to detect rare variants desired. Next generation sequencing technologies have the potential to decrease this cost, but offer sequence output designed for unbiased sequencing of entire genomes on the order of megabases and gigabases. A candidate locus is more often on the scale of kilobases, however, which makes next-generation sequencing inefficient in terms of both cost and information content (e.g. sequencing one individual's G6PD gene [15 kb] via Illumina would yield data with sequencing depth on the order of millions of reads and cost the same as sequencing the entire *P. falciparum* genome [23 Mb] at 10x depth). Until the cost of sequencing is reduced significantly, the development of target enrichment or sample pooling strategies will be critical to making fine-mapping experiments feasible.

In the actual fine-mapping association scan itself, there are also a number of challenging issues that have bearing on the ability to resolve signal from noise. One is the presence of allelic heterogeneity, where multiple alleles at a single gene may confer independent or even additive risks for disease susceptibility, any one of which may have only a small effect, but in concert, may exert quite a large effect. Another problem concerns the patterns of LD around a locus of interest in different populations. For example, in African populations, where LD between markers falls off rapidly with physical distance, the power to detect any single causal variant in a GWAS is low if marker density across the genome is low, but this lack of long-range LD may prove useful in precisely identifying a causal variant when marker density is sufficiently high (e.g. in fine-mapping), as has been demonstrated for the SNP associated with HbAS in West Africa [173, 174].

Choosing relevant phenotypes Another challenge in an association study, especially an expensive and time-consuming GWAS, is deciding on the appropriate clinical and molecular phenotypes to test. With respect to clinical phenotypes, this involves choosing precise and useful definitions for case and control groups that allow for maximal power to detect a genetic association. A control group may be a representative sample of the population at large, matched for possible confounding variables (e.g. age, sex, town) to the cases. Alternately, one can gain slightly more power by choosing a control group that is directly opposite to the case group, i.e. a group in whom disease prevalence is less than in the general population. Case definitions must be similarly precise, and should avoid subjective criteria that might vary between clinicians and study sites. When feasible and appropriate, clinical subphenotypes should also be examined, as different manifestations of the same ostensible disease may involve varying pathophysiological processes with unique genetic determinants.

A further adjunct to an association scan based on case-control status, involves consideration of molecular or cellular phenotypes, both quantitative and qualitative. Such phenotypes might include fasting lipid levels in a study of coronary artery disease, or fetal hemoglobin levels in a study of thalassemia. These

studies of disease-associated molecular phenotypes offer the potential to more precisely elucidate the mechanism of protection (e.g. if a single SNP affects both susceptibility to myocardial infarction and levels of LDL), but also can reveal genetic modifiers of important molecular phenotypes that might go undetected in an association scan of case-control status, but still have biological relevance.

1.4 Using new tools to address outstanding questions about G6PD and malaria

This dissertation is broadly concerned with understanding genetic diversity at the *G6PD* locus and how that diversity is translated into variation at the level of molecular and clinical phenotypes, especially those relevant to severe malaria. I bring to bear modern genetic tools in investigating age-old questions surrounding the malaria protection hypothesis. Gaining a better understanding of host protective polymorphisms and malaria pathophysiology generally are prerequisites to the creation of rational drugs and vaccines, each critical to adequate management of malaria in the future.

1.4.1 Defining global genetic variation at G6PD

I begin in the first two results chapters with a description of how I assessed global variation at *G6PD* and developed a new methodological approach to polymorphism discovery.

- **Chapter 3** – I begin by describing the process of variant discovery and validation in twelve populations in Africa and southeast Asia, and the merging of that data with information from the literature and public resources to construct a database of known and novel G6PD variants.
- **Chapter 4** – Next, I discuss the development of a new tool for population-wide SNP discovery, harnessing the throughput of next-generation sequencing via sample pooling.

1.4.2 Understanding variation in G6PD molecular phenotypes

In the next two chapters, I consider how genetic variation at *G6PD* translates into biology at a molecular level.

- **Chapter 5** – Here, I consider many of the variants discovered and validated earlier, and test each in turn for association with G6PD enzyme activity in Kenya, looking specifically for genetic modifiers of quantitative G6PD enzyme activity and qualitative G6PD deficiency risk.
- **Chapter 6** – This chapter is largely methodological and concerns the development of a novel technique which enables high-throughput typing of enzyme deficiency in individual RBC using fluorescence.

1.4.3 Fine-mapping association study of severe malaria in the Gambia

Finally, I consider whether genetic variants at *G6PD* are associated with severe malaria.

- **Chapter 7** – In the final results chapter, I present a fine-mapping association study of severe malaria and its constituent clinical subphenotypes in the Gambia, in populations exhibiting known allelic heterogeneity at the locus.

Chapter 2

Materials and Methods

2.1 Overview of Key Methodology

2.1.1 Massively parallel sequencing

The massively parallel sequencing (MPS) work was performed on the Illumina/Solexa platform [164] (Supplemental Methods B). Several MPS technologies exist [175], and each has notable advantages including the pyrosequencing-based 454/Roche platform, which can generate read lengths of several hundred base pairs (bp) in length, and the ligase-based ABI/SOLiD platform, which exhibits enhanced accuracy (less noise because two bases are interrogated per ligation reaction) and the ability to sequence homopolymer tracts. I chose Illumina principally for its high throughput (hundreds of megabases of sequence generated per run) and ease of sample preparation (the other two platforms involve a difficult technique called emulsion PCR).

In brief, input DNA was sheared, size-selected via electrophoresis, and ligated to adapter sequences, which then acted as primer templates for subsequent amplification reactions and as mate-partners for binding to adapter sequences on the flow cell surface. Further isothermal amplification performed in situ generated millions of isogenic clusters, each of which produced a single sequencing read. As illustrated in Figure 2.1, each sequencing cycle involved a single base pair elongation step, followed by image acquisition via CCD camera, and finally, dye and terminator cleavage to allow the next cycle to occur. This process can repeat itself for anywhere between 35 to 150 times, generating millions of sequence reads of 35-150 bp (at the time of our experiment, reliable single-end cycle/read length was 36 bp, due to nascent strand asynchrony and/or misincorporation errors). Sequencing library preparation was performed according to a modified Illumina protocol [176] by M. Quail and co-workers at Wellcome Trust Sanger Institute.

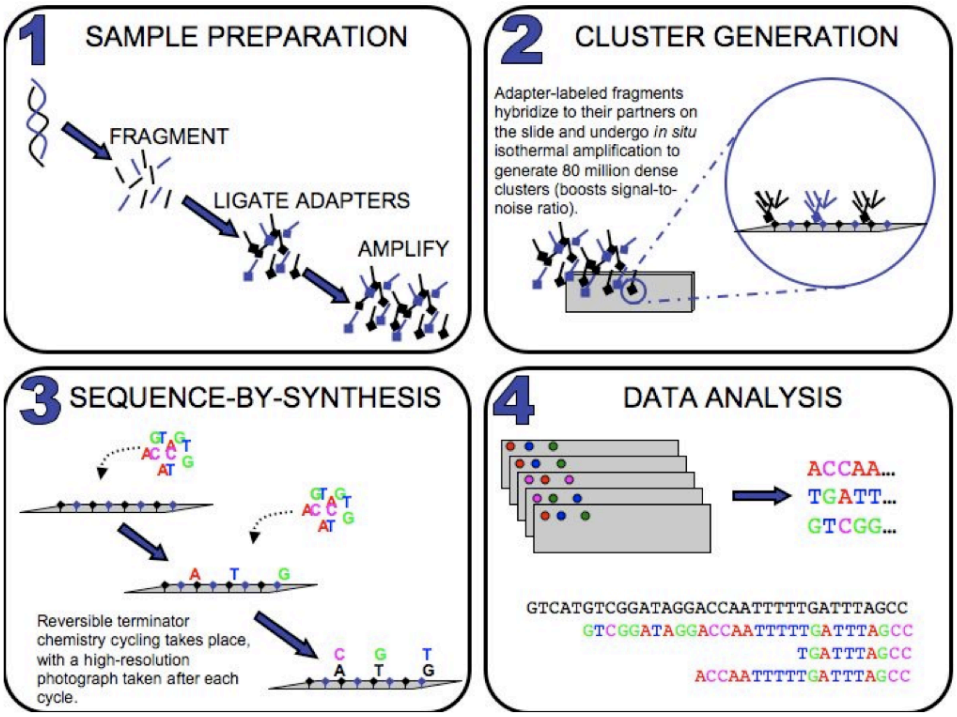


Figure 2.1: Schematic of MPS on the Illumina/Solexa platform.

Technical issues that arise with the Illumina platform are numerous, but for the present purpose, important considerations relate to the library preparation and the sequencing chemistry. The former is important in that adaptor ligation and gel extraction during size selection can both be biased towards enriching high GC content sequence, as adaptor ligation and gel extraction columns both favor dsDNA which is thermodynamically more favored with higher GC content [176, 177]. Sequencing chemistry limitations affect quality of base calls after several sequencing cycles, and thus, read length. Three steps must happen in parallel for all oligonucleotides on a cluster: (1) removal of the reversible terminator, (2) removal of the fluorophore, and (3) incorporation of a new nucleotide. As this sequence of processes becomes more and more asynchronous, signal to noise at a cluster decreases with increasing cycle number, and limits the quality of longer reads. The short read length, in turn, also limits the ability to do *de novo* sequencing of a genome, and map reads to low complexity reference sequences.

2.1.2 Multiplexed genotyping

The mass spectrometry-based Sequenom iPLEX platform [178, 179] was used for multiplexed genotyping reactions (Supplemental Methods: Chapters 3, 5, and 7). Manufacturer's instructions (http://www.sequenom.com) were followed except that the template DNA was first prepared by whole-genome amplification via primer extension preamplification [PEP] reaction [180] on a thermal cycler (MJ Research) using 96-well plates (Greiner) as shown in Table 2.1.

Table 2.1: Primer extension pre-amplification reaction. Reactions were performed using *Taq* polymerase (Bioline) and N15 random hexanucleotide primers (Operon) in a reaction volume containing: 2.5 μl of 10x reaction buffer (Bioline), 0.25 μl *Taq* polymerase (5 U/μl), 0.6 μl dNTPs (8 mM), 1.25 μl Mg^{2+} (50 mM), 1.1 μl N15 primers (200 μM), 5 μl template DNA (1 ng/μl), and 14.3 μl H_2O.

Denaturation	
94°C	03:00
Main Cycling (50 cycles)	
94°C	01:00
37°C	02:00
+0.1°C/sec until 55 °C	
55°C	04:00
End	
72°C	06:00

The PEP product served as the template for multiplexed PCR reactions in Thermofast 384-well plates (ABgene), conducted using HotStarTaq polymerase (Qiagen) and standard PCR conditions (Table 2.2). Unincorporated dNTPs were subsequently hydrolysed by shrimp alkaline phosphatase (SAP) treatment (Table 2.3).

Table 2.2: iPLEX PCR conditions.

iPLEX PCR Reaction Components	*µl per well*
multiplex primer cocktail (100µM)	0.68
$MgCl_2$ (25mM)	0.65
dNTPs (25mM)	0.2
buffer (10x)	1.25
H_2O	2.82
Taq (5 U/µl)	0.4
Master Mix Volume (µl)	6
DNA (1 ng/µl)	4
Total Volume	10

iPLEX PCR Thermal Cycling Protocol		*T (°C)*	*Time*
	Activation	94	15:00
45x	Denaturation	94	0:20
	Annealing	56	0:30
	Extension	72	1:00
	Final Extension	72	3:00
	End	15	15:00

Table 2.3: iPLEX SAP conditions.

iPLEX SAP Reaction Components	*per well (µl)*
H_2O	1.53
buffer (10x)	0.17
SAP (1.7 U/µl)	0.30
Master Mix Volume	2.00
PCR Product Volume	5.00
Total Volume	7.00

iPLEX SAP Thermal Cycling Protocol	*T (°C)*	*Time*
Digestion	37	40:00
Inactivation	85	05:00
End	15	15:00

These PCR products then served as templates for the single-base extension reaction that interrogated the variant alleles, a reaction that employed modified dideoxy-NTPs (ddNTPs) that prevent elongation at the 3' end of an oligonucleotide, and a unique thermal cycling protocol with a nested internal loop designed to maximize extension efficiency (Table 2.4).

Table 2.4: iPLEX extension reaction conditions.

iPLEX Extension Reaction Components	
	per well (μl)
H_2O	0.66
buffer (10x)	0.2
iPLEX termination mix - ddNTPs	0.2
multiplex primer cocktail (10 μM)	0.9
Taq	0.04
Master Mix Volume	2
Post-SAP PCR Product Volume	7
Total Volume	9

iPLEX Extension Reaction Thermal Cycling Protocol			
		T (°C)	*Time*
	Activation	94	0:30
40x [	Denaturation	94	0:05
5x [	Annealing	52	0:05
	Extension	80	0:05
	Final Extension	72	3:00
	End	15	15:00

The reaction mixture was then desalted via cation-exchange resin, spotted onto a SpectroCHIP preloaded with acidic matrix (3-hydroxypicolinic acid) via nanodispenser, and subjected to MALDI-TOF mass spectrometry. Finally, the mass spectra generated were translated into genotypes by examining the ratio of the peaks associated with either allele at a biallelic site. These cluster plots were further curated and re-annotated manually at each SNP for each plate using Spectro-TYPER software.

Limitations of the Sequenom technology chiefly revolve around multiplex design and SNPs that exhibit poor genotype clustering. For this study, multiplex design was a stepwise process (see Supplemental Methods: Chapter 3), where priority-ranked lists of SNPs were sequentially input to design software (Spectro-DESIGNER) to allow for the creation of multiplex reactions with, ideally, a high number of SNPs of high priority that were likely to yield high performance quality. The most obvious difficulty is that sometimes a SNP of interest will not be retained in the multiplex-fitting algorithm. Additionally, highly polymorphic sequence becomes problematic with respect to finding appropriate PCR and extension primers, as variation present at primer binding sites decreases the efficiency of PCR and extension reactions due to poor primer binding. This can lead to genotyping errors and patterns of nonrandom missingness for individuals with SNPs adjacent to a particular interrogated SNP. To circumvent this problem, a custom Perl script and an online Sequenom tool were used to mask the SNP flanking sequence for any polymorphisms present, using variants present in either dbSNP or my in-house G6PD database (see Chapter 3). Finally, after the wet lab work is complete, cluster membership must be assigned based on mass spectra, and low signal strength resulting from inefficient extension can lead to inaccurate or insensible genotype calls.

2.1.3 G6PD enzyme activity assay

The assay referred to in Supplemental Methods: Chapters 5 and 6 is a modified version of the 'macro' assay given in Randox Laboratories G6PD kit (PD 410), with positive (PD 2618) and negative (PD 2617) controls used where necessary. The general method is given here, though the specific volumes used in Supplemental Methods: Chapters 5 and 6 can be found in the relevant sections. It is possible to substitute RPMI 1640 for the Randox reaction buffer (Reagent R1).

Whole blood (10 ul) was washed three times in 100 ul normal saline (0.154 M NaCl in distilled water) with centrifugation at 3000 rpm for 5 minutes. Washed RBC (50% Hct, 10 ul) were lysed by incubation with 20 ul of digitonin (Reagent R4 in PD 410) for 15 minutes at 4°C, with the hemolysate isolated via centrifugation at 3000 rpm for 10 minutes at 4°C. Five ul of hemolysate was incubated with 300 ul of buffer (Reagent R1, 31.7 mM triethanolamine, 3.2 mM EDTA in H_2O, pH=7.6) and 10 ul μl NADP+ (Reagent R2, 0.34 mM in H_2O) at 37°C for 5 minutes. Finally, 5 ul of glucose-6-phosphate (Reagent R3, 0.58 mM in H_2O) was added and spectrophotometry readings were taken, with duplicate measurements taken for each sample. Catalytic rate was assessed by the formation of reaction product NADPH, which absorbs light at 340 nm (Figure 2.2).

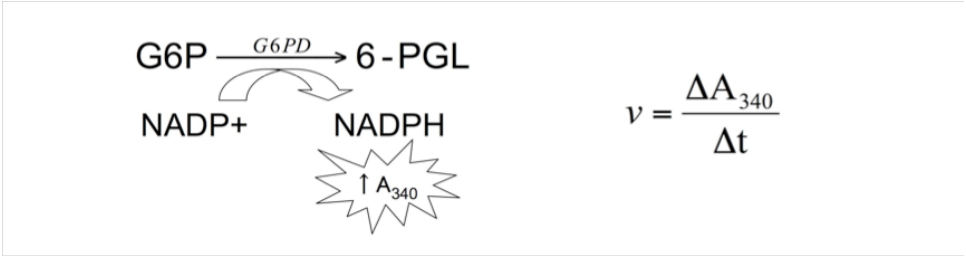


Figure 2.2: Schematic of G6PD enzyme activity principle.

The change in absorbance at 340 nm over time was measured using a spectrophotometer, with readings taken every 10 seconds for at least 3 minutes (see Supplemental Methods: Chapters 5 and 6 for exact times used) and activity normalized to hemolysate hemoglobin content by recording absorbance at 540 nm:

$$\text{G6PD activity} = \frac{\frac{\Delta OD340}{\Delta t}}{OD540} \tag{2.1}$$

Factors that can artificially inflate the activity measured include high reticulocyte or white blood cell content, which underlines the importance of using healthy controls (anemia can lead to secondary reticulocytosis) and thorough RBC washing procedures, respectively. Additionally, aberrant low activity measurements can occur with older blood, so all testing was done within 3-5 days after collection. As G6PD exhibits catalytic cooperativity, and degradation of active enzyme can occur over time, the use of older blood samples might artificially amplify differences between, for example, normal individuals and heterozygotes.

2.1.4 Data analysis pipelines

For all genotyping studies conducted, similar methods were used to process, validate, and analyze the data after genotypes were downloaded from the MalariaGEN database. These methods generally included some combination of custom scripts written in the Perl and R [181] programming languages, along with software packages designed by others, including PLINK [182] and snpMatrix [183].

2.1.4.1 Housekeeping and data reorganization in Perl

The Perl programming language was used to perform some basic housekeeping and quality control on the data, as Perl is known for its ease of text-processing and ability to interface with object-oriented bioinformatic tools such as the BioPerl module suite developed by EMBL [184]. Such housekeeping included, for example, the removal of duplicate sample entries and the standardization of genotype data using HapMap conventions (i.e. using 'NN' to encode missing genotypes and using diploid format for male genotypes on chromsome X). Perl was also used to merge data from different sources to create appropriate files for downstream data analysis. This included using a sample key file which mapped the anonymous identifiers associated with genotype data to study code identifiers associated with phenotype data.

Mapping clinical data Using these study code identifiers, relevant phenotype data could be obtained, including case-control status, severe malaria clinical subphenotype status, ethnicity, and clinical gender. These identifiers also facilitated retrieval of past genotype data associated with the samples stored in a central MalariaGEN database, which, in the present case included important genotypes at G6PD202, G6PD376, and HbS that were obtained as part of a general MalariaGEN quality control (QC) process for all incoming DNA samples.

Genetic gender Another part of the QC pipeline involved gender curation by 'genotyping' three fixed nucleotide differences between the two copies of the amelogenin gene found on the X and Y chromosomes, where male samples show both *AMELX* and *AMELY* nucleotide peaks, and females are monomorphic for the *AMELX* peak. This system allowed the curation of clinical gender (CG), i.e. that gender which was obtained at the field site health centers, using the genetic gender (GG) inferred from the amelogenin variants. In our case, we assigned gender in the following order:

1. if both CG and GG were missing, an initial guess was made using the study code suffix for trio parents (M=mother, F=father);
2. if either CG or GG was missing, the other was used to assign gender;
3. if both CG and GG were nonmissing, but discordant, gender was set to missing.

The final output from Perl was a series of PED files and 'subject support' files. PED is a standard file format in genetic association studies (both family-based and case-control) where the first six columns hold identifiers, gender, and phenotype, while columns 7 and beyond hold genotype data:

1. family ID
2. individual ID
3. paternal ID

4. maternal ID

5. sex

6. phenotype [case-control status]

7. genotype data

The 'subject support' file is simply a flat table holding additional sample metadata (e.g. HbS genotype, ethnicity, clinical subphenotype, etc.). Both types of files were made for case-control, trio, and combined datasets, with only complete trios retained in a trio PED file.

2.1.4.2 Quality control using PLINK

The PED files created in Perl were then passed to PLINK, a software package designed for genome-wide association studies [182], which has a variety of simple tools to perform data quality checks. The process began by obtaining general information on the samples and SNPs that were genotyped. This included, for each SNP, information on allele frequency and deviation from Hardy-Weinberg equilibrium. Additionally, for each SNP and for each individual sample, data was obtained on missingness (the percent of undefined genotypes), and for trio datasets, information on the percent of Mendelian errors (child genotype inconsistent with parental genotypes).

The QC data was then used to inform the selection of appropriate SNP and sample inclusion criteria. For missingness and Mendelian errors, the empirical cumulative distribution function was assessed for each measure to decide *ad hoc* cutoffs that sensibly removed outlying samples or SNPs. These cutoffs varied between datasets for a variety of reasons that included SNP genotyping assay quality and sample quality. For Hardy-Weinberg equilibrium (HWE), an absolute cutoff of $P < 10^{-}5$ (exact test) was set, using controls of similar ethnicity, in order to remove SNPs with likely high genotyping error. For allele frequency, thresholds were defined for MAF specific to each dataset. All of these cutoffs were designed to retain high quality, high information content data, but came at the expense of decreased sample size and SNP coverage. After deciding upon appropriate inclusion criteria, finished PED files were created and then passed to R and fastPHASE.

2.1.4.3 Pre-analysis procedures in R

Upon passing the data to R, further checks were performed on the data using the snpMatrix package [183], including:

a. allele frequency differences between males and females

b. heterozygosity versus call rate

c. call rate in cases vs. controls

d. extreme violation of HWE using z-test

Additionally, information from tables in my internal G6PD database (e.g. inferred functional consequences, ancestral alleles) was merged with the MAP file, the PED partner file which holds physical location data for each SNP. Finally, prior to analysis, genotype data was recoded based on ancestral allele identity, any remaining heterozygous genotypes found in males were coded as missing, and samples for which gender or phenotype (case-control status) were undefined were removed.

2.1.4.4 Haplotype phasing in fastPHASE

Using Perl, a PED file was first translated into a format compatible with fastPHASE [185], with separate PED files used for different ethnic groups. This program infers haplotypes using a hidden Markov model approach, where individual haplotypes are modeled as clusters or mosaics of the most common haplotypes in a population. While a number of parameters for the model may be altered, the approach that worked best was to use default settings, and let the program search for the optimal number of clusters (K) by searching the set $K = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$ for the value of K which minimized overall error rate (as assessed by imputation of randomly masked genotypes). Since data were on chromosome X, haploid male genotypes represented a known set of haplotypes that were be used to improve and assess phasing accuracy in females. The latter was performed by simply comparing the frequency distributions for all haplotypes in males and females using a χ^2 test for equality of proportions.

2.1.5 Sample collection

The samples used in Chapters 3, 5, and 7 were collected and processed as part of consortial activities by the Malaria Genomic Epidemoiology Network (MalariaGEN), an organization founded in 2005 with the support of the Bill and Melinda Gates Foundation and the Wellcome Trust as part of the Grand Challenges in Global Health initiative [186]. Its goal is to harness the tools of genetic epidemiology to better understand the disease at the level of the host (pathogenesis and host protective mechanisms), the parasite (virulence and drug resistance), and the vector (ecology and insecticide resistance).

In this dissertation, I describe the use of samples from several different MalariaGEN study sites and protocols. While MalariaGEN itself sought approval from an institutional review board (IRB) at Oxford (OXTREC 020-06: Malaria Genomic Epidemiological Network: Learning from the human genome how protective immunity against malaria works (D. Kwiatkowski)), individual study sites were responsible for collecting the blood samples and providing necessary local ethics approval from an appropriate IRB, with assistance from MalariaGEN where necessary. DNA was extracted at these partner sites, and shipped to Oxford where each sample was assigned a standard MalariaGEN sample tracking code to uniquely identify it for further laboratory work. The sections of this thesis that use MalariaGEN samples or resources include:

- Chapter 3 – Samples were collected as part of MalariaGEN Consortial Project 3 (CP3), a project designed to survey genetic variation at four loci previously associated with protection from malaria, including *HBA*, *HBB*, *G6PD*, and *SLC4A1*. Study sites participating in this project include those in Burkina Faso, Cameroon, Ghana, The Gambia, Kenya, Mali, Malawi, Nigeria, Papua New Guinea, Sudan, Tanzania, and Vietnam.

- Chapter 5 – Samples were collected as part of a birth cohort study (SSC1192: The genomic epidemiology of childhood diseases in Kilifi, Kenya (T. Williams)) run by run by Thomas Williams, an investigator at the Kenya Medical Research Institute-Wellcome Trust Research Programme in Kilifi, Kenya. This study was conducted with permission of the director of the institute (Kevin Marsh), and is funded by the Wellcome Trust (076934) and the European Union EVIMALAR FP7 Network of Excellence. Samples were processed for enzyme activity and DNA extractions in Kilifi, but were genotyped in Oxford by the MalariaGEN resource centre based at the Wellcome Trust Centre for Human Genetics.

- Chapter 7 – Samples were collected as part of MalariaGEN Consortial Project 1, a project designed to conduct large-scale genetic association studies of severe malaria. The study site discussed here is at the Medical Research Council Unit in the Gambia, with severe malaria cases recruited at Royal Victoria Hospital in Banjul. Clinical protocols there were run by Muminatou Jallow, Fatou Sisay-Joof, and Margaret Pinder. Cases were recruited under protocol SCC670/630: Identification of host factors that predispose to severe malaria (D. Kwiatkowski), while controls were recruited under the following protocols:

  - SCC1029v2: Investigating mechanisms of protective immunity against malaria: recruitment of controls for genome-wide analysis of malaria resistance genes (M. Jallow)
  - SCC872: Re-organisation of DNA collections at MRC The Gambia (G. Sirugio)
  - SCC885: Epidemiology, virology, immunology and clinical outcome of post-natal cytomegalovirus (CMV) infection in early life (K. Flannagan)
  - SCC948: A strategy to immunize young infants against measles: Phase I trials (K. Flannagan)
  - SCC954: Effect of EBV infection on antibody responses to malaria and to vaccines (K. Flannagan)
  - L2004.29v1: Sukuta birth cohort ongoing recruitment (K. Flannagan)

The work presented in Chapters 4 and 6 is not directly associated with MalariaGEN. In chapter 4, the samples used are publicly available from the International HapMap consortium, and details on its recruitment and sample collection process has been described elsewhere [160]. In chapter 6, the samples used were obtained as part of an existing National Institute of Allergy and Infectious Diseases (NIAID) study run by Rick Fairhurst designed to examine the protective effect of hemoglobinopathies against malaria, and subjects were recruited with permission from the local IRB at the University of Bamako as

well as an institutional IRB at NIAID (NCT00669084: Innate and Acquired Resistance to Plasmodium Falciparum Malaria in Mali (R. Fairhurst)).

2.2 Supplemental Methods: Chapter 3 – Characterizing global genetic variation at *G6PD*

2.2.1 Cataloging known variation at *G6PD*

Cataloging known variation at the locus was a necessary first step in designing downstream association studies. The most important known variants at *G6PD* are those nonsynonymous variants identified in previous cDNA sequencing of G6PD deficient individuals. As the majority of these are sporadic mutations found in Western countries that do not contribute significantly to population prevalence in Africa and southeast Asia, greater emphasis was placed on the minority of functional variants that are most likely to be common, and so I made special note of variants found in more than one individual or more than one published study in MalariaGEN populations of interest (Table 2.5). The full list of variants found in clinical case studies or population surveys was merged with: (a) public data from dbSNP and Ensembl, and (b) novel variants identified during resequencing (see below) and organized into a database. Full details of the literature search and database assembly can be found in the corresponding results chapter.

Table 2.5: Common functional SNP table. Shown here, for each common functional SNP, is: the colloquial identifier, any associated dbSNP identifier, the genomic position (hg19), observed alleles (on the [+]-strand, according to UCSC Genome Browser), ancestral allele (according to the Ensembl EPO pipeline), and G6PD exon-intron position (according to UCSC genome annotation coordinates), amino acid change (according to UCSC genome annotation coordinates), and G6PD deficiency class (as defined by the WHO; see Introduction for more details).

Name	dbSNP ID	Position	Observed Alleles	Ancestral Allele	G6PD - Exon/Intron	Amino Acid Change	Class
Kaiping	NULL	153760472	C/T	C	E12	Arg463His	2
Canton	NULL	153760484	C/A	C	E12	Arg459Leu	2
Union	NULL	153760605	G/A	G	E11	Arg454Cys	2
Chinese-5	NULL	153761184	G/A	G	E9	Leu342Phe	3
Chatham	NULL	153761205	C/T	C	E9	Ala335Thr	3
G6PD_Betica	NULL	153761240	A/G	A	E9	Leu323Pro	3
Viangchan	NULL	153761337	C/T	C	E9	Val291Met	2
G6PD_MexicoCity	NULL	153762340	C/A	C	E7	Arg227Leu	3
Vancouver_2	NULL	153762605	G/A	G	E6	Arg198Cys	2
G6PD_Med	NULL	153762634	G/A	G	E6	Ser188Phe	2
G6PD_Santamaria	NULL	153762655	T/A	T	E6	Asp181Val	2
Chinese-3	NULL	153762704	T/C	T	E6	Asn165Asp	2
Mahidol	NULL	153762710	C/T	C	E6	Gly163Ser	3
Chinese-4	NULL	153763476	C/A	C	E5	Gly131Val	3
Vanua_Lava	NULL	153763485	A/G	A	E5	Leu128Pro	2
G6PD_376	rs1050829	153763492	T/C	T	E5	Asn126Asp	4
G6PD_202	rs1050828	153764217	C/T	C	E4	Val68Met	3
Songklanagarind	NULL	153764223	A/T	A	E4	Phe66Ile	2
Orissa	NULL	153764383	G/C	G	E3	Ala44Gly	3
Honiara	NULL	153774272	T/C	T	E2	Ile33Met	1
Gaohe	NULL	153774276	T/C	T	E2	His32Arg	2

2.2.2 Resequencing at *G6PD*

Resequencing was performed as part of an exon sequencing project called ExoSeq at Wellcome Trust Sanger Institute (WTSI) by teams led by K. Kivenen and A. Coffey using samples from MalariaGEN CP3 (refer to section 2.1.5 for protocol details). A total of 288 individuals comprised the sample set, with

48 individuals each from Burkina Faso, Cameroon, Nigeria, Kenya, Vietnam, and Papua New Guinea. While cost constraints prevented the addition of more individuals or populations from Africa, this set should be generally reflective of common variation in east and west Africa, as the majority of African ethnic groups in our studies are of shared Bantu ancestry.

A total of 28 amplicons approximately 600-800 bp in length were designed via automated algorithm (i.e. in silico PCR), covering about two-thirds of the 15 kb *G6PD* gene, with full coverage of all exons, but a few gaps in noncoding sequence, specifically in introns 1 and 2. Capillary sequencing of the PCR products yielded chromatograms that were input into an automated SNP calling pipeline, which was supplemented with manual calls from examination of individual sequence traces. Using custom scripts written in Perl, genotypes for each individual derived from sequencing were validated by pairwise comparison with genotype data derived from the Sequenom platform in the same individuals.

2.2.3 Multiplexed genotyping

The Sequenom iPLEX platform (reviewed in detail above) was used to genotype 135 SNPs in over 2000 samples from twelve countries as part of MalariaGEN CP3, including Burkina Faso, Cameroon, Ghana, The Gambia, Kenya, Mali, Malawi, Nigeria, Papua New Guinea, Sudan, Tanzania, and Vietnam, with sample sets gender- and ethnicity-balanced for each country. The goal was to broadly survey genetic variation at the *G6PD* locus across these study sites as preparation for downstream association analyses, with a special focus on functional variants known to be associated with G6PD deficiency. The principle for assay design using the Sequenom software was to prioritize inclusion of these functional variants, but also of variants identified during resequencing, which introduced ascertainment bias with respect to functional consequence and allele frequency, respectively.

The assay design consisted of sequential input of three lists of desired SNPs to the Sequenom assay design software to build four multiplex genotyping reactions. The first list contained 25 high priority SNPs, i.e. those of known functional or population genetic significance, in rank order of importance. The second list contained the first list, and then added 72 additional variants (ranked according to function and allele frequency), which consisted of both SNPs identified during resequencing and also other literature-described variants. Finally, the third list contained lists 1 and 2, but added to this 211 variants from dbSNP and Ensembl, ranked according to proximity to *G6PD*. For each list, polymorphisms and their flanking sequences were force-coded on the (+) strand, SNP-masked using a custom Perl script, and double-checked against the UCSC reference sequence. Using standard program settings and allowing for a maximum of 40 SNPs per well, four multiplex reactions were designed, interrogating a total of 135 SNPs across the locus.

Details of the laboratory techniques are described earlier (see Overview above), and a list of PCR and

extension primers used are shown in Table S2.2. All cluster plots were curated manually, with individual cluster plots checked for strength of signal and ability to reliably distinguish heterozygotes from homozygotes. A total of 104 SNPs met cluster plot QC standards, with 78 of these being polymorphic in at least one subpopulation (Table 2.6).

Table 2.6: Multiplex performance summary – Supplemental Methods: Chapter 3. This table details the assay design and curation process of polymorphic variants in the genotyping study referred to in Supplemental Methods: Chapter 3.

	Initial	*Passed Assay Design*	*Passed Cluster Check QC*	*Polymorphic in 1+ subpopulation*
High Priority	25	25	25	15
Medium Priority	72	61	47	42
Low Priority	211	49	32	21
TOTAL	308	135	104	78

2.2.4 Data curation

Quality control was first performed on genotype data. Genotype data was curated using custom Perl and R scripts, and the genetic association toolkit in PLINK [182], as reviewed in detail earlier. Here, the specific criteria for exclusion of individual samples and SNPs were:

1. Individuals: (a) ambiguous gender, or (b) missingness $> 35\%$
2. SNPs: (a) DAF $< 0.1\%$, or (b) missingness $> 20\%$

The criteria for missingness is looser here than in subsequent chapters because this was still at the pilot phase of assay design, i.e. there was a higher prior probability of more call failures here. The dataset was also stratified by country and then assessed for deviation from HWE semi-manually in R using z-test in the snpMatrix package [183], but no outliers were observed. A total of 1828 of 2100 individual samples, and 51 of 78 SNPs were retained after the entire QC process.

2.3 Supplemental Methods: Chapter 4 – Detection and quantitation of population genetic variation using pooled massively parallel sequencing of whole-genome amplified DNA.

2.3.1 Sample preparation

First, multiple displacement amplification [187] was used to whole-genome amplify genomic DNA from each individual sample. It uses the highly processive Φ29 DNA polymerase and random hexamer primers to synthesize random genomic fragments of approximately 10 kb in length during an isothermal amplification process. The MDA reaction was carried out in quadruplicate using two different commercial kits ('Repli-G Midi' [Qiagen] and 'GenomiPhi HY' [GE Healthcare]), to minimize representational bias and/or artifacts from any individual reaction. All reactions were performed in a thermal cycler (MJ Research) using 96 well plates (Greiner).

The procedure used for Genomiphi was as follows:

1. Heat Denaturation: 9 ul sample buffer (proprietary) + 1 ul DNA (10 ng) was heated at 95°C for 3 min, then cooled to 4°C.

2. Amplification: 10 ul denaturation reaction product + 9 ul amplification buffer (proprietary) + 1 ul enzyme mix (10 ng) was heated at 30°C for 2 h.

The procedure used for Repli-G was as follows:

1. Chemical Denaturation: 2 ul Buffer D2 (Reagent DLB, proprietary) + 15 ul DNA (2-3 ng/ul) in TE buffer (10 mM Tris-HCl, 0.5 mM EDTA, pH=8.0) was mixed thoroughly and incubated at room temperature for 3 min.

2. Neutralization: 3 ul Buffer N1 (stop solution, proprietary) was added and mixed thoroughly to stop the denaturation reaction.

3. Amplification: 20 ul denaturation reaction product + 29 ul amplification buffer + 1 ul enzyme mix (10 ng) was heated at 30°C for 2 h.

Next, long-range PCR was performed on the amplified MDA product with 'Platinum *Taq*' (Invitrogen), using a step-down extension step in order to maximize amplification specificity. As before, all reactions were performed in a thermal cycler (MJ Research) using 96 well plates (Greiner). The PCR reaction conditions and primer pairs are shown in Tables 2.7 and 2.8.

Table 2.7: PCR conditions. Reactions were performed using Platinum *Taq* (Invitrogen, 11304-029) in a reaction volume containing: 2.0 μl of buffer (10x), 0.2 μl *Taq* polymerase, 2.0 μl dNTPs (2 mM), 0.8 μl Mg^{2+} (50 mM), 4.0 μl primers (20 μM), 2.5 μl template DNA (50-100 ng/μl), and 9.0 μl H_2O. T_a = annealing temperature, i = touchdown cycle number.

Denaturation	
98°C	03:00
Touchdown Cycling (10 cycles)	
95°C	00:30
$T_a + 10 - i$ °C	00:30
68°C	10:00
Main Cycling (25 cycles)	
95°C	00:30
T_a °C	00:30
68°C	10:00

Table 2.8: PCR primers. Annealing temperature and amplicon length are shown for all primer pairs.

Forward primer	*Reverse primer*	*Amplicon* (bp)	T_A (°C)
CGATCTGTCGCTATTGCTCA	GGTGGTCACCTCACTCCACT	4592	60
TGAATTCTTTGGCATTGCAG	AATGCGGGGGTAGTAAAGCA	5482	62
CCTCCACACTGCTCCTTCTC	GGAAGATTTGTGGCCACTGT	5135	58
GGTCACGGGGGCTGGTAATG	AGCGTGGGGTTGCTGTGACT	5496	70
AACTAAACAGCTGGACGCGG	CCAGTCCCCCAAAGCGCATA	5478	80
GGTGACAGGGAGCTAAGCTGAGC	CCCGGAAACGGTCGTACACT	5410	64
TGGCGGACTCAGACTTCTCT	GAGACTCTTCGCCCAGTACG	4603	58
ATGGGAACCCAGTGAGAGTG	GGAGGGTTAAACGAGGAAGC	4518	62

The PCR reaction products were then cleaned of primers and salts using QIAquick chromatography columns (Qiagen), using the following procedure:

1. Bind: The 20 ul PCR reaction was mixed with 100 ul of Buffer PB (proprietary), applied to the QIAquick column, and centrifuged at 13000 rpm for 30 sec, with flow-through discarded after spin.

2. Wash: 750 ul of Buffer PE (proprietary) was applied to the QIAquick column, and centrifuged at 13000 rpm for 30 sec, with flow-through discarded after spin.

3. Elute: Sample was eluted from the QIAquick column using 50 ul TE buffer (10 mM Tris-HCl, 0.5 mM EDTA, pH=8.0), with centrifugation at 13000 rpm for 60 sec.

PCR products were then quantitated using PicoGreen (Invitrogen), which assesses double-stranded DNA concentration using an intercalating fluorophore:

1. Standards were prepared using phage lambda DNA (included with the kit) diluted in TE buffer (10 mM Tris-HCl, 0.5 mM EDTA, pH=8.0), with eight different concentrations used (0, 50, 100, 200, 600, 1000, 1400, and 2000 ng/ml).

2. Samples were prepared by creating 1:100 dilutions of DNA in 200 ul of TE buffer.

3. 50 ul of both standards and samples were plated in duplicate on a 96 well black plate (Greiner), meaning that only 8 standards + 40 samples were assayed per plate.

4. 50 ul of dilute PicoGreen reagent (25 ul PicoGreen stock solution in 5 mL distilled water) was added and incubated for 5 min at room temperature in darkness.

5. A plate reader (Tecan) was used to measure fluorescence with excitation at 485 nm and emission at 535 nm. Finally, a custom Excel macro was used to derive a standard quantitation curve from the fluorescent intensity data for the standards using linear regression, and this equation was then used to estimate DNA concentration in the samples.

Amplicons were then combined in equimolar amounts for each individual to generate target amplicon libraries, which represent amplified *G6PD* sequence for an individual. Finally, target libraries from each individual were pooled in equimolar amounts to generate six pools of different size (N = 2, 4, 8, 16, 32, 40), each containing a total of 0.5 μg of DNA. A schematic of the sample preparation workflow is shown in Figure 2.3. Pre-processing and massively parallel sequencing were then performed according to a modified Illumina protocol [176], as described above.

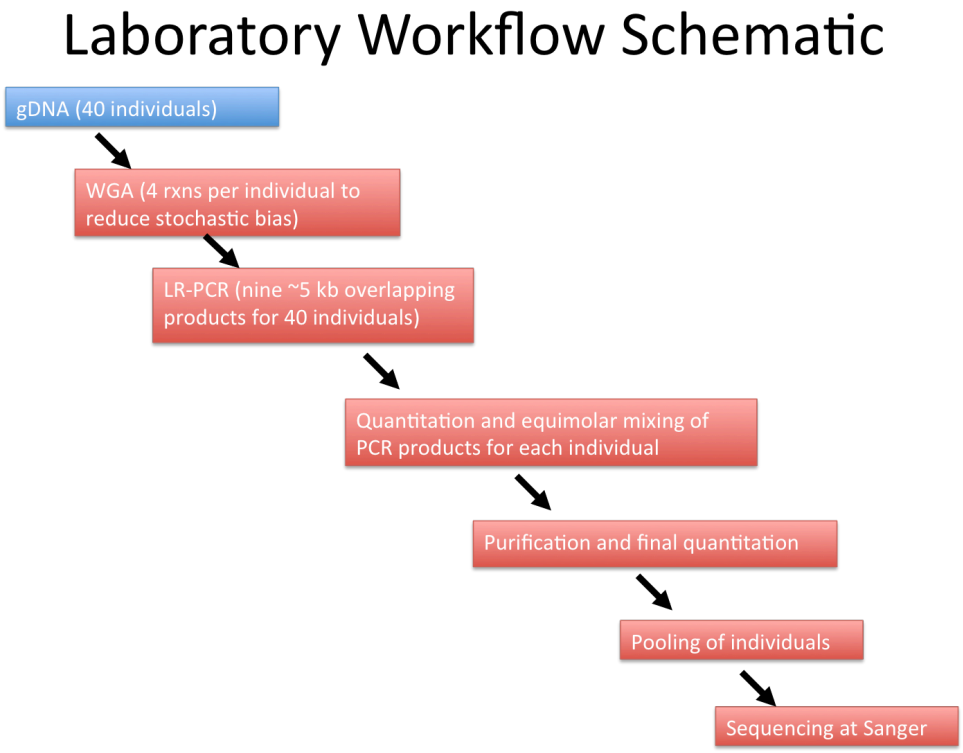


Figure 2.3: Schematic of lab workflow for Illumina pooling experiment.

2.3.2 Mapping sequence reads

Sequencing was performed on an Illumina Genome Analyser (at the WTSI), with approximately 6 million single-end reads generated per run. Each 36 bp read was then mapped to reference sequence using the program MAQ [188]. The reference sequence used was the chromosome X assembly contig 'AF277315.6' on which *G6PD* lies. The program was run using default settings via the 'easyrun' Perl script pipeline included with the software, allowing 2 mismatches per read alignment, which effectively limited SNP detection sensitivity in regions of extremely high SNP density, but enhanced alignment specificity and read quality (more mismatches might be an artifact of poor base calls rather than actual SNPs). Using the 'pileup' command, a condensed representation of the data was generated at each nucleotide position in the 'pileup' format, which included information on read depth, base calls, base and map quality scores, and cycle number.

2.4 Supplemental Methods: Chapter 5 – Genetic determinants of glucose-6-phosphate dehydrogenase activity in Kenya

2.4.1 Sample collection and handling

Samples were collected as part of an ongoing birth cohort study run by Kenya Medical Research Institute (KEMRI) and the Wellcome Trust in Kilifi, Kenya (refer to section 2.1.5 for protocol details). Blood was drawn from 2100 infants (less than 6 months old) via heel prick into capillary tubes, and was stored on ice until arrival at the main laboratories on the same day. Upon arrival, samples were stored at 4°C and processed for activity measurements and DNA extraction with 4-5 days of receipt. Ethics approval was granted by the KEMRI review board, and written informed consent obtained from parents prior to enrollment.

2.4.2 G6PD activity assay

Gross G6PD activity was measured in duplicate with the G6PDH kit (Randox Laboratories, PD 410), using a modified protocol as described in detail earlier, using 96 well black plates (Costar). The specific procedure used here for these heel prick samples was:

1. Hemolysate preparation: 10 ul washed RBC (50% Hct) + 25 ul digitonin (Reagent R4 in PD 410) was incubated at 4°C for 15 min. Hemolysate was isolated via centrifugation at 3000 rpm for 10 minutes at 4°C.

2. G6PD Reaction: 5 ul hemolysate + 10 ul NADP (Reagent R2, 0.34 mM) + 5 ul glucose-6-phosphate (Reagent R3, 0.58 mM) + 300 ul reaction buffer (Reagent R1, 31.7 mM triethanolamine, 3.2 mM EDTA, pH=7.6) was mixed thoroughly with pipette at room temperature. The reaction mixture was then placed immediately on plate reader for activity measurements.

Readings were taken over a 3 min time course on a UV/VIS plate reader (Tecan) at 340 nm and 540 nm, which assesses the absorbance of NADPH and hemoglobin, respectively (note that Hb readings do not measurably change over time). The latter Hb measurement was especially important as a means of correcting for variation in G6PD activity between samples due to variance in absolute G6PD enzyme amounts, with Hb concentration being a surrogate measure.

2.4.3 Sequencing and genotyping.

Capillary sequencing of 288 African and Asian samples (48 from Kenya) across 15 kb at the *G6PD* locus identified a total of 72 SNPs, including 49 novel variants. These variants, along with known variants identified through a comprehensive literature search and those found in public databases (e.g. dbSNP, Ensembl), were compiled into a relational database. The resequencing and database compilation efforts are covered in detail in Supplemental Methods: Chapter 3 and its corresponding results chapter.

Genotyping was performed on the Sequenom iPLEX platform (reviewed above) as part of a larger effort

to perform fine-mapping association studies at *G6PD* in a global cohort as part of MalariaGEN Consortial Project 1. Data was used from genotyping assay performance in Supplemental Methods: Chapter 3 as a guide to the assembly of a variant panel for this study and the work described in Supplemental Methods: Chapter 7. The assay design procedure was analogous to that presented earlier, but with fewer SNPs input into the design algorithm, as choice was largely dictated by those SNPs with strong observed strong assay performance. Overall, sixty-eight SNPs, 24 of 24 high priority (including 17 nonsynonymous variants), 27 of 32 medium priority, 17 of 25 low priority variants were able to be designed into three multiplexes. The laboratory techniques are described in detail in the methods overview section. Multiplex primers and assay design are summarized in Table S2.4.

2.4.4 Data curation

Quality control was first performed on both phenotype and genotype data. Inclusion criteria for the enzyme activity phenotype included removal of samples with negative values for OD340, and phenotype data was also checked for any batch effects or extreme outliers. Genotype data was curated using custom Perl and R scripts, and the genetic association toolkit in PLINK [182], which was reviewed in detail earlier. Here, the specific criteria for exclusion of individual samples and SNPs were:

1. Individuals: (a) ambiguous/undefined gender, or (b) missingness $> 10\%$
2. SNPs: (a) DAF $< 1\%$, (b) deviation from HWE in females ($P < 10^{-5}$), or (c) missingness $> 5\%$

A total of 1828 of 2100 individual samples, and 30 of 68 SNPs were retained after this QC process.

2.4.5 Genetic modeling

All analyses were performed in R [181]. Single site association tests were performed via multiple linear regression ('lm' function in R with Gaussian error model), with G6PD enzyme activity (Y, continuous) modeled as a function of genotype (X_1, continuous), gender (X_2, categorical), and in controlled tests, G6PD202 genotype (X_3, categorical):

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{2.2}$$

Genotypes represented derived allele counts (i.e. 0=ancestral homozygote, 1=heterozygote, 2=derived homozygote), and were coded differently according to three different models of inheritance:

1. Additive: males = 0/2, females = 0/1/2
2. Dominant: males = 0/1, females = 0/1/1
3. Recessive: males = 0/1, females = 0/0/1

All haplotypes present at frequency greater than 0.01 were tested for association with G6PD enzyme activity. Haplotypes were inferred using fastPHASE using default parameters (as described earlier).

Haplotype tests used the same model as in Equation 2.2 except that each individual was scored 0 or 1 for X_1 based on the absence or presence of the test haplotype (one haplotype was chosen at random for each female to equalize male and female contribution to the association tests, as recommended by Clayton [189]). Qualitative test odds ratios were calculated via logistic regression ('glm' function in R with binomial error model), with the odds of G6PD deficiency state (Y, continuous) modeled as a function of genotype (X_1, continuous)

$$Y = e^{\beta_1 X_1} \tag{2.3}$$

2.5 Supplemental Methods: Chapter 6 – A novel cytofluorometric assay for detection and quantification of G6PD deficiency

2.5.1 Sample collection

Forty-two samples were collected by venipuncture from children at a community health center in Kenieroba, Mali (refer to section 2.1.5 for protocol details). Blood was collected in the morning, stored in acid-citrate-dextrose solution, and transported to the laboratory in Bamako on ice on the same day. The buffy coat and plasma were isolated via centrifugation and removed. The red cell pellet was washed three times in RPMI 1640, and stored at 50% Hct at 4°C prior to assay. Blood samples were processed for biochemical and cytofluorometric activity measurements within four days of their receipt. All samples were obtained with informed consent in accordance with protocols at University of Bamako and the National Institutes of Health.

2.5.2 G6PD activity assay

Gross G6PD activity was measured in duplicate with the G6PDH kit (Randox Laboratories, PD-410), using a modified protocol as described in detail earlier. The specific procedure for these blood samples was:

1. Hemolysate preparation: 10 ul RBC (50% Hct) + 20 ul digitonin (Reagent R4 in PD 410) was incubated at 4°C for 5 min.

2. G6PD reaction: 2 ul hemolysate + 4 ul NADP (Reagent R2, 0.34 mM) + 2 ul glucose-6-phosphate (Reagent R3, 0.58 mM) + 120 ul reaction buffer (Reagent R1, 31.7 mM triethanolamine, 3.2 mM EDTA, pH=7.6) was incubated at room temperature. This reaction mixture was then placed immediately on plate reader for activity measurements.

Readings were taken over a 30 min time course[1] on a UV/VIS plate reader (Perkin Elmer) at 340 nm and 540 nm, which assess the absorbance of NADPH and hemoglobin, respectively (NB: Hb readings do not measurably change over time). The hemoglobin measurement was especially important as a means of correcting for variation in G6PD activity between samples due to variance in absolute G6PD enzyme amounts, with Hb concentration being a surrogate measure.

2.5.3 Cytofluorometric assay

This assay represents a cytofluorometric readout of the classical methemoglobin reduction test (MRT). First, the MRT was performed (steps 1 and 2), using slightly modified published protocols [190, 191], and then the cytofluorometric method (steps 3 through 5):

[1] The longer time course here was used after observing that the reaction rate remained linear for more than 1 h.

1. Oxidation of Fe^{2+} to Fe^{3+}: 100 ul washed RBC (5% Hct) + 100 ul sodium nitrite (0.125 M) was incubated at RT for 20 min. The RBC were then washed 3x using phosphate-buffered saline (PBS), with centrifugation at 1800 rpm for 1 min, followed by resuspension in 100 ul PBS.

2. Reduction of Methemoglobin (Fe^{3+}): All of mixture #1 + 18 ul glucose (0.28 M in PBS) + 6 ul Nile blue sulfate (0.01%) was incubated at 37°C for 90 min in open microcentrifuge tubes.

3. Formation of Cyanmethemoglobin: All of mixture #2 + 2.5 ul KCN (0.4 M) was incubated at RT for 5 min.

4. Reaction with Hydrogen Peroxide: 5 ul mixture #3 + 100 ul hydrogen peroxide (3% in PBS) was agitated vigorously by hand, and then washed 2x in PBS, with centrifugation at 3000 rpm for 3 min.

5. Cytometry/Microscopy: Immediately, cell activity measurements were obtained on a flow cytometer or fluorescence microscope. Strong fluorescence was generated with excitation either in the UV (e.g. DAPI filter cube) or green (e.g. FITC filter cube) range. Details of the measurement parameters on both instruments can be found in the corresponding results chapter.

2.6 Supplemental Methods: Chapter 7 – G6PD deficiency exerts contrasting effects on two major clinical presentations of severe malaria in West Africa

2.6.1 Study cohort

The study cohort included 2650 severe malaria cases and 3300 controls recruited at Royal Victoria Hospital in Banjul. Cases were confirmed by blood smear, and designated as 'severe' according to established WHO guidelines. Designation of clinical subphenotypes in severe cases was as follows:

- severe malarial anemia (SMA) – hematocrit $\leq 15\%$ or hemoglobin ≤ 5 g/dl
- cerebral malaria (CM) – Blantyre coma score ≤ 3 in the absence of recent convulsions, hypoglycemia or meningitis.

These subphenotypes were not exclusive of one another and were both present in a subset of individuals. A total of 1538 cases had defined clinical subphenotypes: 540 (SMA), 871 (CM), and 127 (both). Control samples were representative of the population as they were derived from cord blood. Informed consent was obtained prior to sample collection, and study approval was given by the Oxford Tropical Research Network Ethics Committee and the Gambia Government/MRC Joint Ethics Committee (refer to section 2.1.5 for protocol details).

2.6.2 Genotyping.

The same Sequenom multiplex reactions were used here as in Supplemental Methods: Chapter 5, although this sample set did exhibit slightly poorer cluster performance, as reflected in the slightly less stringent inclusion criteria. The laboratory techniques are described in detail in the methods overview section. Multiplex primers and assay design are summarized in Table S2.4.

2.6.3 Data curation

Quality control was first performed on both phenotype and genotype data. Inclusion criteria for case-control phenotype was simply whether or not case-control status was defined. Genotype data was curated using custom Perl and R scripts, and the genetic association toolkit in PLINK [182], which is reviewed in detail earlier. Here, the specific criteria for exclusion of individual samples and SNPs were:

1. Individuals: (a) ambiguous gender[2], or (b) missingness $> 20\%$
2. SNPs: (a) DAF $< 1\%$, (b) deviation from HWE in females ($P < 10^{-5}$), or (c) missingness $> 10\%$

In addition to the amelogenin SNPs, the X chromosome data itself was used in PLINK to recode gender according to an inbreeding coefficient test based on observed and expected heterozygosity (H):

$$F_{IS} = (H_{obs} - H_{exp})/H_{exp} \tag{2.4}$$

[2]see Overview section earlier for detailed discussion of gender assignment

When this metric was unable to confidently impute a gender, pre-imputed gender (i.e. the sex assigned by the standard gender-resolving process described in the Overview section) was retained. A total of 1828 of 2100 individual samples, and 30 of 68 SNPs were retained after the entire QC process.

2.6.4 Genetic modeling

All analyses were performed in R [181]. All tests were performed via multiple logistic regression ('glm' function in R with binomial error model), with odds of being a case (Y, continuous) modeled (Equation 2.5) as a function of:

1. genotype (X_1, continuous),
2. gender (X_2, categorical)
3. curated ethnicity (X_3, categorical)
4. HbS genotype (X_4, categorical)

$$Y = e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4} \tag{2.5}$$

Genotypes represented derived allele counts (i.e. 0=ancestral homozygote, 1=heterozygote, 2=derived homozygote), and were coded according to three different models of inheritance:

1. Additive: males = 0/2, females = 0/1/2
2. Dominant: males = 0/1, females = 0/1/1
3. Recessive: males = 0/1, females = 0/0/1

All haplotypes present at frequency greater than 0.01 were tested for association with severe malaria. Haplotypes were inferred using fastPHASE using default parameters (as described earlier). Haplotype tests used the same model as in Equation 2.5 except that each individual was scored 0 or 1 for X_1 based on the absence or presence of the test haplotype (one haplotype was chosen at random for each female to avoid bias). Analysis of covariance was used to test for an interaction between sex and genotype, where a significant coefficient for the interaction term in the model suggests that odds ratios in males and females are not equal:

$$Y = e^{\beta_1 X_1 * \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4} \tag{2.6}$$

Certain basic tests of association were conducted with the snpMatrix package [183], including the 2 d.f. general genotypic and the 1 d.f. Cochran-Armitage trend tests with Mantel extension. The Mantel strata were curated ethnicity, and all basic tests were conducted in HbS AA individuals[3].

[3]While logistic regression allowed for including both ethnicity and HbS strata in the model, these basic tests allow only one stratification; the strong protective effect of HbS was thus controlled by only including HbS AA individuals.

A genotype was defined for composite G6PD deficiency trait according to the sum of all deficiency alleles present at G6PD968, G6PD 542, and G6PD202:

1. Additive: males = 0/2, females = 0/1/2
2. Dominant: males = 0/1, females = 0/1/1
3. Recessive: males = 0/1, females = 0/0/1

Thus, a 202T male that had only ancestral alleles at the other two sites, was coded as a 2 in the additive model. If a female had a single deficiency allele at any of the three sites, she was coded as a 1 in the additive model, and if she had two deficiency alleles at a single site, she was coded as a 2. A female could also be a 'compound heterozygote' of sorts, i.e. have two different derived alleles at two different deficiency loci (e.g. 968G and 202T), and in fact, nine of these were identified.

2.7 Supplemental Information

2.7.1 Custom scripts

2.7.1.1 Perl

- duplicate_remover.pl
  - Function: This script takes a genotype 'table format' file (i.e. has whitespace at the first line, unique identifiers in the first column, no white space within cells, etc.) and assembles a hash, keeping the best of duplicate entries (i.e. that with the least missing data).
  - Input/Output: Input – genotype 'table format' file. Output – genotype 'table format' file printed to STDOUT.
  - Workflow: Cycles through each row entry in the table eliminating duplicates by retaining the entry with the fewest missing genotype data.
  - Usage: perl duplicate_remover.pl
- factory_to_PED.pl
  - Function: This script takes a genotype table format file from MalariaGEN Factory, along with metadata from other table format files, and generates files for downstream analysis including PED and subject support files.
  - Input/Output: Input – 'table format' files holding data for main genotypes (-geno), other genotypes (-cov), phenotypes (-pheno), the first six columns of a PED file (-ped), and a key file (-key) linking identifiers in the PED and genotype files to the phenotype files. Output – PED and subject support files printed to file for full case-control data, full trio data, and also partial case-control data (without the trio cases).
  - Workflow: Imports the input files as hash tables, standardizes it (genotypes are coded according to HapMap convention), performs quality control checks (gender curation, trio checks), and then populates output PLINK-compatible PED files and subject support files using the data collected by cross-mapping identifiers (discussed in detail in section 2.1.4.1).
  - Usage: perl factory_to_PED.pl -geno genotypes_no_dups.txt -cov cov.txt -key key.txt -pheno pheno.txt -ped factory_ped_tableformat.txt
- g6pd_sequenom_flank.pl
  - Function: This script takes a FASTA sequence and SNP masks it using SNPs provided by the user.
  - Input/Output: Input – FASTA sequence file and SNP coordinate file with two columns: (1) holds coordinates; (2) holds observed alleles separated by a divider. Output – Sequenom input file with two columns: (1) coordinates; (2) string holding upstream and downstream flanking sequence seperated by the observed alleles for SNP of interest (e.g. AAG[G/T]CCT).

- Workflow: Stores the FASTA sequence as a string. Stores the SNP coordinate file as a hash, with key=coordinate, and value=observed alleles. Masks the FASTA sequence using IUPAC SNP coding, and then cycles through the SNP coordinate file to pull out substrings from the masked FASTA.
- Usage: perl g6pd_sequenom_flank.pl

- g6pdDB_dev_001.pl
  - Function: This script populates the three tables of the g6pdDB database, using a Perl wrapper to perform mySQL scripting.
  - Input/Output: Input – three files for the core, func, and data tables. Output – mySQL database file.
  - Workflow: Uses DBI to log onto g6pdDB database, and populates tables using the 'mysqlimport' function on the three table files.
  - Usage: perl g6pdDB_dev_001.pl -core core.txt -func func.txt -data data.txt

- pileup_parse.pl
  - Function: This script takes a MAQ pileup file and generates summary data for reference and alternate read calls at each position (see section 4.3 for details).
  - Input/Output: Input – MAQ pileup sequence file. Output – summary data at each position in a table printed to STDOUT.
  - Workflow: Cycles through each line in the MAQ pileup file, and stores mean data on reference and alternate calls (cycle number, base and map qualities, observed strand, read depth) for each position.
  - Usage: perl pileup_parse.pl –pileup out.pileup –range 6906:35906

2.7.2 R

- R_solexapool_wkst_2.R
  - Function: This script takes the output of pileup_parse.pl and generates a list of putative SNPs that meet user-defined criteria.
  - Input/Output: Input – output file from pileup_parse.pl. Output – dataframe (table) holding list of putative SNPs and summary data for each.
  - Workflow: First, cycles through each row (position) in the input data and does some housekeeping functions and calculations using func_transform:
    a. Remap – Remap to new coordinate system if user provides remapping coordinates (for the first and last position in the data).

b. PCR redundancy – Determine PCR amplicon redundancy (AKA 'tiling overlap') at each position if the user provides primer coordinates.

c. Read depth (normalized) – Determine normalized read depth relative to PCR amplicon coverage.

d. AAF – Determine maximum nonreference allele frequency (i.e. the most likely alternate allele frequency).

e. Error rate – Calculates a trimmed mean of alternate allele frequency calls (i.e. the mean of the two alternate call frequencies that are least likely to represent the minor allele).

f. Noise statistic – Calculates a test statistic comparing the observed number of reads from the most frequent alternate allele with the expected value based on the total number of alternate allele reads. As this test statistic is chi-square distributed under the null hypothesis, the -log10 P value of the test statistic on chi-square with 1 d.f is recorded.

g. Strand balance ratio – Obtains the ratio of (+)-strand to (-)-strand calls for REF and ALT alleles.

h. Cycle number difference – Obtains the absolute value of the difference in cycle number for REF and ALT alleles.

i. Nucleotide transition – Records the observed nucleotide transition from REF->ALT.

j. GC content – Records the %GC content in 10 bp windows.

Next, the script takes the transformed dataframe and subsets it using func_snp_filter which has flags for user-defined criteria for cutoff values, but otherwise acts to trim the data according to the basic criteria of:

1. AAF – User must define a minimum AAF that the putative SNP must demonstrate.
2. Noise statistic – Noise statistic for a putative SNP must not be in the upper quartile of the empirical distribution of noise statistic values across the data set.
3. Base quality – Base quality for a putative SNP must be in the upper decile of the empirical distribution of base quality values across the data set.
4. Cycle number difference – Cycle number difference for a putative SNP must be in the upper tertile of the empirical distribution of cy_diff values across the data set.

– Usage: Open and edit script within R console.

- R_cp3_wkst_20110106.R, R_CP1_wkst_20110125.R, R_G6PD_biochem.R

  – Function: These analysis scripts performed analyses mentioned in Chapters 3, 5, and 7.

  – Input/Output: Input – PED and subject support files generated by factory_to_PED.pl. Output – analysis tables and graphs as presented in thesis.

- Workflow: First, imports data from PED and subject support files, merges it with data from g6pdDB database and phenotype metadata, and then performs analyses listed using functions in the basic R library, along with the snpMatrix package (these are detailed in the Supplemental Methods sections for the respective chapters).
- Usage: Open and edit script within R console.

- R_mali_cytochem.R
  - Function: This script takes a directory containing standard FCS format flow cytometry data files and calculates for each file the estimated proportion of G6PD+ and G6PD- RBC present.
  - Input/Output: Input – FSC files. Output – analysis dataframe (table).
  - Workflow: First, imports data from FCS files using the flowCore package, and imposes a morphological gate on the data using the rectangleGate function, accepting only those events meeting $5.5 < FSC.H < 6.5$ and $4.0 < SSC.H < 6.0$. Next, uses the raw data from the FL1.H channel and determines any maxima present in the empirical distribution using kernel density estimation of the probability density function. When two maxima are found for the G6PD+ and G6PD- populations, the average FL1 intensity of the two maxima determines FL1 gate. When only one maximum is found, the gate is placed 0.5 log units away, as this represents half the average distance (1 log unit) between G6PD+ and G6PD- cells in our data. The proportion of G6PD+ cells are those with FL1 intensity greater than the gate value, and vice-versa for G6PD- cells.
  - Usage: Open and edit script within R console.

2.7.3 Linkage disequilibrium

Linkage disequilibrium (LD) describes a pattern of nonrandom association between alleles at two loci, specifically the condition where the observed haplotype frequencies are significantly different than those expected based on allele frequencies. Linkage equilibrium (LE) describes the situation where observed haplotype frequencies are equal to the product of constituent allele frequencies. It is similar to Hardy-Weinberg equilibrium (HWE) which is satisfied when genotype frequencies are equal to the product of constituent allele frequencies. There are both formal statistical tests and informal descriptive statistics to assess pairwise LD between two markers [192, 167].

2.7.3.1 Formal hypothesis testing

Formal tests are based on the independence of haplotype counts in a contingency table of alleles, using either the chi-square or Fisher's exact test to assess the difference between observed and expected haplotype frequencies. Across two biallelic sites, the 2x2 matrix of possible haplotypes would be:

Expected haplotype counts (E) are a function of the total number of haplotypes present (N) and the corresponding allele frequencies (P), e.g. $E_{AB} = N \times P_A P_B$. Observed haplotype counts are directly

	A	a
B	AB	aB
b	Ab	ab

observed for all but the double heterozygote individuals, in whom phase is ambiguous. Various procedures have been devised to estimate these unobserved haplotype counts, including those that utilize EM algorithms and hidden Markov models [193, 194, 185].

The likelihood of LD versus LE can be summarized with a log-odds (LOD) score:

$$LOD = log_{10}\frac{\text{odds(LD)}}{\text{odds(LE)}} \tag{2.7}$$

with odds(LD) = 1 - odds(LE), and odds(LE) being the P-value from the formal tests. LOD is approximately normally-distributed with $|LOD| > 2$ (i.e. 100-fold greater odds of LD than LE) conventionally regarded as strong evidence for LD.

2.7.3.2 Informal descriptive statistics

Two important descriptive measures of LD are D' and r^2. Both measures depend on the fundamental metric D, which describes the difference in the observed (P_{AB}) and expected ($P_A P_B$) haplotype frequencies:

$$D = P_{AB} - P_A P_B \tag{2.8}$$

The condition for LE is met if $D = 0$, i.e. if the frequency of haplotype AB is the same as that given by random assortment of alleles A and B. Note that the sign of D is arbitrary, thus: $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$.

D' is frequency-independent and sensitive to recombination events One derives D' by normalizing D to its maximum value if $D > 0$ or to its minimum value if $D < 0$:

$$D' = D/D_{max} \quad \text{if } D > 0\text{; where } D_{max} = \min(P_A P_b, P_a P_B)$$
$$D' = D/D_{min} \quad \text{if } D < 0\text{; where } D_{min} = \max(P_A P_B, P_a P_b)$$

Unlike D, D' is frequency-independent and equal to 1 unless all four haplotypes are present. Thus, it is particularly sensitive to recombination events, and does not exhibit much dynamic range.

r^2 is frequency-dependent and is the square of the correlation coefficient between two alleles The r^2 metric depends on both D and allele frequencies:

$$r^2 = \frac{D^2}{P_A P_a P_B P_b} \tag{2.9}$$

It formally represents a correlation coefficient between alleles at two loci, and is thus inversely related to the relative power one has to detect a true disease association when using a marker SNP in lieu of a true causal variant. In other words, if the power of a study when the causal variant itself is typed is P, then the power of the marker SNP is related to the extent of LD between it and the causal SNP and is P/r^2.

Factors that influence LD A variety of population genetic phenomena can influence LD, and should be considered when drawing inferences (e.g. regarding selection) based on genomic patterns of LD [192]. For example, populations bottlenecks and genetic drift can lead to the loss of rare haplotypes, thus increasing LD. Additionally, population stratification can artificially inflate LD– an extreme example would be a situation where two populations may be fixed for opposite haplotypes (e.g. AB and ab) and have D' = 0 on their own, but have spurious D' = 1 when considered together. Finally, structural changes at the level of DNA sequence such as inversions, which lead to decreased recombination rate, and gene conversion events, which reduce haplotype diversity, can both serve to increase LD.

2.7.4 Phylogenetic analysis

Phylogenetic analysis encompasses a range of methodologies used to estimate the ancestral relationships between individuals, populations, or species [195, 196]. A commonly used way to represent these relationships is with a phylogenetic tree. The topology, or shape, of such trees consists of edges, or connections between organizational taxonomic units (OTUs), and nodes, which represent the branch points between descendant OTUs and their shared most recent common ancestor.

A few major classes of methods are:

1. Parsimony – Direct sequence data is used to construct a tree which minimizes the number of changes (e.g. mutations, recombinations) across all nodes.

2. Distance matrix methods – The original data is condensed into a matrix of distances between OTUs, which is then used to generate a tree using a particular algorithm (e.g. neighbor-joining or unweighted pair group using arithmetic mean) that estimates topology and distances.

3. Maximum likelihood – The original data itself is used to construct the most-likely tree given the observed data. An advantage to this method is that the fullest representation of the data is considered (which contrasts with distance matrix methods) and the full tree space is explored, with the likelihood of the data observed calculated for each possible tree configuration.

In Chapter 3, gene frequency data is used to infer phylogeny via a maximum likelihood method for continuous characters using the 'contml' program in PHYLIP. The principles of this approach were first described by Cavalli-Sforza and Edwards [197], with a restricted maximum likelihood estimation procedure later developed by Felsenstein [198], and rely on the notion that differences in allele frequencies that develop between populations are a function of genetic drift and can be modeled via random walk (akin to Brownian motion). Briefly, allele frequency data for n alleles in n-dimensional space can be used to derive 'chord' distances, which are included with times since divergence T in a likelihood function (L), where values of T are then derived which maximize $\log(L)$, the log-likelihood of the observed data given the parameter estimates.

2.7.5 Principal components analysis

Principal components analysis (PCA), a technique developed by Pearson [199], facilitates reduction of a multivariate dataset to a lower dimensional space which represents the major components of variance across the data. Each principal component (PC) tries to explain as much variance as possible, with the stipulation that adjacent PCs must be uncorrelated with, or orthogonal, to one another. Thus each PC explains only as much variance as possible that has not been explained by preceding PCs. The proportion of variance explained by a particular PC can be obtained by taking the ratio of its eigenvalue to the sum of eigenvalues for all PCs [200].

As an example, a mean-normalized dataset (i.e. each variable is standardized by mean subtraction) in three dimensions can be reduced by finding the line that passes through the origin [0, 0, 0] and minimizes the distance (sum of squares) between each point and the line. This line represents the first principal component (PC). A second PC axis can also be drawn, which must also pass through the origin, but also be orthogonal to the first line, while also minimizing the distance between each point and the line. Each line can now thought of as an axis with arbitrary units, where each datum in the dataset can be projected along either axis at the point of intersection for the line joining point and axis.

Strictly speaking, the use of gene frequency data from diverse populations often violates the assumption that each variable in the dataset be approximately normally distributed, but it is still commonly utilized as a descriptive tool for understanding population differentiation [201], and more recently as a quantitative covariate used in lieu of self-reported ethnicity in correcting for population stratification in genetic association studies [202].

Table S2.1: Multiplex definitions table – Supplemental Methods: Chapter 3 (1 of 2).

Mplex	SNP	PCR Primer 1	PCR Primer 2	Extension Primer	Mass	Alleles
216	rs2230037	ACGTTGGATGCATCAGCAAGACACTCTCTC	ACGTTGGATGCAGAAGACGTCCAGGATGAG	GCTCCCTGACGCCTA	4489	G/A
217	hG6PD_968	ACGTTGGATGTGGGGAACCCCGATGGAGA	ACGTTGGATGACGACGGCTGCAAAAGTGG	CCACCAAAGGGTACC	4531	C/T
218	G6PD_X_b36_153416679	ACGTTGGATGAGGTAGAAGAGGCGGTTGG	ACGTTGGATGATGATGCAGCCTCCTACCAG	TGTGACCCCAGGTGG	4609	A/G
219	hG6PD_542	ACGTTGGATGAGATCTGGTCCTCACGGAAC	ACGTTGGATGATCATCGTGGAGAAGCCCTT	TGGTTGGACAGCCGG	4649	T/A
220	G6PD_X_b36_153415014	ACGTTGGATGAGAACCACCTACTGCAGATG	ACGTTGGATGTTCTCATCACGGACGTCATC	aGAGAAGCCCGCCTCC	4836	T/A
221	rs2515904	ACGTTGGATGAGCTGCGTTACCCCCTTGA	ACGTTGGATGTGGGAGGGCGTCTGAATGAT	CGGGGAGTGAGGATCA	5011	C/G
222	hG6PD_680	ACGTTGGATGTTTGCCAACAGGATCTTCGG	ACGTTGGATGCCTTGAAGGTGAGGATAACG	ccGCCCCATCTGGAACC	5076	G/T
223	G6PD_X_b36_153413799	ACGTTGGATGTCATCCTGGACGTCTTCTGC	ACGTTGGATGATAGCCCACAGGTATGCAGG	GCCAGATGCACTTCGTG	5186	G/A
224	G6PD_X_b36_153427466	ACGTTGGATGCCTGGCTTTTAAGATTGGG	ACGTTGGATGGCCTTCCATCAGTCGGATAC	GCACCCATGATGATGAA	5203	C/T
225	rs5986990	ACGTTGGATGAAGTCACCACCTCTGAGCAC	ACGTTGGATGCTGGGACATGACAACTTGGG	TGACCTGGACTGGCAGT	5226	G/A
226	rs2230036	ACGTTGGATGAAGGCCCTGAACGAGCGCAA	ACGTTGGATGTGCACTGCTGGTGGAAGATG	ggCGAGGTGAGGCTGCA	5316	T/C
227	G6PD_X_b36_153413666	ACGTTGGATGTCAATCTGGTGCAGCAGTGG	ACGTTGGATGACCGGCCTCCCAAGCCATA	GCAGTGGGGTGAAAATA	5324	C/T
228	rs2515905	ACGTTGGATGAGTGTCTGAGCTATCAGACC	ACGTTGGATGGGTGATGCTATCACTGAATC	CCCTACACAGCCAAGCAC	5382	G/A
229	G6PD_X_b36_153412566	ACGTTGGATGATTGTGAGACCAGCTTGAGG	ACGTTGGATGTTGCTGGAAGTCATCTTGGG	ctccGTGGGGTCACCTTCC	5732	T/C
230	G6PD_X_b36_153414531	ACGTTGGATGCATTCTCTCCCTTGGCTTTC	ACGTTGGATGAGGACCACATTGTTGGCCTG	tGCTTTCTCTCAGGTCAAG	5770	T/C
231	rs762515	ACGTTGGATGTGGCACAGGCTTGAATCTCG	ACGTTGGATGCTACTCAGGATCACTACTGG	gatGGGCTCTTCTGTCTGT	5817	T/C
232	G6PD_X_b36_153414758	ACGTTGGATGGACAGTCACTATGTGACCAG	ACGTTGGATGGGTCCATCTCGAGTCTATTC	gggCAGGGAAGGCCATTGC	5919	G/A
233	G6PD_X_b36_153417405	ACGTTGGATGTCACTCTGTTTGCGGATGTC	ACGTTGGATGATGGCCTTCTGCCCGAAAAC	gtgcGGCGGGAACGGGCAT	5935	A/G
234	G6PD_X_b36_153428979	ACGTTGGATGTCAGACTTCTCTCCGGAGC	ACGTTGGATGAAGTAAAAACACAAGCCCCG	ggagaCGGCTTCCCGAGTTC	6134	C/T
235	rs12393550	ACGTTGGATGAAAGTGGCCAGGGAGGAGTG	ACGTTGGATGTTCTACGACATCGGGCCCTG	aaCCGCAGCACAGGTGAGGA	6185	G/A
236	G6PD_X_b36_153426354	ACGTTGGATGCCTGGATGCACTCAAGAATC	ACGTTGGATGACTAGTTTTGATAGAGCTCC	CCATTAACTTACCTTGTAACC	6300	C/A
237	rs2515906	ACGTTGGATGTGGCTCAAGCGTGTAATCTC	ACGTTGGATGAGGCTGGTCTTGAACTTCTG	cccCGTGTAATCTCAGCACTT	6317	G/A
238	G6PD_X_b36_153424232	ACGTTGGATGTTCAGGCTGACGTGTGTTTC	ACGTTGGATGCTCTTTGTGGGATCTATGCG	GACGTGTGTTTCAGAAGATAC	6485	T/C
239	G6PD_X_b36_153413991	ACGTTGGATGAGATGATGACCAAGAAGCCG	ACGTTGGATGTCTGTAGGGCACCTTGTATC	gggtGAGCTGGACCTGACCTA	6487	G/A
240	G6PD_X_b36_153430003	ACGTTGGATGGGTTTGTTTGTTTTTGAGAG	ACGTTGGATGAGAGTGAGCTGAGATCATGC	gGTTTTTGAGAGATGGAGTCT	6547	C/T
241	G6PD_X_b36_153426720	ACGTTGGATGCTCTGCCTCTTCACTAACAC	ACGTTGGATGCGAAATTGAGGCAAAGAGAC	GATACAAACACTTAGCTTTTCT	6668	G/A
242	G6PD_X_b36_153429686	ACGTTGGATGTGCGTAAAAGCCAATCTCCC	ACGTTGGATGAGCACAGTTGGAGAGGAAGG	caaGGCCTACAAGGCCCTGTAC	6689	G/T
243	G6PD_X_b36_153422839	ACGTTGGATGTGAGGTAAGACGGACAGTAG	ACGTTGGATGTCAGCTTAGCTCCCTGTCAC	tggcGGGGTAAGGCAGGCCGCT	6857	C/A
244	rs28534384	ACGTTGGATGAAAGTCTTGGAACATGTCCG	ACGTTGGATGGTGCATCAGGGTAGCCTTC	ccccgTCGTGGCCTTTGGAACAC	6976	G/C
245	G6PD_X_b36_153426256	ACGTTGGATGCTGAGCTCAGTGGTTCTTTG	ACGTTGGATGCAAAGGTGTCAAGGAGTAGG	gaTTTGTTCACATAGAACAACGC	7032	T/C
246	rs5986992	ACGTTGGATGAGGGAGGAGTAGAAACAGTC	ACGTTGGATGTTCTGGAGCACTGGCCAAG	gaacGTAACCTGGGTCCTGAATC	7049	C/A
247	rs762517	ACGTTGGATGACGTGACCTCAAGTGCTTAC	ACGTTGGATGCTGGCCCTTTTTCTCACTTC	AAAAAGAGTGGTTCAAAAATAAG	7162	A/G
248	G6PD_X_b36_153424319	ACGTTGGATGACAGCGATTGGGTACTTCTG	ACGTTGGATGTTACTTGCCGTGGTTCCCTG	TCCCTGCCCTCCTCA	4400	G/A
249	rs2071429	ACGTTGGATGTTTGCCCTCCCTCCCTGTGT	ACGTTGGATGGCAGCAGTGGGGTGAAAATA	CCTCCCAAGCCATAC	4442	G/A
250	rs6643678	ACGTTGGATGTTGTGGCTGAATTCTTTGGC	ACGTTGGATGAAGGCTGGAGCCCTGTTTTC	TTGGCATTGCAGGCA	4608	A/G
251	G6PD_X_b36_153415799	ACGTTGGATGATCTCCTTGCCCAGGTAGTG	ACGTTGGATGCGGCTGTCCAACCACATCT	GGTAGTGGTCGATGC	4664	A/G
252	G6PD_X_b36_153417417	ACGTTGGATGTGTTCCGGGATGGCCTTCTG	ACGTTGGATGTCACTCTGTTTGCGGATGTC	TCTGCCCGAAAACACC	4795	T/A
253	G6PD_X_b36_153414378	ACGTTGGATGTGCCCCGCGGGTCCACCAC	ACGTTGGATGTCGAAGGCATCACCTACCAT	TTTTGCAGCCGTCGTC	4839	G/A
254	G6PD_X_b36_153413623	ACGTTGGATGTTTTCACCCCACTGCTGCAC	ACGTTGGATGCACCTGCCATAAATATAGGG	gTGAGCTGGAGAAGCC	4971	G/A
255	G6PD_X_b36_153416656	ACGTTGGATGTTGGTGACGGCCTCGTAGAC	ACGTTGGATGAGCGCCTCAACAGCCACAT	GCCAGGTAGAAGAGGC	4980	A/G
256	rs7887673	ACGTTGGATGTGTGGCTTCCTCATCCCCTC	ACGTTGGATGGGCTTCTTAACTAGCACTTC	cGGGGCCAGAGTTCACC	5196	A/G
257	G6PD_X_b36_153411172	ACGTTGGATGTCCTGCCTCCACCCAAGTAG	ACGTTGGATGAGGTGGAAGACCATCCACTG	gGCTGAAGGGACAGTTG	5316	C/G
258	G6PD_X_b36_153413455	ACGTTGGATGAGGCAGACGAGCTGATGAAG	ACGTTGGATGGTTCACCCACTTGTAGGTGC	GATGAAGAGAGTGGGTT	5354	C/A
259	G6PD_X_b36_153415904	ACGTTGGATGTGATCCTCACTCCCCGAAGA	ACGTTGGATGAGGGCTTCTCCACGATGATG	CTCCGGGCTCCCAGCAGA	5446	T/C
260	G6PD_X_b36_153424389	ACGTTGGATGAGAAGTACCCAATCGCTGTC	ACGTTGGATGGCACGTGATCCTAGAATCTG	gGCTCAATGAACGGAGAA	5582	C/T
261	rs2472393	ACGTTGGATGAGCACCTGGAACAGCTCCG	ACGTTGGATGTAGGATCACGTGCCCTCAAG	gggcACAGCTCCGGCTCCC	5751	T/C
262	G6PD_X_b36_153412691	ACGTTGGATGACACCAAACATGGTCATGGC	ACGTTGGATGTGCTTCCTCTGTGGCCACTG	cCTGGACATGGGGTCAGAA	5878	C/T
263	G6PD_X_b36_153423752	ACGTTGGATGAACTAAGGCCCAGAGAAGTG	ACGTTGGATGATGATGCTGGAGAGCTTCAC	tcTCAATCCCAGTCTCTTCC	5939	C/T
264	G6PD_X_b36_153427470	ACGTTGGATGGCGATGCCTTCCATCAGTC	ACGTTGGATGCTGGGAGATACTCACCGATG	aTTCCATCAGTCGGATACAC	6061	T/C
265	G6PD_X_b36_153417577	ACGTTGGATGTGGGACACACTTACCAGATG	ACGTTGGATGTCTAACCACACACCTGTTCC	TGGGGTAGATCTTCTTCTTG	6145	C/G
266	G6PD_X_b36_153426402	ACGTTGGATGTGGCAAAGATGGTCATACAG	ACGTTGGATGTACACGTGGTCTCAGTTTTC	TGGTCATACAGTAATGAGCA	6165	C/G
267	G6PD_X_b36_153426313	ACGTTGGATGAGCAGGAGCTCTATCAAAAC	ACGTTGGATGTGAACAAAGAACCACTGAGC	gACTAGTTAGTTCTTCCACAT	6371	G/A
268	rs60589782	ACGTTGGATGCCACGCCCAGCTAATTTTTG	ACGTTGGATGGTCAAGAGATCAAGACCATC	gcgcGAGATGGGGTTTCATCA	6502	C/T
269	rs28579420	ACGTTGGATGTGCTCTCCCAGGATTGATTG	ACGTTGGATGTTACAGCTGTGATGGTTTAG	taTCTCCATTATCTCAAGTAGC	6644	T/C
270	G6PD_X_b36_153412620	ACGTTGGATGCTCAAGCTGGTCTCACAATG	ACGTTGGATGGACCATGTTTGGTGTCAGAG	CTGGTCTCACAATGACAATATG	6718	C/A
271	rs12861291	ACGTTGGATGTTTCAGAACAGCCCCTTCTC	ACGTTGGATGCCAACATGAAGGATGGGAAC	gtccgGGTTTCCTCCCTATCTTG	6948	G/C
272	rs4898389	ACGTTGGATGCCTCCCTCAACTTAGATACC	ACGTTGGATGATGTTCCCAGAACTCAGTCC	CTTAGATACCAATTAGAGATCAC	7016	G/A

Table S2.2: Multiplex definitions table – Supplemental Methods: Chapter 3 (2 of 2).

Mplex	SNP	PCR Primer 1	PCR Primer 2	Extension Primer	Mass	Alleles
273	rs62617844	ACGTTGGATGGCTCTAAATAGAGTGATGTC	ACGTTGGATGCCCCGAGATAACTTTGCCAT	gagcTGTCTTTTGTTTCCAAAGT	7036	G/A
274	rs4256746	ACGTTGGATGAACTCGACATTTCTACTTGG	ACGTTGGATGTGGGTAGATCTGGGTTTTGG	ggTCTACTTGGTTATCTAGTAGC	7061	G/A
275	G6PD_X_b36_153413678	ACGTTGGATGACCGGCCTCCCAAGCCATA	ACGTTGGATGTCAATCTGGTGCAGCAGTGG	TCAGCGACGAGCTCC	4538	C/A
276	G6PD_X_b36_153413788	ACGTTGGATGATAGCCCACAGGTATGCAGG	ACGTTGGATGTCATCCTGGACGTCTTCTGC	GGGGCCGGCAGCTGG	4675	A/G
277	G6PD_X_b36_153415828	ACGTTGGATGATCATCGTGGAGAAGCCCTT	ACGTTGGATGAGATCTGGTCCTCACGGAAC	CTGTCCAACCACATCT	4761	G/A
278	G6PD_X_b36_153416165	ACGTTGGATGAATGAGGAAGCCATGGCTTG	ACGTTGGATGTTCCCCTCCTGCCTCATGCT	aGCCACTCAGACCCAC	4780	G/C
279	G6PD_X_b36_153414937	ACGTTGGATGATGCACACCCCAGCTCAGT	ACGTTGGATGATGACGTCCGTGATGAGAAG	gGTGCCTCGTCACAGA	4882	C/T
280	rs762516	ACGTTGGATGTTCAGTAGAACCAAGGGACG	ACGTTGGATGTGTAAGCACTTGAGGTCACG	AGGGACGACAGAAGTA	4988	C/T
281	rs5986875	ACGTTGGATGAGAACTCAGACAAGGGTGAC	ACGTTGGATGCTGTTGGCAAATCTGCAGGG	cAGGTCAGTTCCTCCAC	5106	G/A
282	G6PD_X_b36_153411188	ACGTTGGATGCAGGACAATCGCTTGAACCC	ACGTTGGATGCAGAGCTGAAGGGACAGTTG	cGAAGACCATCCACTGG	5164	G/A
283	G6PD_X_b36_153416473	ACGTTGGATGACCACTCTATGAGCGTGTCC	ACGTTGGATGAACGCTGCCACCTTGTGGTC	gtATCTGGGCAGCATCC	5186	T/C
284	G6PD_X_b36_153423083	ACGTTGGATGAAAGCAGTTGGGAGCTTCTC	ACGTTGGATGTTGTGTCCTTGGGCAGATCC	tCTTCTCATCTTCCCCAC	5296	A/G
285	rs6643680	ACGTTGGATGTGGCATCTAGCAGGTACAGG	ACGTTGGATGTCAGTGCCATCGGCACCTG	cGCTCCACACCCTACAAC	5334	T/C
286	G6PD_X_b36_153429253	ACGTTGGATGTTCAGGCTCCCCCCTCTTTT	ACGTTGGATGAGTAACCTGGGTCCTGAATC	aCCCCCTCTTTTTTGGGC	5392	C/T
287	G6PD_X_b36_153414077	ACGTTGGATGATCTTCCACCAGCAGTGCAA	ACGTTGGATGCATCATCTTGGTGTACACGG	CTGGTGATCCGCGTGCAG	5532	G/C
288	rs61042368	ACGTTGGATGTCCTAGAGCTTCACTGCTCG	ACGTTGGATGGCCAGAATTCAGTCTGCTTC	ccaaTCGCCTCCACTAGAA	5702	G/A
289	hG6PD_plus202	ACGTTGGATGATGGCCTTCTGCCCGAAAAC	ACGTTGGATGTCACTCTGTTTGCGGATGTC	GCCCGAAAACACCTTCATC	5702	T/C
290	G6PD_X_b36_153412734	ACGTTGGATGAAGTGAGGAGGCTCCTCAGT	ACGTTGGATGTCATTTGCTTCCCTGAGGAC	cgcaACAGAGGAAGCACCC	5801	A/G
291	rs5986877	ACGTTGGATGACCTTGCTCTTCCCCCTTC	ACGTTGGATGGCAGAAGAGCTCTGCTTTAG	ccccCTCAGGCACCCATCT	5919	G/C
292	rs12837628	ACGTTGGATGTTTGAGACCCAGTCTCACTC	ACGTTGGATGAGGAGAATCGCTCAAACCTG	gtccCGATCTTGGCTCACCG	6045	C/G
293	G6PD_X_b36_153416596	ACGTTGGATGAACGGCAAGCCTTACATCTG	ACGTTGGATGCTCTTCTACCTGGCCTTGC	cagCTGGCTCATGCAGGACT	6118	C/T
294	rs1050831	ACGTTGGATGCCGTGCCACCCGTAGGCAG	ACGTTGGATGTTGCGGATTTAATGGCAGGG	aagaAGCAGCTGGGACCCCT	6136	G/A
295	rs6655250	ACGTTGGATGGAGAGGTTCAATGGACTCAC	ACGTTGGATGATGCCTTTGCTCATCCTTCG	cccacTGGACTCACAGTTCCA	6311	G/A
296	rs12837311	ACGTTGGATGCGTAAATAGTTGTTTCACTG	ACGTTGGATGACAGAGTTGGACTGTGTCTC	ggatTTGTTTCACTGTCTTGC	6409	C/A
297	rs5945202	ACGTTGGATGCCCAAACATCTCAGTGGAG	ACGTTGGATGTCTGAGATCGGAGAAGAACC	tAGTAAAGAATAACAAGGCAG	6520	G/A
298	rs28458996	ACGTTGGATGAAGGGCGCATTGCATTAAGC	ACGTTGGATGTCTCCATTATCTCAAGTAGC	aggACAGCTGTGATGGTTTAG	6541	C/G
299	rs7879049	ACGTTGGATGTTTACCACAGCCCATGGCAC	ACGTTGGATGGCTCAGGGGAGTGAGAAAAG	aggggCATCCTGGCATAGAGCA	6809	G/A
300	rs11797720	ACGTTGGATGTTCGAAAGCCTACCAAGGAG	ACGTTGGATGTGTCTTTCCTATGAGTGCAG	aaccaAATTGTCCAGCTTGTAGA	7032	T/C
301	rs35228794	ACGTTGGATGGGAAGGTTTTTGTTTTTGGG	ACGTTGGATGTCTGGTGCATTGCAGATTGG	TTGTTTTTGGGTAATGATTGTCA	7106	A/G
302	G6PD_X_b36_153424456	ACGTTGGATGCAGCAGATTCTAGGATCACG	ACGTTGGATGTTCTTCCACGCCTGATGCCT	ACCCTGCCTACCTCC	4409	C/T
303	G6PD_X_b36_153416019	ACGTTGGATGTGCATCATTCAGACGCCCTC	ACGTTGGATGATCTGGGAACACAAGGCACG	CCAGAACCTCCTCGC	4458	C/T
304	G6PD_X_b36_153416108	ACGTTGGATGACCCAGCATGAGGCAGGAG	ACGTTGGATGAACACCTTCCGGGAAGCCTT	GCATGCCAGCAATGC	4562	G/A
305	G6PD_X_b36_153416670	ACGTTGGATGAGGTAGAAGAGGCGGTTGG	ACGTTGGATGATGATGCAGCCTCCTACCAG	CGGTTGGCCTGTGAC	4600	C/A
306	G6PD_X_b36_153414129	ACGTTGGATGAAGGCCCTGAACGAGCGCAA	ACGTTGGATGATCACCAGCTCGTTGCGCTT	gCCATGATGTGGCCGG	4938	G/A
307	G6PD_X_b36_153412861	ACGTTGGATGTTAATGGCAGGGCATTGAGG	ACGTTGGATGCCGTGCCACCCGTAGGCAG	TCCCAGCTGCTGCGTCT	5113	A/G
308	G6PD_X_b36_153414709	ACGTTGGATGTGCTTGGCTCTGCTCACCCT	ACGTTGGATGTCAGAATAGACTCGAGATGG	ttGACCCTCCAGGACCA	5115	C/T
309	rs12389569	ACGTTGGATGTTGGAGCTCACAGGAGCACT	ACGTTGGATGGTGAGTCACCCTATGGAATG	GAGCACTGAGGCCCTTG	5211	G/A
310	rs1894260	ACGTTGGATGATCTCTCAAGGTTTCACCCC	ACGTTGGATGAAAGCTCTGAGTACACTGCG	TGGGTAGGAGCTGGAAA	5340	T/C
311	rs1050757	ACGTTGGATGTATTGGCCACTGGGTCTCAC	ACGTTGGATGATTCTCTCACGTGGGTGCTC	GGGTGGGAGGGAGGGAC	5422	T/C
312	G6PD_X_b36_153415894	ACGTTGGATGTCCCGAAGGGCTTCTCCAC	ACGTTGGATGTGATCCTCACTCCCCGAAGA	GGCTTCTCCACGATGATG	5491	C/T
313	rs56191948	ACGTTGGATGAGTAACCTGGGTCCTGAATC	ACGTTGGATGAGGCTCCCCCCTCTTTTTTG	ccccCCCGCCTTGGCCAGT	5662	T/C
314	rs6643679	ACGTTGGATGAAGGCTGGAGCCCTGTTTTC	ACGTTGGATGTTGTGGCTGAATTCTTTGGC	AGCCCTGTTTTCACCCCTT	5666	G/A
315	G6PD_X_b36_153427408	ACGTTGGATGTCATCATCATGGGTGCATCG	ACGTTGGATGTTGGAAAAGCTGAGGCATGG	CCCAATCTTAAAAGCCAGG	5766	T/C
316	G6PD_X_b36_153415121	ACGTTGGATGTCCCCATGCCCTTGAACCAG	ACGTTGGATGAGGTGGTTCTGCATCACGTC	GCCCTTGAACCAGGTGAAC	5798	T/C
317	G6PD_X_b36_153411257	ACGTTGGATGGGGTTCAAGCGATTGTCCTG	ACGTTGGATGAAAATTTGCCAGGCGTGGTG	gaCTCCTGAGTAGCTGGGA	5869	C/A
318	rs7881468	ACGTTGGATGAAGTGTTGGGATTACAGGCG	ACGTTGGATGCAGGATTTGGTTTGCTGATG	ggtgCAGGCGTGAGCCACG	5895	A/G
319	rs28470352	ACGTTGGATGGGTGAGAAAATGTTGGTGGC	ACGTTGGATGAAGAGCGAGACTCCGACTTA	ccgTTGTTTGCACTTCACCA	6019	A/T
320	rs2004651	ACGTTGGATGGACATGGGAAATAGGGAAGG	ACGTTGGATGTGGGTATCCAAACAGGAAGC	GAAATAGGGAAGGTGATAAC	6263	G/T
321	rs34193178	ACGTTGGATGATGCCCAGTTCTGCCTTGCT	ACGTTGGATGCACTTTTGCAGCCGTCGTC	ctctAAGGCATCACCTACCAT	6310	C/G
322	rs5986997	ACGTTGGATGATTCTACTGGGATGGGACAC	ACGTTGGATGTGCACAGGAATATCTAGCCC	aaTGGCTTGTCTTCTACATTT	6377	T/C
323	rs7053878	ACGTTGGATGGCAACAGAGGAGATGCTGTC	ACGTTGGATGACGGTCTGCTCCATAAGAGG	tAGATGCTGTCTCAAAAAAAA	6446	A/T
324	rs743544	ACGTTGGATGGGGTTGTCTGTTTCTATTCG	ACGTTGGATGGCTAACCCAAAGAAAAGATG	GTCTTTTTCAAACACCAATTTC	6619	G/A
325	rs2515910	ACGTTGGATGGACAGCTTTTGAAACGAGGG	ACGTTGGATGCGGAGGCAGTGAAATGACAA	ccgatGGCCCAGGGTCCTTAGG	6752	G/A
326	rs3205649	ACGTTGGATGCTGCTCTCCTACTGCATTCC	ACGTTGGATGACACCAAGGACCTGCACTG	TAATCAATAAAAACTGAGGGAA	6808	C/T
327	rs11156578	ACGTTGGATGCAACACCCGTAAAGGGTCTG	ACGTTGGATGTGGCCTTCCTCTATCTCAAC	cggtTCTGTGCTGAGGTGGATT	6828	A/G
328	rs60030796	ACGTTGGATGCGATTAACTGGACATTGTGC	ACGTTGGATGGAGTGGTCTTGGCCAAAAAG	GGAAAATATAATGAATTCGAGAAA	7466	G/A

Table S2.3: Multiplex definitions table – Supplemental Methods: Chapters 5 and 7 (1 of 2).

Mplex	SNP	PCR Primer 1	PCR Primer 2	Extension Primer	Mass	Alleles
1197	rs5986990	ACGTTGGATGAAGTCACCACCTCTGAGCAC	ACGTTGGATGCTGGGACATGACAACTTGGG	ACCTGGACTGGCAGT	4593	G/A
1197	G6PD_X_b36_153416679	ACGTTGGATGAGGTAGAAGAGGCGGTTGG	ACGTTGGATGATGATGCAGCCTCCTACCAG	TGTGACCCCAGGTGG	4609	A/G
1197	rs2071429	ACGTTGGATGTTTGCCCTCCCTCCCTGTGT	ACGTTGGATGGCAGCAGTGGGGTGAAAATA	cCCTCCCAAGCCATAC	4731	G/A
1197	G6PD_X_b36_153415014	ACGTTGGATGAGAACCACCTACTGCAGATG	ACGTTGGATGTTCTCATCACGGACGTCATC	cGAGAAGCCCGCCTCC	4812	T/A
1197	G6PD_X_b36_153413799	ACGTTGGATGTCATCCTGGACGTCTTCTGC	ACGTTGGATGATAGCCCACAGGTATGCAGG	CCAGATGCACTTCGTG	4857	G/A
1197	rs2230037	ACGTTGGATGCATCAGCAAGACACTCTCTC	ACGTTGGATGCAGAAGACGTCCAGGATGAG	tgGCTCCCTGACGCCTA	5122	G/A
1197	G6PD_X_b36_153412566	ACGTTGGATGATTGTGAGACCAGCTTGAGG	ACGTTGGATGTTGCTGGAAGTCATCTTGGG	ggGTGGGGTCACCTTCC	5218	T/C
1197	G6PD_X_b36_153415799	ACGTTGGATGCGGCTGTCCAACCACATCT	ACGTTGGATGATCTCCTTGCCCAGGTAGTG	GTGAGGACCAGATCTAC	5219	G/A
1197	G6PD_X_b36_153414378	ACGTTGGATGTGCCCCGCGGGTCCACCAC	ACGTTGGATGTCGAAGGCATCACCTACCAT	cccTTTGCAGCCGTCGTC	5402	G/A
1197	hG6PD_680	ACGTTGGATGTTTGCCAACAGGATCTTCGG	ACGTTGGATGCCTTGAAGGTGAGGATAACG	TCGGCCCCATCTGGAACC	5420	G/T
1197	rs2472393	ACGTTGGATGAGCACCTGGAACAGCTCCG	ACGTTGGATGTAGGATCACGTGCCCTCAAG	ccgcACAGCTCCGGCTCCC	5671	T/C
1197	rs2515905	ACGTTGGATGAGTGTCTGAGCTATCAGACC	ACGTTGGATGGGTGATGCTATCACTGAATC	CCCCTACACAGCCAAGCAC	5672	G/A
1197	rs762515	ACGTTGGATGCTACTCAGGATCACTACTGG	ACGTTGGATGTGGCACAGGCTTGAATCTCG	CATGTCTGTCTTGCCTGATA	6074	C/T
1197	rs2515904	ACGTTGGATGTGGGAGGGCGTCTGAATGAT	ACGTTGGATGAGCTGCGTTACCCCCTTGA	ggaGTCTGAATGATGCAGCT	6197	G/C
1197	rs1734792	ACGTTGGATGTTAGTCCATATGTCAGTTC	ACGTTGGATGTAAGGAGGGAAAGTGGAAGG	TCCATATGTCAGTTCCCCTCT	6283	C/A
1197	G6PD_X_b36_153426720	ACGTTGGATGCGAAATTGAGGCAAAGAGAC	ACGTTGGATGCTCTGCCTCTTCACTAACAC	ggGGCAAAGAGACAGTGTTCA	6544	A/G
1197	G6PD_X_b36_153428979	ACGTTGGATGTCAGACTTCTCTCCGGAGC	ACGTTGGATGAAGTAAAAACACAAGCCCCG	acccaGCCGGCTTCCCGAGTTC	6632	C/T
1197	G6PD_X_b36_153427466	ACGTTGGATGCTGGGAGATACTCACCGATG	ACGTTGGATGTTTTCCAGGGCGATGCCTTC	gtGATGCACCCATGATGATGAA	6783	C/T
1197	G6PD_X_b36_153429686	ACGTTGGATGTGCGTAAAAGCCAATCTCCC	ACGTTGGATGAGCACAGTTGGAGAGGAAGG	tccctGCCTACAAGGCCCTGTAC	6920	G/T
1197	G6PD_X_b36_153426354	ACGTTGGATGCCTGGATGCACTCAAGAATC	ACGTTGGATGACTAGTTTTGATAGAGCTCC	TTGCCATTAACTTACCTTGTAACC	7238	C/A
1197	G6PD_X_b36_153426256	ACGTTGGATGCTGAGCTCAGTGGTTCTTTG	ACGTTGGATGCAAAGGTGTCAAGGAGTAGG	ggaTTTGTTCACATAGAACAACGC	7361	T/C
1197	G6PD_X_b36_153424232	ACGTTGGATGTTCAGGCTGACGTGTGTTTC	ACGTTGGATGCTCTTTGTGGGATCTATGCG	gatgcCGTGTGTTTCAGAAGATAC	7408	T/C
1197	G6PD_X_b36_153414709	ACGTTGGATGTGCTTGGCTCTGCTCACCCT	ACGTTGGATGTCAGAATAGACTCGAGATGG	gctaCCTGAGGACCCTCCAGGACCA	7597	C/T
1197	rs60030796	ACGTTGGATGCGATTAACTGGACATTGTGC	ACGTTGGATGGAGTGGTCTTGGCCAAAAAG	AGGAAAATATAATGAATTCGAGAAA	7779	G/A
1197	rs766420	ACGTTGGATGGGCTTTTGTGAGAATACCAG	ACGTTGGATGGAATAATCTGTTAGGTTCAG	TCATGTTCCCTTTTATTCCTGGCCTTA	8133	G/C
1197	rs28470352	ACGTTGGATGGGTGAGAAAATGTTGGTGGC	ACGTTGGATGCAAAAGAGCGAGACTCCGAC	ATTCTTTGCTTTTGTTTGCACTTCACCA	8462	A/T
1198	rs763737	ACGTTGGATGACCTTAGCAACCCTGGTTTC	ACGTTGGATGGTTCCCAGCCTTCAGGAATG	TTGCTCTCCAGCCCT	4455	G/A
1198	G6PD_X_b36_153417417	ACGTTGGATGTGTTCCGGGATGGCCTTCTG	ACGTTGGATGTCACTCTGTTTGCGGATGTC	CTGCCCGAAAACACC	4491	T/A
1198	rs2230036	ACGTTGGATGTGCACTGCTGGTGGAAGATG	ACGTTGGATGAAGGCCCTGAACGAGCGCAA	CGGCCACATCATGGAA	4875	C/T
1198	hG6PD_542	ACGTTGGATGAGATCTGGTCCTCACGGAAC	ACGTTGGATGATCATCGTGGAGAAGCCCTT	GTGGTTGGACAGCCGG	4978	T/A
1198	G6PD_X_b36_153412734	ACGTTGGATGTCATTTGCTTCCCTGAGGAC	ACGTTGGATGAAGTGAGGAGGCTCCTCAGT	CCACTCCCCAAGTCACC	5020	G/A
1198	rs12389569	ACGTTGGATGTTGGAGCTCACAGGAGCACT	ACGTTGGATGGTGAGTCACCCTATGGAATG	ttGCACTGAGGCCCTTG	5177	G/A
1198	G6PD_X_b36_153415904	ACGTTGGATGTGATCCTCACTCCCCGAAGA	ACGTTGGATGAGGGCTTCTCCACGATGATG	taCGGGCTCCCAGCAGA	5180	T/C
1198	G6PD_X_b36_153413678	ACGTTGGATGACCGGCCTCCCAAGCCATA	ACGTTGGATGTCAATCTGGTGCAGCAGTGG	CCCTCAGCGACGAGCTCC	5406	C/A

Table S2.4: Multiplex definitions table – Supplemental Methods: Chapters 5 and 7 (2 of 2).

Mplex	SNP	PCR Primer 1	PCR Primer 2	Extension Primer	Mass	Alleles
1198	G6PD_X_b36_153416656	ACGTTGGATGTTGGTGACGGCCTCGTAGAC	ACGTTGGATGAGCGCCTCAACAGCCACAT	AGGCCAGGTAGAAGAGGC	5623	A/G
1198	G6PD_X_b36_153413455	ACGTTGGATGAGGCAGACGAGCTGATGAAG	ACGTTGGATGGTTCACCCACTTGTAGGTGC	TGATGAAGAGAGTGGGTT	5659	C/A
1198	G6PD_X_b36_153414531	ACGTTGGATGAGGACCACATTGTTGGCCTG	ACGTTGGATGCATTCTCTCCCTTGGCTTTC	cCTGAGATGCATTTCAACA	5772	C/T
1198	rs762516	ACGTTGGATGTTCAGTAGAACCAAGGGACG	ACGTTGGATGTGTAAGCACTTGAGGTCACG	aaAAGGGACGACAGAAGTA	5928	C/T
1198	rs915941	ACGTTGGATGGTCTATGGTCATGACGTCTG	ACGTTGGATGGCTTATATAGGCGGTCTGTC	cccGCGTCCACCCGTCTTCG	5966	C/A
1198	G6PD_X_b36_153412861	ACGTTGGATGTTAATGGCAGGGCATTGAGG	ACGTTGGATGCCGTGCCACCCGTAGGCAG	cGGTCCCAGCTGCTGCGTCT	6061	A/G
1198	G6PD_X_b36_153426313	ACGTTGGATGTGAACAAAGAACCACTGAGC	ACGTTGGATGAGCAGGAGCTCTATCAAAAC	GCCAGCACTGAGGGGCGCAC	6153	A/G
1198	rs4898389	ACGTTGGATGATGTTCCCAGAACTCAGTCC	ACGTTGGATGCCTCCCTCAACTTAGATACC	ttCCCAGACCCTATCAGAATC	6310	A/G
1198	rs915942	ACGTTGGATGACCGCAAGAGACGACAACTG	ACGTTGGATGCTATATAAGCCATGCGCAGG	cccACGACAACTGCGCTTACT	6311	G/A
1198	rs5986997	ACGTTGGATGATTCTACTGGGATGGGACAC	ACGTTGGATGTGCACAGGAATATCTAGCCC	gaTGGCTTGTCTTCTACATTT	6393	T/C
1198	G6PD_X_b36_153414758	ACGTTGGATGGGTCCATCTCGAGTCTATTC	ACGTTGGATGGACAGTCACTATGTGACCAG	tAACAAGCTGAGGCCCAGAGA	6473	A/G
1198	G6PD_X_b36_153412620	ACGTTGGATGCTCAAGCTGGTCTCACAATG	ACGTTGGATGGACCATGTTTGGTGTCAGAG	CTGGTCTCACAATGACAATATG	6718	C/A
1198	rs7053878	ACGTTGGATGACGGTCTGCTCCATAAGAGG	ACGTTGGATGGCAACAGAGGAGATGCTGTC	CTTCATCCCCTTATTTAAAAAAA	6926	T/A
1199	G6PD_X_b36_153424319	ACGTTGGATGACAGCGATTGGGTACTTCTG	ACGTTGGATGTTACTTGCCGTGGTTCCCTG	TCCCTGCCCTCCTCA	4400	G/A
1199	hG6PD_968	ACGTTGGATGTGGGGAACCCCGATGGAGA	ACGTTGGATGACGACGGCTGCAAAAGTGG	CCACCAAAGGGTACC	4531	C/T
1199	G6PD_X_b36_153414937	ACGTTGGATGATGCACACCCCAGCTCAGT	ACGTTGGATGATGACGTCCGTGATGAGAAG	GTGCCTCGTCACAGA	4553	C/T
1199	G6PD_X_b36_153417405	ACGTTGGATGTCACTCTGTTTGCGGATGTC	ACGTTGGATGATGGCCTTCTGCCCGAAAAC	GGCGGGAACGGGCAT	4683	A/G
1199	G6PD_X_b36_153415828	ACGTTGGATGATCATCGTGGAGAAGCCCTT	ACGTTGGATGAGATCTGGTCCTCACGGAAC	CTGTCCAACCACATCT	4761	G/A
1199	rs5986875	ACGTTGGATGAGAACTCAGACAAGGGTGAC	ACGTTGGATGCTGTTGGCAAATCTGCAGGG	AGGTCAGTTCCTCCAC	4817	G/A
1199	rs12393550	ACGTTGGATGTTCTACGACATCGGGCCCTG	ACGTTGGATGAAAGTGGCCAGGGAGGAGTG	AGCGCCCAGGGTGCGC	4908	A/G
1199	G6PD_X_b36_153416019	ACGTTGGATGTGCATCATTCAGACGCCCTC	ACGTTGGATGATCTGGGAACACAAGGCACG	GGCCAGAACCTCCTCGC	5116	C/T
1199	G6PD_X_b36_153413623	ACGTTGGATGTTTTCACCCCACTGCTGCAC	ACGTTGGATGCACCTGCCATAAATATAGGG	ATTGAGCTGGAGAAGCC	5259	G/A
1199	rs5986992	ACGTTGGATGAGTAGAAACAGTCCAGCTCG	ACGTTGGATGTTCCTGCTCCGCGCTTCTG	TAACCTGGGTCCTGAATC	5475	C/A
1199	G6PD_X_b36_153411172	ACGTTGGATGTCCTGCCTCCACCCAAGTAG	ACGTTGGATGAGGTGGAAGACCATCCACTG	GAGCTGAAGGGACAGTTG	5629	C/G
1199	G6PD_X_b36_153414077	ACGTTGGATGCATCATCTTGGTGTACACGG	ACGTTGGATGATCTTCCACCAGCAGTGCAA	GTGTACACGGCCTCGTTGG	5836	C/G
1199	G6PD_X_b36_153423083	ACGTTGGATGAAAGCAGTTGGGAGCTTCTC	ACGTTGGATGTTGTGTCCTTGGGCAGATCC	GAGCTTCTCATCTTCCCCAC	5964	A/G
1199	rs61042368	ACGTTGGATGGCCAGAATTCAGTCTGCTTC	ACGTTGGATGTCCTAGAGCTTCACTGCTCG	CCTCAAAGCTTCCTCGAAGT	6037	A/G
1199	G6PD_X_b36_153427408	ACGTTGGATGTTGGAAAAGCTGAGGCATGG	ACGTTGGATGTCATCATCATGGGTGCATCG	TGAGGCATGGAGCAGGCACT	6207	C/T
1199	rs762513	ACGTTGGATGGGCTGTATACATTTACGCTG	ACGTTGGATGAATATGGTAGCCACTTGCCC	CAAACAGAATTCACCATTTAG	6382	G/A
1199	rs7879049	ACGTTGGATGTTTACCACAGCCCATGGCAC	ACGTTGGATGGCTCAGGGGAGTGAGAAAAG	AGCACATCCTGGCATAGAGCA	6424	G/A
1199	rs743548	ACGTTGGATGATCATAGGTGGGAGAGGAGC	ACGTTGGATGCAATCCTACCTCCTGATCAC	GGAGCTCTGAAGAGCCCGGCAAG	7124	A/G
1199	rs743545	ACGTTGGATGTAAGTTGGAGGCTGCAGCAT	ACGTTGGATGTGACAACAGCCACCAGTCC	CAGCATGGGCACCTGGGGACAGA	7124	G/A
1199	rs766419	ACGTTGGATGGGCCATGCATAGAGAAAAGC	ACGTTGGATGAATGTCTACAAGCCCCACTC	GCATAGAGAAAAGCAGAAACGTCT	7428	C/T
1199	rs5986877	ACGTTGGATGACCTTGCTCTTCCCCCTTC	ACGTTGGATGGCAGAAGAGCTCTGCTTTAG	CCCTTCCTCCTCTCAGGCACCCATCT	7714	G/C

Chapter 3

Characterizing global genetic variation at *G6PD*

In this first results chapter, I present an approach to characterizing and organizing genetic variation at the *G6PD* locus. I describe the creation of a locus-specific database for *G6PD*, variant discovery via targeted resequencing, and global frequency variation in novel and known polymorphisms. In surveying the landscape of genetic variation at *G6PD* in different malaria-endemic areas, preparatory groundwork is laid for genetic association studies of molecular and clinical phenotypes presented in Chapters 5 and 7.

3.1 Abstract

The extent of genetic variation at the locus encoding glucose-6-phosphate dehydrogenase (G6PD) is of considerable interest to investigators conducting population genetic or disease association studies at *G6PD*. Here I describe a three-part approach to improving understanding of genetic diversity at the locus in malaria-endemic populations from twelve countries (10 African, 2 Asian): Burkina Faso, Cameroon, Ghana, The Gambia, Kenya, Mali, Malawi, Nigeria, Papua New Guinea, Sudan, Tanzania, and Vietnam. First, I present a new locus-specific database for *G6PD* that includes most known variants at the locus identified in previous resequencing studies, organizes previous nomenclature and coordinate systems, and interfaces with the modern bioinformatic tools and data found at Ensembl. Next, I discuss resequencing of the gene in 288 individuals from six of the populations, where 67 SNPs were identified in total, including 49 novel variants, one of which is a sporadic nonsynonymous polymorphism (G1187C; Pro396Ala). Finally, via dense genotyping of 51 SNPs in 1671 samples from all twelve study populations, I observed significant functional allelic heterogeneity at *G6PD* in both West Africa and Papua New Guinea, and also limited functional haplotype diversity for the 202T and 968G alleles, suggestive of a recent origin for the two most common G6PD deficiency variants in Africa. It is expected that the resources and data presented here will be of utility for other researchers interested in genetic diversity at *G6PD* in these populations and how it informs studies of human demographic history and disease susceptibility.

3.2 Introduction

3.2.1 Genetics of the *G6PD* locus

The *G6PD* gene is located on the long arm of the X chromosome, in a region of high gene density and high levels of linkage disequilibrium (LD), several hundred kilobases away from genes for color blindness and hemophilia [73]. It has long been known to be a gene that exhibits high levels of functional genetic diversity. Natural functional variation in the gene was initially described using a combination of protein electrophoresis and thermodynamic/kinetic studies, where several hundred distinct protein variants associated with G6PD deficiency were estimated to exist [103]. These variants were given colloquial names that usually referred to the location of discovery by convention, e.g. 'G6PD Seattle' or 'G6PD Mediterranean.' With the advent of DNA sequencing, the initial count of distinct protein variants has proven to be an overestimate. Nonetheless, there are still over one hundred G6PD functional variants associated with varying levels of G6PD deficiency that have been identified, several of which are rather common, and in some areas, multiple non-rare functional variants co-exist, such as west Africa and southeast Asia. These polymorphisms are generally nonsynonymous variants that were identified in cDNA sequencing of affected individuals [73]. It is important to note that this existing variation data suffers from ascertainment bias with respect to its focus on coding sequence, and as yet undiscovered noncoding variants may also play an important role in phenotypes associated with G6PD deficiency.

Population genetic studies at *G6PD* have examined the genetic architecture surrounding these functional variants, with the aim of providing corroboration of the malaria protection hypothesis. Such work has documented low levels of genetic diversity on functional allele haplotypes, including reduced microsatellite [121] and nucleotide [122] diversity, and long stretches of haplotype homozygosity several hundreds of kilobases in span [123], findings which are unusual for genetic variants present at high frequency and suggestive of a rapid rise in gene frequency in the recent past. Unbiased, genome-wide scans for positive selection, however, have not yet identified *G6PD* as a selection target [126, 127], which may be reflective of unusually high LD in the region generally [161], which itself may be affected by gene proximity-dependent patterns of mutation pressure on the X chromosome [203]. Further investigation of haplotype diversity with denser marker coverage, including variants specific to malaria-endemic populations, may prove useful in resolving the issue of positive selection at the *G6PD* locus.

3.2.2 Clinical implications of functional genetic diversity

One underlying premise for this dissertation concerns the idea that functional genetics at the *G6PD* locus cannot be fully encapsulated by the effect of a single nonsynonymous variant, as is commonly assumed. An underestimation of the effects of other variants at the locus could lead to enhanced Type II error in association studies where only one variant is studied, and thus erroneously lead to the conclusion that G6PD deficiency is not involved in a particular clinical phenotype or lead to conflicting results for studies conducted in populations of different genetic background. As there is a long history of conflicting

results in association studies looking at G6PD deficiency and severe malaria, it seemed plausible that unrecognized functional allelic heterogeneity may play a role in the between-study heterogeneity that has been observed.

3.2.3 A worldwide survey of genetic variation at *G6PD*

I set out to further investigate the malaria protection hypothesis, and genetic variation at *G6PD* generally, with an eventual aim toward conducting case-control association studies of severe malaria in a diverse worldwide cohort. A preliminary to this sort of fine-mapping genetic association work is understanding genetic variation at the locus in different populations where different functional mutations may predominate. Here I present an approach to these questions in three parts: (1) assembly of a database of known functional variants, (2) resequencing of the gene in several hundred individuals from sub-Saharan Africa and southeast Asia, and (3) dense genotyping of variants present at the locus to validate novel resequencing variants and provide empirical insights into population genetics at the locus.

3.3 Materials and Methods

3.3.1 Literature search

Expert-authored review articles on G6PD variants included those by Beutler and Vulliamy [103, 204, 94]. Controlled PubMed searches were conducted by concatenating the term 'glucose 6 phosphate dehydrogenase' with country or region names pertinent to MalariaGEN Consortial Project 3 (see Chapter 2 for details regarding MalariaGEN and its consortial projects). For example, for Vietnam this included: (a) the country itself and any aliases – 'Vietnam', 'Viet Nam,' and 'French Indochina'; (b) the geographic region in which it lay – 'Asia', 'southeast Asia'; (c) any adjacent countries or areas of shared ancestry – 'Cambodia', 'Laos', 'Thailand', 'Burma', and 'China'. Specific studies that proved useful to complement the review articles included work by Hirono and Beutler [205], Vulliamy et al. [206], Ganczakowski et al. [207], Tishkoff et al. [121], Saunders et al. [122, 124], Sabeti et al. [123], Verrelli et al. [208], and Hue et al. [101]. A total of 133 literature-described SNPs were identified, including 107 nonsynonymous polymorphisms. Twenty-two of these variants were observed in more than one published report, and are therefore likely common.

3.3.2 Database assembly

Each variant found in the literature survey and the dbG6PD database [209] was first validated through manual and semi-automated means, with mapping validity checks for various coordinate systems (genomic, cDNA, peptide), and functional consequence (i.e. specific amino acid change for a nonsynonymous variant), using a combination of custom Perl scripts and web-based tools available at the UCSC Genome Browser [210] and Galaxy [211] websites. Exon-intron location was determined using boundary coordinates from Ensembl transcript 'G6PD-201.' Ancestral alleles were inferred using Ensembl's EPO pipeline, which uses multiple sequence alignment of chimpanzeee, macaque, and orangutan sequence to infer ancestral allele identity. The reference sequence used was *Homo sapiens* assembly contig 'AF277315.6', and all coordinates are shown for build GRCh37/hg19. Peptide and cDNA coordinates are relative to the peptide sequence start site at chrX:153775085. Clinical severity of known deficiency variants was categorized according to WHO guidelines on a scale from 1 (severely deficient) to 4 (not deficient) as previously reported [204]. The BioPerl Ensembl API [184] was used to verify the aforementioned data, and also add to it by providing all predicted functional consequences for each variant, which in several cases involved multiple transcripts in the same gene, and even multiple genes (*IKBKG* and *G6PD* share a common promoter).

3.3.3 Capillary sequencing

Resequencing at *G6PD* was performed on a total of 288 population-representative samples, with 48 samples each coming from six countries, including Burkina Faso, Cameroon, Nigeria, Kenya, Vietnam, and Papua New Guinea. Overlapping amplicons were generated by PCR to cover approximately 15 kb at the locus, with a small gap in coverage at intron 1 (Figure 3.1). Capillary sequence traces were analyzed

as part of the ExoSeq pipeline at the Wellcome Trust Sanger Institute, with SNPs called using an in-house algorithm called ExoTrace. As this algorithm was found to exhibit poor specificity, extensive manual curation of individual traces was necessary to retain only high confidence SNPs. Pairwise validation with genotype data from the same individuals (see below) was performed using custom Perl scripts.

3.3.4 Multiplexed genotyping

Study cohort The sample set was a cohort comprising MalariaGEN Consortial Project 3 [186], and included 1671 population controls from twelve malaria-endemic countries, including Burkina Faso, Cameroon, Ghana, The Gambia, Kenya, Mali, Malawi, Nigeria, Papua New Guinea, Sudan, Tanzania, and Vietnam (Table 3.1). All samples collected were obtained after informed consent, and appropriate permissions to conduct this study were obtained from the Oxford Tropical Research Committee (U. of Oxford), the Medical Research Council (UK), and regional and national ethics boards at individual study sites (refer to section 2.1.5 for protocol details).

Table 3.1: Samples used in genotyping study. The number of samples from each country included in the CP3 genotyping study are shown stratified by gender. Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Ghana (GH), The Gambia (GM), Kenya (KE), Mali (ML), Malawi (MW), Nigeria (NG), Papua New Guinea (PG), Sudan (SD), Tanzania (TZ), and Vietnam (VN).

	BF	CM	GH	GM	KE	ML	MW	NG	PG	SD	TZ	VN
Male	68	28	49	133	66	36	54	29	126	58	112	49
Female	76	30	56	137	89	55	40	30	125	69	107	49
TOTAL	144	58	105	270	155	91	94	59	251	127	219	98

Assay design Of the variants listed in the master database (including those identified via resequencing), 308 met various inclusion criteria prioritizing function, proximity to *G6PD*, and estimated frequency in our dataset. These were then clustered into three priority groups (Table 3.2):

- High (25) – solely SNPs of known functional or population genetic significance, with strong population- or region-specific evidence
- Medium (72) – primarily variants identified during resequencing (see above), with a few lower priority literature variants
- Low (211) – solely dbSNP/Ensembl variants

Table 3.2: Priority scoring of SNPs in pilot phase.

	Initial	*Met Inclusion Criteria*	*Priority 1*	*Priority 2*	*Priority 3*
Known Functional/Significant	105	40	25	15	0
Novel Resequencing	60	57	0	57	0
Other (e.g. dbSNP)	410	211	0	0	211
TOTAL	575	308	25	72	211

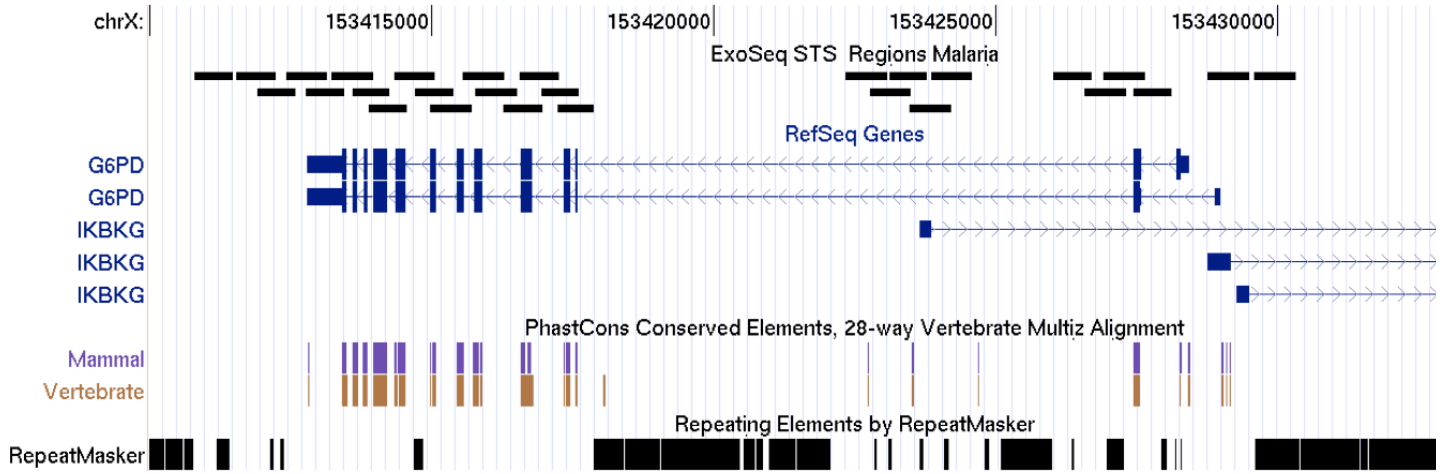


Figure 3.1: MalariaGEN CP3 capillary sequencing coverage. This UCSC Genome Browser graphic illustrates PCR amplicon coverage (top track, black) for capillary resequencing at *G6PD*. The other tracks show genes annotated by RefSeq (blue), conserved sequence as inferred by PhastCons (purple and brown), and repetitive sequence elements marked by RepeatMasker (bottom track, black). *IKBKG*=inhibitor of NF-kB kinase γ. Figure generated by K. Kivinen, WTSI.

Each list was internally ranked and input sequentially to Sequenom MASSarray iPLEX assay design software, such that multiplexes were built around higher priority SNPs first. Polymorphisms and flanking sequences submitted were force-coded on the (+) strand, SNP-masked using a custom Perl script, and double-checked against the UCSC reference sequence. Ultimately, four multiplexed genotyping reactions were designed, assaying a total of 135 SNPs across the locus, the majority clustered in a 35 kb window flanking *G6PD*, with 11 additional SNPs located 50-70 kb telomeric to the gene. There is a gap in SNP coverage in the region 40 kb telomeric to *G6PD* due to the presence of a segmental duplication in the region due to doubly-mapping SNPs (Figure 3.2).

Laboratory workflow Genotypes were obtained by the Sequenom MASSarray iPLEX platform (described in detail in Chapter 2). A total of 135 SNPs were assayed, and of these, 104 met manual cluster plot curation QC standards, with 78 of these being polymorphic in at least one subpopulation (Table 2.6).

3.3.5 Genetic analyses

Custom scripting was done in Perl to standardize data format, perform a few basic analyses, and generate custom-formatted files for PLINK [182], which was used to curate the data set. SNPs were removed if they exhibited minor allele frequency (MAF) < 0.001 or missingness > 0.20. Individuals were excluded if they exhibited missingness > 0.35. A total of 51 polymorphisms and 1671 individuals met these inclusion criteria. Using the statistical software R [181] and the 'snpMatrix' package [183], we performed further checks on the data, including examining individual population data for extreme deviation from Hardy-Weinberg equilbrium ($P < 10^{-5}$), the relationship between individual call rate and heterozygosity, and gender differences in MAF. Wright's F_{ST} was calculated using the following formula [212]:

$$F_{ST} = \frac{(p_1 - p_2)^2}{(p_1 + p_2)(2 - [p_1 + p_2])} \tag{3.1}$$

Principal components analysis (PCA) on mean-centered derived allele frequency (DAF) data [213, 201] was performed using the 'prcomp' function in R. Phylogenetic trees were calculated using PHYLIP [214]. First, bootstrap replicates (N = 20) of DAF data stratified by country or ethnic group membership were generated, with each replicate sampling a certain number of individuals at random (80 per country, 20 per ethnic group). Next, these allele frequency tables were used to generate consensus trees using maximum likelihood estimation (in PHYLIP, 'contml'), via a method which assumes genetic distance between subpopulations is a function of genetic drift. Trees were rendered using TreeViewX [215]. Haplotypes were inferred using fastPHASE [185] under default settings, with known male haplotypes used as a reference input to improve phasing accuracy in females. Pairwise difference between haplotype sequences (PWD) was calculated by counting the number of allele differences found on a pair of randomly chosen haplotypes, with means calculated via bootstrap replication (N=1000).

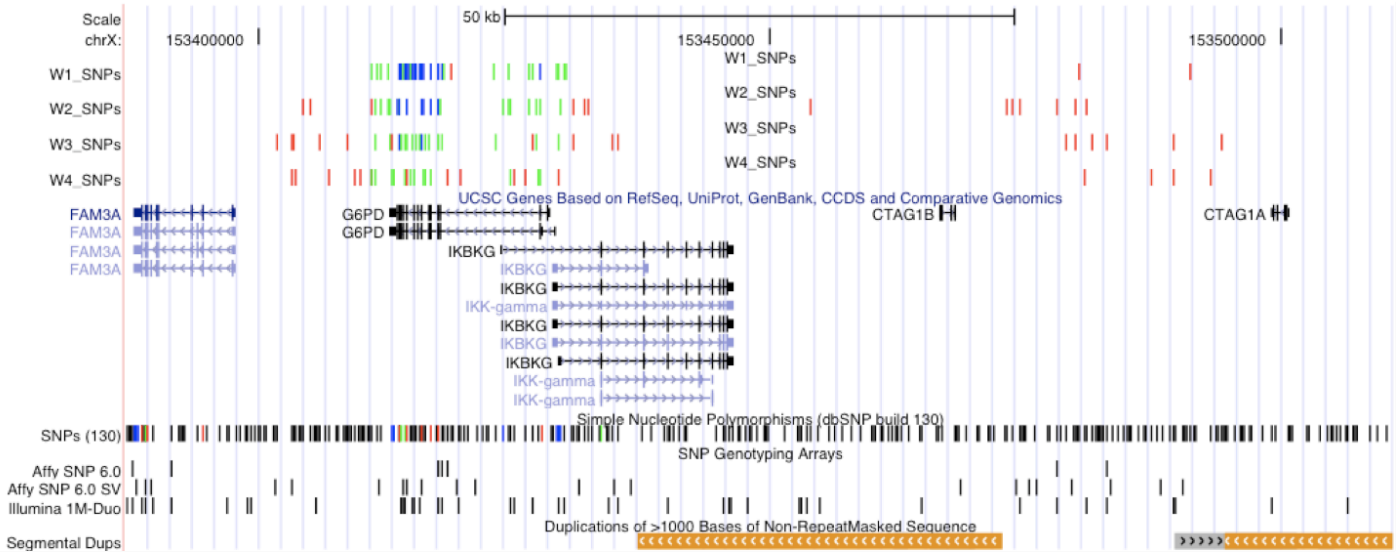


Figure 3.2: Genomic location of pilot phase SNPs. Variants included in the pilot phase study are shown in the top track and color-coded by priority – high (blue), medium (green), and low (red). Also shown are: genes annotated by UCSC Genome Browser (middle track, blue and black), SNPs found in dbSNP 130 and on common genotyping arrays (lower tracks), and segmental duplications (bottom track, orange). *IKBKG*=inhibitor of NF-kB kinase γ.

3.4 Results

3.4.1 Assembling and curating a database of *G6PD* variants

To finely map genetic association with a particular phenotype of interest, one would ideally type all known variation at and around a given candidate locus. Over the past fifty years, natural variation at *G6PD* has been extensively documented throughout the world, by hematologists and population geneticists alike. I have mined this data from the literature, collated it with information from other public repositories, and used modern bioinformatic tools to assemble an internally- and externally-consistent database of all known variants at *G6PD*. This provides an important supplement to the population-specific resequencing work, as the latter was limited in several ways, including: (a) having low sensitivity to detect rare variants due to small sample size, (b) exhibiting incomplete sequencing coverage owing to gaps, and (c) lacking enrichment for functional variant discovery as the resequencing panel comprised population-representative samples as opposed to solely G6PD deficient individuals.

3.4.1.1 Literature search

I first performed a systematic review of the literature, beginning with several expert-authored review articles [103, 204, 94] and the locus-specific online database, dbG6PD [209]. This was then supplemented by country- and region-specific PubMed searches to find existing resequencing and dense genotyping surveys (see Methods for full list). The majority of these 133 literature-described variants are likely functional as most are nonsynonymous polymorphisms detected via cDNA sequencing of individuals with deficient enzyme activity. As most of these are also sporadic mutations, validation evidence was also noted, and specifically whether the variant had been observed in multiple published accounts. This step provided a qualitative estimate of allele frequency for such variants, and allowed for the prioritization of polymorphisms likely to be relatively common in our data set, as these would be the ones for which we would be best-powered to detect associations with biochemical and clinical phenotypes of interest.

3.4.1.2 Data curation and standardization

After merging the above list with uniquely-mappable variants listed in dbSNP and Ensembl that flank *G6PD* ($\pm$50 kb), I set about curating and standardizing the data set. This process was especially important for known literature-described deficiency variants, most of which did not have entries in public databases and were identified before the human genome was assembled, and were thus described using somewhat arbitrary coordinate systems. In order to be able to design primers for the Sequenom genotyping assays, it was necessary to firmly localize each of these variants.

First, each polymorphism was mapped to the current genome build (GRCh37/hg19) using contextual clues from the literature, such as predicted amino acid change, cDNA position, restriction enzyme site sequence, etc. I also determined exon/intron location for all SNPs located at *G6PD* itself, as well as cDNA and peptide sequence position for all coding SNPs. Next, I verified the two alleles observed (there

were only biallelic SNPs in our list), determined which one was ancestral (according to Ensembl's EPO pipeline), and forced allele coding to the positive strand. I determined the functional consequence of each SNP for every transcript with which it was associated (as Xq28 is a gene-dense region, several SNPs were associated with more than one transcript); for example, whether it was coding (nonsynonymous, synonymous), splice-site, etc. Importantly, multiple approaches were taken, with an initial manual and semi-automated approach later validated and supplemented by an automated Perl- and mySQL-based database construction process using the BioPerl suite of bioinformatic tools [184] and web-based tools available at the UCSC Genome Browser [210] and Galaxy [211] websites.

Regarding nomenclature for the named G6PD variants, there were two main issues to consider. First, many named G6PD variants are genetically-identical, but were once thought to be electrophoretically- and biochemically-distinct and were isolated independently in different parts of the world, and many different aliases can be associated with the same genetic variant (the opposite can also be true, e.g. 'A-' can refer to a cDNA mutation at either position 202 or 968). Second, the established convention includes all nonsynonymous alleles as a haplotype in describing certain variants. So the major 'A-' African variant is described as 202T/376C, as 202T invariably occurs on a 376C background. These issues were addressed here by using one unique primary identifier for each SNP, and a list of all alias identifiers associated with that SNP, along with any dbSNP identifiers (e.g. 'rs' IDs). Thus, a database query for the 'A-' alias identifier will yield five results: four for the 202T, 376C, 542A, 680A, and 968G alleles, and one for the 376C background that all lie upon.

3.4.2 Variant discovery

As a preliminary to fine-mapping association studies at *G6PD* across the MalariaGEN sample set, efforts at variant discovery were undertaken. Resequencing at *G6PD* was performed on 288 samples from six countries (48 samples each) as part of MalariaGEN Consortial Project 3, including Burkina Faso, Cameroon, Nigeria, Kenya, Vietnam, and Papua New Guinea. As it was economically unfeasible to obtain sequence data for all MalariaGEN study sites, these sample sets were chosen to represent west Africa (Burkina Faso, Cameroon, Nigeria), east Africa (Kenya), and the two relatively genetically-divergent populations of Vietnam and Papua New Guinea in southeast Asia. The Nigerian cohort was also useful as an internal control, as it consisted of HapMap Yoruba parents for whom corroborating data was available from three sources, including HapMap, the 1000 Genomes Project, and the Illumina pooling experiment described in Chapter 4. Sample sets were gender-balanced and representative of the population at large. Although a modest increase in sensitivity to detect less common variants was possible if only females were typed (as males are hemizygous, i.e. have only one X chromosome), males were included to allow for the possibility of gender-specific differences in allele frequency. Sequence was generated using PCR amplicons tiling approximately 15 kb at the locus, with a gap in coverage at intron 1 (Figure 3.1) due to failed PCR amplification. At the time of analysis, emphasis was not placed on

identifying insertion-deletion polymorphisms (indels) present in the sequence data for two main reasons. First, the ExoSeq pipeline was unable to reliably call these variants, and second, the downstream genotyping platform used in subsequent association studies (Sequenom iPLEX) would be unable to type such indel variants. This limitation precluded our analysis of this important class of SNPs which represents about 10% of all known single nucleotide variation in the genome.

The quality of data generated was assessed via multiplexed genotyping of thirty-two SNPs in the same samples (see below). A total of 31 of 32 (96.9%) variants detected via resequencing demonstrated heterozygosity via genotyping in one or more populations. There was also strong agreement between individual genotype calls obtained by the two methods, as mean pairwise concordance was 97.5%, with discordant outliers traceable to specific sample sets (Figure 3.3), which may be indicative of poor sample quality or a batch effect.

A total of 67 SNPs were identified, 49 of which were novel (Table 3.3, Figure 3.4). While the total number of SNPs identified is far less than the number present in dbSNP in the same region (about 200), this finding is unsurprising for a variety of reasons. First, many of the SNPs present in dbSNP have poor validation evidence, having only been observed by one study, or found through bioinformatic means via comparison of sequences deposited in RefSeq, and many of these also map to more than one location in the genome. Second, the populations sampled for sequencing and genotyping in dbSNP are more genetically divergent than the populations included in this study, where four of six groups were of shared Bantu ancestry. Interestingly, though, as the populations of interest here are underrepresented in dbSNP, these ascertainment differences also explain why so many novel variants were discovered in this project, highlighting the importance of population-specific variant discovery. Finally, as this study only included 48 individuals per population, it was not well-powered to detect rare variants which represent the majority of polymorphic sites in the genome [162].

Overall, mean DAF was greater in African samples than in Asian samples, as many variants were monomorphic in Vietnam and Papua New Guinea. Several of the SNPs appeared to be private to either Africa or Asia, with primarily rare variants found in Papua New Guinea (Table 3.4). The pattern of genetic diversity overall was consistent with that observed genome-wide, where African populations generally exhibit greater nucleotide diversity.

Several of the polymorphisms may be potential regulatory variants, including four SNPs found in the 3' untranslated region (UTR) and one found in the 5' untranslated region (UTR). The 5' UTR is especially interesting as several lines of evidence seem to indicate that it is a shared promoter sequence for both *G6PD* and *IKBKG*. First, it is a site of DNAse hypersensitivity, i.e. a region of open chromatin (euchromatin) associated with high transcriptional activity. Additionally, it is associated with certain

modified histones such as H3K27Ac (histone H3 acetylated at lysine 27) known to be associated with active promoters. Finally, it is the site of several transcription factor binding sites as determined by ChIP-seq (chromatin immunoprecipitation followed by sequencing). Such evidence has been made possible by harnessing the throughput of next-generation sequencing technologies (discussed in Chapter 4) to generate unbiased, fine-resolution maps of all sequence associated with any number of transcription factors and other histone markers of active transcription. Polymorphisms found at such sites will be more likely to have the most direct and specific effects on gene expression, though no expression quantitative trait loci (eQTLs) have yet been identified at the locus, suggesting that much remains to be elucidated regarding known regulatory variants that might impact gene expression [216]. It is also worth noting that if a bidirectional promoter does indeed exist at the locus, this may have implications for gene expression regulation by noncoding RNAs [217]. A gene graphic illustrating some of the known features of gene expression at *G6PD* is shown in Figure 3.5.

Coding variants were found in exons 6 and 10, and included three synonymous variants and one sporadic nonsynonymous variant at position chrX:153760883 (G1187C, Pro396Ala) found in one individual in Papua New Guinea. This latter mutation is likely functional as it involves a similar residue change as that found in severe deficiency variant G6PD Bari (Pro396Leu).

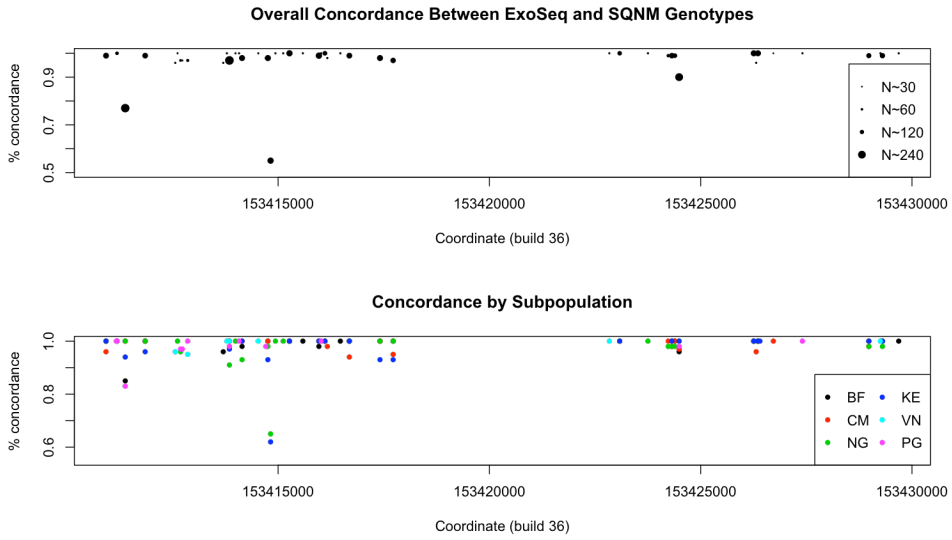


Figure 3.3: Concordance between resequencing and genotyping data. Overall (top) and country-stratified (bottom) percent pairwise concordance between genotypes obtained by capillary resequencing versus those obtained Sequenom genotyping. Size of points in top panel is proportionate to the number of pairwise genotype comparisons made (N). Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Kenya (KE), Nigeria (NG), Papua New Guinea (PG), and Vietnam (VN).

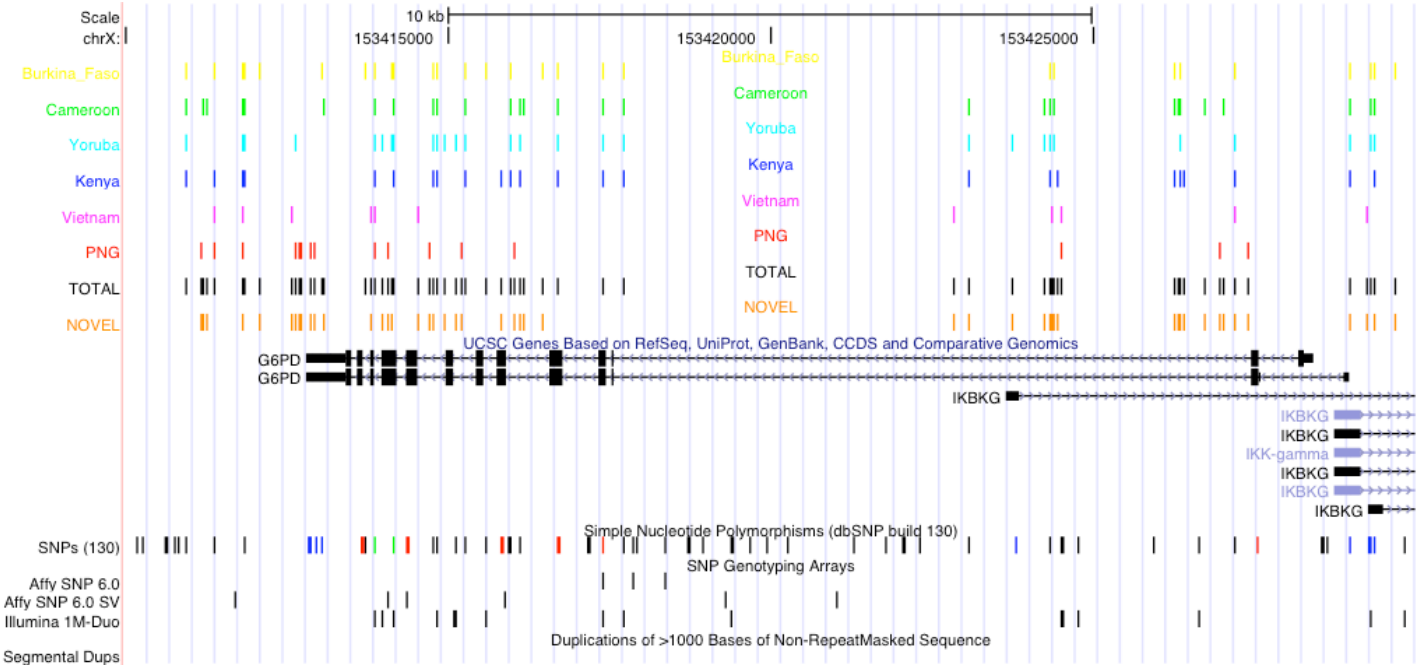


Figure 3.4: MalariaGEN CP3 capillary sequencing polymorphisms. This UCSC Genome Browser graphic shows the genomic location of polymorphisms discovered during resequencing, with variants color-coded by country. Also shown are: genes annotated by UCSC Genome Browser (middle track, blue and black), SNPs found in dbSNP 130 and on common genotyping arrays (lower tracks), and segmental duplications (bottom track, none shown).

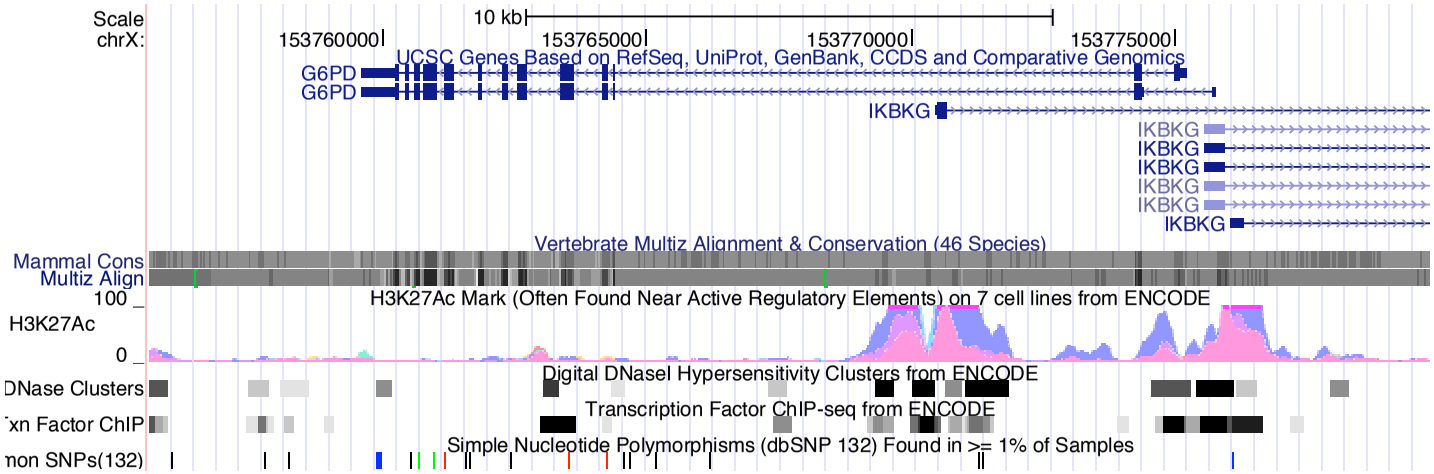


Figure 3.5: Transcriptional regulation at *G6PD*. This UCSC Genome Browser graphic shows the *G6PD* and *IKBKG* genes (top track) along with conserved sequence inferred by PHAST (middle tracks) and features associated with active regulatory elements as annotated by the ENCODE project, including H3K27 (histone 3, lysine 27) acetylation marks, transcription factor ChIP-seq (chromatin immunoprecipitation followed by sequencing), and DNase hypersensitivity sites. Also shown are common SNPs found in dbSNP 132 (bottom track).

Table 3.3: Variants identified in resequencing study. A total of 67 variants were identified, including 49 novel variants (with prefix 'NT'). Asterisks indicate colorNewChangeColor nonsynonymous variants. Exon (E) and intron (I) locations are also indicated, along with variants identified upstream (US) and downstream (DS) of extitG6PD. Position is specified in GRCh37/hg19 coordinates. Alleles shown are on the (+)-strand as given by UCSC Genome Browser, with ancestral allele inferred by Ensembl's EPO pipeline. Refer to section 3.3.2 for coordinate and exon mapping details.

Name	Position	Observed Alleles	Ancestral Allele	G6PD - Exon/Intron
rs12389568	153757733	G/A	G	DS
rs12389569	153757734	G/A	G	DS
NT_011726.13_4578452	153757978	C/G	C	DS
NT_011726.13_4578468	153757994	G/A	G	DS
NT_011726.13_4578537	153758063	C/A	C	DS
rs2515906	153758189	G/A	A	DS
NT_011726.13_4579089	153758615	G/A	G	DS
rs12393550	153758660	G/A	G	DS
NT_011726.13_4579361	153758887	T/C	T	DS
NT_011726.13_4579846	153759372	C/T	C	DS
NT_011726.13_4579900	153759426	C/A	C	DS
NT_011726.13_4579907	153759433	C/T	C	DS
NT_011726.13_4579971	153759497	C/T	C	DS
NT_011726.13_4580014	153759540	G/A	G	DS
NT_011726.13_4580141	153759667	G/A	G	E13
NT_011726.13_4580192	153759718	A/T	A	E13
rs1050757	153759858	T/C	C	E13
NT_011726.13_4580339	153759865	C/T	C	E13
NT_011726.13_153413060	153759866	C/T	C	E13
NT_011726.13_4581068	153760594	G/A	G	I11
rs2230037	153760654	G/A	G	E11
NT_011726.13_4581271	153760797	G/A	G	E10
** NT_011726.13_4581357	153760883	G/C	G	E10
NT_011726.13_4581409	153760935	G/A	G	E10
** Viangchan	153761337	C/T	C	E9
nt3042	153761515	C/T	C	I8
NT_011726.13_4582038	153761564	G/A	G	I8
rs5986990	153761628	G/A	G	I8
NT_011726.13_4582217	153761743	T/C	T	I8
NT_011726.13_4582401	153761927	T/C	T	I7
NT_011726.13_4582483	153762009	G/A	G	I7
rs2515905	153762075	G/A	G	I7
rs5986875	153762392	G/A	G	I6
NT_011726.13_4583104	153762630	G/A	G	E6
rs2515904	153762771	G/C	G	I5
NT_011726.13_4583299	153762825	C/T	C	I5
NT_011726.13_4583388	153762914	G/A	G	I5
NT_011726.13_4583445	153762971	G/C	A	I5
NT_011726.13_4583753	153763279	T/C	T	I5
** G6PD_376	153763492	T/C	T	E5
** G6PD_202	153764217	C/T	C	E4
rs762515	153764528	T/C	T	I2
NT_011726.13_4590119	153769645	C/A	C	I2
NT_011726.13_4590363	153769889	G/A	G	I2
NT_011726.13_4591032	153770558	C/T	C	I2
NT_011726.13_4591512	153771038	T/C	T	I2
NT_011726.13_4591599	153771125	A/G	A	I2
NT_011726.13_4591624	153771150	G/A	G	I2
NT_011726.13_4591669	153771195	T/C	C	I2
NT_011726.13_4591736	153771262	C/T	C	I2
rs2472393	153771296	C/T	T	I2
NT_011726.13_4593536	153773062	C/T	C	I2
NT_011726.13_4593593	153773119	G/A	G	I2
NT_011726.13_4593634	153773160	A/C	A	I2
NT_011726.13_4593682	153773208	G/C	G	I2
NT_011726.13_4594000	153773526	A/G	A	I2
NT_011726.13_4594246	153773772	G/C	G	I2
NT_011726.13_4594297	153773823	G/A	G	I2
rs7064060	153773998	C/T	C	I2
NT_011726.13_4594481	153774007	G/A	G	I2
NT_011726.13_4594688	153774214	T/C	T	I2
NT_011726.13_4596259	153775785	T/C	C	US
NT_011726.13_4596533	153776059	C/T	C	US
rs5986992	153776107	C/A	C	US
NT_011726.13_4596651	153776177	G/T	G	US
NT_011726.13_4596966	153776492	G/T	G	US
NT_011726.13_4597283	153776809	C/T	C	US

Table 3.4: Resequencing allele frequency summary. Derived allele frequency distribution for variants identified during resequencing is shown, with novel variants are listed in parentheses. Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Kenya (KE), Nigeria (NG), Papua New Guinea (PG), and Vietnam (VN).

	BF	CM	NG	KE	VN	PG
DAF > 0.05	19 (6)	18 (9)	16 (6)	19 (9)	5 (2)	1 (0)
DAF < 0.05	9 (6)	11 (9)	12 (10)	6 (4)	7 (7)	15 (13)
Total	28 (12)	29 (18)	28 (16)	25 (13)	12 (9)	16 (13)
Mean DAF	0.101	0.059	0.056	0.067	0.009	0.018

3.4.3 Mapping global patterns of variation at *G6PD*

Studying genetic architecture at *G6PD* in diverse populations is of interest for both population genetic and disease association studies. I analysed genotypes at 51 polymorphic loci for 1671 individuals from twelve MalariaGEN partner sites in Burkina Faso, Cameroon, Ghana, The Gambia, Kenya, Mali, Malawi, Nigeria, Papua New Guinea, Sudan, Tanzania, and Vietnam. Using this data, genetic and haplotypic diversity were analysed, both within and between populations across our data set.

3.4.3.1 Multiplex genotyping assay design and quality control

Typed polymorphisms were ascertained according to a priority weighting scheme that took into consideration functional significance, validation evidence, and genomic proximity to *G6PD* (Table 3.2). This prioritization scheme was adopted as part of an effort to develop a set of multiplex genotyping assays at *G6PD* to be taken forward into large case-control association studies of severe malaria as part of MalariaGEN Consortial Project 1 [186]. The highest priority was given to nonsynonymous variants previously associated with G6PD deficiency trait (i.e. named variants such as '202/A-' or 'Viangchan') and other polymorphisms of known population genetic significance. The next highest priority was given to SNPs that had a strong likelihood of being polymorphic in at least one population in our sample set, based on evidence from resequencing or genotyping surveys. Finally, multiplexes were filled out using remaining SNPs that were closest to the gene itself.

Four multiplex genotyping reactions were designed to type 135 SNPs. A total of 104 of these met cluster plot quality control standards, of which 78 were polymorphic in at least one of the study populations, including 15 high, 42 medium, and 21 low priority variants. Further quality control filters were applied at the analysis stage, where SNPs were excluded if they exhibited MAF < 0.001 or missingness > 0.20, and individuals were excluded if the exhibited missingness > 0.35. A total of 51 polymorphisms and 1671 individuals met these criteria for inclusion (Table 3.5).

Table 3.5: Polymorphic variants in genotyping study. A total of 51 variants met inclusion criteria in the genotyping study ($MAF > 0.001$ and missingness < 0.20). Asterisks indicate colorNewChangeColor non-synonymous variants, two of which, G6PD376 (A+) and rs34193178 (Mira d'Aire), are bystander mutations not associated with G6PD deficiency (i.e. WHO class 4). Exon (E) and intron (I) locations are also indicated, along with variants identified upstream (US) and downstream (DS) of *G6PD*. Position is specified in GRCh37/hg19 coordinates. Average DAF for combined African populations (AFR), Papua New Guinea (PG) and Vietnam (VN) are shown at right.

Name	Position	Observed Alleles	Ancestral Allele	G6PD - Exon/Intron	AFR	PG	VN
rs28470352	153753490	T/A	T	DS	0.344	0.000	0.000
rs61042368	153755336	G/A	G	DS	0.124	0.000	0.000
rs1894260	153755989	C/T	C	DS	0.055	0.029	0.034
rs12389569	153757734	G/A	G	DS	0.069	0.000	0.000
G6PD_X_b36_153411172	153757978	C/G	C	DS	0.001	0.006	0.000
G6PD_X_b36_153411188	153757994	G/A	G	DS	0.003	0.000	0.000
rs12393550	153758660	G/A	G	DS	0.338	0.000	0.000
G6PD_X_b36_153412861	153759667	G/A	G	E13	0.000	0.032	0.014
G6PD_X_b36_153413623	153760429	G/A	G	E12	0.080	0.000	0.000
rs2071429	153760508	A/G	A	I11	0.923	0.116	0.155
** Union	153760605	G/A	G	E11	0.000	0.019	0.000
rs2230037	153760654	G/A	G	E11	0.309	0.074	0.150
rs2230036	153760953	C/T	C	E10	0.120	0.000	0.000
** rs34193178	153761160	C/G	C	E9	0.005	0.000	0.000
** hG6PD_968	153761240	A/G	A	E9	0.007	0.000	0.000
G6PD_X_b36_153414709	153761515	C/T	C	I8	0.002	0.011	0.000
G6PD_X_b36_153414758	153761564	G/A	G	I8	0.125	0.000	0.000
rs5986990	153761628	G/A	G	I8	0.007	0.000	0.000
G6PD_X_b36_153414937	153761743	T/C	T	I8	0.007	0.000	0.000
rs2515905	153762075	G/A	G	I7	0.164	0.000	0.000
rs5986875	153762392	G/A	G	I6	0.012	0.000	0.000
** hG6PD_542	153762655	T/A	T	E6	0.003	0.000	0.000
rs2515904	153762771	G/C	G	I5	0.164	0.000	0.000
G6PD_X_b36_153416019	153762825	C/T	C	I5	0.000	0.045	0.000
G6PD_X_b36_153416108	153762914	G/A	G	I5	0.040	0.000	0.000
** Vanua Lava	153763485	A/G	A	E5	0.000	0.013	0.000
** hG6PD_plus376	153763492	T/C	T	E5	0.341	0.000	0.000
** hG6PD_plus202	153764217	C/T	C	E4	0.111	0.000	0.000
rs762515	153764528	T/C	T	I2	0.361	0.000	0.000
rs762516	153764663	C/T	C	I2	0.163	0.000	0.000
rs762517	153764734	G/A	G	I2	0.164	0.000	0.000
rs743544	153765166	G/A	G	I2	0.000	0.095	0.224
G6PD_X_b36_153423083	153769889	G/A	G	I2	0.020	0.000	0.000
G6PD_X_b36_153424232	153771038	T/C	T	I2	0.007	0.000	0.000
G6PD_X_b36_153424319	153771125	A/G	A	I2	0.357	0.000	0.000
G6PD_X_b36_153424389	153771195	T/C	C	I2	0.888	1.000	1.000
rs2472393	153771296	C/T	T	I2	0.070	0.923	0.590
G6PD_X_b36_153426256	153773062	C/T	C	I2	0.113	0.000	0.000
G6PD_X_b36_153426354	153773160	A/C	A	I2	0.046	0.000	0.000
G6PD_X_b36_153426720	153773526	A/G	A	I2	0.008	0.000	0.000
G6PD_X_b36_153427408	153774214	T/C	T	I2	0.000	0.013	0.000
G6PD_X_b36_153428979	153775785	T/C	C	US	0.498	1.000	1.000
rs5986992	153776107	C/A	C	US	0.010	0.000	0.000
G6PD_X_b36_153429686	153776492	G/T	G	US	0.003	0.000	0.000
rs7887673	153824899	G/A	G	US	0.027	0.000	0.000
rs5986997	153827549	C/T	C	US	0.004	0.000	0.000
rs4898389	153827637	A/G	A	US	0.927	0.077	0.313
rs5986877	153828269	C/G	C	US	0.919	0.077	0.137
rs7879049	153829693	A/G	A	US	0.374	0.074	0.137
rs7053878	153834100	T/A	T	US	0.061	0.717	0.082
rs60030796	153836171	A/G	A	US	0.103	0.000	0.000

3.4.3.2 Population genetic studies

Across the data set, a wide range of derived allele frequencies (DAF) was observed, with some derived alleles approaching fixation in Asia and rare in Africa, and vice-versa (Figure 3.6).

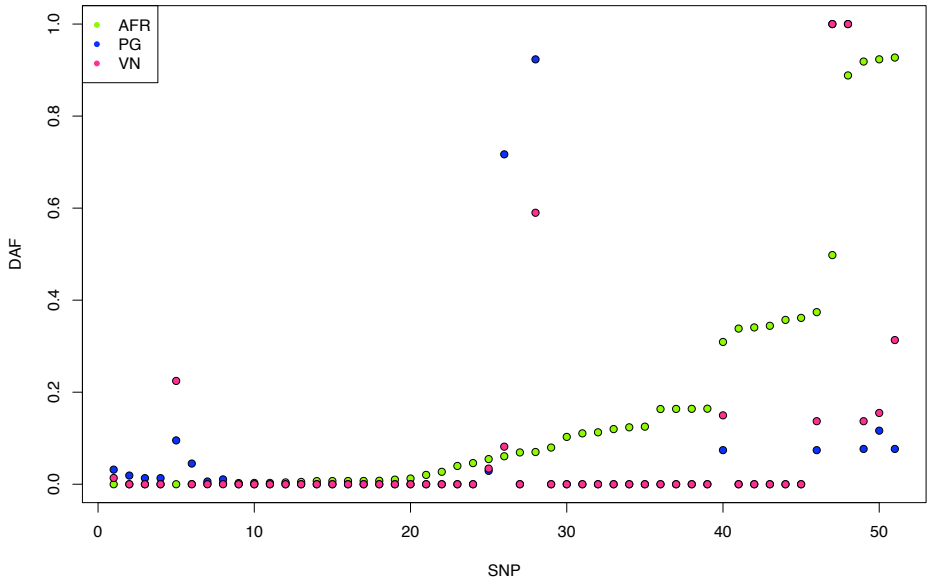


Figure 3.6: Minor allele frequencies stratified by subpopulation. Minor allele frequencies for all 51 SNPs surveyed at the *G6PD* locus, stratified by subpopulation, and ordered by DAF in African subpopulation.

A variety of approaches were taken to further characterize this population variation, including measuring Wright's F-statistic (F_{ST}) in each population, multivariate decomposition using principal components analysis, and the construction of phylogenetic trees. Wright's F_{ST} quantitates the proportion of allele frequency variance explained by subpopulation membership. In measuring F_{ST}, one can see largely two clusters, one for the African subpopulations and another for the Asian ones (Figure 3.7), suggesting that population differentiation at this locus mirrors that seen across the genome as a whole, lending weight to the serial founder theory of human migration [218].

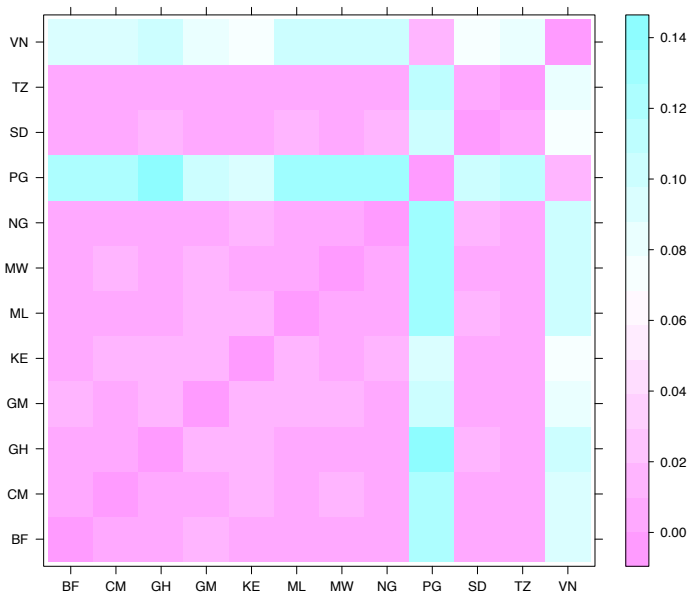


Figure 3.7: Heatmap of mean pairwise F_{ST} values. Mean pairwise F_{ST} between the 12 subpopulations in the study. Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Ghana (GH), The Gambia (GM), Kenya (KE), Mali (ML), Malawi (MW), Nigeria (NG), Papua New Guinea (PG), Sudan (SD), Tanzania (TZ), and Vietnam (VN).

I also used allele frequency data at each SNP in principal components analysis (PCA, reviewed in Chapter 2: Supplemental Information), a technique which is useful in summarizing multidimensional data sets in terms of the major components of their variance. Here, I found the same two clusters as before, with the first principal component sharply separating African from Asian populations, and the second component distinguishing between Vietnam and Papua New Guinea (Figure 3.8). Complementary results were obtained when examining particular ethnic groups, with those from Vietnam (Kinh) and Papua New Guinea (Madang, Alotau, Rabaul, and Ilahita) sharply differentiated from African groups along the first principal component, though subtle variation within sympatric groups was also readily apparent.

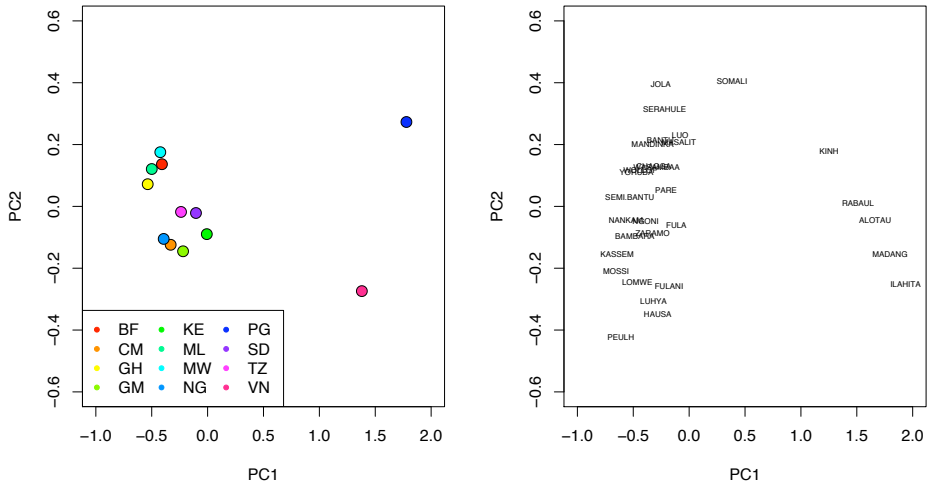


Figure 3.8: Principal components analysis using allele frequency data. Allele frequency data across all 51 SNPs was used to distinguish clusters of relatedness via PCA, using country- (left) and ethnicity-stratified (right) datasets. Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Ghana (GH), The Gambia (GM), Kenya (KE), Mali (ML), Malawi (MW), Nigeria (NG), Papua New Guinea (PG), Sudan (SD), Tanzania (TZ), and Vietnam (VN). All ethnic groups are African except for those from VN (Kinh) and PG (Madang, Alotau, Rabaul, and Ilahita).

Finally, allele frequency data was also used to generate maximum likelihood estimates for phylogenetic relationships (techniques reviewed in Chapter 2: Supplemental Information) at *G6PD* between different subpopulations. The consensus cladograms obtained (Figure 3.9) mirror the PCA results in showing a clear separation between Asian and African populations, at the level of both country and ethnic group membership, and further, generally comport with language group relations – for example, all three ethnic groups speaking the Nilo-Saharan language (Luo, Somali, and Masalit) clustered with one another in the tree.

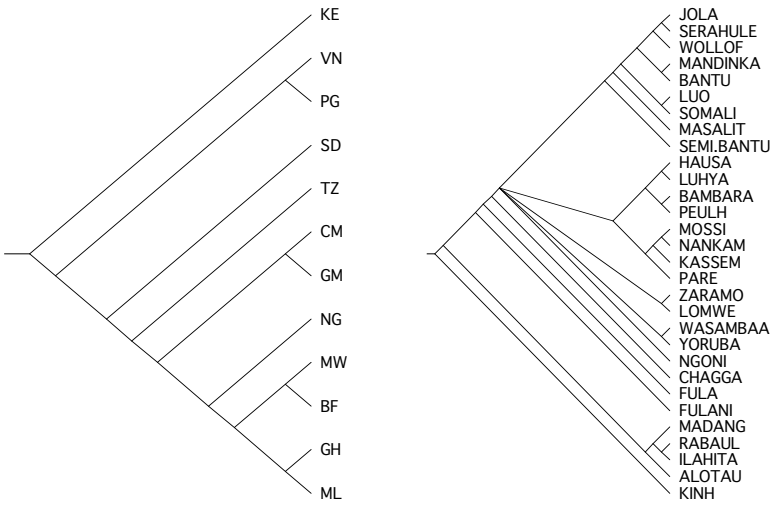


Figure 3.9: Phylogenetic trees. Maximum likelihood cladograms were constructed using allele frequency data under a model of genetic drift using PHYLIP. Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Ghana (GH), The Gambia (GM), Kenya (KE), Mali (ML), Malawi (MW), Nigeria (NG), Papua New Guinea (PG), Sudan (SD), Tanzania (TZ), and Vietnam (VN). All ethnic groups are African except for those from VN (Kinh) and PG (Madang, Alotau, Rabaul, and Ilahita).

I also examined the prevalence and genetic architecture of known functional alleles at *G6PD*, i.e. those with a strong clinical association to G6PD deficiency trait. In all countries except Vietnam, at least one of these alleles was present at frequency greater than 1%, and while the 202T allele was the dominant functional allele in most African countries, significant allelic heterogeneity was present in the West African countries of the Gambia and Mali, as well as in Papua New Guinea (Figure 3.10).

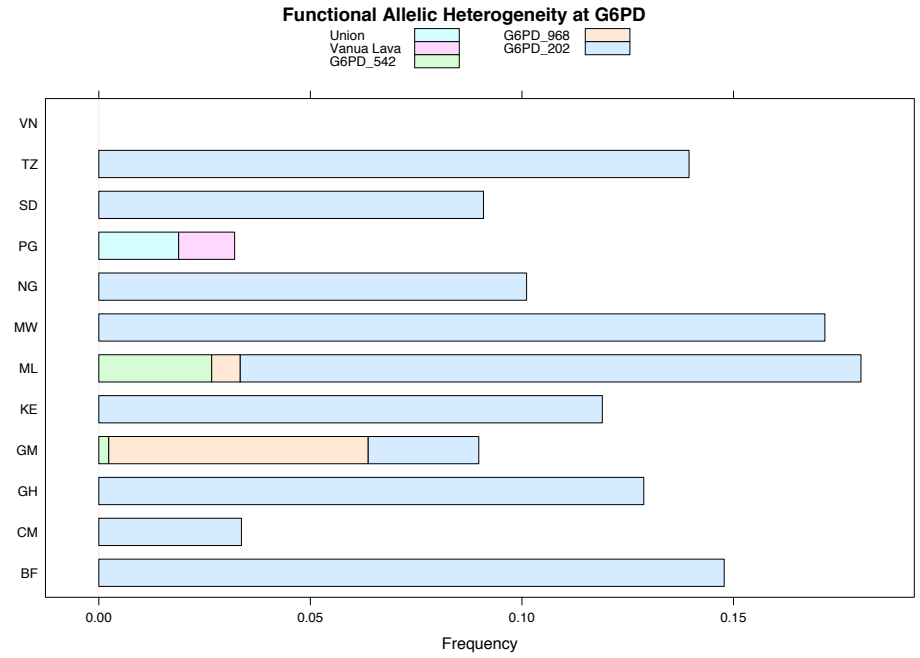


Figure 3.10: Allele frequency summary for known functional variants at G6PD. Allele frequency data is shown for the five known nonsynonymous variants associated with G6PD deficiency that were found in our dataset, with the Gambia, Mali, and Papua New Guinea exhibiting allelic heterogeneity. Countries are coded as follows: Burkina Faso (BF), Cameroon (CM), Ghana (GH), The Gambia (GM), Kenya (KE), Mali (ML), Malawi (MW), Nigeria (NG), Papua New Guinea (PG), Sudan (SD), Tanzania (TZ), and Vietnam (VN).

I further examined the haplotype structure upon which G6PD functional alleles lie, specifically looking at the two common deficiency alleles in the dataset, 202T ('G6PD A-') and 968G ('G6PD Betica'), both of which exist on a 376C derived allele background ('G6PD A+'). In comparing these haplotypes to ancestral haplotype sequence (376T, which is termed 'G6PD B' by convention), one sees greater haplotype diversity in B than A+, and near complete homogeneity among functional allele haplotypes, which is evident both visually (Figure 3.11) and numerically, in a quantitative comparison of the average number of pairwise differences in haplotype sequence (B = 5.06, A+ = 3.44, 202T = 0.15, 968G = 0.21). These results indicate that B is somewhat older than A+, and that the two functional alleles are relatively young, as comparatively less time has elapsed to allow for mutation and recombination to break down LD with neighboring sites. The recent origin of these functional alleles coupled with their high frequency may suggest a selection process or a demographic change such as a bottleneck or rapid population expansion.

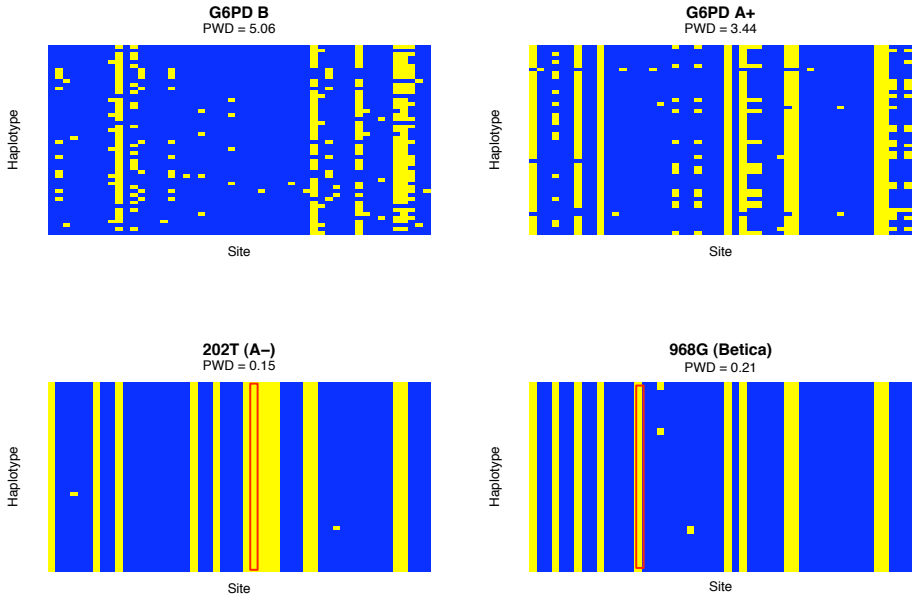


Figure 3.11: Functional haplotype diversity at *G6PD*. Haplotypes bearing common known nonsynonymous alleles associated with G6PD deficiency (202T and 968G [red outline]) are compared with their shared progenitor (376C, A+) and the ancestral background sequence (B). Fifty haplotypes were sampled at random for B, A+, and 202T, and all 26 haplotypes are shown for 968G. Mean number of pairwise sequence differences between haplotypes (PWD) obtained by bootstrap replication (N=1000) is shown at the top of each panel. Sites are shown across the x-axis and individual haplotypes piled along the y-axis, with ancestral (blue) and derived (yellow) alleles differentiated by color.

3.5 Discussion

The work presented here set out to organize, add to, and better understand the picture of global genetic diversity at the *G6PD* locus, a crucial first step in conducting disease association and population genetic studies. The locus-specific database discussed here should be a useful general resource for researchers interested in *G6PD*, while the novel variants described and allele frequency data obtained through re-sequencing and dense genotyping at the locus adds depth to the understanding of genetic variation at *G6PD* generally, and functional allelic diversity specifically.

I anticipate that the database will grow to include more information on predicted protein changes for nonsynonymous variants from publicly-available bioinformatic tools. Such programs include older, sequence-based prediction methods such as PolyPhen or SIFT [219], which offer qualitative predictions of predicted functional consequence, though these are often of poor positive and negative predictive value as they rely heavily on generalized heuristics to make assumptions about ternary and quaternary structural changes from primary peptide sequence. As both three-dimensional structural and comparative genomics data exist for G6PD [220, 221], newer, structure- and evolutionary-based approaches such as that described by Bao and Cui [222] or found in PolyPhen-2 [223] can offer better predictive power, though no algorithm can obviate the need for detailed study of relevant molecular phenotypes as presented in Chapter 5. Nonetheless, it will be interesting to see what these modern algorithms will contribute to our understanding of how nonsynonymous variation might affect protein folding, as many common variants (e.g. G6PD 202) are thought to destabilize tertiary and quaternary structure of the enzyme [98], and still others (e.g. G6PD 376) appear to be bystander mutations with little or no independent effect [224, 94].

I also hope to incorporate some measure of population-specific relative allele frequency data, to help researchers better plan genetic association studies that achieve adequate power in the presence of allelic heterogeneity. The importance of accounting for allelic heterogeneity has already been shown to be relevant to detecting a malaria-protective signal in West Africa [143], and will undoubtedly be important to association studies in similarly heterogenous areas such as southeast Asia and Oceania as well [103]. Lamentably, many earlier surveys that identified G6PD protein electrophoretic variants were performed before the advent of DNA sequencing and PCR [90], so the DNA mutations for many deficiency variants may still be unknown, which justifies further targeted resequencing work.

The sequencing and genotype data discussed here represents a starting point – it was limited in its focus on variants present directly at the gene itself, and is most relevant for researchers studying the same populations. The limitations were largely imposed by a combination of economic and methodological limitations. Ideally, and perhaps in the near future, it will be cost-effective to have whole genome sequence data for all individuals in an association study, and thus be able to detect the effects of even rare SNPs and structural variants (i.e. insertions, deletions, copy number variants), but at present

only a piecemeal approach is possible, and both the sequencing and genotyping technologies used in the present survey are greatly limited with respect to throughput. One possibility, presented in Chapter 4, is the utilization of massively-parallel sequencing to accomplish the goals of variant discovery at greatly reduced time and cost, but even these improved sequencing modalities still require validation in the form of genotyping. The Sequenom platform is limited to a maximum of forty SNPs per well, and this theoretical maximum is rarely achieved when there are high-priority SNPs which must of necessity be present in a multiplex, thus constraining the available search space for other multiplex primers. Another technology that was considered was the Illumina GoldenGate custom bead array, though this was deemed cost-ineffective for the relatively small number of SNPs being genotyped.

In the case of G6PD deficiency, where there is documented allelic heterogeneity it is especially important to capture the majority of functional genetic variation in a population, and the data from Vietnam and Papua New Guinea suggest a shortcoming in this regard. In future work, it may be worth conducting targeted resequencing studies in individuals with documented G6PD deficiency in order to enrich for discovery of known functional alleles. Interestingly, the data from Africa shown here suggests that 202T ('A-') is the dominant functional allele throughout much of the MalariaGEN study sites, including areas in which association studies of severe malaria have demonstrated a clear protective role for the variant, such as Nigeria [138] and Mali [19], as well as those where evidence is lacking, such as Kenya [39]. If G6PD deficiency were indeed causal with respect to protection from severe malaria, then given a similar 202T allele frequency in each country, one would expect each study to be equivalently powered to detect a protective effect. However, as no other known functional allele was present, and no novel nonsynonymous variants were identified during resequencing, it is possible that other differences in host genetic background or environmental influences between these populations may account for the discrepancies found in these association studies. One possibility, discussed in Chapter 5, is the idea that genetic modifiers can alter the biochemical penetrance of the 202T allele, and thus have an effect on the distribution of G6PD deficiency trait in different populations, thus complicating interpretation of association studies in which only the 202/A- variant was genotyped. Another possibility, explored in Chapter 7, is the idea of environmental influence in the form of the predominance of specific subtypes of severe malaria in different parts of Africa, and the possibility that G6PD deficiency has a different relationship with respect to the different subtypes.

Also missing in this study was a more complete survey of potential long-range interactors present in promoter and enhancer regions that might affect G6PD activity by altering transcript levels. Interestingly, though the sequencing data at intron 1 is somewhat limited owing to repetitive sequence, we found a number of novel variants at putative regulatory regions shared with *IKBKG* (i.e. regions that are hypersensitive to digestion with DNase and colocalize with certain histone marks indicative of active regulatory elements [UCSC Genome Browser, data not shown]). Finally, with regard to population

genetics, while the present study confirms notions of a serial founder effect in human migration out of Africa and a recent origin for two African functional variants, the limited genomic window of focus did not permit long-range haplotype testing for selection [127], and this study was underpowered to be able to study homogeneity on other functional haplotypes, such as those from southeast Asia. Future work should ideally not focus on any single functional variant, as this underestimates not only the prevalence of G6PD deficiency, but also the extent of allelic heterogeneity underlying expression of the trait in many areas of the world.

3.6 Acknowledgments and Contributions

Literature search and database assembly were performed by SS. Capillary sequencing was performed by teams at the Wellcome Trust Sanger Institute led by Katja Kivinen (Team 112) and Allison Coffey (ExoSeq Team), with some assistance in sequence trace annotation provided by SS. Multiplexed genotype assays were designed by SS, with assistance from Kirk Rockett and Kathrin Schuldt, and performed by SS and Christina Hubbart. All data analysis, including writing of custom scripts, was conducted by SS.

Chapter 4

Detection and quantitation of population genetic variation using pooled massively parallel sequencing of whole-genome amplified DNA

In the previous chapter, I discussed the use of conventional capillary sequencing to survey global genetic variation at the G6PD locus. As this process is both time- and resource-limited, I set about developing novel methodology to address the same question, but at substantially reduced cost and time. Here I present proof-of-principle for the applicability of pooled massively parallel sequencing to sensitively detect and accurately quantitate population genetic variation.

4.1 Abstract

Massively parallel short read DNA sequencing technologies (MPS) have dramatically increased throughput at greatly reduced cost compared to conventional capillary sequencing, making possible the routine sequencing of entire genomes and transcriptomes. While the cost and throughput advantages are easily realized by sequencing projects that assay whole genomes on the order of mega- or gigabase pairs of DNA, investigators are often interested in deep resequencing of targeted genomic regions (e.g. a single gene) on the order of kilobases. Methods to harness MPS throughput capability towards more focused resequencing projects in a cost-effective manner will be essential for follow-up analysis in genome wide association studies, as a complete picture of population sequence variation at candidate loci is required in order to accurately identify causal variants associated with disease.

I describe here a simple, cost-effective approach where pooled MPS facilitated detection and quantitation of low frequency genetic variation at the *G6PD* locus in forty unrelated Yoruba individuals from Ibadan, Nigeria (International HapMap Consortium). Comparison of the results with individual genotype and sequence data from the same individuals demonstrated the sensitivity (92%) and specificity

(74%) of this method and its potential applicability for other focused resequencing projects, especially those where sample DNA quantity may be limited.

4.2 Introduction

Second-generation sequencing technologies use massively parallel sequencing (MPS) of short sequence read length to generate throughput on the scale of gigabases at one-hundredth the cost of conventional capillary sequencing. They have greatly enhanced the potential of molecular genetics, with independent investigators now able to produce data that was once only able to be generated at large sequencing centers. While the completion of the draft sequence of the human genome [159] required billions of dollars and took over a decade to complete, resequencing an entire human genome can now be completed in a matter of weeks [164]. Efforts by consortia such as 1000 Genomes Project [165] have sought to fully capture whole genome diversity in humans and other organisms, and even in complete microbial ecosystems [225]. Additionally, applications of MPS such as RNA-seq or ChIP-seq [226, 227] have the potential to supersede existing array-based technologies in cataloguing epigenetic, transcriptional, and structural variation in genomes.

But the minimal ascertainment bias that makes these new applications so powerful is also what limits the utility of MPS for more targeted sequencing work. When one's region of interest is orders of magnitude below the capacity of the platforms, the systems are inherently inefficient in terms of both cost and information content. With the increasing importance of genome-wide association studies (GWAS) in medical genetics, investigators have focused on refining the signal at putative disease-associated loci, and cataloguing genetic variation at target loci in specific populations of interest are critical to statistical imputation methods that seek to pinpoint causal variants. Pre-processing enrichment and pooling strategies are thus needed to circumvent platform-inherent inefficiencies and bring down the total cost of functional/causal variant discovery.

Target enrichment strategies include hybridization-based approaches and polymerase chain reaction-based (PCR) approaches. The former includes molecular inversion probe [228] and hybrid capture methods [229, 230], and takes advantage of programmable microarrays to generate hundreds of thousands of short oligonucleotides to capture target sequences. While such approaches allow great breadth of sequence capture, enabling enrichment of large targets on the order of megabases (e.g. exomes), this comes at a cost of limited sensitivity and specificity of target capture. Additionally, the high cost of custom arrays and hybridization equipment and steep sample quantity demands (tens of micrograms) can be prohibitive for many investigators. Uniplex, multiplex, and emulsion variations on PCR [231], while time-intensive and limited in throughput by amplicon and multiplex size, allow greater specificity and sensitivity of target enrichment, and are thus ideal for applications where the target region in on the order of kilobases.

Pooling strategies have been increasingly exploited to reduce per-sample sequencing cost and address the problem of post-enrichment capacity excess on MPS platforms. One set of methods involves 'DNA

barcoding' of individual samples via adaptation of PCR primer [231] or flow cell adaptor sequence [232, 233, 234] with short sequence tags (< 10 bp), such that the sample identity of a particular sequencing read may be obtained. As barcoding is expensive, laborious, and biased in target coverage, some researchers have tried untagged pooling strategies, either pooling genomic DNA before enrichment [235, 236], or enriching individual samples of genomic DNA (e.g. via PCR) to generate target libraries that were then pooled [237, 238].

Presented here is a set of simple laboratory and informatic methods where, starting with limited quantities of genomic DNA, I was able to reproducibly catalogue and quantitate low frequency genetic variation at the *G6PD* locus in a HapMap Yoruba cohort via untagged, pooled MPS (Illumina GA1) on PCR-enriched, whole-genome amplified DNA.

4.3 Materials and Methods

Workflow overview Whole genome amplified genomic DNA (gDNA) was used as a template for generating 5-6 kb long-range PCR amplicons tiling 29 kb across the *G6PD* locus for each of 40 individuals. For each individual, a target library consisting of amplicons combined in equimolar quantities was created to minimize variance in sequencing depth. Six gender-balanced pools of individuals were created (N = 2, 4, 8, 16, 32, and 40) and sequenced in duplicate on separate flow cell lanes using the Illumina GA1.

Sample set The sample set consists of 40 unrelated Yoruba (YRI) individuals (20 male, 20 female) from Ibadan, Nigeria. These samples have been extensively genotyped by the International HapMap Consortium[161] and resequenced at several candidate loci (including *G6PD*) by the MalariaGEN consortium [186].

Sample preparation Genomic DNA for each individual was amplified via multiple displacement amplification [187], performed in quadruplicate using two different commercial kits ('Repli-G' [Qiagen] and 'Genomiphi' [GE Healthcare]), to minimize representational bias and/or artifacts from any individual reaction.

This amplified product served as template for long-range PCR reactions using 'Platinum *Taq*' (Invitrogen). Eighteen 5-6 kb overlapping amplicons tiled across roughly 29 kb at/around *G6PD* locus were designed on SNP- and repeat-masked genomic sequence using SNPmasker [239] and Primer3 [240]. Eight of these amplicons (Figure 4.1) proved robust in initial testing and were produced successfully for the full sample set (population sampling coverage = 100%), with only one gap in sequence coverage of 692 bp across the region (sequencing coverage = 97.7%). PCR reaction conditions and primer pairs are shown in Tables 2.7 and 2.8. PCR products were quantitated using PicoGreen (Invitrogen), and amplicons were then combined in equimolar amounts for each individual to generate amplicon libraries.

Amplicon libraries from each individual were then pooled in equimolar amounts to generate six pools of different size, with the number of individuals in each pool being N = 2, 4, 8, 16, 32, or 40. Each pool contained a total of 0.5 μg of DNA. Pre-processing and sequencing were then performed according to a modified Illumina protocol [176].

Mapping sequence reads Each single-end, 36 cycle sequencing run yielded approximately 6 million sequence reads. These reads were then mapped to a 147 kb reference sequence (chromosome X assembly contig 'AF277315.6') with the program MAQ [188] using default settings via the 'easyrun' Perl script pipeline included with the software, and allowing for only 2 mismatches per alignment, a setting designed to balance alignment specificity and SNP detection sensitivity concerns. The map file created by MAQ was then used to create a pileup sequence file (Figure S4.1), which shows, for each position along the

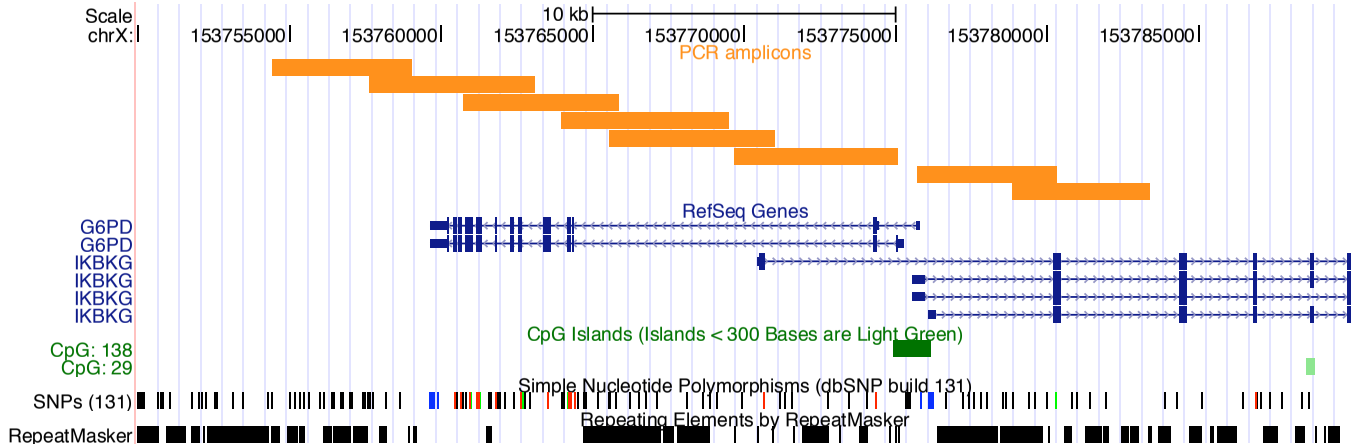


Figure 4.1: Amplicon coverage at *G6PD* locus. There are eight 5-6 kb PCR amplicons (orange) tiling across 29 kb about the *G6PD* locus (blue), with one gap in coverage at a CpG island (green). SNPs found in dbSNP 131 and repetitive sequence marked by RepeatMasker are also shown (bottom tracks). *IKBKG*=inhibitor of NF-kB kinase γ.

reference sequence, information on read depth, base calls, base quality scores, map quality scores, and cycle number.

Transforming and parsing pileup sequence Each line of input in a pileup sequence file is an ASCII-encoded summary of information from all sequencing reads probing a single position along the reference sequence. Pileup files were read into custom Perl and R scripts (available upon request), and the following information was extracted or calculated for each base position along the reference:

a. general information –

   i. sequencing depth (i.e. how many sequencing reads probe a particular position), both raw and normalized by 'PCR redundancy' (i.e. how many amplicons tile across a particular position)

   ii. the reference and most likely alternate allele (and observed nucleotide transition, e.g. "A->G")

   iii. GC content (calculated in 10 bp windows)

b. base call frequencies – what proportion of sequencing base calls matched the reference allele, and what proportion were alternate calls A, C, G and T

c. reference and alternate allele information – for both the reference and 'most likely' alternate allele (the alternate allele with maximum base call frequency at a particular position), information was stored for:

   i. average base quality score[1]

   ii. average map quality score[2]

   iii. average sequencing cycle number contributing to the call (and the difference between reference and alternate, which I termed 'cy_diff = cycle number difference')

   iv. strand support – number of reads probing positive and negative strands (and the ratio of the number of positive strand reads to the total number of reads, which I termed 'strand balance')

d. noise statistic – calculated by comparing the observed number of reads (O) from the most frequent alternate allele with the expected value (E) based on the total number of alternate allele reads at a particular position. It is a test statistic (U) that is distributed as chi-square (1 d.f.) under the null hypothesis (Figure S4.2):

$$U = \frac{(O-E)^2}{E} \tag{4.1}$$

[1] $BaseQ = -10log\Pr\{\text{base call is incorrect}\}$
[2] $MapQ = -10log\Pr\{\text{mapping assigment is incorrect}\}$

4.4 Results

4.4.1 Sequence quality and depth

Critical to polymorphism detection in this experiment was the availability of sequencing data of high quality, minimal systematic bias, and sufficient sequencing read depth.

The distributions of base and map quality (BaseQ and MapQ) scores were centered about their maximum possible values (BaseQ [median] = 59.5; MapQ [median] = 100.7), suggesting an overall high quality of reliable sequence obtained. Additionally, greater than 70% of sequence reads were successfully mapped during each sequencing run on average, consistent with mapping yields observed by others [236, 238], and providing read depth (RD) on the order of $4 - 6 \times 10^3$ (Table S4.1). Uniformity of read depth was also strong, with over 90% of sites exhibiting RD within twofold range of the median. Finally, per-site alternate allele frequency was found to be reproducible across duplicate sequencing runs, with Pearson correlation coefficients ranging from 0.93 to 0.99.

Several systematic biases affecting read depth were observed, the most prominent of which was PCR amplicon redundancy (number of amplicons tiling across a site), which accounted for 59% of total variance in read depth ($P < 2 \times 10^{-16}$, one-way ANOVA; Figure S4.3A and B). GC content also played a role, with both low and high extremes of GC content demonstrating markedly decreased read depth (Figure S4.3C). Additionally, it was found that sites with lower read depth tended to exhibit greater variance in mapping qualities (Figure S4.3D). Finally, in explaining obvious discontinuities in the read depth profile, it was clear that both highly repetitive sequence (Figure S4.3E) and PCR primer sequence (Figure S4.3F) had an outsize influence, as one might expect, given imperfect informatic (mapping algorithms) and experimental (post-PCR clean up) methods, respectively.

Sufficient sequence read depth is perhaps the most critical element to the sensitivity of polymorphism detection in a pooled MPS experiment. The dependency of sensitivity on read depth is governed by three factors:

a. the length of the target region (L)

b. the number of chromsomes in the pool (N_c)

c. the desired alternate allele frequency (AAF) detection limit for a particular SNP

The length of the target region (L) is important because with a fixed number of short reads in a flow cell channel ($R_N = 6 \times 10^6$ reads) and a limit on read length ($R_L = 36$ bp), the quantity of mappable sequence obtained is also constrained ($S = R_N \times R_L = 216$ Mb). The mappable sequence can be divided by the length of the amplified target region (L) to yield the maximum expected per-site read depth ($RD = \frac{S}{L}$). Thus, the longer the target region, the lower RD will be, given technical constraints on the

amount of mappable sequence that can obtained. So, for example, with the *G6PD* target region having $L = 29000$ bp, one would theoretically expect approximately 7400 reads to probe/tile a given site on average, which is in agreement with the experimental observation of $RD \approx 4 - 6 \times 10^3$ (as only 70% of reads generated were successfully mapped).

Owing to these constraints on read depth, there is also a limit to alternate allele detection sensitivity for a given quantity of mappable sequence. The power $(1 - \beta)$ to detect a given variant at a specified AAF given a specified N_c and RD can be defined as $Pr_A \times Pr_B$, where:

$Pr_A = 1 - (1 - AAF)^{N_c}$ probability of sampling the variant chromosome in a sample of size N_c

$Pr_B = 1 - (1 - 1/N_c)^{RD}$ probability of sampling the variant read given RD and pool frequency $1/N_c$

The terms Pr_A and Pr_B are antagonistic, reflecting the fact that power to detect variants at a particular frequency AAF is limited by N_c even at increasing RD as the limiting factor becomes the number of chromosomes pooled, and similarly, when $N_c > RD$, the available read depth is not sufficient to adequately sample all N_c chromosomes (Figure S4.4).

In the dataset considered here, with an effective $N_c = 80$, and $RD \approx 4 - 6 \times 10^3$, each individual chromosome[3] contributing to the pool was sequenced to approximately 50x depth, a redundancy that provided sufficient power to detect singleton mutations.

4.4.2 Polymorphism detection algorithm

Polymorphism detection algorithms must balance dueling concerns of sensitivity and specificity, and thus are ultimately user-defined and somewhat subjective. My heuristic algorithm works by filtering a list of putative SNPs based on user-adaptable threshold criteria for a set of quality control metrics. The metrics considered in the algorithm are:

a. AAF

b. Noise statistic

c. Base quality score (BaseQ)

d. Cycle number difference

e. *optional* – Map quality score (MapQ), read depth, strand support

The first filter is AAF, which would ideally be the only necessary filter, but as 8% of the sites in my dataset demonstrated $AAF > 1\%$, it is likely that using only an AAF threshold would have yielded many false positive hits, given that only 1% of all sites in the genome are thought to be polymorphic. In order

[3] In sample pooling prior to sequencing, males and females were treated equivalently, thus rendering males effectively 'diploidized'.

to minimize Type I error and false discovery rate, one must choose the AAF threshold relative to the base call/mismatch error rate (E), which serves as a practical check on the lower limit for SNP detection. The greater E is, the lower the specificity in detecting rare variants. Practically speaking, this means one cannot reliably detect variants at $AAF \leq E$ without concurrent increases in false discovery rate. In my experiments, I measured per-site base call error rate as the mean of the two least likely alternate allele frequencies (i.e. the alleles that were neither the major nor the putative minor allele), and found an average value across the datasets of $E = 1.95 \times 10^{-3}$, suggesting that the maximum chromosome pool size to allow singleton detection was $N_c \leq 500$.

To help distinguish true polymorphism from base call errors, I defined a noise statistic U (defined above), which quantitates the extent to which the observed number of reads (O) from the most frequent alternate allele at a site differs from the expected number (E) based on the total population of alternate allele reads. The principle, which borrows from concepts in information entropy, is that a site with a spurious polymorphism will demonstrate greater disagreement between O and E, and thus have higher U.

Two other important, interrelated filters were base quality and cycle number difference. As base quality is inversely correlated with cycle number (in later cycles, there is decreased signal to noise ratio, due to nascent strand asynchrony and incomplete fluorophore/reversible terminator cleavage), this is an important limitation of sequence read length on the Illumina platform. Alternate allele calls exhibited significantly greater variance in cycle number compared to reference allele calls ($P < 2.2 \times 10^{-16}$, F-test), and base quality was concomitantly diminished at higher cycle numbers (Figure S4.5). Cycle number difference was defined as the difference between the average reference and alternate allele cycle number at a site. Unlike an absolute cycle number cutoff, this measure allowed for regional variation in reference cycle number, e.g. permitting SNP calls at the extremes of an amplicon.

In order to evaluate any interdependency or redundancy in the primary filter set, pairwise Spearman rank correlation tests were performed on all four metrics. An ellipse plot of the correlation matrix (Figure S4.6) showed little correlation between individual measures, validating the independent utility of each to the filtering algorithm. Additionally, simulations were performed to assess the properties of each metric with respect to sensitivity and specificity of the algorithm as a whole (Figure S4.7). Alternate allele frequency had the greatest dynamic range of impact on algorithm performance, so in choosing default settings for the algorithm as a whole, I sought to maximize sensitivity with respect to this filter, while prioritizing specificity for the other three metrics (Table 4.1). A graphic demonstrating the role of each algorithm component in distinguishing true SNPs from spurious ones in a 1 kb region containing two well-known SNPs representing the common G6PD deficiency haplotype ("A-") found in Africa is shown in Figure 4.2.

Although I found the aforementioned four filters to adequately balance sensitivity and specificity for

my data set, included in the algorithm is a set of optional filters which may prove useful to other investigators. These include: (a) a minimum MapQ score for alternate allele calls, (b) a read depth range for allowed SNP sites, and (c) a 'strand balance ratio' range for the allowed proportion of alternate allele calls that come from the positive strand (which ideally would be 50%).

Table 4.1: SNP filtering algorithm criteria.

Category	*Default criteria*	*Example* (run_ID = 1012_s_8)
Alternate allele frequency	min: at least one chromosome	0.0125
Noise statistic	max: 75% quantile	2.67
Base quality (alternate allele)	min: 90% quantile	55.6
Cycle number difference (alternate allele)	max: 33% quantile	3.9
Strand balance ratio	range: 0.25 to 0.75	0.25 - 0.75
Read depth (corrected)	range: bounded by 1% and 99% quantiles	969 - 7288
Map quality (alternate allele)	min: 50% quantile (median)	75.6

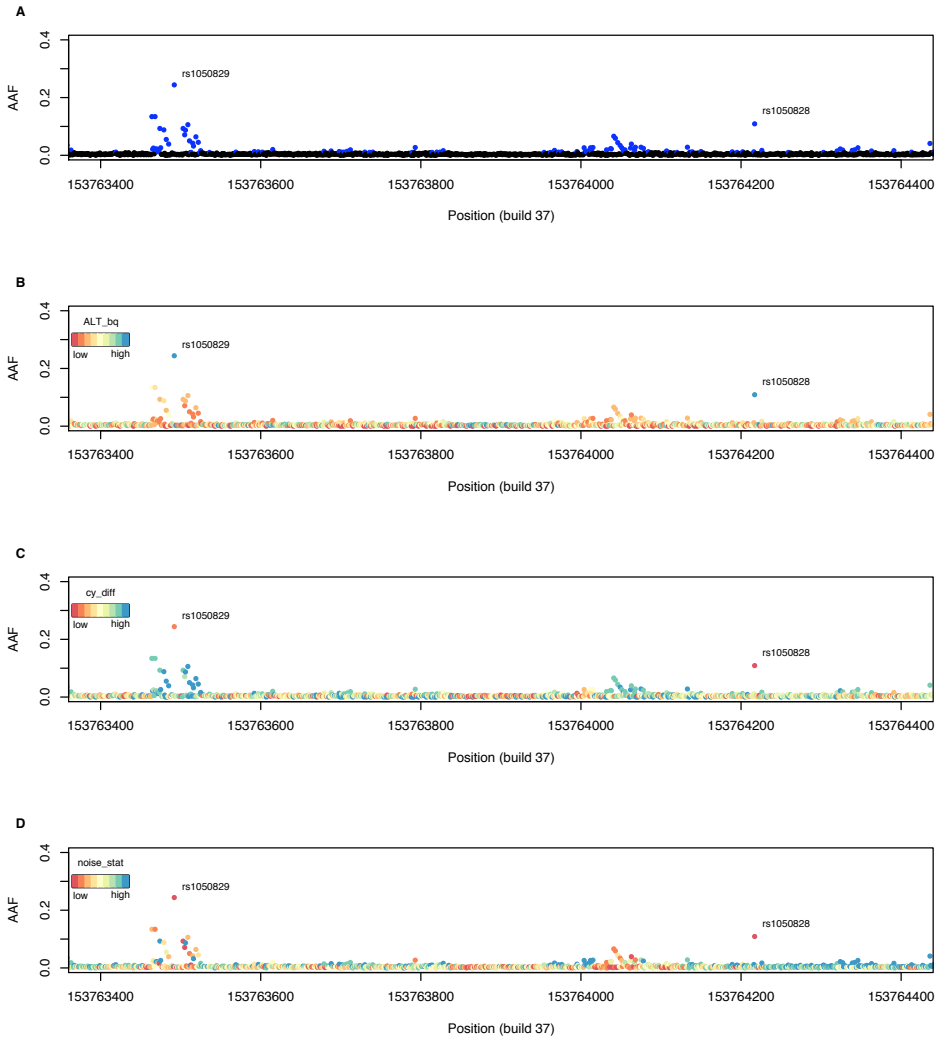


Figure 4.2: Demonstration of filtering algorithm. In this example, across 1 kb of sequence, 106 sites exhibit alternate allele frequency (AAF) greater than or equal to 1% (a – blue). However, when one examines these putative SNPs with respect to algorithm metrics, it becomes clear which of these are likely to be true variants, and which are likely false positive hits. A real SNP will have high base quality (b – blue), low cycle number difference (c – red), and low noise statistic values (d – red). This stretch of sequence contains two well-known SNPs (rs1050828 = "G6PD 202", rs1050829 = "G6PD 376") which together define the common enzyme deficiency haplotype ('A-') seen in sub-Saharan Africa.

4.4.3 Performance characteristics

Overall, across nearly 30 kb of sequence, my approach identified 66 putative SNPs in the $N = 40$ pool, including 23 of 25 known variants (sensitivity = 92.0 %). Additionally, when compared to known genotype data (from HapMap and MalariaGEN), the expected allele frequencies exhibited strong correlation with the observed AAF from pooled MPS ($R^2 = 0.93$, Figure 4.3). Similarly strong correlation was seen across technical duplicates and in different pool sizes (Figure S4.8). Comparing the putative SNP list with those from an in-house database of *G6PD* variants (compiled primarily from dbSNP 131 and MalariaGEN; see Chapter 3) suggested a maximum false discovery rate of about 26% (17/66), though finding even 66

variants across 30 kb of sequence would only reflect a SNP density at the locus of 1 per 450 bp, which comports with the current estimate of human genome SNP density of 1 per 300 bp. Interestingly, even at large pool sizes, specificity seemed to remain steady at around 70-80%, suggesting that increased pool size was not associated with a concomitant rise in the false discovery rate (Figure 4.4). A full summary of SNP detection algorithm performance and alternate allele frequency distributions for all runs is shown in Table S4.2 and Figure 4.4 respectively.

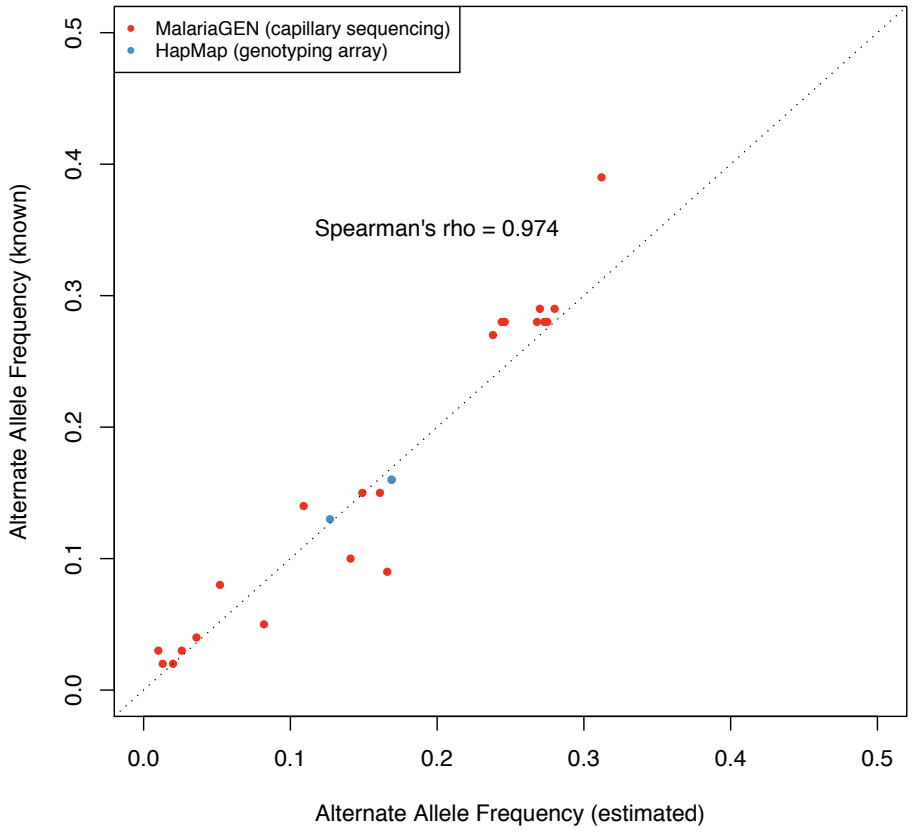


Figure 4.3: Estimated versus known alternate allele frequency for N=40 pool. Estimated alternate allele frequency (AAF) derived from the proportion of alternate allele calls probing a site in pooled MPS is plotted against known AAF derived from genotype or capillary sequence data from the same individuals. The $y = x$ diagonal (dotted line) is shown for reference.

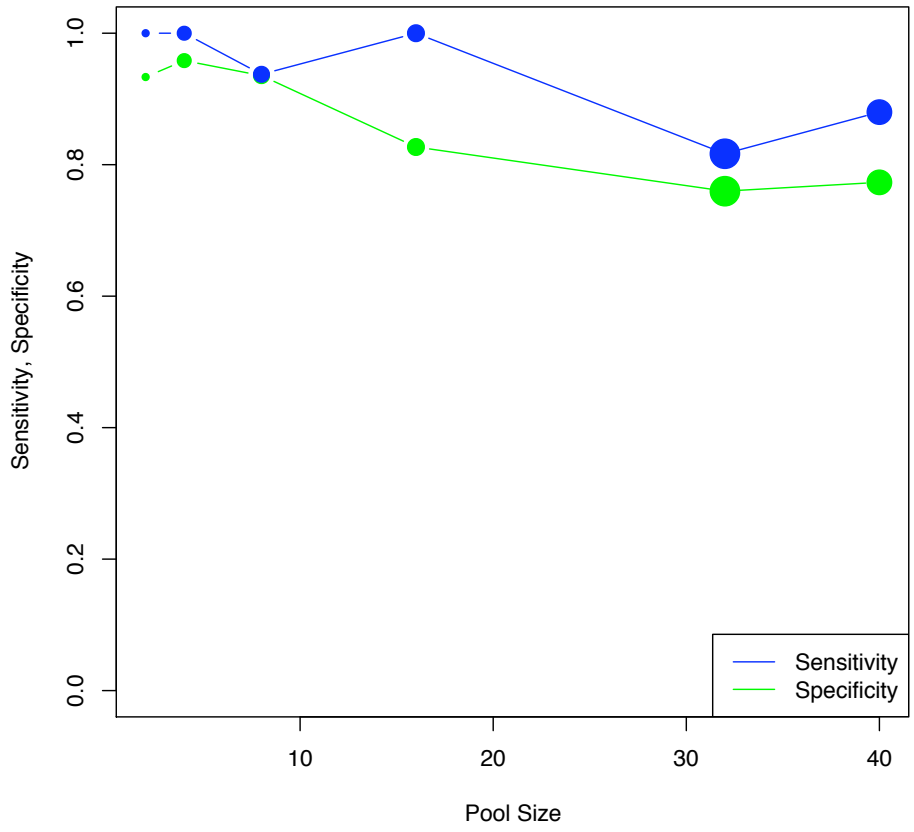


Figure 4.4: Sensitivity and specificity across different pool sizes. Sensitivity and specificity of the method as calculated by comparison with known genotype (HapMap, MalariaGEN) and variation (dbSNP 131) data. The size of the circle denotes the relative number of detectable SNPs at each N.

4.5 Discussion

Approaches to harness the considerable capacity of massively parallel sequencing technologies toward targeted resequencing projects are an important adjunct to large-scale genomics work. The set of laboratory and informatic methods presented here enabled rapid detection and quantitation of population genetic variation at the *G6PD* locus in a cohort of forty individuals. The approach is simple, fast, cost-effective, reproducible, and employs common laboratory techniques. Additionally, the present study demonstrated the feasibility of using whole genome amplified DNA, making this method especially useful to investigators working with limited quantities of genomic DNA, e.g. clinical or archeological specimens.

While target enrichment strategies based around programmable microarrays are well-suited for large scale targets (e.g. exomes), their considerable cost and limited specificity make PCR-based approaches especially appealing to researchers interested in accurate and comprehensive deep resequencing at smaller regions of interest on the order of kilobases. And with any enrichment strategy, researchers must also grapple with how best to maximize sequencing throughput while maintaining adequate sequencing depth, both in terms of per-sample cost and information content. The smaller the target region, the greater the need to exploit sample pooling strategies to minimize per-sample sequencing redundancy. My approach to the questions of MPS target enrichment and cost efficacy involved deep resequencing of PCR-amplified untagged sample pools. The principle merits of this approach lie in simplicity and reproducibility– it employs common laboratory methods, interfaces seamlessly with existing informatics tools, and was found to be consistent across multiple sequencing runs. The chief limitations of my method lie in the target enrichment process, principally the time and cost associated with optimizing long-range PCR reactions and achieving accurate quantitation to minimize coverage bias. One could significantly reduce this time and cost by pooling gDNA prior to target enrichment and/or using multiplexed PCR, though in the latter case one would likely sacrifice sequence depth uniformity as a result.

Several unique features of this work may provide useful insights for investigators interested in SNP detection via pooled MPS.

First, I demonstrated here the feasibility of pooled MPS using whole genome amplified DNA. While a parallel experiment using genomic DNA was not conducted, the quantity (66 SNPs across 30 kb = 1 per 450 bp) and AAF distribution of putative SNPs identified for each pool suggest no inflation in false positive hits, and are consistent with both whole genome estimates of SNP density (1 per 300 bp) and AAF distribution[161], as well as those from other candidate gene MPS pooling experiments in the literature [236, 238]. Thus, even with the additional set of multiple displacement amplification steps present in my laboratory protocol, any error and/or biases introduced seem to be minimal. My approach then may prove especially useful for investigators working with older archived samples whose biological import may be immeasurable but whose quantity is in short supply.

Additionally, owing to the massive sequencing depth achieved, I was able to obtain accurate estimates of base call (mismatch) error rate E on the Illumina platform, which was estimated at 0.19% overall (E = mean of the two least frequent allele calls at a site). Interestingly, the rate was about 50% greater at sites where the reference allele is A/T (0.23%) versus C/G (0.15%), suggesting that there may be sequence context-dependent constraints on the specificity of pooled MPS in detecting rare genetic variation, a phenomenon I also observed empirically in local 'hotspots' enriched for putative SNPs with low base qualities and high noise statistic values.

Finally, given the particular samples used (HapMap Yoruba) and the genomic region considered (*G6PD*), I was able to use existing genotype and sequence data to accurately gauge the sensitivity and specificity of the technique. In the largest pool ($N = 40$), 92% (23/25) of known variants were detectable, with a maximum false discovery rate of only 26% (17/66). I was also able to provide an assessment of how pool size impacts SNP detection and quantitation. As expected, sensitivity and specificity were maximal at the lowest pool sizes given the constrained polymorphism search space, but interestingly specificity remained relatively steady as pool size increased from $N = 16$ to $N = 40$, which suggested that the algorithm components designed to maximize specificity were working properly. However, I do expect that at higher N, where the singleton AAF detection threshold starts to approach E, a concomitant increase in false discovery rate may be inevitable.

As researchers gain increasing power to pinpoint causal genetic variants with newer, more comprehensive technologies, an integral part of the follow-up experiments to genome-wide association studies (GWAS) will involve deciphering the role of low frequency variants in common disease. Combined target enrichment/pooling methods such as the approach described here will thus be essential for rapidly and cost-effectively assaying genetic variation at candidate loci in large case-control cohorts, thereby accelerating the discovery pipeline and moving one step closer to clinical translation.

4.6 Acknowledgments and Contributions

Experimental design was the work of S.S. All laboratory work prior to Illumina sequencing was performed by S.S. Illumina sequencing was the work of a team led by Michael A. Quail at the Wellcome Trust Sanger Institute. Genotype data from the HapMap, 1000 Genomes, and MalariaGEN (see Chapter 3 for individual contributions) was used where indicated. All data analysis, including writing of custom scripts, was conducted by S.S.

4.7 Supplemental Information

Table S4.1: Summary statistics from alignment. Seven pools were sequenced in duplicate. Shown here are: the proportion of sequencing reads generated that successfully mapped to the reference sequence; the median read depth, MapQ and BaseQ across all sites; uniformity (percent of sites within two-fold range of median RD) and reproducibility (Pearson correlation coefficient across duplicate runs).

	Pool size	Reads mapped (%)	Read depth (median)	BaseQ (median)	MapQ (median)	Uniformity	Reproducibility
962_s_3	2	75.5	5119	58.8	97.6	0.91	0.98
962_s_5	2	76.0	5162	58.6	97.0	0.91	0.98
962_s_6	4	75.5	5256	59.3	100.4	0.91	0.99
962_s_7	4	74.5	5035	59.2	99.7	0.91	0.99
1012_s_1	8	76.5	5550	59.7	101.9	0.92	0.97
1012_s_2	8	77.0	5695	59.8	102.9	0.92	0.97
1012_s_3	16	74.5	5094	59.5	101.0	0.91	0.99
1012_s_5	16	75.0	5047	59.5	100.9	0.91	0.99
1012_s_6	32	74.0	5284	60.1	104.5	0.91	0.94
1012_s_7	32	74.0	5265	60.1	104.6	0.90	0.94
1012_s_8	40	66.0	4493	59.6	101.4	0.91	0.93
1264_s_8	40	65.0	3748	61.1	113.1	0.90	0.93

Table S4.2: Algorithm settings and performance at different pool sizes. Specific algorithm criteria used in each sequencing run are shown here (see Table 4.1 for generalized algorithm settings) along with the total number of SNPs detected, and the sensitivity and specificity of each run.

	Pool size	AAF (min)	ALT_bq (min)	cy_diff (max)	noise_stat (max)	SNPs called	Sensitivity	Specificity
962_s_3	2	0.250	53.4	5.5	3.5	15	1.00	0.93
962_s_5	2	0.250	53.4	5.4	3.6	15	1.00	0.93
962_s_6	4	0.125	54.7	4.9	3.1	24	1.00	0.96
962_s_7	4	0.125	53.9	5.4	3.3	24	1.00	0.96
1012_s_1	8	0.062	55.6	3.9	3.1	31	0.94	0.94
1012_s_2	8	0.062	55.8	4.1	3.0	31	0.94	0.94
1012_s_3	16	0.031	54.9	4.0	2.9	52	1.00	0.83
1012_s_5	16	0.031	55.1	3.8	2.9	52	1.00	0.83
1012_s_6	32	0.016	56.6	3.4	2.7	64	0.80	0.77
1012_s_7	32	0.016	56.6	3.4	2.7	65	0.83	0.75
1012_s_8	40	0.012	55.6	3.9	2.7	66	0.92	0.74
1264_s_8	40	0.012	59.0	3.6	1.7	51	0.84	0.80

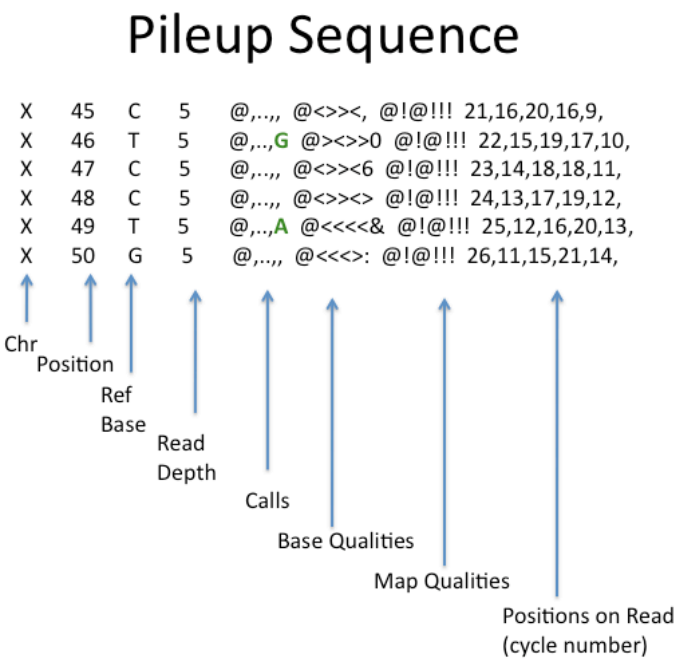


Figure S4.1: Pileup sequence schematic. The pileup sequence file generated by MAQ is a condensed summary of sequence read calls, ASCII-encoded quality scores, and cycle numbers for each site along the reference sequence.

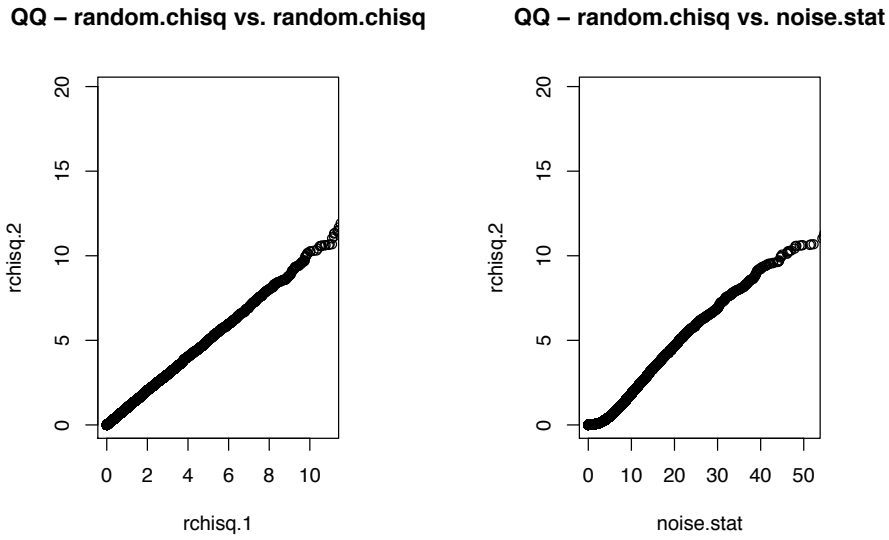


Figure S4.2: Quantile-quantile plot for noise statistic. The noise statistic behaves as a chi-square distributed variable, as is clear from a QQ plot against a random chi-square variate (right). A QQ plot of two random chi-square variates is shown for comparison (left).

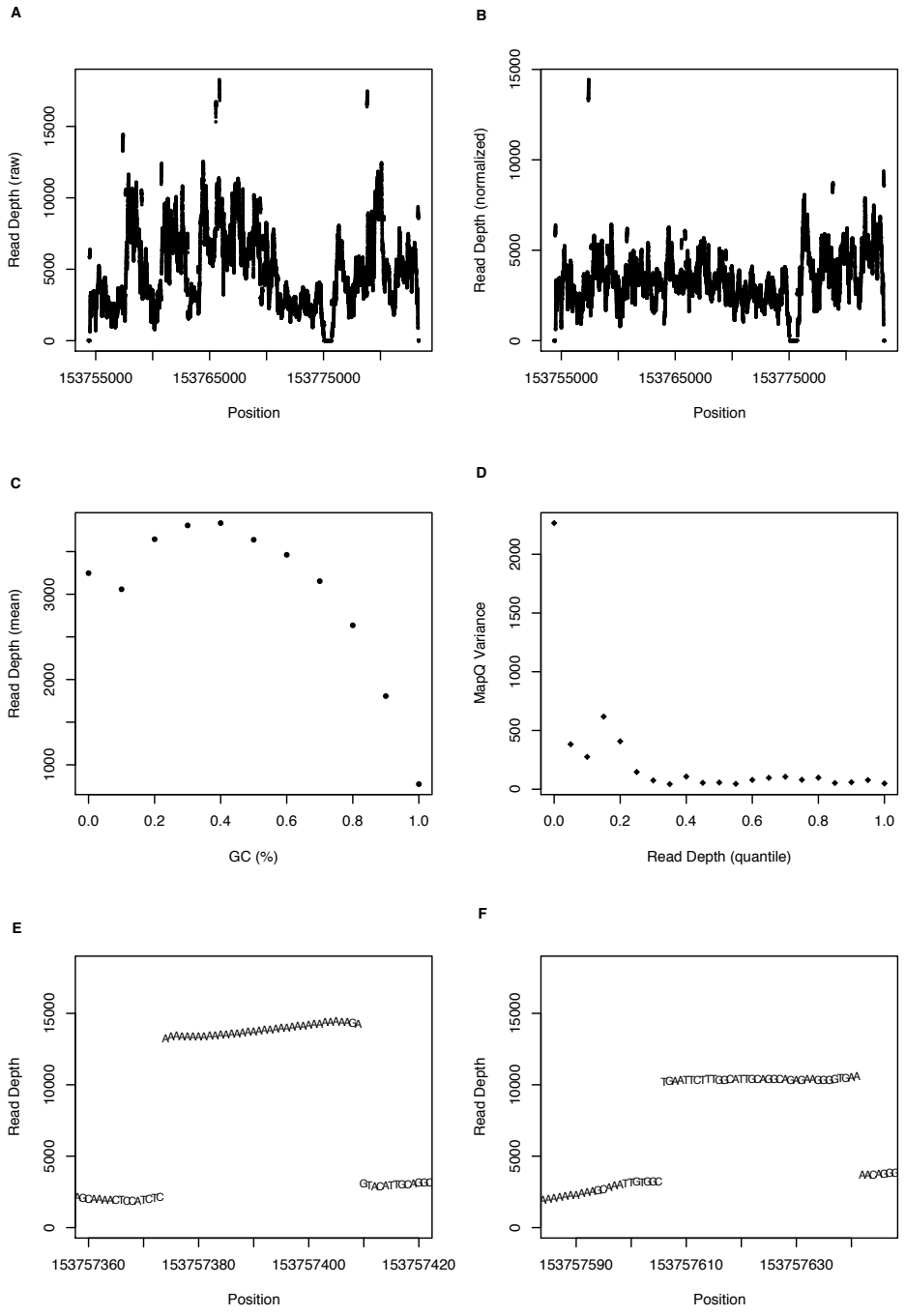


Figure S4.3: Systematic biases in sequencing read depth. Several systematic biases in sequencing read depth were observed. As expected, variance in raw read depth (a) was significantly reduced after normalizing by the number of amplicons tiling a site (b). GC content also demonstrates a read depth bias disfavoring extreme values of GC% (c). Additionally, variance in mapping qualities was markedly increased at lower read depths (d). Discontinuities in the read depth profile could often be traced to highly repetitive sequence elements (e) or primer sequence (f).

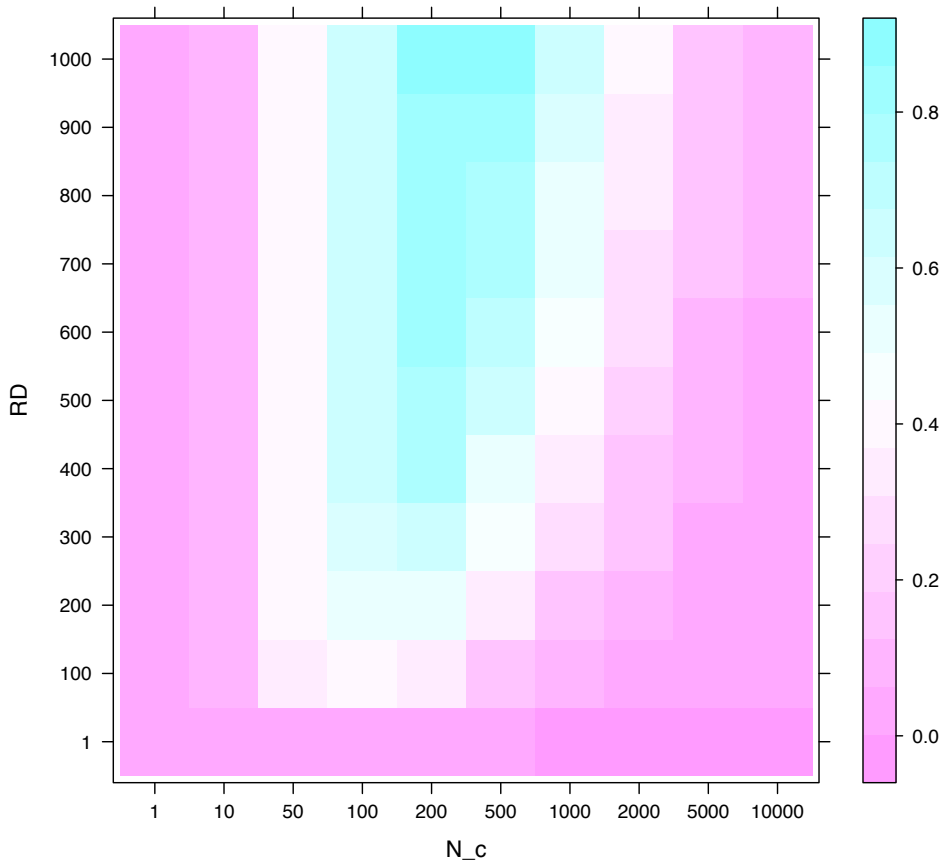


Figure S4.4: Power to detect a variant by pooled MPS as a function of N_c and RD. Heatmap showing the relationship between read depth (RD), the number of chromosomes sampled and pooled (N_c), and the sensitivity to detect a variant present at $AAF = 0.01$

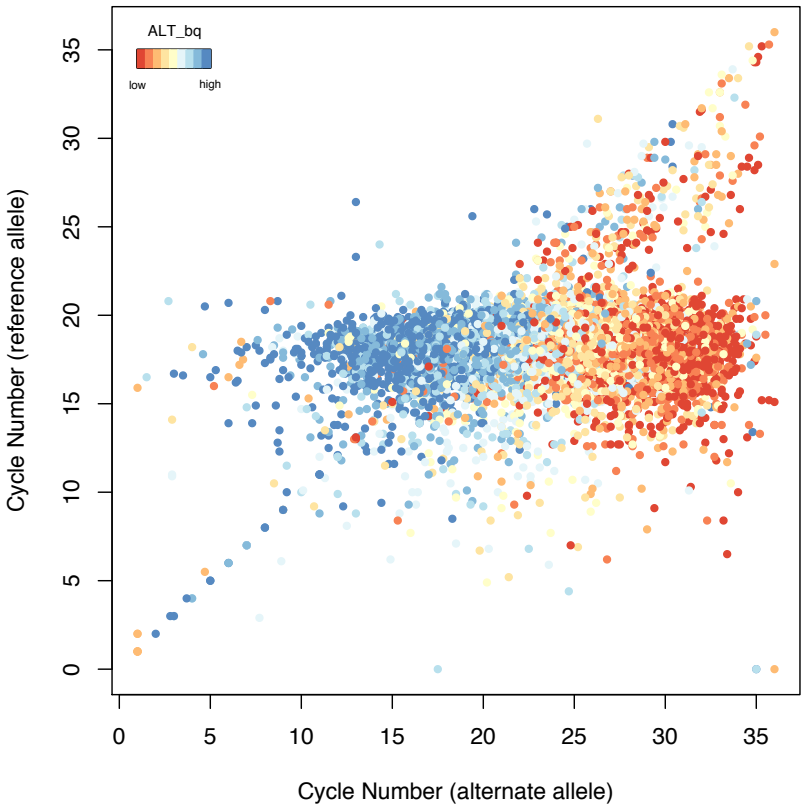


Figure S4.5: Relationship between reference and alternate cycle number and base quality. Cycle number distribution is narrow for the reference allele, while that for the alternate allele is broader and skewed toward higher values, where base qualities tend to be lower.

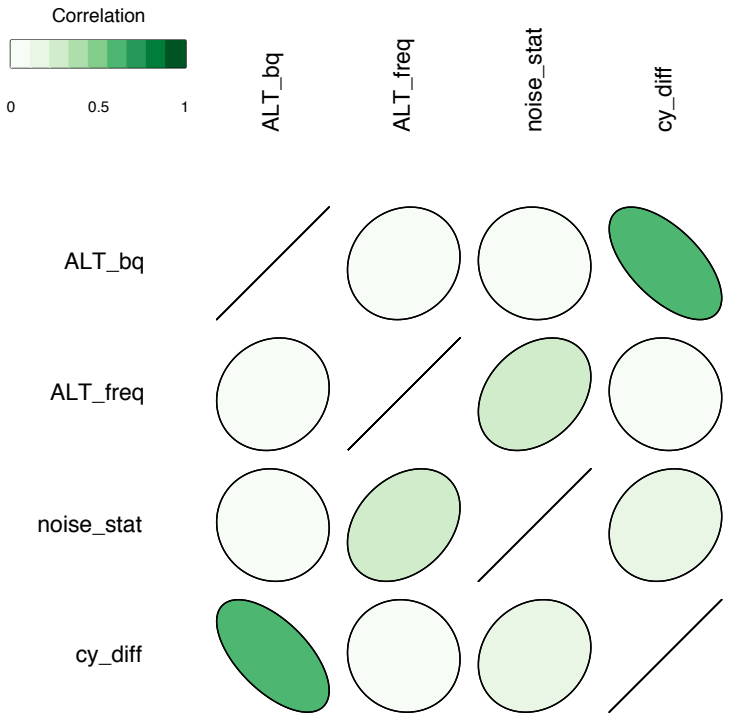


Figure S4.6: Ellipse plot of correlation matrix for algorithm metrics. Pairwise Spearman rank correlation coefficients between the four principle metrics in the algorithm are summarized graphically by both color and elliptical curvature. Positive correlations are centered about the line $y = x$, while negative ones are centered about $y = -x$. As the coefficients range from -0.61 to 0.33, one can reasonably assume that each metric is behaving relatively independently and making a nonredundant contribution to the SNP detection algorithm. ALT_freq = alternate allele frequency, ALT_bq = alternate allele base quality, noise_stat = noise statistic, cy_diff = cycle number difference.

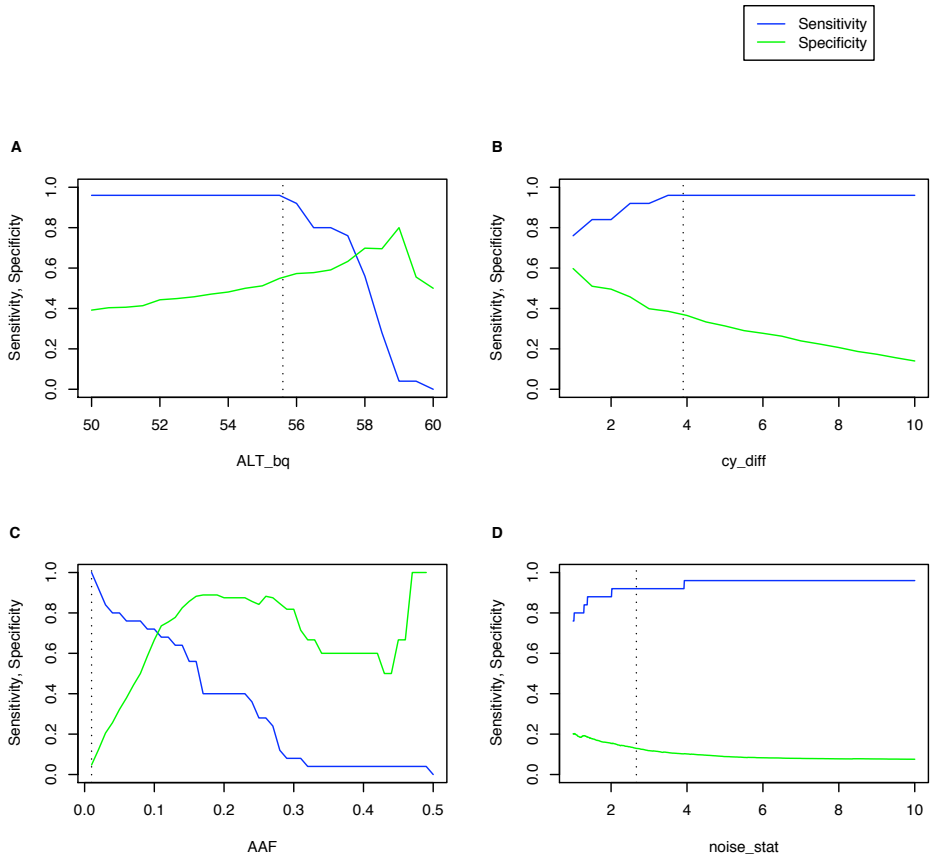


Figure S4.7: Algorithm performance simulations. The SNP detection algorithm was run on an N=40 dataset using only the filter indicated in each plot, with sensitivity (blue) and specificity (green) assessed at stepwise gradations of the given metric. This provided insight into the independent behavior of algorithm components with respect to performance metrics and helped inform decisions around setting default values (dotted line) for algorithm filters. AAF = alternate allele frequency, ALT_bq = alternate allele base quality, noise_stat = noise statistic, cy_diff = cycle number difference.

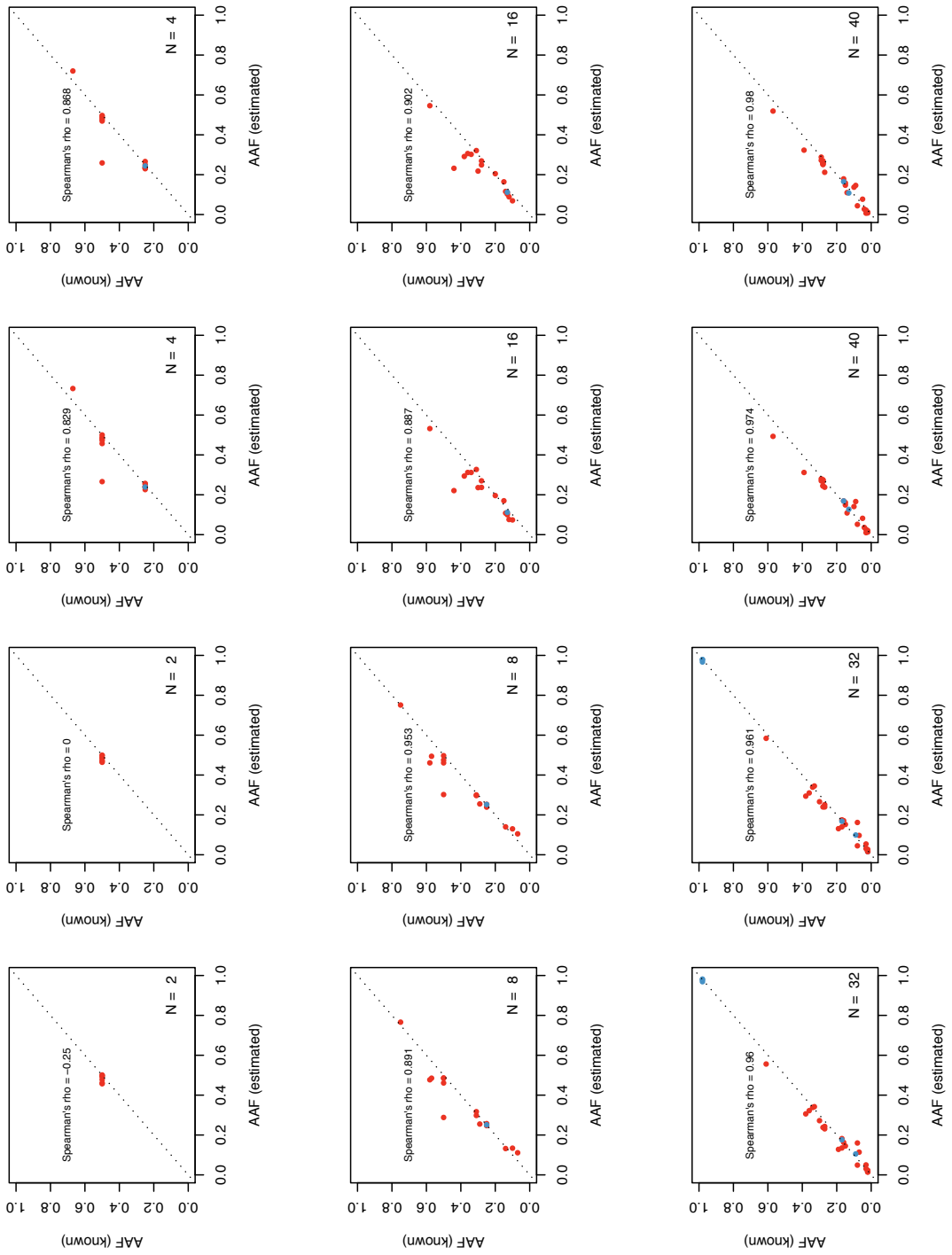


Figure S4.8: Estimated versus known alternate allele frequency for all pools. The strength of correlation between AAF values from pooled MPS and those from other methodologies remains consistent across duplicate experiments and at different pool sizes. In each plot, the $y = x$ diagonal (dotted line) is shown for reference. Technical duplicates are shown in adjacent plots.

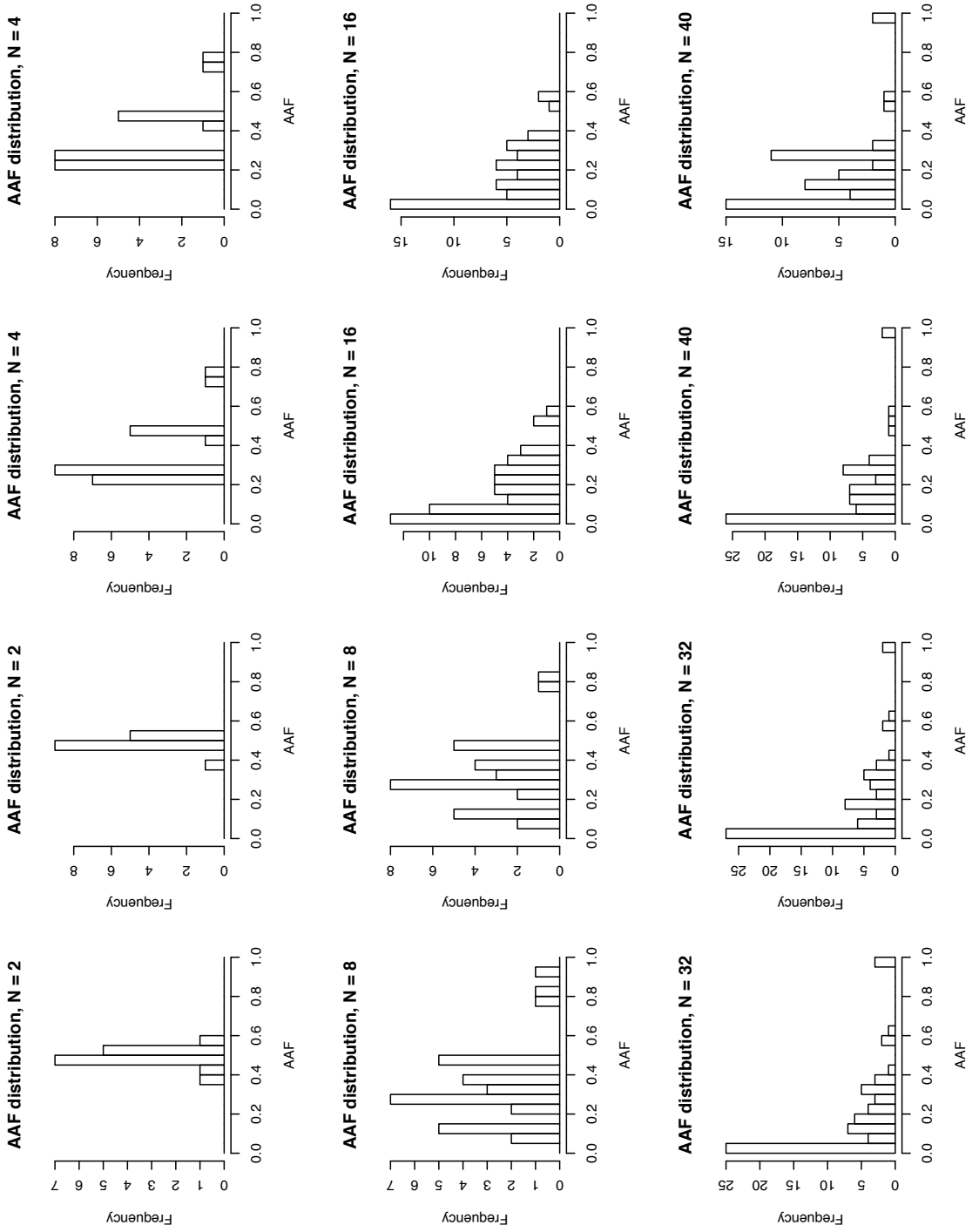


Figure S4.9: Alternate allele frequency distribution for filtered SNP list across different pool sizes. Alternate allele frequency distribution for the filtered SNP list at the highest pool sizes mirrors those in whole-genome surveys. Technical duplicates are shown in adjacent plots.

Chapter 5

Genetic determinants of glucose-6-phosphate dehydrogenase activity in Kenya

In the previous two chapters, I focused on strategies for characterizing population genetic variation at the *G6PD* locus. In the following two chapters, I will examine the relationship between that genetic variation and biochemical function. As genetic variance is more predictive of proximal molecular phenotypes than of distal clinical phenotypes, studying the link between genetics and molecular phenotypes is a necessary adjunct to understanding the results of clinical association studies. I present here a fine-mapping genetic association study of G6PD enzyme activity in Kenya that revealed major and minor genetic modifiers of enzyme activity, each of which was found to play a role in defining G6PD deficiency state in an individual.

5.1 Abstract

Glucose-6-phosphate dehydrogenase (G6PD) deficiency, the most common enzyme disorder in man, is associated with increased risk for neonatal jaundice and oxidant-induced hemolysis, but is also thought to be protective against malaria. There is substantial unexplained variance in genetic association with these important clinical phenotypes, however, as clinical association studies of severe malaria have yielded conflicting results, and determinants of susceptibility to hemolysis by drugs such as primaquine remain poorly understood. One possibility is that our knowledge of the genetic determinants of molecular phenotypes, such as G6PD enzyme activity, that underlie these clinical phenomena is incomplete, and that association studies which assay only one genetic variant will underestimate the magnitude or statistical significance of a true effect in the presence of unrecognized functional allelic diversity.

We systematically examined this possibility in Kenya by first surveying genetic variation at G6PD via resequencing, and then conducting a fine-mapping association study of erythrocyte G6PD enzyme activity in 1828 children across 30 polymorphisms at or around the locus. We confirmed a strong functional role for the 202T allele (A-/rs1050828; $P = 1.5 \times 10^{-200}$), and also identified other common variants that

exert smaller effects independent of 202T. The effects of these common variants were correlated with one another, and haplotype analyses suggested that each variant tags one of two haplotype motifs that are opposite in sequence identity and effect on G6PD activity. While smaller in magnitude than that for 202T, these effects are also of biological and possible clinical significance – for example, the 376C allele (A+/rs1050829) significantly augmented 202CT heterozygote risk for deficiency trait (OR = 2.11; 95% CI 1.17 - 3.83; P = 0.014).

These results suggest that multiple alleles at the G6PD locus play a role in determining G6PD deficiency state and therefore merit consideration in genetic association studies of clinical phenotypes including severe malaria and drug sensitivity.

5.2 Introduction

G6PD deficiency is a trait that is thought to be both harmful and beneficial. While clinical penetrance is generally mild, G6PD deficiency can predispose to hemolytic anemia and neonatal jaundice, generally in the context of oxidant stressors generated by certain foods, medications, and infections. Counterbalancing these deleterious effects is the notion that G6PD deficiency confers strong protection against malaria, and that such protection has conferred a selective advantage to G6PD-deficient individuals in malaria-endemic regions, as reflected in the close geographical correlation between the global distribution of G6PD deficiency and that of historical malaria endemicity.

Despite many decades of research, however, our understanding of both positive and negative clinical phenotypes associated with G6PD deficiency remains incomplete. Association studies of severe malaria have disagreed over the strength and specificity of protection, especially with respect to gender, with past studies demonstrating protection for only males [19], only females [136], both [39, 143], or neither [174]. Additionally, there is geographical heterogeneity in association testing results as well, with the only compelling evidence thus far coming from west Africa. Similarly, the precise determinants of the hemolytic response to oxidative stressors remain unknown. One especially important drug reaction is that involving the antimalarial primaquine [91, 92]. It is currently the only medication capable of radical cure of *Plasmodium vivax* infection and one of few that acts against the *Plasmodium falciparum* gametocyte; it therefore has high value in eliminating reservoirs of infection and reducing transmission rates, tasks that are crucial to the success of any malaria elimination program. While G6PD deficiency alleles are strongly associated with increased risk to primaquine-induced hemolysis, there exists significant unexplained variance in this association, especially for mosaic female heterozygotes [241].

One challenging aspect to designing and interpreting clinical association studies lies in how G6PD deficiency is defined. In the past, investigators used one of a variety of enzyme activity assays (discussed in Chapter 6) to qualitatively define the level of enzyme deficiency in an individual. Later, this was supplemented with protein electrophoresis, which provided a means to define G6PD genotype for an individual, lending categorical clarity to the more qualitative activity assay data. Coupled with kinetic and thermodynamic studies, activity assays and electrophoresis facilitated the identification of several hundred seemingly distinct protein variants throughout the world that were associated with G6PD enzyme deficiency. With the development of DNA sequencing, however, it has been found that the extent of protein variation is not mirrored in the genetic data, where just over 100 sequence polymorphisms have been identified, mostly nonsynonymous variants within the gene itself. Two explanations could account for this discrepancy. The first, and perhaps most plausible, is that inter-laboratory variation in experimental technique led to an overinflated estimate of G6PD enzyme diversity. Another explanation, however, is that the observed protein diversity is real, and that genetic diversity at the level of *G6PD* coding sequence only partially explains variation in enzyme activity.

Indeed, while many genetic association studies at *G6PD* implicitly assume that one nonsynonymous variant has both full positive and negative predictive power for G6PD deficiency trait, this approach belies the considerable allelic heterogeneity present at the locus [143, 103, 207]. G6PD deficiency is perhaps best understood as a family of heritable traits exhibiting widely varying prevalence, as well as differential biochemical and clinical penetrance, all factors which influence the power of an association test [143]. For example, demonstrating the association of a milder variant with drug-induced hemolysis might require a much larger sample size than would be required for a more severe variant. Additionally, as mentioned earlier, it is possible that unrecognized genetic modifiers of G6PD activity have yet to be identified, and that these may play a role in determining an individual's G6PD deficiency status.

To inform understanding of how genetic variation at *G6PD* impacts clinical outcomes, its more immediate relationship with molecular phenotypes is considered here. I draw upon *G6PD* resequencing data from Chapter 3 in order to survey the landscape of *G6PD* genetic variation in Kenya, and then test thirty polymorphisms for association with G6PD enzyme activity.

5.3 Materials and Methods

Sample collection. Capillary blood was drawn by heel prick from 2100 infants (less than 6 months old) recruited into an ongoing prospective birth cohort study run by Kenya Medical Research Institute (KEMRI) and the Wellcome Trust in Kilifi, Kenya. Samples were stored at 4°C and processed for activity assays and DNA extraction within four days. Ethics approval was granted by the KEMRI review board, and written informed consent obtained from parents prior to enrollment (refer to section 2.1.5 for protocol details).

Enzyme activity assay. The assay used was a modified version of the 'macro' assay given in Randox Laboratories G6PD kit (PD 410), using positive (PD 2618) and negative (PD 2617) controls where necessary. Ten microliters of whole blood was washed three times in 100 μl normal saline. Washed RBC were lysed by incubation with 25 μl of digitonin for 15 minutes at 4°C. Five microliters of the hemolysate was incubated with 300 μl of buffer (31.7 mM triethanolamine, 3.2 mM EDTA, pH=7.6) and 10 μl NADP+ (0.34 mM) at 37°C for 5 minutes. Finally, five microliters of glucose-6-phosphate (0.58 mM) was added and spectrophotometry readings taken, with duplicate measurements taken for each sample. Catalytic rate was assessed by the formation of reaction product NADPH, which absorbs light at 340 nm. The change in absorbance at 340 nm was measured over a 3 minute time course using a plate spectrophotometer (Tecan). Activity was normalized to hemolysate hemoglobin content by recording absorbance at 540 nm:

$$\text{G6PD activity} = \frac{\frac{\Delta OD340}{\Delta t}}{OD540} \tag{5.1}$$

Sequencing and genotyping. Resequencing was performed as part of MalariaGEN Consortial Project 3 (see Chapter 2.1.5 for an overview of MalariaGEN consortial projects) and an exon sequencing project called ExoSeq at Wellcome Trust Sanger Institute (WTSI). A total of 288 individuals comprised the sample set, with 48 individuals each from Burkina Faso, Cameroon, Nigeria, Kenya, Vietnam, and Papua New Guinea. A total of 28 amplicons approximately 600-800 bp in length were designed, covering about two-thirds of the 15 kb *G6PD* gene, with full coverage of all exons, but a few gaps in noncoding sequence, specifically in introns 1 and 2 (see Chapter 3 for further details).

Genotyping was performed on the Sequenom MASSarray iPLEX platform. A total of three multiplex genotyping reactions were designed, assaying a total of 68 SNPs. The SNPs input to the design software included those discovered during resequencing (N=72), as well as nonsynonymous variants from the literature (N=25), and several hundred variants within 1 Mb of *G6PD* extracted from dbSNP and Ensembl databases. Inclusion in the multiplex reactions was priority-weighted according predicted functional consequence and estimated derived allele frequency (DAF) criteria. Template DNA used for genotyping was genomic DNA amplified by primer extension pre-amplification. Genotype calls were manually curated by examining cluster plots to identify ambiguous genotypes.

Data curation and analysis. Enzyme activity data was filtered to keep only individuals with positive absorbance readings at both 340 and 540 nm, and checked for batch effects and extreme outliers. Genotype data was filtered using custom Perl and R scripts, along with PLINK [182] and snpMatrix [183]. Individuals exhibiting greater than 10% missing data or ambiguous gender (based on chromosome X genotypes) were excluded. SNPs were removed if they showed greater than 5% missing data, DAF $<$ 1%, or extreme deviation from Hardy-Weinberg equilibrium ($P < 10^{-4}$, Figure S5.1). In total, 1828 of 2100 individuals and 30 out of 68 SNPs met these inclusion criteria (Table S5.1).

Statistical analyses were conducted in R [181]. Single marker tests of association for each derived allele were conducted via multiple linear regression. Covariates included sex, and in the controlled tests, G6PD202 genotype. Three different modes of inheritance are considered– in the additive model, male genotypes were coded 0 or 2, and females were 0, 1, or 2; in the dominant and recessive models, males were coded 0/1, and females were coded 0/1/1, or 0/0/1, respectively. Haplotypes were inferred using fastPHASE under default settings with known male haplotypes used as a reference input to reduce search space and minimize phasing error rate in females (Figure S5.2). All haplotypes with frequency $>$ 1% were tested for association; individuals were coded 0 or 1 based on absence/presence of the test haplotype, with females represented by one haplotype selected at random. Odds ratios (OR) for qualitative tests were calculated via logistic regression. Effect sizes (ES) are presented here with 95% confidence interval in brackets. Allele sharing distance was calculated using a Euclidean distance metric for each individual from both H+ and H- haplotypes. An individual was categorized as either H+ or H- if their distance metric falls into the lowest quintile of the respective empirical distribution.

5.4 Results

5.4.1 Variants studied

The variants studied here were ascertained using a combination of sources, including MalariaGEN resequencing efforts, published accounts in the literature, and online variation databases. Resequencing was carried out as part of an ongoing laboratory effort to catalogue global genetic variation at the *G6PD* locus (as discussed in Chapter 3). A total of 15 kb at the *G6PD* locus was sequenced in a worldwide panel comprised of 288 African and Asian individuals, including 48 from Kenya. This work identified a total of 72 SNPs, including 49 novel variants. Published accounts in the literature yielded an additional 25 nonsporadic, nonsynonymous variants associated with G6PD deficiency in Africa and Asia, including 4 that have been identified in sub-Saharan Africa [103]: the 202/A-, 542, 680, and 968 polymorphisms (number refers to cDNA coordinate). Finally, several hundred polymorphisms (mostly noncoding SNPs) within 1 Mb of *G6PD* were also identified in dbSNP and Ensembl to supplement the list.

This list of several hundred SNPs was then priority-weighted according to predicted functional consequence (e.g. nonsynonymous, splice site, etc.), estimated allele frequency, and proximity to *G6PD* in order to decide which would be prioritized for inclusion in genotyping assays. A total of 68 were typed in the current study using the Sequenom MASSarray iPLEX platform. Of the SNPs typed, 30 met final study inclusion criteria, which included having derived allele frequency (DAF) greater than 0.01 and less than 5% missing data. It should be noted that most excluded SNPs failed based on DAF criteria, as many were private to west Africa or Asia (this genotyping assay was shared between several projects in the laboratory). This included the 542, 680, and 968 SNPs, which were monomorphic (DAF = 0%) in Kenya. All SNPs included in this association study are shown in Table S5.1.

5.4.2 The 202 polymorphism is a major determinant of G6PD activity

An initial association scan was performed across thirty SNPs in 1828 individuals with only gender included as a covariate. These tests yielded several significant associations (Figure 5.1, Table S5.2), with the maximum effect size (ES) and lowest P value recorded at 202/A- (ES = -0.22 [-0.23 - -0.21], $P = 1.5 \times 10^{-200}$), a nonsynonymous SNP (rs1050828; C>T; Val68Met) known to be associated with G6PD enzyme deficiency. The ES for the 202T allele corresponds to a roughly 40% reduction in enzyme activity for heterozygotes, and 80% for hemizygotes and homozygotes, consistent with previous estimates [100, 242, 243]. Since linkage disequilibrium with 202T was highly correlated with signal strength at all other sites, the association testing was repeated, this time including 202/A- genotype as a covariate. In these controlled tests, there was strong evidence of a causal role for 202T, as all other association signals dropped off by many orders of magnitude (Figure 5.2, Table S5.3), though twelve SNPs still retained significance at $P < 10^{-3}$, suggesting a role independent of 202/A-.

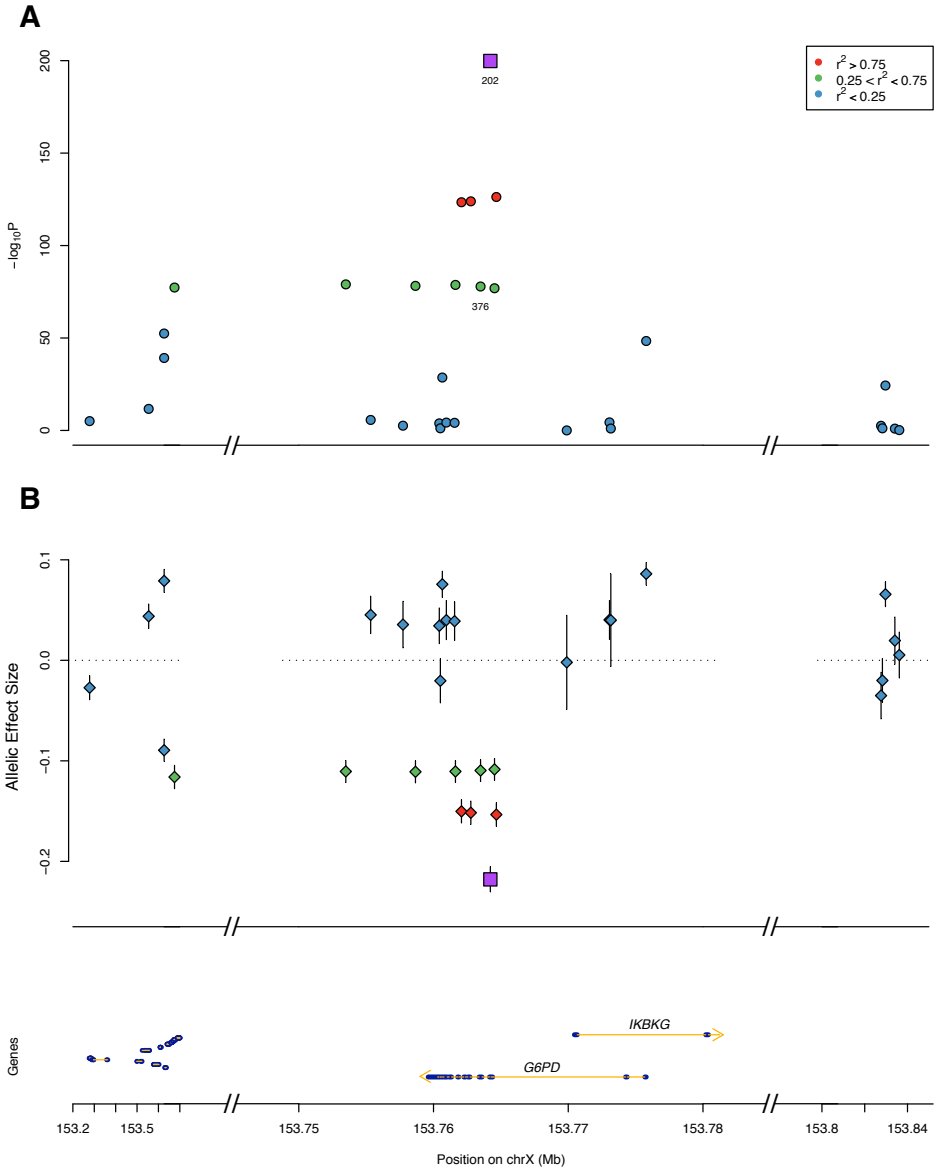


Figure 5.1: Fine mapping association study of G6PD activity. (a) P-values and (b) estimated allelic effect sizes from unbiased association tests (covariate: gender), with different colors (see key in panel A) indicating strength of linkage disequilibrium with G6PD202 (purple square). Gene graphic shows exons (blue) and introns (yellow) for genes annotated at UCSC Genome Browser (hg19); *IKBKG*=inhibitor of nuclear factor kappa-B kinase subunit gamma.

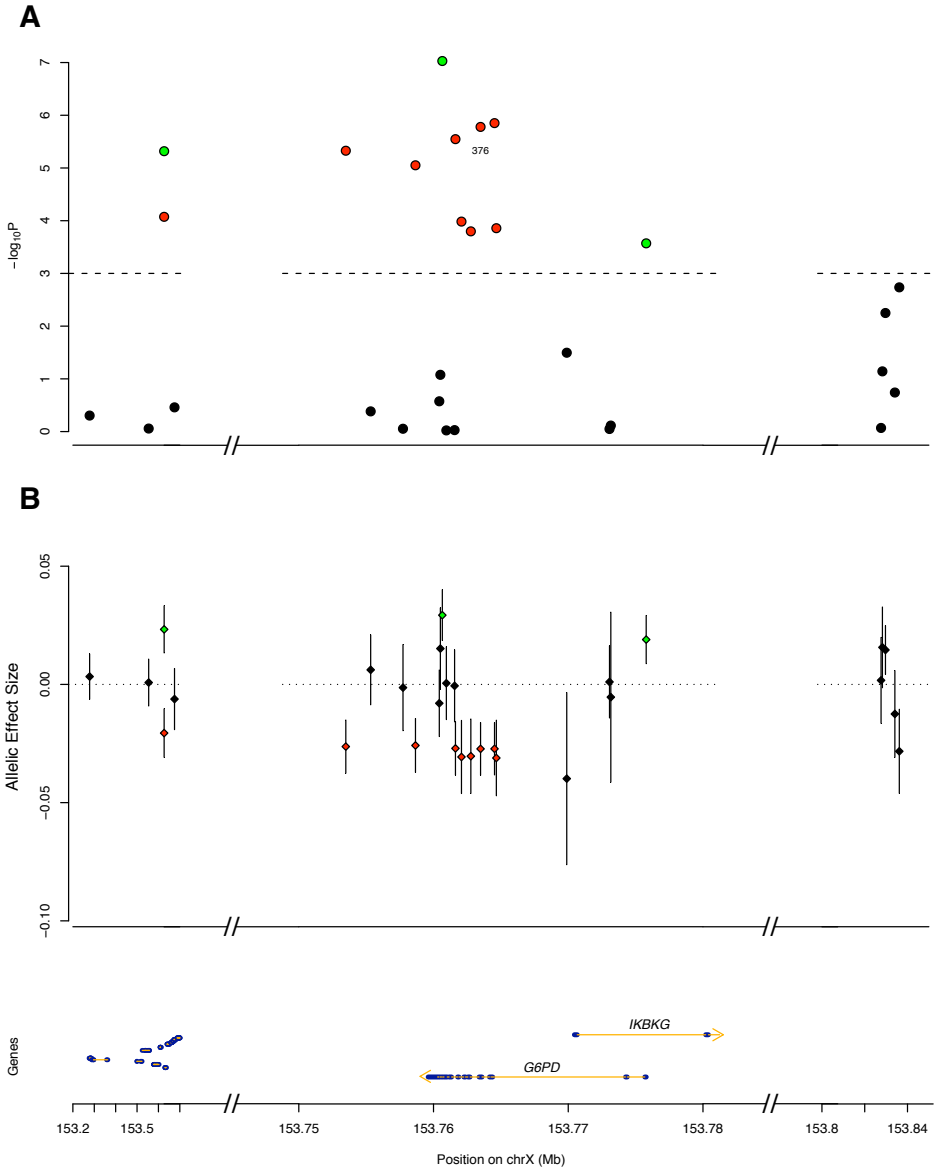


Figure 5.2: Association testing for effects independent of G6PD202. (A) -Log10 P-values and (B) estimated allelic effect sizes from 'controlled' association tests (covariates: gender, G6PD202 genotype). The twelve polymorphisms exhibiting significant effects at $P < 0.001$ are colored according to whether they exhibited positive ES (green) or negative ES (red). Gene graphic shows exons (blue) and introns (yellow) for genes annotated at UCSC Genome Browser (hg19); *IKBKG*=inhibitor of nuclear factor kappa-B kinase subunit gamma.

5.4.3 Other variants exhibit smaller effects independent of 202

As these twelve SNPs exhibited similar P values and ES magnitudes (Figure 5.2, Table S5.3), and were all common variants (DAF > 0.25) at a locus exhibiting a strong pattern of linkage disequilibrium (Figure S5.3), I considered the parsimonious hypothesis that they were all related signals. Upon examination of all haplotypes across these twelve sites, it was found that the two most frequent haplotypes (H+ = 19%; H- = 21%) were: (a) exactly opposite in allele identity (Figure 5.3C), with H+ possessing exactly three derived alleles at sites with positive ES and H- possessing exactly nine derived alleles at the

sites with negative ES (Figure 5.2B), and (b) significantly different in mean enzyme activity in 202CC individuals (H+: 0.57 vs. H-: 0.47; $P = 9.1 \times 10^{-5}$, t-test, Figure 5.3C). In association testing of all full-length and twelve-site haplotypes, it was observed that all haplotypes significant at $P < 0.025$ similarly exhibited differential ES direction and resembled either H+ or H- in sequence identity (Figure 5.3A and B). Additionally, extensive allele sharing with either H+ or H- haplotype explained significant variance independent of G6PD202 ($P = 4.3 \times 10^{-6}$, two-way ANOVA, Figure S5.4). Lastly, it was noted that SNPs tagging the H+ (rs2230037) and H- (G6PD376) haplotypes, in concert with G6PD202, were sufficient to describe a minimal adequate genetic model of G6PD activity that could explain 41% of variance in G6PD activity (Table 5.1).

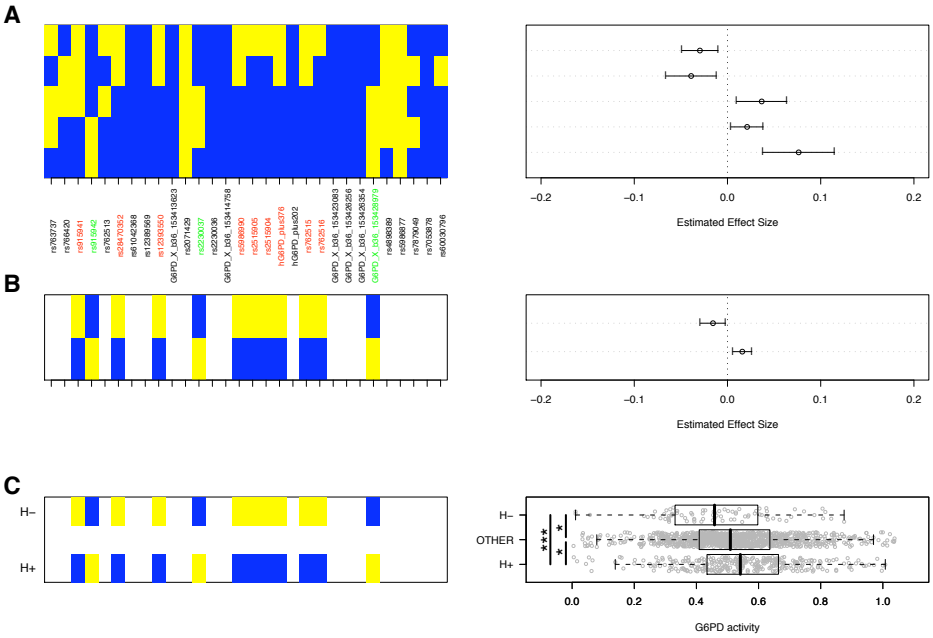


Figure 5.3: Haplotype-based association testing for effects independent of G6PD202. Results of association testing of (A) full-length and (B) partial/twelve-site haplotypes are shown for all haplotype tests suggestive of association ($P < 0.025$, covariates: gender, G6PD202 genotype), with haplotypes defined by patterns of ancestral (blue) and derived (yellow) alleles. SNP labels for the twelve sites that comprise the partial haplotypes are colored according to whether they exhibited positive ES [green] or negative ES (red) in single SNP controlled tests (Figure 5.2). (C) G6PD activity distribution in wild-type 202CC homozygotes as stratified by carriage of H+, H-, or neither haplotype. Pairwise t-test results are shown via asterisk (*: $P < 0.05$ and ***: $P = 9.1 \times 10^{-5}$). In box-and-whisker plot, center of box represents median, edges of box represent upper and lower quartile bounds, and whiskers represent either minimum and maximum values or 1.5 times the interquartile range, whichever is less extreme.

Table 5.1: Minimal adequate genetic model of G6PD activity. A linear model containing both G6PD202 as well as SNPs tagging the H+ (rs2230037) and H- (G6PD376) haplotypes explained significantly more variance than a model containing G6PD202 alone ($P < 0.001$), and could not be improved via the inclusion of additional SNPs.

SNP	Allelic ES (95% CI)	% change
G6PD202	-0.198 (-0.213 - -0.183)	-37.9
rs2230037(H+)	+0.022 (+0.011 - +0.034)	4.3
G6PD376 (H-)	-0.017 (-0.030 - -0.005)	-3.3

5.4.4 Multiple variants contribute to risk of G6PD deficiency state

Qualitative G6PD activity status was also examined, using the intermodal trough of the bimodal activity distribution in males as an *ad hoc* cutoff to distinguish normal from deficient (Figure 5.4A, Figure S5.5). A total of 162 males and 97 females could be categorized as deficient, including 20% of 202CT heterozygotes, in whom the H- tag SNP, G6PD376, was associated with both reduced enzyme activity (ES = -0.06 [-0.10 - -0.03], $P = 0.0009$, Figure 5.4B) and increased risk for deficiency trait (OR = 2.11 [1.12 - 3.84], $P = 0.014$). Overall, using the clearer distribution profile in males, it was estimated that the prevalence of G6PD deficiency in this study population is 18%, and that 202T accounts for 85% of this figure, which suggests that there may be rare functional variants this study was underpowered to discover, consistent with observations elsewhere in east Africa [243].

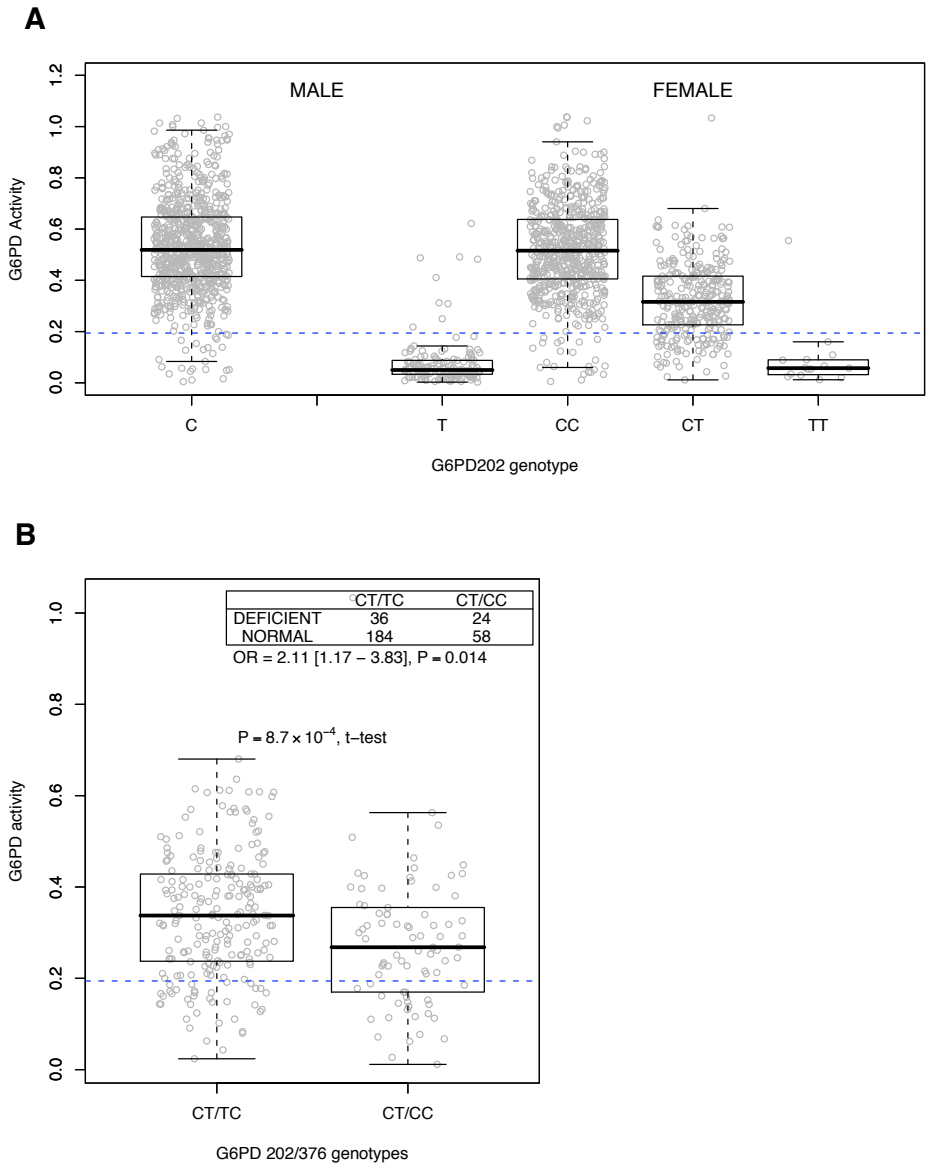


Figure 5.4: Genetic modifiers of G6PD activity. (A) G6PD activity distribution stratified by genotype at G6PD202. (B) G6PD activity in 202CT heterozygotes stratified by genotype at G6PD376 (NB: the 202T derived allele always exists on a 376C derived allele background), with contingency table (inset) for qualitative test of association at G6PD376 in 202CT heterozygotes. Also shown in both plots is the reference line (blue) dividing normal vs. deficient for qualitative tests of G6PD deficiency. In box-and-whisker plots, center of box represents median, edges of box represent upper and lower quartile bounds, and whiskers represent either minimum and maximum values or 1.5 times the interquartile range, whichever is less extreme.

5.5 Discussion

This study provides new insights into the genetic basis for quantitative and qualitative variation in G6PD enzyme activity. I found that the largest proportion of variance in G6PD activity could be explained by a single polymorphism, 202/A-, a well-studied coding variant thought to be a target of recent positive selection [121, 123], but one that has yielded conflicting results in previous association studies of severe malaria. Complicating this simple picture, however, is the finding that common variants at the locus modulate G6PD activity independently of 202/A-, and may have clinical significance in augmenting by two-fold the risk of G6PD deficiency state in 202CT heterozygote females. Taken together, the data suggest that multiple genetic variants at the *G6PD* locus impact enzyme activity, and that multi-site approaches might help clarify association studies of clinical phenotypes associated with G6PD deficiency, including protection from malaria and drug-induced hemolysis.

The 202/A- polymorphism is one of the most well-studied alleles in the host susceptibility genetics of malaria. While the finding of strong association at 202/A- is not surprising, the present study is the first to systematically consider the possibility that other SNPs, even those found several hundred kilobases away from *G6PD*, might be even more strongly associated with G6PD deficiency, and that 202/A- is only indirectly associated with deficiency, a finding which might explain discrepancies in previous association studies. Although the number of SNPs surveyed here is relatively small and does not include rare variants, I did not observe any SNPs more strongly associated with G6PD deficiency than the 202 variant, strongly suggestive of a causal role. Additionally, while all known functional variants in sub-Saharan Africa were genotyped as part of the study, only 202 was found to be polymorphic in Kenya (DAF = 17%). This allelic homogeneity contrasts with the situation in some areas of West Africa [143], and suggests that given a similar allele frequency and biochemical penetrance, clinical association studies of malaria in each region should be equivalently powered if indeed the G6PD deficiency trait is directly responsible (and not simply a marker) for protection. The strongest evidence implicating G6PD deficiency in malaria protection would ideally be consistent with an additive genetic model, with hemizygous males and homozygous females exhibiting similar clinical phenotypes, and heterozygous females displaying intermediate phenotypes.

This study also demonstrates the power of a haplotype-based approach in resolving multiple correlated association signals in a region. Though two haplotype motifs were found that were common and of significant effect independent of 202/A-, I was unable to further localize the signal, even when examining all possible two- and three-locus haplotypes (data not shown), suggesting that these haplotypes are merely markers for the true, untyped causal variants. Interestingly, among the twelve variants of significant effect independent of 202/A- were 376/A+, a well-known nonsynonymous variant, and rs915942 and G6PD_X_b36_153428979, predicted 5' UTR splice-site variants within *RPL10* (a large-subunit ribosomal

gene) and *G6PD*, respectively. While one can only speculate as to the precise nature of the structure-function relationship mediating these effects, the biology of the locus makes for interesting possibilities for interaction between 202/A- and the H+/H- variants, including the possibility of cell-cell interactions synergistically impacting enzyme activity distribution in mosaic females.

The current investigation was limited in focus to the *G6PD* gene itself. Given its fundamental role in a wide variety of biomolecular processes, it seems likely that G6PD activity would be influenced by a variety of other genes, both directly and indirectly. These might include genes which affect the expression or stability of allosteric regulators such as p53 [76] or dihydroepiandrosterone [75], or other genes in the same biosynthetic pathways, such as 6-phosphogluconate dehydrogenase or glutathione reductase. Additionally, if decreased G6PD protein stability in the erythrocyte is the molecular basis for enzyme deficiency as has been suggested [98, 96], then a host of factors associated with red cell senescence, including oxidant stress, ubiquitination pathways, or even splenic function may also play a role in influencing the gross G6PD enzyme activity generated by a heterogeneous population of RBC. Only whole-genome genotype or sequence data will be able to consider the full range of additional genetic and epistatic influences on G6PD enzyme activity and should be strongly considered for future work, as should studying G6PD activity distribution at the single-cell level, especially to understand variability in heterozygote mosaic females. I discuss the latter in further detail in Chapter 6.

At one level, the work presented here reveals how common, standing genetic variation superimposed on a Mendelian trait can significantly impact expression of that trait. More important, however, is how this finding translates into our understanding of how G6PD deficiency is linked to important clinical manifestations. Specifically, it will be worth considering whether multi-site genetic testing for G6PD deficiency is more predictive of clinical outcome than single SNP testing. For example, in considering a public health intervention such as mass primaquine administration as part of a malaria elimination campaign, it may be important to err on the side of sensitivity in diagnostic genetic testing to reduce morbidity associated with preventable episodes of acute hemolytic anemia. Finally, at a more general level, our work highlights the importance of measuring intermediate molecular phenotypes that can bridge the gulf between variance at the genetic level and variance in clinical phenotypes. This is especially important for association studies of G6PD deficiency given its considerable allelic heterogeneity, varying biochemical penetrance, and a long history of inconsistent findings.

5.6 Acknowledgments and Contributions

I thank all the children and their families who participated in this study. S.S. designed the study in collaboration with Tom Williams, who works at KEMRI-Wellcome Trust Programme in Kilifi, Kenya. S.S. designed and validated the enzyme activity and genotyping assays. Alex Macharia, Johnstone Makale, and Sophie Uyoga performed the enzyme activity assays at KEMRI-Wellcome Trust Programme in Kilifi, Kenya. S.S., Rachel Craik, and Christina Hubbart performed genotyping at the Wellcome Trust Centre for Human Genetics in Oxford, UK. Capillary sequencing data was generated as part of MalariaGEN CP3 (discussed in Chapter 3) at the Wellcome Trust Sanger Institute in Hinxton, UK. S.S. conducted all statistical analyses.

Supplemental Information

Table S5.1: SNP summary table. Shown here for each SNP assayed is: the genomic position (hg19/GRCh37), observed alleles (on the [+]-strand, according to UCSC Genome Browser), ancestral allele (according to the Ensembl EPO pipeline), G6PD exon-intron position (according to UCSC genome annotation coordinates), and derived allele frequency stratified by sex.

Name	Position (build 37)	Observed Alleles	Ancestral Allele	G6PD - Exon/Intron	DAF (male)	DAF (female)
rs763737	153278307	G/A	G	DS	0.611	0.607
rs766420	153554404	G/C	G	DS	0.361	0.321
rs915941	153626649	C/A	C	DS	0.481	0.462
rs915942	153626738	G/A	G	DS	0.421	0.429
rs762513	153675171	A/G	A	DS	0.310	0.311
rs28470352	153753490	T/A	T	DS	0.391	0.391
rs61042368	153755336	G/A	G	DS	0.111	0.110
rs12389569	153757734	G/A	G	DS	0.067	0.072
rs12393550	153758660	G/A	G	DS	0.390	0.387
G6PD_X_b36_153413623	153760429	G/A	G	E12	0.129	0.113
rs2071429	153760508	A/G	A	I11	0.925	0.921
rs2230037	153760654	G/A	G	E11	0.253	0.265
rs2230036	153760953	C/T	C	E10	0.104	0.098
G6PD_X_b36_153414758	153761564	G/A	G	I8	0.105	0.103
rs5986990	153761628	G/A	G	I8	0.391	0.392
rs2515905	153762075	G/A	G	I7	0.269	0.259
rs2515904	153762771	G/C	G	I5	0.265	0.256
hG6PD_plus376	153763492	T/C	T	E5	0.400	0.395
hG6PD_plus202	153764217	C/T	C	E4	0.162	0.183
rs762515	153764528	T/C	T	I2	0.403	0.400
rs762516	153764663	C/T	C	I2	0.262	0.253
G6PD_X_b36_153423083	153769889	G/A	G	I2	0.018	0.012
G6PD_X_b36_153426256	153773062	C/T	C	I2	0.105	0.101
G6PD_X_b36_153426354	153773160	A/C	A	I2	0.014	0.021
G6PD_X_b36_153428979	153775785	T/C	C	US	0.488	0.484
rs4898389	153827637	A/G	A	US	0.935	0.928
rs5986877	153828269	C/G	C	US	0.926	0.918
rs7879049	153829693	A/G	A	US	0.323	0.323
rs7053878	153834100	T/A	T	US	0.063	0.072
rs60030796	153836171	A/G	A	US	0.073	0.068

		DAF		Additive		Dominant		Recessive	
	Position	Male	Female	P	Effect Size	P	Effect Size	P	Effect Size
hG6PD_plus202	153764217	0.162	0.183	1.5e-200	-0.22 (-0.23 - -0.21)	8.1e-163	-0.30 (-0.32 - -0.28)	2.8e-147	-0.44 (-0.47 - -0.41)
rs762516	153764663	0.262	0.253	5.5e-127	-0.15 (-0.17 - -0.14)	9.7e-110	-0.23 (-0.25 - -0.21)	1.1e-96	-0.30 (-0.32 - -0.27)
rs2515904	153762771	0.265	0.256	1.2e-124	-0.15 (-0.16 - -0.14)	1.5e-107	-0.23 (-0.24 - -0.21)	2.0e-95	-0.29 (-0.32 - -0.27)
rs2515905	153762075	0.269	0.259	3.7e-124	-0.15 (-0.16 - -0.14)	1.4e-108	-0.23 (-0.24 - -0.21)	1.2e-93	-0.29 (-0.31 - -0.26)
rs28470352	153753490	0.391	0.391	1.0e-79	-0.11 (-0.12 - -0.10)	6.4e-69	-0.18 (-0.20 - -0.16)	2.9e-65	-0.20 (-0.22 - -0.18)
rs5986990	153761628	0.391	0.392	1.9e-79	-0.11 (-0.12 - -0.10)	1.2e-68	-0.18 (-0.20 - -0.16)	6.1e-65	-0.20 (-0.22 - -0.18)
rs12393550	153758660	0.390	0.387	6.4e-79	-0.11 (-0.12 - -0.10)	1.1e-69	-0.18 (-0.20 - -0.16)	1.2e-62	-0.20 (-0.22 - -0.17)
hG6PD_plus376	153763492	0.400	0.395	1.4e-78	-0.11 (-0.12 - -0.10)	4.0e-68	-0.18 (-0.20 - -0.16)	3.0e-64	-0.20 (-0.22 - -0.17)
rs762513	153675171	0.310	0.311	5.0e-78	-0.12 (-0.13 - -0.10)	9.8e-73	-0.18 (-0.20 - -0.17)	2.3e-56	-0.21 (-0.23 - -0.18)
rs762515	153764528	0.403	0.400	1.2e-77	-0.11 (-0.12 - -0.10)	1.3e-68	-0.18 (-0.20 - -0.16)	3.6e-62	-0.19 (-0.21 - -0.17)
rs915941	153626649	0.481	0.462	3.7e-53	-0.09 (-0.10 - -0.08)	3.4e-47	-0.15 (-0.17 - -0.13)	3.0e-43	-0.15 (-0.17 - -0.13)
G6PD_X_b36_153428979	153775785	0.488	0.484	4.2e-49	+0.09 (+0.07 - +0.10)	4.5e-41	+0.14 (+0.12 - +0.16)	1.2e-42	+0.15 (+0.13 - +0.17)
rs915942	153626738	0.421	0.429	6.4e-40	+0.08 (+0.07 - +0.09)	2.2e-33	+0.13 (+0.11 - +0.15)	8.1e-35	+0.14 (+0.12 - +0.17)
rs2230037	153760654	0.253	0.265	2.4e-29	+0.08 (+0.06 - +0.09)	3.4e-25	+0.11 (+0.09 - +0.13)	1.9e-24	+0.15 (+0.12 - +0.17)
rs7879049	153829693	0.323	0.323	5.0e-25	+0.07 (+0.05 - +0.08)	1.3e-20	+0.10 (+0.08 - +0.12)	3.1e-22	+0.13 (+0.10 - +0.15)
rs766420	153554404	0.361	0.321	2.2e-12	+0.04 (+0.03 - +0.06)	4.0e-11	+0.07 (+0.05 - +0.09)	2.5e-10	+0.08 (+0.06 - +0.11)
rs61042368	153755336	0.111	0.110	2.1e-06	+0.05 (+0.03 - +0.06)	8.6e-06	+0.06 (+0.04 - +0.09)	5.1e-05	+0.09 (+0.05 - +0.13)
rs763737	153278307	0.611	0.607	9.0e-06	-0.03 (-0.04 - -0.02)	5.5e-06	-0.05 (-0.08 - -0.03)	2.2e-04	-0.04 (-0.06 - -0.02)
G6PD_X_b36_153426256	153773062	0.105	0.101	4.3e-05	+0.04 (+0.02 - +0.06)	1.2e-04	+0.06 (+0.03 - +0.08)	5.3e-04	+0.08 (+0.03 - +0.13)
rs2230036	153760953	0.104	0.098	5.5e-05	+0.04 (+0.02 - +0.06)	1.8e-04	+0.05 (+0.03 - +0.08)	4.7e-04	+0.08 (+0.04 - +0.13)
G6PD_X_b36_153414758	153761564	0.105	0.103	7.3e-05	+0.04 (+0.02 - +0.06)	2.5e-04	+0.05 (+0.02 - +0.08)	5.3e-04	+0.08 (+0.03 - +0.13)
G6PD_X_b36_153413623	153760429	0.129	0.113	1.5e-04	+0.03 (+0.02 - +0.05)	1.4e-04	+0.05 (+0.03 - +0.08)	3.3e-03	+0.06 (+0.02 - +0.10)
rs12389569	153757734	0.067	0.072	2.6e-03	+0.04 (+0.01 - +0.06)	1.3e-02	+0.04 (+0.01 - +0.08)	1.9e-03	+0.09 (+0.03 - +0.14)
rs4898389	153827637	0.935	0.928	3.0e-03	-0.04 (-0.06 - -0.01)	7.1e-02	-0.05 (-0.11 - +0.00)	1.3e-03	-0.05 (-0.09 - -0.02)
rs2071429	153760508	0.925	0.921	6.8e-02	-0.02 (-0.04 - +0.00)	2.5e-01	-0.03 (-0.08 - +0.02)	5.0e-02	-0.03 (-0.06 - -0.00)
rs5986877	153828269	0.926	0.918	7.1e-02	-0.02 (-0.04 - +0.00)	2.6e-01	-0.03 (-0.08 - +0.02)	5.1e-02	-0.03 (-0.06 - +0.00)
G6PD_X_b36_153426354	153773160	0.014	0.021	8.9e-02	+0.04 (-0.01 - +0.09)	7.9e-02	+0.06 (-0.01 - +0.12)	3.2e-01	+0.06 (-0.06 - +0.18)
rs7053878	153834100	0.063	0.072	1.0e-01	+0.02 (-0.00 - +0.04)	1.7e-01	+0.02 (-0.01 - +0.06)	9.9e-02	+0.05 (-0.01 - +0.10)
rs60030796	153836171	0.073	0.068	6.5e-01	+0.01 (-0.02 - +0.03)	8.6e-01	-0.00 (-0.04 - +0.03)	1.8e-01	+0.04 (-0.02 - +0.09)
G6PD_X_b36_153423083	153769889	0.018	0.012	9.3e-01	-0.00 (-0.05 - +0.04)	8.7e-01	-0.01 (-0.08 - +0.06)	9.6e-01	+0.00 (-0.11 - +0.11)

Table S5.2: Initial association test results. Shown here for each SNP surveyed is: genomic position (hg19/GRCh37), DAF stratified by sex, P values and effect sizes (with 95% CI) for initial, uncontrolled association tests under three different genetic models.

Table S5.4: Association test results of full-length and partial haplotypes for effects independent of G6PD202. Shown here for each haplotype surveyed is: allele identity across all sites (0=ancestral, 1=derived), population frequency, P values and effect sizes (with 95% CI) for G6PD202-controlled association tests.

Haplotype	Frequency	P	Effect Size
000100000010000000000000101000	0.010	0.0001	+0.076 (+0.038 - +0.115)
__01_0__0__1__0000_00___1_____	0.186	0.0024	+0.016 (+0.006 - +0.026)
101011001010001111011000011000	0.043	0.0029	-0.030 (-0.049 - -0.010)
011001001010001001010000011101	0.021	0.0049	-0.039 (-0.066 - -0.012)
111010000011000000000000111000	0.021	0.0080	+0.037 (+0.010 - +0.064)
100100000011000000000000111100	0.055	0.0171	+0.021 (+0.004 - +0.038)
__10_1__1__0__1111_11___0_____	0.207	0.0249	-0.016 (-0.029 - -0.002)
__10_1__1__0__1001_10___0_____	0.100	0.0284	-0.015 (-0.028 - -0.002)
010100000011000000000000111100	0.011	0.0525	+0.036 (-0.000 - +0.073)
000100000011000000000000111100	0.010	0.0596	+0.037 (-0.001 - +0.076)
000100000011000000000000111000	0.015	0.0761	-0.029 (-0.061 - +0.003)
__10_0__0__1__0000_00___1_____	0.062	0.1173	+0.013 (-0.003 - +0.029)
100000000011000000000000111000	0.037	0.1411	-0.016 (-0.037 - +0.005)
000100100010110000000010011100	0.038	0.1480	+0.015 (-0.005 - +0.036)
110100000011000000000000111100	0.018	0.2054	+0.019 (-0.011 - +0.049)
__01_0__0__0__0000_00___0_____	0.092	0.2459	+0.008 (-0.006 - +0.022)
__00_0__0__1__0000_00___1_____	0.011	0.2500	+0.022 (-0.015 - +0.059)
100100000011000000000000111000	0.015	0.2505	-0.019 (-0.052 - +0.013)
010100100010110000000010011100	0.011	0.2525	+0.022 (-0.015 - +0.058)
__10_0__0__0__0000_00___1_____	0.080	0.2864	-0.008 (-0.022 - +0.007)
000101001010001111011000011000	0.023	0.2936	-0.014 (-0.040 - +0.012)
100100000110000000000000111000	0.022	0.2959	+0.014 (-0.013 - +0.041)
100100000000000000000000100010	0.014	0.3055	+0.018 (-0.016 - +0.051)
111000000000000000000000100010	0.014	0.3236	-0.017 (-0.050 - +0.016)
001011001010001111111000011000	0.018	0.3267	-0.016 (-0.047 - +0.016)
__00_1__1__0__1001_10___0_____	0.015	0.3695	-0.015 (-0.047 - +0.018)
__01_1__1__0__1111_11___0_____	0.036	0.3812	-0.009 (-0.030 - +0.012)
__01_0__0__0__0000_00___1_____	0.081	0.4047	+0.006 (-0.008 - +0.021)
000110100010110000000010011100	0.011	0.4369	-0.015 (-0.053 - +0.023)
111000000011000000000000111100	0.018	0.4512	-0.012 (-0.041 - +0.018)
010101001010001111011000011000	0.011	0.4780	-0.014 (-0.051 - +0.024)
__00_0__0__0__0000_00___1_____	0.064	0.4858	-0.006 (-0.022 - +0.010)
110100000011000000000000111000	0.036	0.5041	+0.007 (-0.014 - +0.028)
101011001010001111111000011000	0.134	0.5169	+0.005 (-0.011 - +0.022)
111001011010001001010000011000	0.016	0.5474	-0.010 (-0.041 - +0.022)
001000000000000000000000110010	0.016	0.6737	-0.007 (-0.038 - +0.025)
__10_0__0__0__0000_00___0_____	0.013	0.7425	+0.006 (-0.029 - +0.040)
000000000110000000000000111000	0.014	0.7539	+0.005 (-0.028 - +0.039)
100100100010110000000010011100	0.013	0.7693	+0.005 (-0.030 - +0.040)
__01_1__1__0__1001_10___0_____	0.015	0.8096	-0.004 (-0.037 - +0.029)
011001011010001001010000011000	0.047	0.8484	-0.002 (-0.020 - +0.017)
111000000000000000000000100000	0.018	0.9575	-0.001 (-0.030 - +0.029)

		DAF		Additive		Dominant		Recessive	
	Position	Male	Female	P	Effect Size	P	Effect Size	P	Effect Size
hG6PD_plus202	153764217	0.162	0.183	1.5e-200	-0.22 (-0.23 - -0.21)	8.1e-163	-0.30 (-0.32 - -0.28)	2.8e-147	-0.44 (-0.47 - -0.41)
rs2230037	153760654	0.253	0.265	9.3e-08	+0.03 (+0.02 - +0.04)	1.3e-07	+0.05 (+0.03 - +0.07)	6.2e-10	+0.08 (+0.05 - +0.10)
rs762515	153764528	0.403	0.400	1.4e-06	-0.03 (-0.04 - -0.02)	1.5e-07	-0.05 (-0.07 - -0.03)	2.0e-11	-0.08 (-0.10 - -0.05)
hG6PD_plus376	153763492	0.400	0.395	1.7e-06	-0.03 (-0.04 - -0.02)	4.4e-07	-0.05 (-0.07 - -0.03)	4.6e-12	-0.08 (-0.10 - -0.06)
rs5986990	153761628	0.391	0.392	2.8e-06	-0.03 (-0.04 - -0.02)	7.0e-07	-0.05 (-0.07 - -0.03)	5.9e-12	-0.08 (-0.10 - -0.06)
rs28470352	153753490	0.391	0.391	4.7e-06	-0.03 (-0.04 - -0.02)	1.0e-06	-0.05 (-0.07 - -0.03)	1.2e-11	-0.08 (-0.10 - -0.06)
rs915942	153626738	0.421	0.429	4.8e-06	+0.02 (+0.01 - +0.03)	2.0e-08	+0.05 (+0.03 - +0.07)	6.4e-12	+0.07 (+0.05 - +0.09)
rs12393550	153758660	0.390	0.387	8.9e-06	-0.03 (-0.04 - -0.01)	4.5e-07	-0.05 (-0.07 - -0.03)	2.0e-10	-0.07 (-0.10 - -0.05)
rs915941	153626649	0.481	0.462	8.4e-05	-0.02 (-0.03 - -0.01)	4.7e-06	-0.04 (-0.06 - -0.03)	8.3e-09	-0.06 (-0.08 - -0.04)
rs2515905	153762075	0.269	0.259	1.0e-04	-0.03 (-0.05 - -0.02)	1.6e-04	-0.05 (-0.08 - -0.02)	2.1e-06	-0.08 (-0.11 - -0.05)
rs762516	153764663	0.262	0.253	1.4e-04	-0.03 (-0.05 - -0.02)	3.2e-04	-0.05 (-0.08 - -0.02)	5.4e-06	-0.08 (-0.12 - -0.05)
rs2515904	153762771	0.265	0.256	1.6e-04	-0.03 (-0.05 - -0.01)	3.5e-04	-0.05 (-0.08 - -0.02)	3.2e-06	-0.08 (-0.12 - -0.05)
G6PD_X_b36_153428979	153775785	0.488	0.484	2.7e-04	+0.02 (+0.01 - +0.03)	9.1e-08	+0.05 (+0.03 - +0.07)	6.7e-13	+0.07 (+0.05 - +0.09)
rs60030796	153836171	0.073	0.068	1.8e-03	-0.03 (-0.05 - -0.01)	6.8e-03	-0.04 (-0.07 - -0.01)	1.0e-01	-0.04 (-0.08 - +0.01)
rs7879049	153829693	0.323	0.323	5.6e-03	+0.01 (+0.00 - +0.02)	7.2e-04	+0.03 (+0.01 - +0.05)	7.1e-06	+0.05 (+0.03 - +0.07)
G6PD_X_b36_153423083	153769889	0.018	0.012	3.2e-02	-0.04 (-0.08 - -0.00)	4.9e-02	-0.06 (-0.11 - -0.00)	1.3e-01	-0.07 (-0.16 - +0.02)
rs5986877	153828269	0.926	0.918	7.2e-02	+0.02 (-0.00 - +0.03)	1.7e-01	+0.03 (-0.01 - +0.07)	9.8e-01	+0.00 (-0.03 - +0.03)
rs2071429	153760508	0.925	0.921	8.4e-02	+0.02 (-0.00 - +0.03)	1.8e-01	+0.03 (-0.01 - +0.07)	9.7e-01	+0.00 (-0.03 - +0.03)
rs7053878	153834100	0.063	0.072	1.8e-01	-0.01 (-0.03 - +0.01)	5.2e-01	-0.01 (-0.04 - +0.02)	3.4e-01	-0.02 (-0.07 - +0.02)
G6PD_X_b36_153413623	153760429	0.129	0.113	2.7e-01	-0.01 (-0.02 - +0.01)	6.1e-01	-0.01 (-0.03 - +0.02)	3.9e-01	-0.02 (-0.05 - +0.02)
rs762513	153675171	0.310	0.311	3.5e-01	-0.01 (-0.02 - +0.01)	5.5e-02	-0.02 (-0.04 - +0.00)	6.9e-02	-0.03 (-0.05 - +0.00)
rs61042368	153755336	0.111	0.110	4.1e-01	+0.01 (-0.01 - +0.02)	2.2e-01	+0.01 (-0.01 - +0.04)	4.8e-01	+0.01 (-0.02 - +0.05)
rs763737	153278307	0.611	0.607	5.0e-01	+0.00 (-0.01 - +0.01)	6.8e-01	-0.00 (-0.02 - +0.02)	7.2e-01	-0.00 (-0.02 - +0.01)
G6PD_X_b36_153426354	153773160	0.014	0.021	7.7e-01	-0.01 (-0.04 - +0.03)	6.0e-01	-0.01 (-0.06 - +0.04)	8.2e-01	-0.01 (-0.11 - +0.09)
rs4898389	153827637	0.935	0.928	8.5e-01	+0.00 (-0.02 - +0.02)	7.9e-01	+0.01 (-0.04 - +0.05)	9.9e-02	-0.02 (-0.05 - +0.00)
rs766420	153554404	0.361	0.321	8.8e-01	+0.00 (-0.01 - +0.01)	3.9e-01	+0.01 (-0.01 - +0.02)	7.4e-02	+0.02 (-0.00 - +0.04)
rs12389569	153757734	0.067	0.072	8.8e-01	-0.00 (-0.02 - +0.02)	7.6e-01	-0.00 (-0.03 - +0.02)	4.4e-01	+0.02 (-0.03 - +0.06)
G6PD_X_b36_153426256	153773062	0.105	0.101	8.9e-01	+0.00 (-0.01 - +0.02)	5.7e-01	+0.01 (-0.02 - +0.03)	9.2e-01	+0.00 (-0.04 - +0.04)
G6PD_X_b36_153414758	153761564	0.105	0.103	9.4e-01	-0.00 (-0.02 - +0.01)	8.0e-01	+0.00 (-0.02 - +0.03)	9.2e-01	+0.00 (-0.04 - +0.04)
rs2230036	153760953	0.104	0.098	9.5e-01	+0.00 (-0.01 - +0.02)	7.0e-01	+0.00 (-0.02 - +0.03)	8.8e-01	+0.00 (-0.04 - +0.04)

Table S5.3: Association test results for effects independent of G6PD202. Shown here for each SNP surveyed is: genomic position (hg19/GRCh37), DAF stratified by sex, P values and effect sizes (with 95% CI) for G6PD202-controlled association tests under three different genetic models.

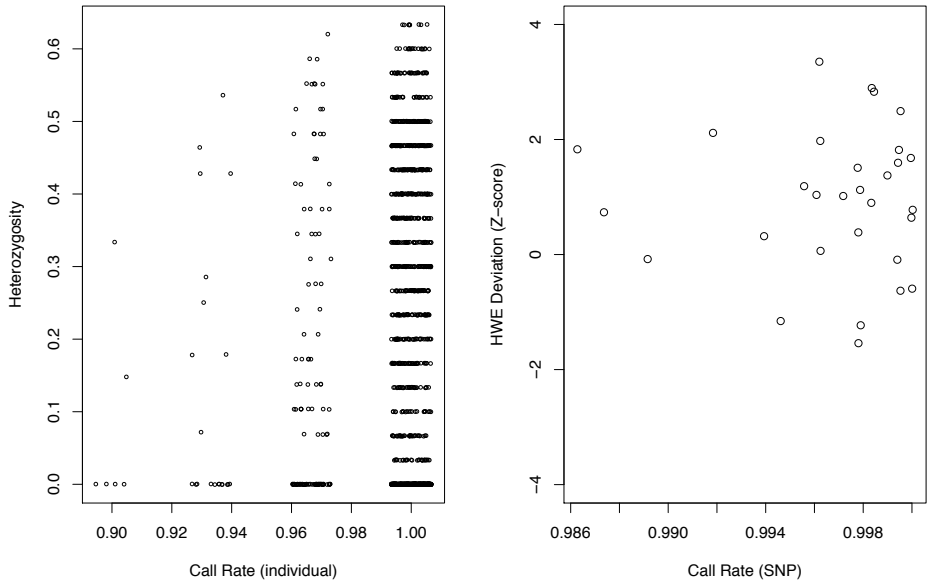


Figure S5.1: Quality control plots. Shown here are plots that examine: (a) whether any individuals with low call rate and high heterozygosity (left), or (b) whether any SNPs with low call rate and extreme deviation from HWE (right) could be found, features indicative of low confidence samples and SNPs, respectively.

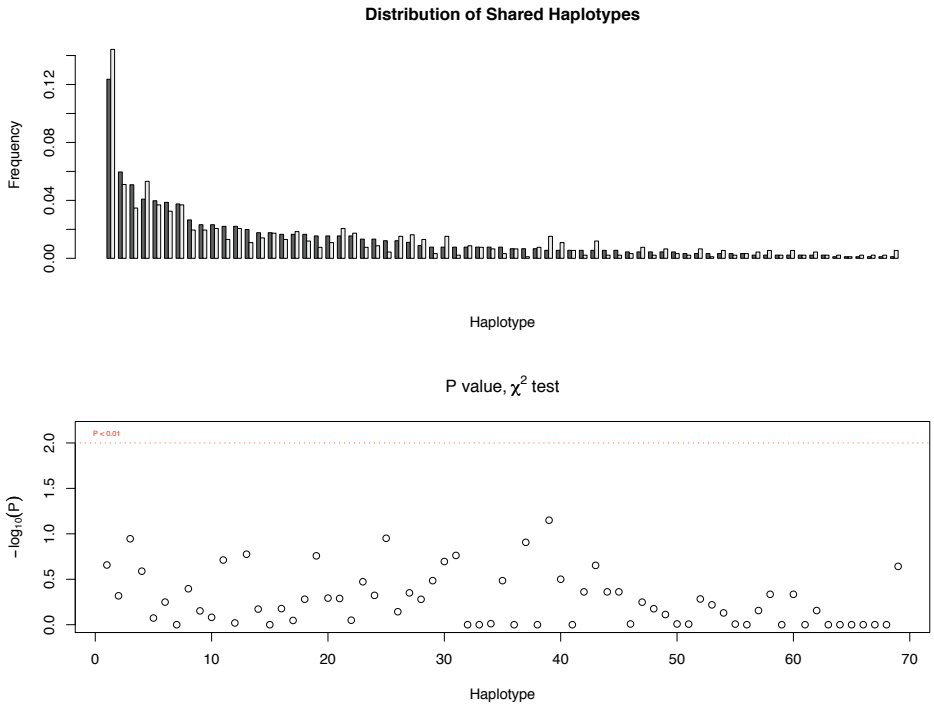


Figure S5.2: Haplotype distribution in males vs. females. The frequency of each haplotype is plotted in decreasing order, stratified by sex (top). No significant differences between male and female haplotype frequencies were found, attesting to highly accurate haplotype phasing (bottom).

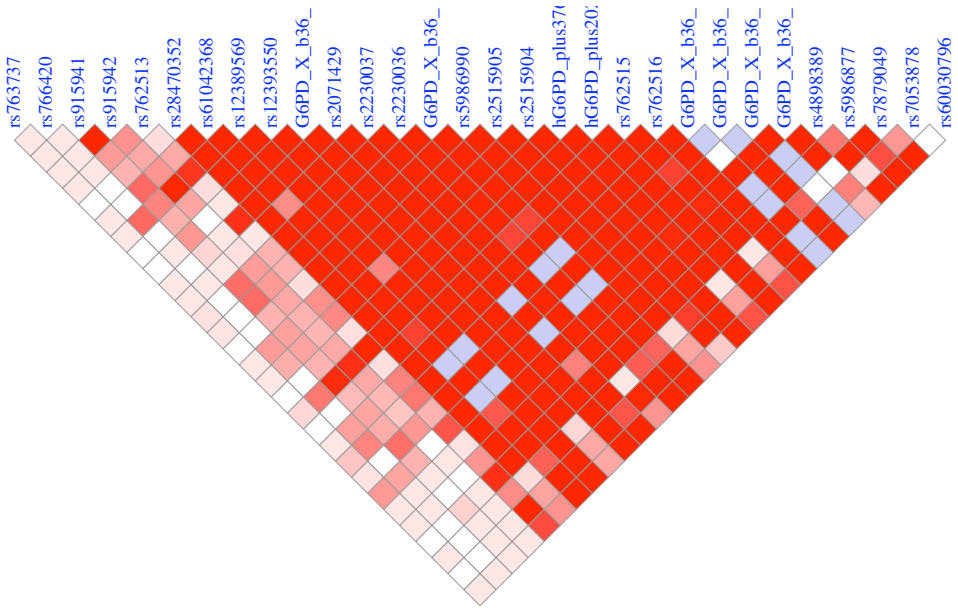


Figure S5.3: Linkage disequilibrium (LD) across region. Pairwise LD matrix is shown in colorimetric form, using the D' and LOD metrics (reviewed in Chapter 2: Supplemental Information). Colors indicate: D' = 1 and LOD $\geq$ 2 (red), D' = 1 and LOD $\leq$ 2 (blue), D' < 1 and LOD $\geq$ 2 (pink), D' < 1 and LOD $\leq$ 2 (white).

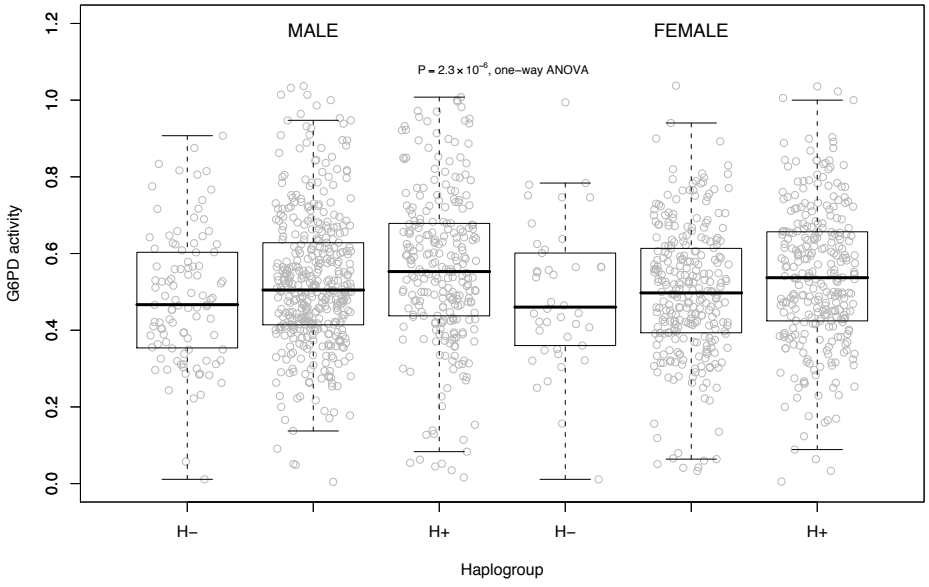


Figure S5.4: Allele sharing with H+ or H- explains variance independent of G6PD202. G6PD activity as stratified by sex and haplogroup, where the latter is defined by categorizing each individual according to extensive allele sharing with either the H+, H-, or neither haplotype. In box-and-whisker plots, center of box represents median, edges of box represent upper and lower quartile bounds, and whiskers represent either minimum and maximum values or 1.5 times the interquartile range, whichever is less extreme.

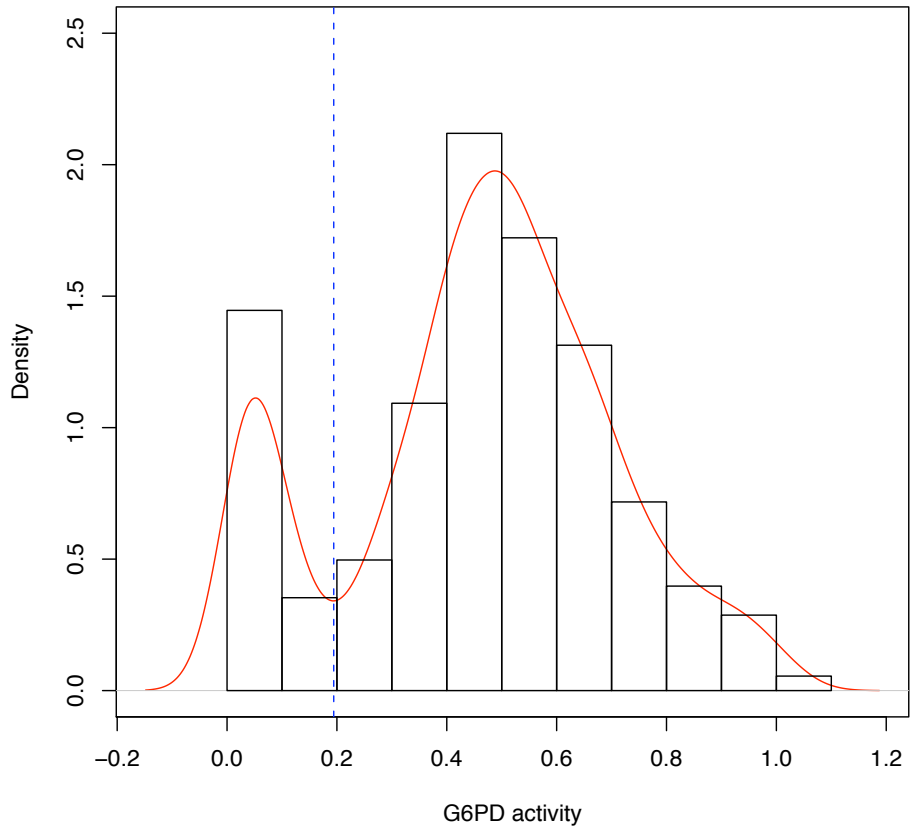


Figure S5.5: Defining qualitative G6PD deficiency. The intermodal minimum in the G6PD activity distribution in males was determined using kernel density estimation of the probability density function (red) and provided our ad hoc cutoff (blue) for distinguishing normal vs. deficient. A histogram of the G6PD activity distribution in males is superimposed for clarity.

Chapter 6

A novel cytofluorometric assay for detecting G6PD deficiency

In the previous chapter, I considered the functional genetics at the *G6PD* locus, conducting a fine-mapping association study looking at gross G6PD enzyme activity. Here, I consider an alternative approach to characterizing G6PD activity variation. In particular, I present a novel method for assessing G6PD activity at the level of an individual red blood cell.

6.1 Abstract

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is an X-linked enzyme disorder affecting hundreds of millions of people worldwide. While clinical manifestation of the trait is generally mild, serious complications can arise, including acute hemolysis and neonatal jaundice, usually occurring in the context of an acute environmental insult from certain drugs, foods, or infections.

Screening and diagnosis of G6PD deficiency is currently performed via genetic or biochemical assay, the former being cost ineffective in populations with significant allelic heterogeneity, and the latter limited in its ability to detect female heterozygotes. Cytochemical assays which measure G6PD enzyme activity at the level of individual red blood cells (RBC) can obviate these shortcomings, but at the expense of added technical complexity and labor.

I describe here a simple, novel cytofluorometric method that extends the classic methemoglobin reduction test, probing G6PD enzyme deficiency at the level of an individual RBC. In preliminary testing of blood obtained from forty-two Malian children, there was good agreement between this method and established techniques such as genotyping and measures of gross enzyme activity. The assay is robust, economical and may be able to serve as a screening method as well as a research tool, especially for high-throughput applications such as flow cytometry. It may thus be useful in studying the differential penetrance of G6PD deficiency trait with regard to important molecular and clinical phenotypes, including primaquine sensitivity and protection from severe malaria.

6.2 Introduction

6.2.1 Glucose-6-phosphate dehydrogenase in the erythrocyte

As reviewed earlier (see Chapter 1), glucose-6-phosphate dehydrogenase (G6PD) is a housekeeping enzyme that catalyses the rate-limiting first step in the pentose phosphate pathway, where it produces nicotinamide adenine dinucleotide phosphate (NADPH), a critical cofactor in a variety of metabolic processes. In the RBC, G6PD activity is especially important as the principal source of reducing power to regenerate antioxidant glutathione, and individuals with deficient G6PD activity are thus at risk for hemolysis precipitated by oxidant stressors found in certain foods or chemotherapeutics, or generated during severe infections. One especially important cause of drug-induced hemolysis is primaquine [91, 92, 241], one of few antimalarial drugs active against both *P. vivax* hypnozoite and *P. falciparum* gametocyte stages. As areas of high malaria transmission also tend to have high prevalence of G6PD deficiency, this drug sensitivity reaction is an important public health concern, especially given recent interest in reducing and eliminating malaria transmission [71, 244].

Despite decades of research, understanding of the molecular pathogenesis of G6PD deficiency in the erythrocyte is incomplete. While many functional allelic variants have been described, it is less well-understood how that genetic heterogeneity translates into diversity at the level of molecular phenotypes, and ultimately to differences in clinical outcome. Since clinical phenotypes such as primaquine-sensitivity and malaria protection are complex and multifactorial in etiology, it is a critical prerequisite that the more proximate link between genetics and molecular phenotypes be clearly elucidated, requiring the development of novel methodology suitable for high-throughput work.

One fundamental molecular phenotype to consider is G6PD enzyme activity, which operates not only at the level of a single RBC, but also in a population of red cells. At the level of a single RBC, while many severe, sporadic variants lead to a defect in the ability of G6PD to form active dimers and tetramers, the means by which most common mutations such as the main African variant G6PD A-, are able to confer deficiency is unknown. As the effect on catalytic rate is minimal, it is thought instead that G6PD deficiency in these cases may result from decreased thermal stability of the protein [98, 96]. According to this theory, nascent RBC emerge from the bone marrow invested with near-normal G6PD activity in deficient individuals, and over time, the RBC aging process leads to a rapid decrease in the enzyme activity levels [95]. As many factors influence red cell senescence, it is also important to think of G6PD enzyme deficiency as a population characteristic, dependent on the age distribution of RBC, where a significant proportion of young reticulocytes can impact the accuracy of a G6PD enzyme activity assay. In females that are heterozygote for deficiency polymorphisms, the picture of cell population variation is even more complicated owing to random X inactivation, which leads to a mosaic population of normal and deficient RBC. Selection processes during or shortly after hematopoiesis, or skewed patterns of X

inactivation [245], which itself may be influenced by age [82], each have the potential to significantly alter the G6PD deficiency status in a heterozygote female.

6.2.2 Detection and quantitation of G6PD deficiency

Various methods have been developed for assessing an individual's G6PD status, and these can be loosely grouped into genetic, gross biochemical, and cytochemical approaches.

Genetic typing. Genetic typing of G6PD variants has a long history which began shortly after the development of gel electrophoresis, with variant G6PD enzymes first identified on the basis of differential migration rate on starch gels [246]. Today, this has been largely superseded by DNA-based genotyping and sequencing, which have allowed identification of hundreds of nonsynonymous mutations associated with G6PD deficiency worldwide [94], including many region-specific common variants.

Biochemical typing. Gross biochemical typing of G6PD involves either direct or indirect measures of enzyme activity in hemolysate. Direct tests are those that assess the enzyme's production of NADPH, which absorbs (340 nm) and fluoresces (450 nm) at characteristic wavelengths of light. These include the quantitative gross activity assay [247], where the initial reaction velocity of the G6PD reaction is measured by examining the rate at which NADPH is produced. A qualitative analogue is the fluorescent spot test [248], a popular screening test in which a drop of blood incubated with G6PD reaction substrates is placed on filter paper and illuminated with UV light, where the presence or absence of fluorescence provides a categorical measure of G6PD activity. Indirect assays of G6PD activity translate NADPH production into a colorimetric readout using chromophores like brilliant cresyl blue [249], resazurin [250], or formazan derivatives [251, 252]; these assays typically involve long incubation periods of several hours (it is likely that they all rely on one or more oxidoreductases with poor kinetic profiles). One indirect test with shorter incubation times and more consistent results is the methemoglobin reduction test (MRT), once a widely-used screening tool in resource-poor areas [190, 191]. In this colorimetric assay, G6PD activity is assessed by treating RBC first with nitrite (converting oxyhemoglobin [red] to methemoglobin [brown]), and then examining the rate of NADPH-dependent methemoglobin reduction in the presence of an appropriate redox catalyst (Nile blue or methylene blue) and substrate (glucose).

Cytochemical typing. Cytochemical typing provides a categorical assessment of G6PD enzyme activity (i.e. 'normal' or 'deficient') at the level of an individual RBC, and thus is the biochemical technique with greatest sensitivity for detecting mosaicism in female heterozygotes, who have a mixed population of normal and deficient RBC due to random X-inactivation [93]. The earliest assay developed, the methemoglobin elution test [253], is a light microscopic technique that extends the aforementioned MRT by treating thin blood smears with peroxide and citric acid prior to conventional staining, with individual RBC labeled according to their relative methemoglobin content. Another more recently developed assay detects the quenching of glutaraldehyde-induced autofluorescence by formazan crystals produced in cells with normal G6PD activity [254]. Owing to their relative difficulty, technical complexity, and often ambiguous results [191], neither technique is particular well-suited to high throughput application.

6.2.3 Methodological limitations

Neither genetic nor gross biochemical approaches capture the full picture of G6PD. While genetic typing offers generally unequivocal interpretation of results, it is imperfectly predictive of both molecular and clinical phenotypes, and typing multiple polymorphisms may not be cost-effective in resource-poor areas of allelic heterogeneity such as West Africa, southeast Asia, or Oceania. Conversely, while gross biochemical approaches offer a functional assessment of G6PD deficiency, they are often difficult to interpret, especially for mosaic female heterozygotes, and insufficiently scalable. Cytochemical methods can thus fill a void, and beyond their utility as a screening technique, can serve as a research tool to understand how genetic polymorphism manifests at the level of G6PD activity in individual RBC.

A straightforward cytochemical typing assay is presented here, termed the 'cytofluorometric method,' which provides a fluorescence readout of the classic MRT at the level of an individual RBC. Upon testing this method on a small cohort of Malian children, it was found to exhibit strong agreement with established methods of genetic and gross biochemical typing. This assay represents a useful addition to the screening and research toolkit for G6PD deficiency, especially in malaria-endemic areas.

6.3 Materials and Methods

Samples. A total of forty-two peripheral blood samples from children ranging in age from one to fifteen years (mean = 5.8) were collected at a rural health center in Kenieroba, Mali, including 22 males and 20 females, with most individuals being of self-identified Malinke ethnicity (79%). Sample collection was nonrandom, as individuals were subject to ascertainment by genotype at G6PD 202. Blood was obtained by venipuncture, anticoagulated using acid-citrate-dextrose, and washed three times in RPMI 1640 prior to storage at 4°C. All samples were obtained with community permission and individual informed consent, collected as part of protocols approved by institutional review boards at the University of Bamako and the National Institute of Allergy and Infectious Diseases (refer to section 2.1.5 for protocol details), and were processed within four days of the receipt.

Genotyping. Genotyping of the G6PD 202 polymorphism was performed via restriction fragment length polymorphism (RFLP) analysis using a nested PCR protocol as described previously [19].

Gross activity assay. Gross activity was measured in duplicate with the G6PDH kit (Randox Laboratories, PD-410), using a modified protocol adapted for high-throughput screening. First, hemolysate was prepared by combining 10 μl of patient red blood cells (RBC, 50% hematocrit) with 20 μl of digitonin. Next, two microliters of this hemolysate was combined with 4 μl NADP (0.34 mM), 2 μl glucose-6-phosphate (0.58 mM), and 120 μl reaction buffer (triethanolamine 31.7 mM, EDTA 3.2 mM, pH 7.6). Reaction rate was assessed using a UV/VIS plate spectrophotometer by observing the rate of production of NAPDH, which absorbs light at 340 nm, over a 30 minute time course. Reaction rate was normalized to total protein content by measuring absorbance of hemoglobin at 540 nm, as follows:

$$\text{G6PD activity} = \frac{\frac{\Delta OD340}{\Delta t}}{OD540} \tag{6.1}$$

Methemoglobin reduction test. The methemoglobin reduction test was performed in duplicate by modification of previously published protocols [190, 191]. One hundred μl of RBC (5% hematocrit) was combined with 100 μl of sodium nitrite (0.125 M), and incubated at room temperature for 20 minutes. Samples were then washed three times with phosphate-buffered saline (PBS), with centrifugation at 1800 rpm for 1 min, and resuspended to 100 μl using PBS. Finally, the washed, nitrite-treated RBC were combined with 18 μl of glucose (0.28 M in PBS) and 6 μl of Nile blue sulfate (0.01%), and incubated at 37°C for 90 minutes in an aerobic environment, i.e. in open microcentrifuge tubes.

Cytofluorometric method. At the completion of the methemoglobin reduction test, 2.5 μl of potassium cyanide (0.4 M) was added to each sample, and incubated at room temperature for 5 minutes. Finally, 5 μl of each sample was added to 100 μl of hydrogen peroxide (3% in PBS), agitated vigorously by hand, and then washed two times in PBS, with centrifugation at 3000 rpm for 3 minutes.

Flow cytometry. Samples were analysed in duplicate using an Accuri C6 flow cytometer, using default factory settings, with 10,000 total events meeting size threshold of FSC-H > 80,000 collected. Samples were excited using an Ar ion laser (488 nm), with fluorescence emission measured in the FL1 channel (533 ± 30 nm).

Fluorescence microscopy. Sample wet mounts were prepared and imaged on an Olympus BX53 fluorescence microscope using the 40x oil immersion objective, a 100 W Hg arc lamp, and a FITC filter cube. All images were taken using a 10 second exposure time and processed identically using ImageJ (NIH).

Data analysis. Custom scripts (available upon request) were written in R to automate data analysis. Gross activity was calculated using the formula above, with coefficients derived from linear modeling functions. Flow cytometry data was first converted from FCS format to raw data using the 'flowCore' package [255], and then subject to morphological gating, with events retained only if they lay in the central 95% of the empirical distribution for both forward scatter (FSC) and side scatter (SSC) parameters. Next, the FL1 fluorescence intensity distribution was analysed to identify and quantify high and low intensity subpopulations. First, maxima in the kernel density estimation function were identified, and then the midpoint between the two maxima was used as an *ad hoc* gate between high and low intensity subpopulations (Figure 6.3). When only one maximum was found (i.e. for homozygotes), the *ad hoc* gate was placed 0.5 log units away (about half the average distance between high and low intensity subpopulations in our dataset).

6.4 Results

6.4.1 Development of cytofluorometric method.

The cytofluorometric method described here can be thought of most simply as a fluorometric readout of a classic test for G6PD deficiency, the methemoglobin reduction test (MRT), at the level of an individual RBC (Figure 6.1).

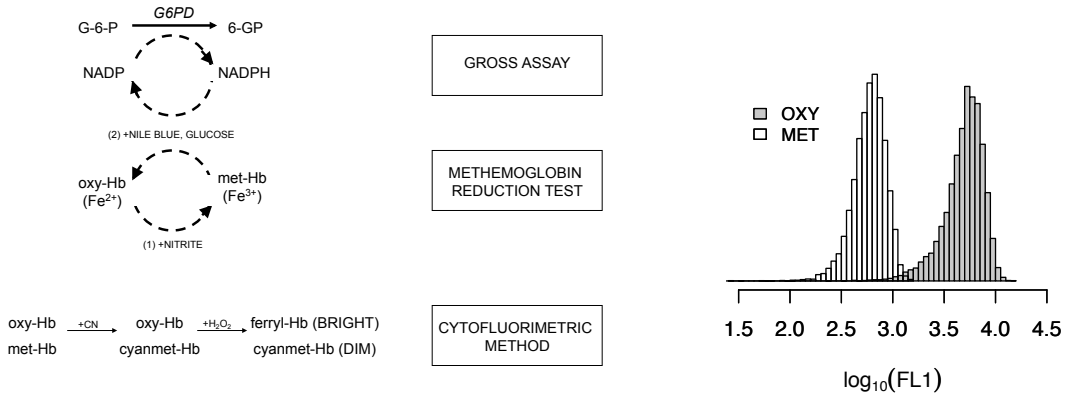


Figure 6.1: Principle of cytofluorometric method. The cytofluorometric method is a fluorometric extension of the methemoglobin reduction test (left). Fluorescence intensity distributions (FL1, right) demonstrate that the sequential reaction of cyanide followed by peroxide in RBC containing predominantly oxyhemoglobin yields a fluorescent product (grey), while little or no fluorescence results in RBC treated with nitrite which contain predominantly methemoglobin (white).

The MRT is a standard colorimetric test for G6PD deficiency [190, 191]. In the first step, RBC were treated with nitrite to oxidize all oxyhemoglobin (oxy-Hb, Fe^{2+}) to methemoglobin (met-Hb, Fe^{3+}). After washing away excess nitrite, glucose (substrate) and Nile blue sulfate (redox catalyst) were added, driving reduction of met-Hb back to oxy-Hb via an NADPH-dependent methemoglobin reductase (NADPH-hemoprotein reductase). The activity of this enzyme is dependent on NADPH, which is produced in the RBC solely by G6PD. The rate of the reductase reaction generally proceeds slowly, but is enhanced by Nile blue, which acts as intermediate electron carrier between NADPH and the reductase. Thus, it is an indirect assay for G6PD deficiency, as normal RBC (G6PD+) are able to quickly reduce met-Hb, while deficient RBC (G6PD-) are not. At the end of the incubation period, then, G6PD+ samples (containing oxy-Hb) were red, G6PD- samples (containing met-Hb) were brown, with heterozygous (G6PD+/-) samples exhibiting an intermediate color. The MRT is sensitive to detection of both mild (class III, e.g. 'G6PD A-') and more severe (class II, e.g. 'G6PD Mediterranean') forms of G6PD deficiency [142, 256], presumably because of a threshold effect where a certain level of G6PD activity is necessary for methemoglobin reduction to proceed at a visually observable rate within the time window allotted.

The cytofluorometric method follows on from the MRT by sequential reaction of cyanide followed by peroxide to generate a fluorescent product inside RBC containing oxy-Hb, but little or no fluorescence from those containing met-Hb. In the first step, only met-Hb reacted with cyanide to generate cyanmethemoglobin (cyanmet-Hb), while oxy-Hb remained unreactive [257]. In the second step, only oxy-Hb reacted with peroxide [258, 259] to generate a fluorescent product (oxy-Hb*), while cyanmet-Hb remained largely nonreactive and nonfluorescent (Figure 6.1). Thus, after these two reactions, G6PD+ cells contained fluorescent oxy-Hb*, while G6PD- cells contained nonfluorescent cyanmet-Hb (Figure 6.2), with fluorescent signal uniformly distributed throughout the cell (Figure 6.3). One cannot say with absolute certainty what the fluorescent adduct(s) generated from these reactions are, but it is likely the major component is derived from conversion of ferrous heme to fluorescent iron-free porphyrin [260, 261, 262, 263], with an additional contribution from lipid peroxidation products such as malonyldialdehyde [264, 265].

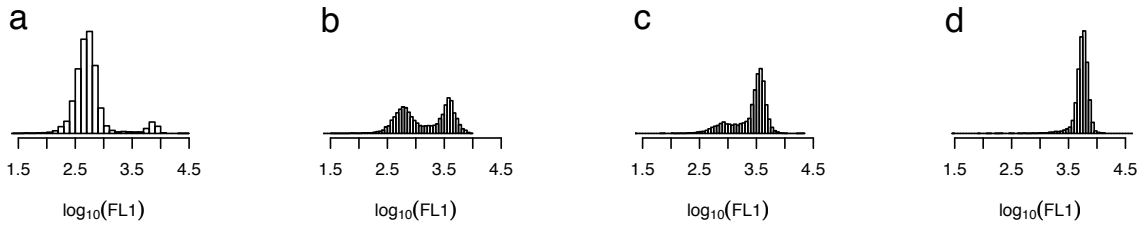


Figure 6.2: Demonstration of cytofluorometric method on artificial mixtures of A+ and A- RBC. Fluorescence intensity distributions (FL1) from artificial mixtures of A+ and A- RBC containing either: (a) 0%, (b) 50%, (c) 75%, or (d) 100% A+ cells. Genotype is defined as follows: 'A+' = 202C (wild-type), 'A-' = 202T (deficient).

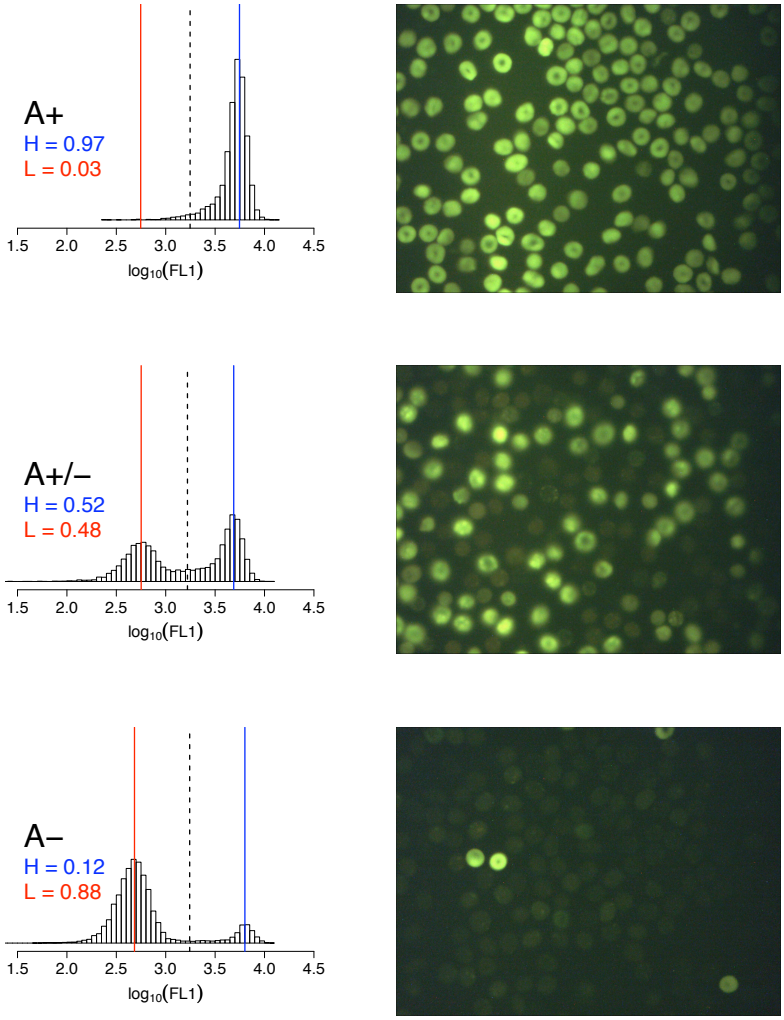


Figure 6.3: Demonstration of cytofluorometric method on RBC from A+, A+/-, and A- individuals. Representative fluorescence intensity distributions (FL1) and fluorescence micrographs from A+ (top), A+/- (middle), and A- (bottom) individuals. Genotype is defined as follows: 'A+' = 202C (wild-type), 'A+/-' = 202CT (heterozygous), 'A-' = 202T (deficient). The estimated proportion of high (blue) and low (red) intensity cells for each individual is indicated. NB: The micrographs show two populations of RBC, one that is bright (G6PD+) and another that is dim (G6PD-).

6.4.2 Correlation with genotype and gross activity data.

The cytofluorometric method was applied to forty-two peripheral blood samples obtained from Malian children as part of an ongoing cohort study. In addition, gross measures of G6PD enzyme activity were obtained using hemolysates derived from the same samples, as was genotype data for the G6PD 202 ('A-', rs1050828, C>T, Val68Met) variant, the most common G6PD deficiency polymorphism in Africa [103]. These latter two 'gold-standard' techniques provide a benchmark against which to evaluate the method.

As expected, there was a strong relationship between the genetic and gross biochemical activity data (the two established techniques), with variation in genotype explaining 78% of variation in gross activity (Figure 6.4A, $P = 5.5 \times 10^{-14}$, 1-way ANOVA). Interestingly, the correlation between genotype and the cytofluorometry data was even stronger, with genotype explaining 89% of variance in the proportion of G6PD+ cells found by cytofluorometry (Figure 6.4B, $P = 1.0 \times 10^{-21}$, 1-way ANOVA). Finally, it was found that the proportion of G6PD+ cells observed by cytofluorometry was well-modeled by gross activity (Figure 6.4C, Spearman's $\rho = +0.88$, $P < 2.2 \times 10^{-16}$), with the relationship remaining significant when considering only heterozygotes, among whom variance in both measures is greatest (Spearman's $\rho = +0.65$, $P = 0.01$).

A few outliers were observed in the data, including two individuals who had homozygous A+ genotypes at G6PD 202 but exhibited mosaicism by cytofluorometry, which may be the result of other, untyped G6PD deficiency mutations that are present at lower frequency in Mali (e.g. G6PD 542, 680, or 968, see Chapter 3, Figure 3.10). The lone heterozygous outlier illustrates the difficulty of detecting certain mosaic females who have seemingly normal enzyme activity, a well-known documented limitation of the MRT [253]. Interestingly, A- males still exhibited a significant population of G6PD+ RBC, which likely reflects the younger population of RBC (e.g. reticulocytes) in which G6PD enzyme levels have not yet declined significantly.

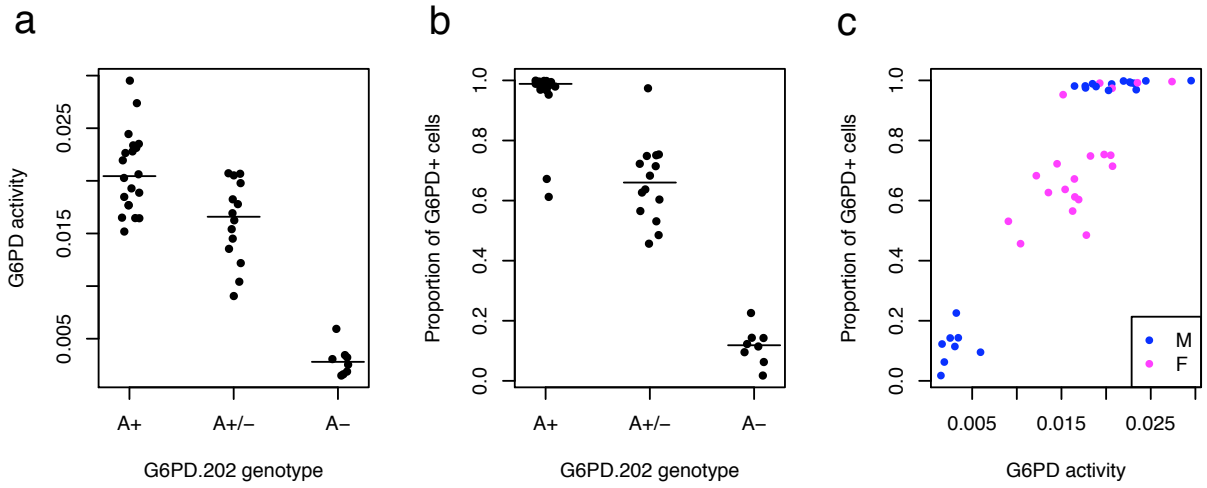


Figure 6.4: Correlation between genotype, gross activity and cytofluorometric data. Concordance between (A) genotype and gross activity, (B) genotype and cytofluorometry, (C) gross activity and cytofluorometry. Genotype is defined as follows: 'A+' = 202C (wild-type), 'A+/-' = 202CT (heterozygous), 'A-' = 202T (deficient). Each point represents the mean of duplicate experiments, with gender indicated by blue (male) and magenta (female) dots. Horizontal line indicates median. G6PD activity units are arbitrary.

6.5 Discussion

The cytofluorometric method described here is a new approach for typing G6PD deficiency at the level of the individual erythrocyte. It provides a fluorescence readout of an established screening method for G6PD deficiency, the methemoglobin reduction test (MRT), utilizing the unique redox chemistry and abundance of hemoglobin in the RBC to generate a differential autofluorescent signal suitable for distinguishing between RBC of normal and deficient activity. In preliminary testing in Malian children, this assay showed strong agreement with other established methods of typing G6PD deficiency, including genotyping and gross biochemical activity. It additionally has performance characteristics which make it useful for clinicians and researchers alike, especially in resource-poor settings.

The literature suggests different possibilities for the nature of the fluorescent signal created when hydrogen peroxide reacts with oxyhemoglobin. Since this peroxide reaction does not occur with cyanmethemoglobin, it seems likely that the major fluorescent adduct created is a protoporphyrin derivative [260, 261, 262, 263] generated from ferryl heme (Fe^{4+}) created as a result of the reaction of peroxide with ferrous heme (Fe^{2+}). Another possible component of the fluorescent signal may derive from lipid peroxidation [264, 265], including fatty acid derivatives such as malonyldialdehyde (MDA). Since MDA generation has been shown to be precipitated by unstable hemoglobins, it is possible that there is a complicated and co-dependent relationship between porphyrin generation and any downstream adducts such as those from lipid peroxidation [263, 266].

Several questions emerging from this work concern the methemoglobin reduction test, upon which the cytofluorometric method is dependent. One issue involves the potential for heritable variation in the activity of the NADPH-dependent methemoglobin reductase enzyme itself to confound assay results. Another known complication in interpreting the results of the MRT concerns cell-cell interaction [191, 253], specifically the finding that soluble factors can traffic from G6PD+ to G6PD- cells, possibly confounding results in mosaic females. While we found that using Nile blue sulfate (instead of methylene blue) and performing the assay at lower hematocrit largely disrupted these interactions (data not shown), there was one A+/- outlier who appeared as a homozygote by cytofluorometry, which may have been a methodological artifact caused by cell-cell interaction during the MRT.

One phenomenon not discussed in detail here, owing to lack of sufficient data, is the observation that the cytofluorometric method not only generated differential fluorescence in RBC based on methemoglobin content, but could also differentially generate ghosted RBC when performed under mild detergent permeabilization. This approach is a logical adjunct to the slide-based methemoglobin elution procedure [259] used in the cytochemical microscopic method [253]. In that procedure, a thin smear of cyanide-treated RBC are prepared and immersed into a mixture of ethanol, citric acid, and hydrogen peroxide, after which the slides are fixed and stained with hematoxylin and erythrosin. Cells containing mainly

met-Hb will appear ghosted (clear), while those containing oxy-Hb will appear normal (pink). While the principle underlying this technique is unclear, it is thought that reaction of oxy-Hb with peroxide protects it from acid elution (ghosting) under conditions of ethanol permeabilization. As cyanmet-Hb does not react with peroxide, it is eluted from the cell, producing ghosts.

Several attempts were made to create a solution-phase analogue to this slide-based technique, and it was found that with a slight alteration of the cytofluorometric method, substituting 0.2% Tween-20 in PBS for plain PBS, cells were sufficiently permeabilized to allow differential ghost formation to occur (even in the absence of citric acid). What this suggests is the possibility for an easy-to-read visual screening test for G6PD deficiency that could be performed in a microcentrifuge tube, where three possible results would indicate the three possible genotypes:

1. G6PD(+): oxy-Hb reacts with peroxide → pellet (pink/red) + supernatant (clear)
2. G6PD(-): cyanmet-Hb does not react with peroxide → pellet (white) + supernatant (pink)
3. G6PD(+/-): two populations of cells → pellet (pink) + supernatant (light pink)

While time and resources did not permit further development of this test, I believe it would be quite useful as a field screening method, particularly for detecting heterozygotes, as the reaction chemistry distinguishes between two populations of cells, where the gross activity assay reflects an ensemble average.

Overall, in spite of the aforementioned concerns, assay performance was found to be robust and reliable, giving consistent results across repeated measurements. Although the focus here was on a high-throughput application of the method using flow cytometry, the technique may be adaptable to additional applications. One direct use would be for investigators interested in methemoglobinemia— to detect methemoglobin in an individual RBC one can simply use the assay independently of the G6PD activity-dependent MRT step. As a screening tool for G6PD deficiency, one need only have a microscope outfitted with a UV lamp to detect G6PD+ and G6PD- cells in a blood sample, with no special fixation or staining procedures required and at a low cost-per-assay. This may be especially well-suited to areas such as southeast Asia and Oceania where there is considerable allelic heterogeneity [103] that makes genotyping cost ineffective. Since the test provides a fine-grained parsing of G6PD activity status at the level of individual RBC, it represents an improvement insofar as information content over the gross biochemical assays from popular screening tests used currently in the field. For example, the older standard, the Beutler fluorescent spot test [248], relies on aggregate production of NADPH from hemolysate, where the fluorescence of NADPH generated is directly observed via UV excitation. A more recently developed test uses water-soluble tetrazolium as a colorimetric readout for NADPH production and is rapidly read, but it still employs hemolysate and lacks sensitivity to detect individuals with milder forms of G6PD deficiency [252]. Finally, a new chromatographic test that uses nitro-blue tetrazolium as the

colorimetric partner compound for NADPH is able to provide rapid results in five minutes, but is likewise dependent on hemolysate and must be performed between 18-25°C, which may limit its utility in tropical environments [267].

Finally, if the ultimate goal is to understand how genetic variation at *G6PD* influences an individual's complex, multifactorial response to primaquine administration or to malaria infection, methods like that described here will be critical in bridging the gap by further elucidating the proximate relationship between genotype and molecular phenotypes.

6.6 Acknowledgments and Contributions

I thank all the children and their families who participated in this study. S.S. designed the study, in collaboration with Rick Fairhurst, an investigator in the Laboratory of Malaria and Vector Biology at the National Institute of Allergy and Infectious Diseases, National Institutes of Health in Bethesda, USA. Seidina Diakite and Karim Traore performed the genotyping work in the lab of Mahamadou Diakite at the Malaria Research Training Centre based at University of Bamako in Bamako, Mali. S.S. designed, validated, and performed all enzyme activity and cytofluorometric assays, including all flow cytometry and microscopy work. S.S. conducted all statistical analyses.

Chapter 7

G6PD deficiency exerts contrasting effects on two major clinical presentations of severe malaria in West Africa

In the final results chapter, I directly consider the central question of this thesis, namely, the relationship between genetic variation at *G6PD* and severe malaria. A fine-mapping association study at the locus is presented here which examines the association with severe malaria and its constituent clinical subphenotypes in the Gambia. Specifically addressed is the importance of considering diversity at the level of both functional genetics and clinical presentation in revealing the complex relationship betweeen G6PD deficiency and severe malaria.

7.1 Abstract

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is a common enzyme disorder that has long been thought to be protective against malaria, though genetic association studies have disagreed with regard to the strength and specificity of a protective effect, which may reflect differences in the host genetic background, environmental influences, or in the particular clinical phenotype(s) considered.

I present here a comprehensive fine mapping case-control association study of severe malaria at the *G6PD* locus in the Gambia. A total of 3300 controls and 2650 severe malaria cases were genotyped at 34 single nucleotide polymorphisms, including three nonsynonymous variants known to be associated with enzyme deficiency in west Africa. The initial scan revealed seven non-independent signals of association with $P < 0.01$, with no single SNP exhibiting uniquely strong statistical significance suggestive of a causal role. In decomposing severe malaria to two of its constituent clinical subphenotypes, severe malarial anemia (SMA) and cerebral malaria (CM), it was found that the protective effects observed were limited to CM, and that the three known deficiency alleles exhibited trends of opposing effect – risk conferred to SMA, and protection with respect to CM. Tests for the composite effect of all G6PD deficiency alleles

showed that G6PD deficiency trait as a whole was significantly associated with protection from CM (additive model: $OR = 0.70$ [0.59 - 0.85], $P = 2.1 \times 10^{-4}$), but trended toward increased susceptibility for SMA, especially with full expression of the trait (recessive model: $OR = 1.44$ [0.99 - 2.09], $P = 0.054$).

These results demonstrate that G6PD deficiency alleles are strongly associated with protection from severe malaria, but in contrast to sickle cell trait, this protection is limited to CM, and furthermore deficiency alleles may predispose to SMA.

7.2 Introduction

7.2.1 G6PD deficiency and the malaria protection hypothesis

The prevalence of G6PD deficiency, the most common enzymopathy in man, is geographically correlated with the historical distribution of malaria [88]. Thus, as malaria is an infectious disease known to be a major contributor to childhood morbidity and mortality in endemic areas [112], and also thought to exhibit highly heritable differences in burden [268, 111], it has long been hypothesized that G6PD deficiency is a target of positive selection owing to partial protection afforded against malaria [110]. Consistent with this idea, population genetic studies have found decreased haplotype diversity associated with deficiency alleles, suggestive of recent selective sweeps in sub-Saharan Africa [121, 123] and southeast Asia [125]. Genetic epidemiological evidence is lacking in clarity, however, with case-control association studies differing with respect to the strength and specificity of a protective effect [136, 39, 19, 143, 243], and even whether such an effect exists at all [269].

7.2.2 Understanding between-study heterogeneity in association tests of G6PD deficiency and malaria

The difficulty in interpreting past studies lies not only in reconciling the different clinical phenotypes considered in each, but also in recognizing different assumptions made in defining the G6PD deficiency trait itself. Additionally, regional variation in environmental pressures or host genetic background may account for some proportion of the variability observed.

7.2.2.1 Clinical phenotypes

Clinical phenotypes considered in association studies of malaria range from measures of malaria poorly correlated with disease burden, such as parasite rate or density [135, 137, 136], to more extreme phenotypes pathognomonic for severe, life-threatening malaria, including profound anemia, respiratory distress, and cerebral manifestations such as convulsions or coma [138, 39, 19, 143]. Additionally, World Health Organization criteria for diagnosis of severe malaria are broad, and include a wide range of symptoms and clinical findings (reviewed in Chapter 1), which vary substantially not only in terms of pathophysiology, but also with respect to prognostic value [21].

7.2.2.2 Defining G6PD deficiency

Interpreting trait carriage is also nontrivial in that G6PD deficiency is both a genetic and a biochemical trait. As an X-linked genetic trait, hemizygous males and homozygous females fully express the deficiency trait, while heterozygous females exhibit incomplete or partial expression due to random X inactivation [93, 253]. As a biochemical trait, enzyme activity can be assayed quantitatively, but overlapping activity distributions for the different genotypic categories make qualitative assessment imperfect [242, 243].

In genetic testing of G6PD deficiency alleles, it is important to note that such alleles can vary with

respect to both population frequency and severity of G6PD deficiency caused, both of which will impact the power of an association test to observe a real effect if one is present. Regarding the former, the lower the allele frequency of a variant, the less power one has to detect a true effect, as small frequency perturbations are more likely to occur by chance when low counts of a particular genotype are observed; testing of low frequency variants therefore requires a greater number of samples to achieve adequate statistical power (Table S7.4). With respect to the severity of a particular G6PD deficiency variant, if the degree of protection or susceptibility conferred is thought to vary in a dose-dependent manner with the extent of G6PD deficiency caused, then one would expect more extreme deficiency variants to generate larger effect sizes, and yield fewer false negative results in association tests. For example, in two recent studies looking at association with *P. vivax* clinical burden, the study considering a more severe variant (G6PD-Mediterranean, class II; [270]) was able to detect a significant protective effect not observed in the study examining a milder variant (G6PD-Mahidol, class III; [125]).

7.2.2.3 Regional influences

The only statistically significant evidence for a protective effect with respect to severe malaria comes from the West African countries of Nigeria, Mali, and the Gambia [138, 39, 19, 143]. It is possible that this localization may be due to regional influences, such as environmental effects or differences in host genetic background, that modulate the effects of G6PD deficiency. For example, differences in host genetic background might affect the biochemical penetrance of G6PD deficiency, and epistatic interactions [141] with other genetic variants exhibiting high population differentiation may influence susceptibility to malaria. Additionally, environmental influences such as diet and more importantly, the ecology of the *Anopheles* vector and the dynamics of malaria transmission should be considered; the latter is especially key given its influence on the clinical manifestations and burden of malaria observed in a region [271].

7.2.3 A comprehensive exploration of the genotype-phenotype space

I sought to clarify the relationship between G6PD deficiency and severe malaria by more fully probing the genotype-phenotype space. I asked specifically: (a) whether genetic variation near the *G6PD* locus was associated with protection from severe malaria, (b) whether that protection was specific for a certain clinical subphenotype, and (c) whether known deficiency alleles could be specifically implicated in protection. To do this, a multi-step approach was taken, beginning with variant discovery via resequencing of the *G6PD* gene in several African cohorts (described in Chapter 3), followed by multiplexed genotyping of over seventy known and novel polymorphisms, and finally a fine-mapping genetic association study in the Gambia, an area of Africa where there is considerable allelic heterogeneity [272, 143] contributing to the prevalence of the G6PD deficiency trait.

7.3 Materials and Methods

Study cohort The case-control cohort was comprised of 3300 controls and 2650 severe malaria cases, recruited at the Royal Victoria Hospital in Banjul, The Gambia. Malaria cases were confirmed by thin blood smear, and designated as 'severe' following WHO criteria, with two partially overlapping clinical syndromes identified in the majority of cases: severe malarial anemia and cerebral malaria. Severe malarial anemia (SMA) was defined here as packed cell volume $\leq 15\%$ or hemoglobin ≤ 5 g/dl, while cerebral malaria (CM) was defined as having a Blantyre coma score ≤ 3 in the absence of other obvious causes, e.g. hypoglycemia or meningitis. Five hundred forty (21%) of our cases met the definition for SMA and 871 (33%) met that for CM, with 127 (5%) cases fitting both definitions (these were excluded in the analysis of each clinical subphenotype). Control samples were derived from cord blood obtained from local perinatal clinics. All samples were collected after obtaining informed consent. Study approval was granted by The Gambia Government/Medical Research Council Joint Ethics Committee and by the Oxford Tropical Research Ethics Committee (refer to section 2.1.5 for protocol details). A demographic summary of the cohort is provided in Table S7.1.

Laboratory methods Resequencing of PCR amplicons tiling across 15 kilobases at the *G6PD* locus was performed in 330 individuals from The Gambia, Burkina Faso, Cameroon, Nigeria, Kenya, Vietnam, and Papua New Guinea, with a total of 72 single nucleotide polymorphisms (SNPs) identified, of which 49 were novel (see Chapter 3 for more details). A literature search yielded an additional 25 nonsynonymous variants known to be associated with G6PD deficiency and identified in sub-Saharan Africa or southeast Asia in more than one individual.

A total of 70 SNPs were assayed via multiplexed genotyping on the Sequenom MASSarray platform using a total of three multiplex reactions, with final SNP inclusion priority weighted according to functional significance and derived allele frequency criteria. Genotyping was performed on genomic DNA that was whole genome amplified via either primer extension pre-amplification or multiple displacement amplification reaction. All genotype calls were manually curated via visual examination of cluster plots and mass spectra, with ambiguous genotypes re-coded as missing.

Statistical analyses Data curation was performed using custom Perl scripts and the PLINK software package [182]. Individual samples were removed if they exhibited missingness ≥ 0.20, and SNPs were excluded if they did not meet DAF ≥ 0.01, missingness ≤ 0.10, or deviated from Hardy-Weinberg proportions (in females) with $P < 10^{-5}$. Genders were curated using chromosome X SNP data, via an inbreeding coefficient test that recodes gender based on observed heterozygosity. Any remaining heterozygous genotypes found in males were coded as missing.

Association testing was conducted using custom scripts written in R [181]. A logistic regression framework was employed, using a generalized linear model assuming a binomial error distribution. Three different genetic models were considered: additive (trend), dominant, and recessive. In the additive model, male genotypes were coded 0/2, with females coded as 0/1/2, as suggested by Clayton [189]. In the dominant and recessive models, males were coded 0/1, and females were coded 0/1/1, or 0/0/1, respectively. Estimated odds ratios were adjusted by including the following covariates: HbS (sickle locus) genotype, self-reported ethnicity, and sex. Analysis of covariance was used to test for sex-specific differences in odds ratio estimates. Additional tests were conducted using the snpMatrix package [183] and included the 2 d.f. general genotypic and the 1 d.f. Cochran-Armitage trend tests, each Mantel-stratified by ethnicity and performed only on HbS AA individuals. All odds ratios are presented here as "OR [95% CI]."

Composite G6PD deficiency trait genotype was defined as the sum of derived alleles present at all deficiency loci (968, 542, and 202) for an individual. For example, in the additive model, a male was coded as 0 or 2, based on the presence or absence of a single derived allele across the three loci, while a female with a total of one derived allele present was coded as 1, two derived alleles as 2 (this included nine 'compound heterozygote' individuals who had single derived alleles at two different deficiency SNPs).

Haplotypes were inferred in females using fastPHASE [185], with default settings, using the directly observed male haplotypes as a reference input. Haplotype phasing accuracy was verified by comparing the haplotype frequency distribution in males and females using chi-square tests. For haplotype association testing, an individual was scored 0/1 for the absence or presence of a particular haplotype. Since females have twice the number of haplotypes as males, one haplotype at random was removed from all females prior to association testing to avoid artificially skewing the contribution of females to the overall testing. Haplotype frequency threshold for testing was 0.01.

7.4 Results

7.4.1 Fine-resolution association mapping at the *G6PD* locus.

After quality control filtering of low quality samples and polymorphisms, individual tests of disease association were conducted for 34 SNPs spanning 600 kb, twenty-three of which localize to a 20 kb region containing the *G6PD* gene itself (Table S7.2). These included four nonsynonymous variants at *G6PD*, which are labeled according to convention based on cDNA sequence position: G6PD968, G6PD542, G6PD202, and G6PD376 (Table S7.3). Three of these mutations are known to be associated with enzyme deficiency, with G6PD376 incidentally related to the others, as it alone does not lead to enzyme deficiency, but the other three deficiency alleles all lie on a 376C derived allele background. The tests conducted include a general 2 d.f. genotypic test and three 1 d.f. tests considering additive (trend), dominant, and recessive models of inheritance. These tests yielded seven SNPs with $P < 0.01$, including one of the known deficiency alleles, 968G ($P_{add} = 0.007$), with 542A ($P_{dom} = 0.012$) also exhibiting modest association (Tables S7.5, S7.6, S7.7 and S7.8). Interestingly, no single SNP stood out from the rest in terms of statistical significance (Figures 7.1 and S7.1), likely reflective of extensive linkage disequilibrium (LD) between the polymorphisms surveyed (Figures S7.7, S7.8, S7.9, and S7.10).

7.4.2 Stratified testing of two major clinical subphenotypes

As severe malaria encompasses several clinical syndromes distinct in pathophysiology and prognostic value, I next examined association with the two major clinical subphenotypes predominant in the Gambia: severe malarial anemia (SMA) and cerebral malaria (CM). Although these subphenotype tests were not equally powered to detect signals of association ($N_{SMA} = 540$, $N_{CM} = 871$), they nonetheless suggested a differential direction of disease association with respect to SMA and CM for all three known *G6PD* deficiency alleles (Figures 7.2, 7.3, S7.2 and S7.3; Table S7.6). Indeed, while these deficiency alleles seemed to confer protection against CM (additive model: $OR_{968G} = 0.71$ [0.56 - 0.89], $OR_{542A} = 0.45$ [0.22 - 0.95], $OR_{202T} = 0.82$ [0.60 - 1.13]), they exhibited a modest trend toward susceptibility for SMA (additive model: $OR_{968G} = 1.10$ [0.89 - 1.37], $OR_{542A} = 1.31$ [0.85 - 2.03], $OR_{202T} = 1.13$ [0.83 - 1.56]). Interestingly, the 542A allele, which is associated with a more extreme enzyme deficiency state (WHO class II, 1-2% residual G6PD activity [273]), yielded more extreme effect sizes for both subphenotypes– greater risk to SMA and greater protection from CM– than either 968G or 202T, which are of milder biochemical penetrance (WHO class III, 10-20% residual G6PD activity [100]).

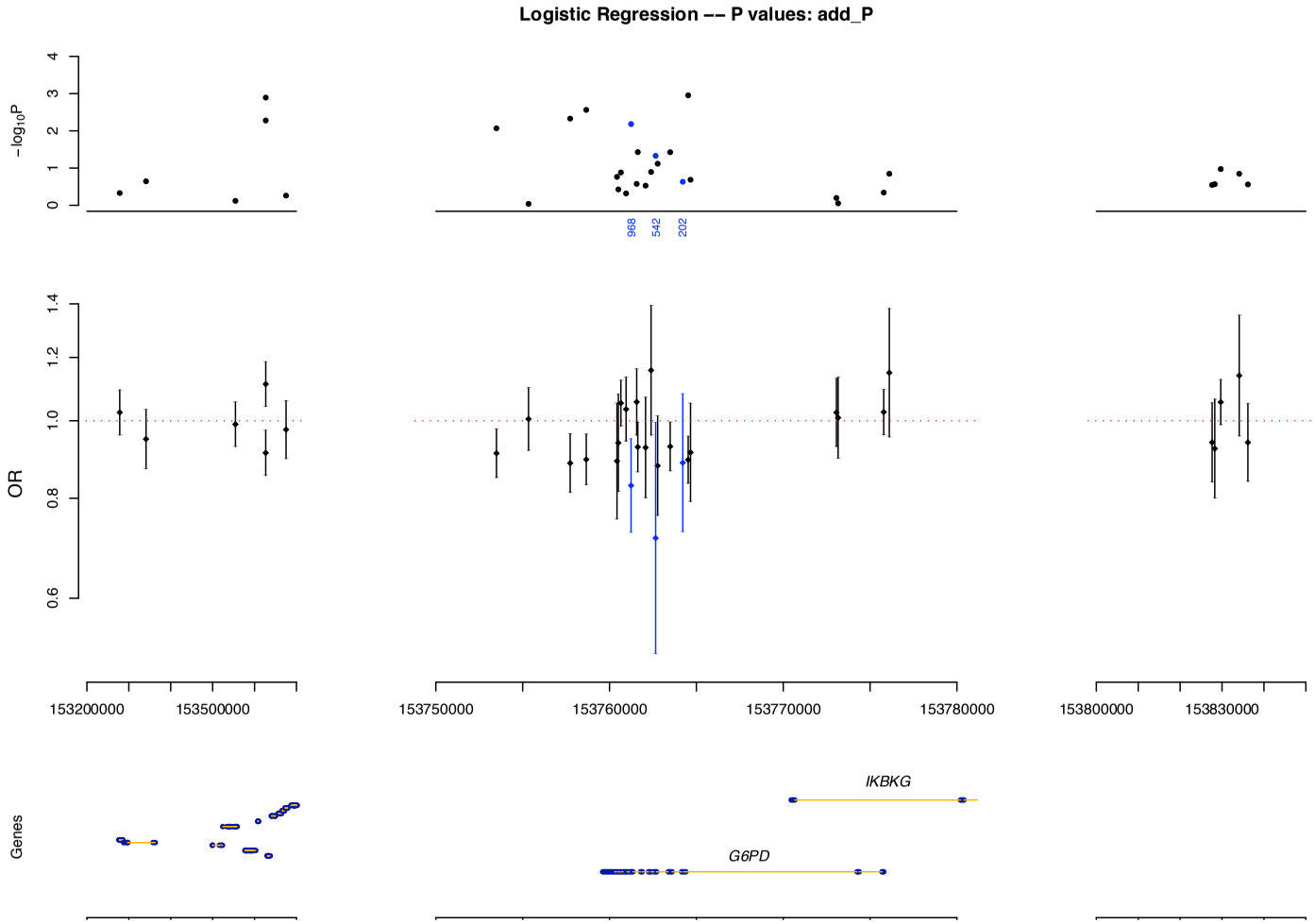


Figure 7.1: Logistic regression association tests – additive (ALL). The results of single SNP association tests for severe malaria (all forms) are shown here, including: statistical significance (top), estimated odds ratios with 95% CI (middle), and gene graphic (bottom). Deficiency-associated SNPs are in blue. Gene graphic shows exons (blue) and introns (yellow) for genes annotated at UCSC Genome Browser (hg19/GRCh37); *IKBKG*=inhibitor of NF-kB kinase γ.

7.4.3 Integrating information from multi-locus haplotypes

Given the strong LD between the polymorphisms in the data set, association tests from individual SNPs cannot be considered to be independent of one another, and it is useful to interpret single SNP association signals in the context of local haplotype structure.

One direct approach to addressing this problem would be to conduct association testing of individual haplotypes. Such testing is highly specific, in that any correlated genotypes will be present on similar haplotypes, but poorly powered to detect signals from causal alleles that lie on genetically diverse backgrounds (i.e. each haplotype that a particular causal allele lies on may exhibit only weak association). One can restrict the haplotype search space by decreasing the number of sites considered, but this increase in statistical power comes at decreased specificity for identifying the true causal variant. As my concern here is resolving the multiple signals observed in single SNP testing, I opted to maximize specificity, and tested all sixteen full-length (34 site) haplotypes present at frequency > 0.01 (542A-bearing haplotypes did not meet this threshold). In these admittedly underpowered tests, one signal of association was found, for a haplotype bearing 968G (the most frequent deficiency allele in the Gambia), which was associated with severe malaria with $OR = 0.64$ [0.49 - 0.85], $P = 0.002$.

Another inferential approach invokes known phylogenetic relationships between haplotypes. The established convention at *G6PD* is to categorize haplotypes with respect to allele identity at G6PD376. Haplotypes bearing 376T (G6PD B) are known to be older than those bearing 376C (G6PD A), the latter forming the base for the even younger deficiency allele haplotypes [206, 121]. In the data presented here, based on increased nucleotide diversity and higher ancestral allele frequencies, it was confirmed that G6PD B is likely older than G6PD A, which is older than the haplotypes bearing known functional alleles, on which most alleles are present at or near fixed allele frequency (DAF = 0 or 1), indicative of a recent time of origin (Figure S7.4).

I re-examined the single SNP test results in the context of these known phylogenetic relationships. Specifically, I considered all sites on a particular deficiency allele haplotype (DAH) that were uniquely associated with that haplotype relative to ancestral sequence, i.e. those sites where the DAF was highly differentiated between the DAH and G6PD B haplotypes ($|\Delta(DAF)| > 0.5$). These DAH-specific sites share three other important characteristics: (1) they include all SNPs exhibiting the strongest LD with the deficiency SNP itself (upper quartile of pairwise r^2 distribution), (2) they encompass all top hits from the single SNP tests ($P < 0.01$), and (3) they include sites which are common to all three DAH, reflecting their common origin from G6PD A. In examining the estimated ORs at these DAH-specific sites, a pattern consistent with susceptibility to SMA and protection from CM was observed, with each signal centered at the *G6PD* gene. In the case of haplotypes bearing 968G and 542A, the functional allele itself exhibited the most extreme effect with respect to both SMA and CM subphenotypes, which

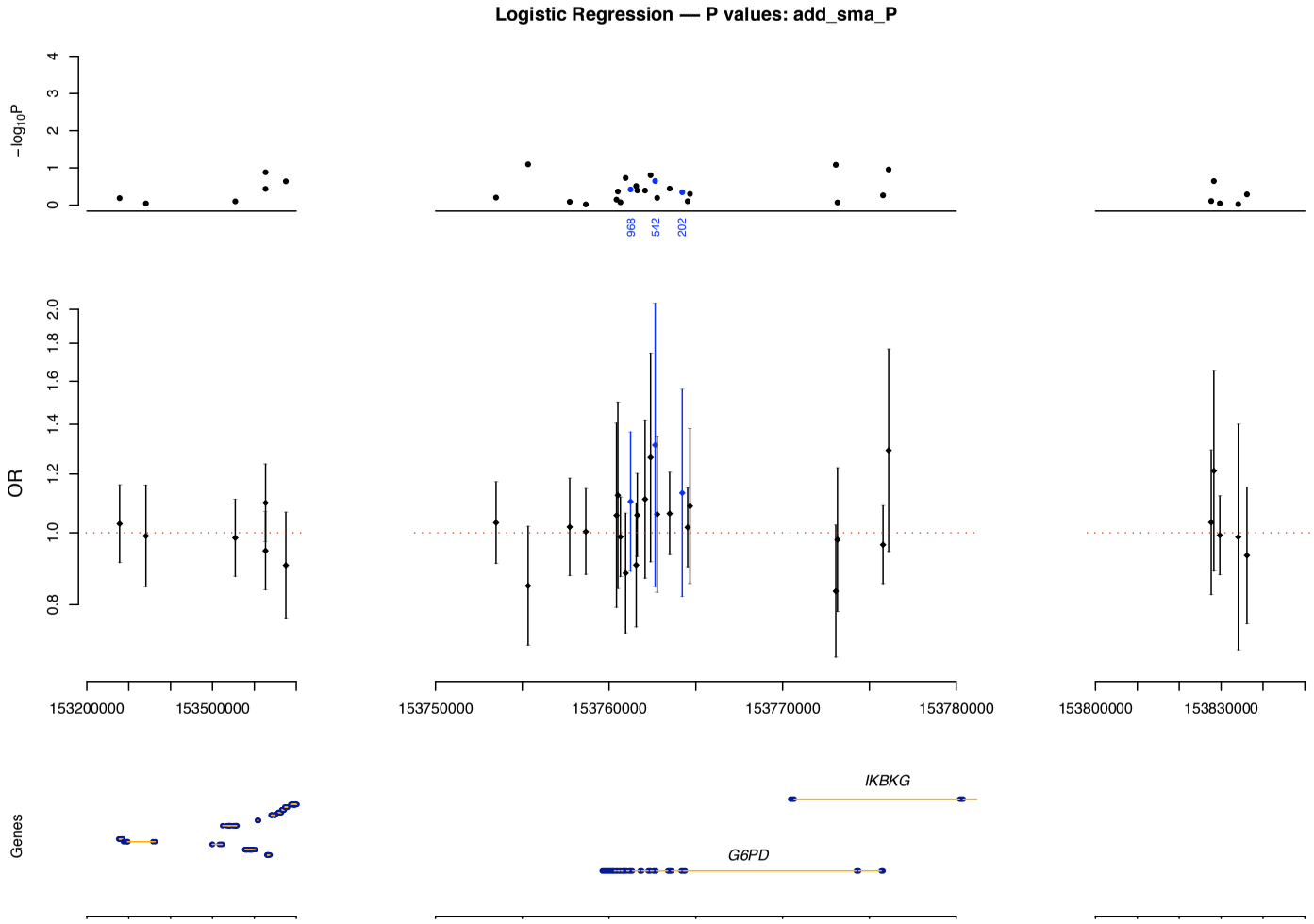


Figure 7.2: Logistic regression association tests – additive (SMA). The results of single SNP association tests for SMA are shown here, including: statistical significance (top), estimated odds ratios with 95% CI (middle), and gene graphic (bottom). Deficiency-associated SNPs are in blue. Gene graphic shows exons (blue) and introns (yellow) for genes annotated at UCSC Genome Browser (hg19/GRCh37); *IKBKG*=inhibitor of NF-kB kinase γ.

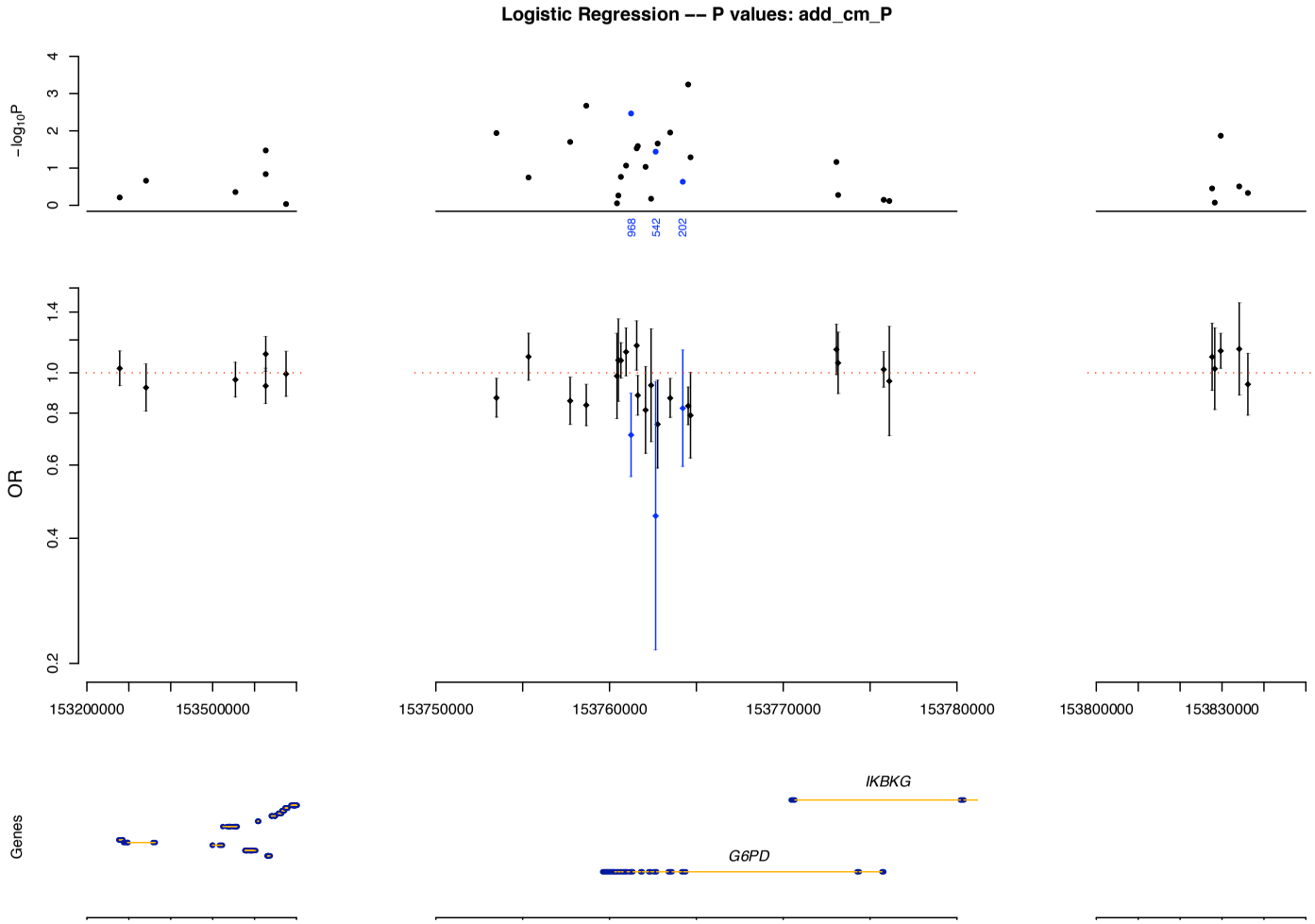


Figure 7.3: Logistic regression association tests – additive (CM). The results of single SNP association tests for CM are shown here, including: statistical significance (top), estimated odds ratios with 95% CI (middle), and gene graphic (bottom). Deficiency-associated SNPs are in blue. Gene graphic shows exons (blue) and introns (yellow) for genes annotated at UCSC Genome Browser (hg19/GRCh37); *IKBKG*=inhibitor of NF-kB kinase γ.

suggested that these G6PD deficiency alleles themselves, apart from other DAH-specific alleles, may exert a causal role with respect to severe malaria (Figures 7.4 and S7.5).

7.4.4 Association of G6PD deficiency trait with severe malaria

Finally, since a similar direction of association with respect to SMA and CM was observed for all three known deficiency variants, and given that each contributes to the overall prevalence of G6PD deficiency trait in the Gambia, it made sense to consider them as a unit and examine the association of G6PD deficiency trait as a whole with severe malaria. Specifically, I tested the hypothesis that possession of one or more copies of derived alleles at any of the known functional sites was associated with susceptibility to severe malaria. A composite genotype representing G6PD deficiency trait status represented the sum of derived alleles across all three deficiency sites, with this genotype scored in the same way as all other genotypes were, namely that hemizygous-deficient males were treated as equivalent to homozygous-deficient females.

The results of these tests indicate that G6PD deficiency trait exerts a strong protective effect with regard to CM ($OR = 0.83$ [0.74 - 0.92], $P = 5.5 \times 10^{-4}$, additive model) but may confer increased risk with respect to SMA, especially in fully-deficient individuals ($OR = 1.44$ [0.99 - 2.09], $P = 0.054$, recessive model). This differential direction of association comported with results from the single SNP tests, was consistent across all genetic models considered, and contrasts with the consistent protective effect across clinical subphenotypes observed for HbS (Table 7.1 and Figure 7.5).

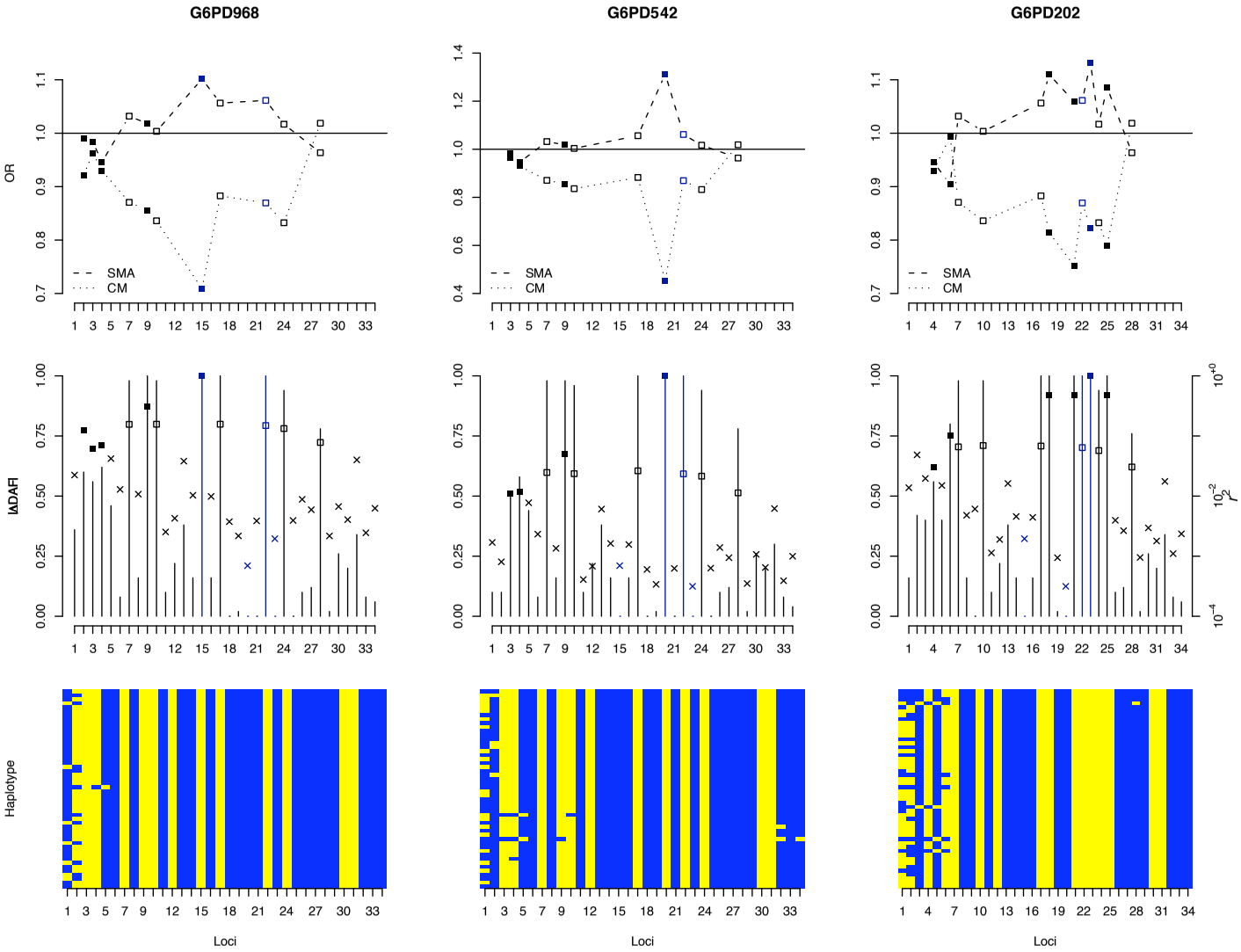


Figure 7.4: Haplotype context. TOP: Shown here are SMA and CM association test OR for all sites on a deficiency allele haplotype (DAH) that uniquely associated with that haplotype relative to ancestral sequence, based on allele frequency differences ($|\Delta(DAF)| > 0.5$). These sites share three other important characteristics: (1) they include all SNPs exhibiting the strongest LD with the deficiency SNP of interest (upper quartile of pairwise r^2 distribution), (2) they also encompass all top hits from the single SNP tests ($P < 0.01$), and (3) there are sites common to all three DAH, reflecting their common origin from G6PD-A. Sites common to all three DAH (open squares) are shown along with sites unique to each DAH (filled squares), with nonsynonymous variants shown in blue (site 22 is G6PD376). MIDDLE: Displayed here are the $|\Delta(DAF)|$ (bars) and pairwise r^2 values (LD metrics are reviewed in Chapter 2: Supplemental Information) with the corresponding deficiency allele (points). BOTTOM: A random sampling of fifty DAH (piled along y-axis) are shown to illustrate haplotype structure, with different colors denoting ancestral (blue) and derived (yellow) alleles.

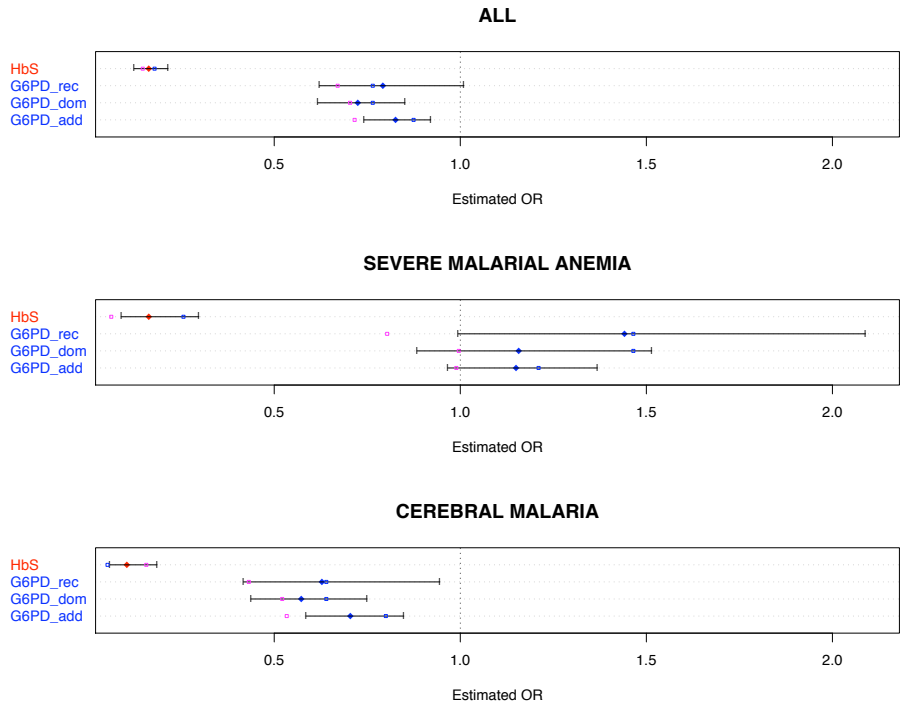


Figure 7.5: Composite G6PD genotype association tests – estimated odds ratios. Results from association tests using composite G6PD genotypes, where composite genotype is the total number of deficiency alleles present at either 968, 542, or 202 in an individual. Estimated OR with 95% CI is shown under three different genetic models (additive, dominant, recessive) for all severe malaria cases (ALL), as well as for each subphenotype (SMA, CM). Also shown are estimated OR from analyses using either only males (blue square) or females (pink square). For comparison, the odds ratio for the HbS allele is also shown.

Table 7.1: Other tests of association.

	Control	Case (SMA, CM)	P- ALL	Estimated RR	P- SMA	Estimated RR	P- CM	Estimated RR
G6PD_add	0.098	0.070 (0.090, 0.060)	5.5e-04	0.83 (0.74 - 0.92)	1.2e-01	1.15 (0.97 - 1.37)	2.1e-04	0.70 (0.59 - 0.85)
G6PD_dom	0.160	0.118 (0.174, 0.104)	9.0e-05	0.72 (0.62 - 0.85)	2.9e-01	1.16 (0.88 - 1.51)	4.9e-05	0.57 (0.44 - 0.75)
G6PD_rec	0.062	0.049 (0.084, 0.045)	6.0e-02	0.79 (0.62 - 1.01)	5.4e-02	1.44 (0.99 - 2.09)	2.6e-02	0.63 (0.42 - 0.95)
HbS	0.133	0.014 (0.006, 0.009)	5.9e-38	0.16 (0.12 - 0.21)	3.5e-09	0.16 (0.09 - 0.30)	1.5e-14	0.10 (0.06 - 0.19)

Sex-specific differences in association In all analyses conducted, I also tested the null hypothesis that adjusted odds ratios for males and females were the same by analysis of covariance, including an interaction term between genotype and sex in the logistic regression model, which was further complemented by repeating all association tests stratified by sex. Though the estimated odds ratios for males and females were often different from one another, these differences were not statistically significant at $P < 0.01$ for any of the top hits, known functional alleles, or any DAH-specific alleles. Only for one SNP, rs2230037, was a marginally significant difference observed between male and female effect sizes ($OR_M = 0.99$ [0.92 - 1.07]; $OR_F = 1.21$ [1.07 - 1.37]; $P_{cov} = 0.006$).

7.5 Discussion

Fine-mapping association study of severe malaria at *G6PD* The malaria protection hypothesis posits that the G6PD deficiency trait has been positively selected in malaria-endemic regions owing to protection conferred against life-threatening manifestations of malaria. The primary aim here was to conduct an agnostic, fine-mapping association study of severe malaria at the *G6PD* locus to determine what, if any signal(s) of association were present, and to see if known deficiency alleles could be specifically implicated in the protective effect. A secondary aim was to see if there were any differences in the strength or direction of association across clinical subphenotypes or between genders, given that such differences might help clarify previous conflicting results from the literature.

The work presented here represents the largest association study of G6PD deficiency to date, and the first to systematically identify and test the variants that surround known deficiency alleles. A number of SNPs associated with protection from severe malaria were identified, as was a subset of SNPs that exhibited differential effects with respect to two important presentations of severe malaria, chief among these being the known G6PD deficiency alleles. In trying to arrive at a parsimonious explanation for the multiple association signals observed, it was found that the deficiency alleles themselves exhibited effect sizes that uniquely stood out among all SNPs specific to deficiency allele haplotypes.

This study provides the somewhat paradoxical result that a single gene can be associated with protection from and susceptibility to a single disease. The apparent paradox can be resolved when one considers the extent to which severe malaria subphenotypes differ from one another. The finding of risk conferred to SMA is not unexpected, then, given that G6PD deficiency is known to generally predispose individuals to hemolytic anemia in the presence of oxidant stresses from infection, drugs, and diet [274, 28]. Additionally, there is the possibility that the increased risk to SMA may simply reflect a methodological issue relating to case classification, where the presence of a chronic, subclinical anemia might predispose G6PD deficient individuals to falling below the WHO-defined threshold for SMA of 5 g/dl. The evidence for such a mild persistent anemia in African G6PD deficient individuals is unclear, however; it is supported by two studies from Tanzania and Malawi [275, 276], but not supported by another study in Nigeria [242].

The other intriguing aspect of the differential association observed concerns natural selection and the malaria protection hypothesis. Clinical epidemiological studies of severe malaria in Kenya and Ghana suggest that SMA has an independent mortality risk of 1-4%, whereas CM has 9-30% mortality risk. At first glance then, it would appear that G6PD deficiency would confer a net selective benefit with respect to malaria, as the stronger signal of association is found with the subphenotype more predictive of severe morbidity and mortality, CM [35]. This should be qualified, however, given that SMA predominates in hyperendemic environments, and thus may on aggregate be just as important a cause of malaria morbidity and mortality as CM [21, 277]. Furthermore, the jaundice and hyperlactatemia that may accompany

SMA are associated with poor prognosis in severe malaria [21]. Additionally, it should be noted that the lower mortality observed for SMA in hospital settings is likely reflective of a cheap, readily-available treatment – transfusion of fresh blood. In rural areas with poor access to health services, the mortality associated with SMA is likely much higher. Finally, while oxidant-induced acute hemolytic anemia is generally mild and self-limiting in European populations, it may be more severe in an African setting where other factors contributing to anemia may be present, such as hookworm infection or nutritional deficiencies [28]. In summary, while the results of this association study are consistent with the malaria protection hypothesis, it seems possible that the risk conferred to SMA represents a non-neglible fitness cost that contributes to balancing selection at the *G6PD* locus.

As this study was limited to just one country, with its own distinct genetic background and environmental exposure history, including its own unique pattern of exposure to malaria, it will be important that further investigation be conducted in as many study populations in as diverse a range of geographic areas as possible. As there is substantial variation in the biochemical penetrance of G6PD deficiency alleles, it will also be useful for future work to include enzyme activity assays when feasible.

Previous association studies of severe malaria Earlier association studies of severe malaria at *G6PD* failed to yield clarity with regard to either the specificity and strength of a protective effect. While much of the attention has centered on the sex-specificity of protection, little has been said about the strength of association for distinct clinical subphenotypes. I decided to re-examine all large association studies of severe malaria, in the hope that I might reconcile my own findings with respect to subphenotype association, but also critically evaluate claims of gender differences.

Gilles et al. [138] performed a modestly powered association study of cerebral malaria in Nigeria, with criteria for inclusion being the presence of high fever, high parasitemia, and either convulsions or coma. Genotype was assigned based on biochemical typing of G6PD enzyme activity, which was validated using three different methods. Though this study included only 200 controls and 100 cases, it nevertheless found a strong protective effect for CM ($OR = 0.31$ [0.10 - 0.79]; $P = 0.007$). Two recent studies examined association with both SMA and CM subphenotypes, each typing the G6PD202 variant ("A-"). Ruwende et al. [39] examined case-control cohorts from the Gambia (421 controls, 556 cases [174 SMA, 382 CM]) and Kenya (292 controls, 250 cases [115 SMA, 135 CM]), while Guindo et al. [19] conducted their study in Mali (2765 controls, 271 cases [223 CM, 48 SMA]). A pooled meta-analysis of these two studies [19] presented strong evidence for protection against CM ($OR = 0.44$ [0.24 - 0.77], $P = 0.002$), but not SMA ($OR = 0.70$ [0.29 - 1.52], $P = 0.48$). Taken together, these three studies suggest that protection from severe malaria is best supported for the CM subphenotype.

The idea of gender differences in association with malaria was first suggested by Luzzatto, Bienzle,

and co-workers [142, 136], who sought to explain balancing selection for an X-linked trait by positing that the heterozygote state, present only in females, was associated with decreased parasite rate and density. In addition to using response variables poorly correlated with disease burden, the specific conclusions drawn in the Bienzle et al. study are incongruous– in males, the nondeficient A allele is found to be protective, and in females, only one type of heterozygote (A-/B but not A-/A) is protected. At the other end of the debate, Guindo et al. have suggested that protection is limited to males, as they fully express the deficiency state in all of their red blood cells. However, their conclusions with respect to gender differences may not be accurate, as their observed female genotype distribution for controls at G6PD202 (TT: 40, CT: 148, CC: 1011) is inconsistent with frequencies expected under an assumption of Hardy-Weinberg equilibrium ($P = 6.6 \times 10^{-22}$, χ^2 test), suggesting cryptic population structure or systematic genotyping error. Given that several other studies have found no significant difference in the protective effect in both males and females [138, 39, 143], it is possible that sex-specific differences in association at *G6PD* have been overstated, as has been the case for many other genes [144].

Formulating new hypotheses for protection Red blood cell polymorphisms have long been thought to confer resistance to malaria, and while strong clinical evidence exists for many variants, including HbS [138, 139, 278], HbC [18, 140], alpha-thalassemia [279, 280], ABO blood group antigens [153, 157], and G6PD, little consensus exists regarding the molecular mechanisms underlying protection.

The most well-studied example, sickle cell trait (HbAS), confers substantial protection against mild as well as severe malaria and each of its constituent subphenotypes [132, 133, 134, 138, 139, 18, 140]. There is consistent *in vitro* evidence demonstrating impaired invasion and growth of parasites in AS erythrocytes [145, 146], an increased propensity for infected AS cells to sickle [147] (enhancing splenic clearance), as well as *in vivo* observations of decreased parasite density in both natural [132, 133, 134] and experimental [132, 281] settings. While HbS may also exert protection through other means, such as modulation of acquired immunity [156] or via impaired rosetting [282] or cytoadherence [155], it seems to have a strong fundamental component involving parasite survival inside the AS red cell itself.

G6PD deficiency, by contrast, seems to only protect against severe malaria, specifically the CM subphenotype. While some have found evidence of invasion preferences [142] or aberrant *in vitro* growth phenotypes [148], others have not [149, 150]. Similarly, evidence of decreased parasite density *in vivo* is mixed for both *P. vivax* [137, 125] and *P. falciparum* [135, 136, 19]. Thus, in contrast to sickle trait, the clinical and experimental evidence for G6PD deficiency suggests that the protection afforded may not be related to parasite survival inside the enzyme-deficient red cell.

Given these differences between HbS and G6PD, it seems all the more important to study clinical

biomarkers and conduct *in vitro* experiments relevant to severe manifestations of malaria. Such investigations are especially critical for G6PD deficiency, where the true protective effect is likely less powerful than for HbS. Indeed, the mechanism of protection may not be erythrocyte-specific– G6PD is an important housekeeping enzyme with a wide distribution in the body, and plays a role in endothelial function (via NO synthase [83, 80]), as well as in the respiratory burst of immune cells (via NADPH oxidase [84]), and G6PD deficiency has been found to be associated with immune-related diseases such as diabetes [104, 105, 106], cardiovascular disease [85, 107, 108], and necrotizing enterocolitis [109]. Ultimately, if knowledge gained from investigations surrounding the malaria protection hypothesis is to inform clinical treatment, a better understanding of the molecular mechanisms as they relate to severe disease will be crucial.

7.6 Acknowledgments and Contributions

I thank all the children and their families who participated in this study. Dominic Kwiatkowski and Kirk Rockett (University of Oxford) designed the case-control study in collaboration with Muminatou Jallow, Fatou Sisay-Joof, and Margaret Pinder, who work at the Medical Research Council laboratories in Banjul, The Gambia. M.J., F.S-J., and M.P. led study participant recruitment at the Royal Victoria Hospital and local health centres near Banjul, The Gambia. K.R., Anna Richardson, Angie Green, Kate Rowlands, Rachel Craik, Christina Hubbart, Alieu Mendy, and other members of the members of the MalariaGEN Resource Centre based at the University of Oxford participated in sample processing and quality control. K.R. and Kate Fitzpatrick curated clinical data. S.S., K.R., Kathrin Schuldt, R.C., and C.H. participated in the design, validation, and performance of genotyping assays at the Wellcome Trust Centre for Human Genetics in Oxford, UK. Capillary sequencing data was generated as part of MalariaGEN CP3 (discussed in Chapter 2) at the Wellcome Trust Sanger Institute in Hinxton, UK. S.S. conducted all statistical analyses. This study was conducted as part of the activities of the MalariaGEN consortium (detailed in Chapter 2), which is funded by the Bill and Melinda Gates Foundation and the Wellcome Trust.

7.7 Supplemental Information

Table S7.1: Demographic summary table. Shown here are the total number of individuals in this study, stratified by ethnicity, sex and case-control status (with cases substratified by subphenotype). Also shown is the mean age (in months) of all cases, as well as those fitting the SMA and CM subphenotype definitions.

	Total	(M, F)	Control	Case	(SMA, CM)	Age	(SMA, CM)
FULA	993	528, 465	662	331	81, 114	44.8	30.9, 50.4
JOLA	923	485, 438	468	455	97, 179	52.6	33.2, 58.2
MANDINKA	1949	1054, 895	1067	882	213, 335	51.3	35.8, 56.1
OTHER	1047	579, 468	593	344	76, 137	53.3	39.0, 60.5
WOLLOF	831	445, 386	517	314	79, 106	46.5	29.0, 53.9
TOTAL	5743	3091, 2652	3307	2326	546, 871	49.7	33.6, 55.8

Table S7.2: Full SNP summary table. Shown here for each SNP are: the genomic coordinate (hg19/GRCh37), the observed alleles (on the [+]-strand, according to UCSC Genome Browser), and the predicted functional consequence per gene transcript.

SNP	Position	Alleles	Gene:Consequence
rs763737	153278307	G/A	IRAK1:INTRONIC
rs1734792	153341060	C/A	MECP2:INTRONIC
rs766420	153554404	G/C	TKTL1:INTRONIC
rs915941	153626649	C/A	RPL10:5PRIME_UTR
rs915942	153626738	G/A	RPL10:SPLICE_SITE,RPL10:INTRONIC,RPL10:5PRIME_UTR
rs762513	153675171	A/G	FAM50A:INTRONIC
rs28470352_A	153753490	T/A	NULL
rs61042368_A	153755336	G/A	NULL
rs12389569_A	153757734	G/A	NULL
rs12393550_A	153758660	G/A	NULL
G6PD_X_b36_153413623_A	153760429	G/A	G6PD:SPLICE_SITE,G6PD:SYNONYMOUS_CODING
rs2071429_A	153760508	A/G	G6PD:INTRONIC
rs2230037_A	153760654	G/A	G6PD:SYNONYMOUS_CODING
rs2230036_A	153760953	C/T	G6PD:SYNONYMOUS_CODING
hG6PD_968_A	153761240	A/G	G6PD:NON_SYNONYMOUS_CODING
G6PD_X_b36_153414758_A	153761564	G/A	G6PD:INTRONIC
rs5986990_A	153761628	G/A	G6PD:INTRONIC
rs2515905_A	153762075	G/A	G6PD:INTRONIC
rs5986875_A	153762392	G/A	G6PD:INTRONIC
hG6PD_542_B	153762655	T/A	G6PD:NON_SYNONYMOUS_CODING
rs2515904_A	153762771	G/C	G6PD:INTRONIC
hG6PD_plus376	153763492	T/C	G6PD:NON_SYNONYMOUS_CODING
hG6PD_plus202	153764217	C/T	G6PD:NON_SYNONYMOUS_CODING
rs762515_A	153764528	T/C	G6PD:INTRONIC
rs762516_A	153764663	C/T	G6PD:INTRONIC
G6PD_X_b36_153426256_A	153773062	C/T	G6PD:INTRONIC,IKBKG:INTRONIC
G6PD_X_b36_153426354_A	153773160	A/C	G6PD:INTRONIC,IKBKG:INTRONIC
G6PD_X_b36_153428979_A	153775785	C/T	G6PD:5PRIME_UTR,IKBKG:5PRIME_UTR,IKBKG:INTRONIC,G6PD:SPLICE_SITE
rs5986992_A	153776107	C/A	IKBKG:5PRIME_UTR,IKBKG:INTRONIC
rs4898389_A	153827637	A/G	AF277315.6:INTRONIC
rs5986877_A	153828269	C/G	AF277315.6:INTRONIC
rs7879049_A	153829693	A/G	AF277315.6:INTRONIC
rs7053878_A	153834100	T/A	NULL
rs60030796_A	153836171	A/G	NULL

Table S7.3: Nonsynonymous SNP summary table. Shown here for each nonsynonymous SNP are: the internal identifier, any associated dbSNP identifier, the colloquial alias, the genomic position (hg19/GRCh37), observed alleles (on the [+]-strand, according to UCSC Genome Browser), amino acid change (according to UCSC genome annotation coordinates), and G6PD deficiency class (as defined by the WHO). See Chapter 1 for more details on WHO classes for G6PD deficiency and see Chapter 2 for more details on coordinate mapping systems used.

SNP	dbSNP ID	Alias	Position	Alleles	Amino Acid Change	Class
G6PD968	NULL	Betica-Selma	153761240	A/G	Leu323Pro	3
G6PD542	NULL	Santamaria	153762655	T/A	Asp181Val	2
G6PD376	rs1050829	A+	153763492	T/C	Asn126Asp	4
G6PD202	rs1050828	A-	153764217	C/T	Val68Met	3

Table S7.4: Sample power calculations for G6PD association studies. Using the CaTS program (A. Skol & G. Abecasis, U. of Michigan), assuming the most conservative effect size reported in the literature (from Clark, et al. [143]), reasonable estimates of MAF across our study sites, and tolerating a Type I error rate of 1/100, we show power calculations for additive and dominant genetic models.

N					*Model*	
cases	controls	MAF	odds ratio	*P*-value	additive	dominant
500	500	0.15	0.75	0.01	0.25	0.17
1000	1000	0.15	0.75	0.01	0.54	0.39
2000	2000	0.15	0.75	0.01	0.88	0.75
500	500	0.10	0.75	0.01	0.17	0.13
1000	1000	0.10	0.75	0.01	0.38	0.30
2000	2000	0.10	0.75	0.01	0.74	0.63
500	500	0.05	0.75	0.01	0.08	0.07
1000	1000	0.05	0.75	0.01	0.19	0.17
2000	2000	0.05	0.75	0.01	0.42	0.38

Table S7.5: Basic tests of association – 2 df genotypic. Shown here are the results of the 2 d.f. general genotypic association tests. For each SNP is listed: genomic position (hg19/GRCh37), as well as P values for association tests for all severe malaria (ALL), or either SMA or CM subphenotype, along with results from sex-stratified tests. Note that 'NA' is the result of insufficient data in the cells of the 2x3 table, which is why males cannot be analyzed alone using this test, as they have an obligate zero count in the heterozygote genotype cells.

SNP	Position	P - ALL	(M, F)	P - SMA	(M, F)	P - CM	(M, F)
rs915942	153626738	6.1e-04	NA, 1.6e-02	3.2e-01	NA, 8.5e-01	2.2e-02	NA, 9.7e-02
rs12389569_A	153757734	1.9e-03	NA, 1.2e-02	8.6e-01	NA, 6.7e-01	1.9e-02	NA, 1.3e-01
rs915941	153626649	3.4e-03	NA, 1.0e-01	4.3e-01	NA, 5.8e-01	2.8e-01	NA, 5.9e-01
rs762515_A	153764528	3.7e-03	NA, 8.2e-02	8.8e-01	NA, 9.3e-01	6.1e-04	NA, 2.0e-02
rs12393550_A	153758660	8.8e-03	NA, 1.2e-01	5.3e-01	NA, 4.9e-01	5.3e-03	NA, 5.4e-02
hG6PD_968_A	153761240	1.5e-02	NA, 3.1e-01	6.1e-01	NA, 7.6e-01	1.1e-02	NA, 6.5e-02
rs28470352_A	153753490	2.5e-02	NA, 1.0e-01	6.9e-01	NA, 6.5e-01	2.5e-02	NA, 1.2e-01
rs7053878_A	153834100	5.4e-02	NA, 1.7e-01	5.7e-01	NA, 2.9e-01	3.4e-01	NA, 4.4e-01
hG6PD_plus376	153763492	6.3e-02	NA, 1.9e-01	2.9e-01	NA, 4.2e-01	1.4e-02	NA, 1.0e-01
rs763737	153278307	8.7e-02	NA, 1.2e-01	3.1e-01	NA, 3.3e-01	7.5e-02	NA, 8.9e-02
rs5986990_A	153761628	1.0e-01	NA, 1.3e-01	5.0e-01	NA, 6.2e-01	6.0e-02	NA, 1.6e-01
rs7879049_A	153829693	3.0e-01	NA, 1.0e+00	8.1e-01	NA, 7.5e-01	5.1e-02	NA, 2.3e-01
rs1734792	153341060	3.1e-01	NA, 3.1e-01	4.6e-01	NA, 4.5e-01	4.9e-01	NA, 6.3e-02
G6PD_X_b36_153414758_A	153761564	3.7e-01	NA, 5.9e-01	5.6e-01	NA, 6.6e-01	7.0e-02	NA, 7.1e-01
rs2230037_A	153760654	3.8e-01	NA, 1.2e-02	9.4e-01	NA, 8.8e-01	3.0e-01	NA, 3.9e-03
rs2071429_A	153760508	3.9e-01	NA, 5.8e-01	6.1e-01	NA, 6.7e-01	6.5e-01	NA, 6.0e-01
rs2515904_A	153762771	4.1e-01	NA, 3.2e-01	6.3e-01	NA, 7.8e-01	1.2e-01	NA, 3.2e-02
G6PD_X_b36_153413623_A	153760429	4.2e-01	NA, 8.0e-01	6.3e-01	NA, 3.3e-01	8.7e-01	NA, 5.7e-01
rs762513	153675171	4.9e-01	NA, 9.9e-02	2.2e-01	NA, 1.4e-01	7.1e-02	NA, 1.8e-02
rs766420	153554404	5.4e-01	NA, 5.2e-01	7.7e-01	NA, 7.7e-01	5.8e-01	NA, 6.7e-01
rs2230036_A	153760953	5.5e-01	NA, 7.5e-01	4.1e-01	NA, 8.7e-01	1.5e-01	NA, 8.2e-01
rs5986877_A	153828269	5.7e-01	NA, 9.1e-01	3.5e-01	NA, 5.9e-01	9.6e-01	NA, 8.0e-01
rs60030796_A	153836171	5.7e-01	NA, 7.4e-01	3.2e-01	NA, 3.7e-01	3.7e-01	NA, 3.9e-01
rs762516_A	153764663	6.3e-01	NA, 6.2e-01	5.3e-01	NA, 6.4e-01	1.9e-01	NA, 6.0e-02
rs4898389_A	153827637	6.6e-01	NA, 8.1e-01	5.2e-01	NA, 4.6e-01	6.0e-01	NA, 2.6e-01
G6PD_X_b36_153426256_A	153773062	7.1e-01	NA, 6.6e-01	2.1e-01	NA, 3.4e-01	4.1e-02	NA, 2.5e-01
rs2515905_A	153762075	7.4e-01	NA, 4.3e-01	5.1e-01	NA, 7.7e-01	3.0e-01	NA, 5.6e-02
G6PD_X_b36_153428979_A	153775785	7.5e-01	NA, 5.9e-02	7.7e-01	NA, 5.5e-01	5.2e-01	NA, 8.7e-02
rs61042368_A	153755336	9.5e-01	NA, 7.7e-01	2.0e-01	NA, 6.3e-01	3.1e-01	NA, 9.9e-01
G6PD_X_b36_153426354_A	153773160	1.0e+00	NA, 9.8e-01	6.8e-01	NA, 7.1e-01	8.4e-01	NA, 7.4e-01
rs5986875_A	153762392	NA	NA, NA	NA	NA, NA	NA	NA, NA
hG6PD_542_B	153762655	NA	NA, NA	NA	NA, NA	NA	NA, NA
hG6PD_plus202	153764217	NA	NA, NA	NA	NA, NA	NA	NA, NA
rs5986992_A	153776107	NA	NA, NA	NA	NA, NA	NA	NA, NA

Table S7.6: Logistic tests of association – 1 df additive. Shown here are the results of the 1 d.f. logistic regression association tests under an additive model. For each SNP is listed: genomic position (hg19/GRCh37), case and control DAF, as well as P values and OR (with 95% CI) for association tests for all severe malaria (ALL), or either SMA or CM subphenotype.

SNP	Position	Control	Case	P - ALL	Estimated OR	P - SMA	Estimated OR	P - CM	Estimated OR
rs762515_A	153764528	0.343	0.311	1.1e-03	0.89 (0.84 - 0.96)	7.9e-01	1.02 (0.90 - 1.1)	5.7e-04	0.83 (0.75 - 0.92)
rs915942	153626738	0.441	0.486	1.3e-03	1.11 (1.04 - 1.18)	1.3e-01	1.10 (0.97 - 1.2)	3.4e-02	1.11 (1.01 - 1.22)
rs12393550_A	153758660	0.308	0.277	2.7e-03	0.89 (0.83 - 0.96)	9.6e-01	1.00 (0.88 - 1.1)	2.1e-03	0.84 (0.75 - 0.94)
rs12389569_A	153757734	0.179	0.167	4.7e-03	0.89 (0.81 - 0.96)	8.2e-01	1.02 (0.88 - 1.2)	2.0e-02	0.86 (0.75 - 0.98)
rs915941	153626649	0.434	0.395	5.3e-03	0.91 (0.85 - 0.97)	3.7e-01	0.95 (0.84 - 1.1)	1.5e-01	0.93 (0.84 - 1.03)
hG6PD_968_A	153761240	0.066	0.054	6.6e-03	0.83 (0.73 - 0.95)	3.8e-01	1.10 (0.89 - 1.4)	3.4e-03	0.71 (0.56 - 0.89)
rs28470352_A	153753490	0.310	0.280	8.5e-03	0.91 (0.85 - 0.98)	6.3e-01	1.03 (0.91 - 1.2)	1.1e-02	0.87 (0.78 - 0.97)
rs5986990_A	153761628	0.312	0.280	3.7e-02	0.93 (0.86 - 1.00)	4.0e-01	1.06 (0.93 - 1.2)	2.6e-02	0.88 (0.79 - 0.98)
hG6PD_plus376	153763492	0.320	0.291	3.7e-02	0.93 (0.87 - 1.00)	3.6e-01	1.06 (0.93 - 1.2)	1.1e-02	0.87 (0.78 - 0.97)
hG6PD_542_B	153762655	0.012	0.007	4.7e-02	0.71 (0.51 - 1.00)	2.2e-01	1.31 (0.85 - 2.0)	3.6e-02	0.45 (0.22 - 0.95)
rs2515904_A	153762771	0.062	0.048	7.6e-02	0.88 (0.76 - 1.01)	6.4e-01	1.06 (0.83 - 1.3)	2.2e-02	0.75 (0.59 - 0.96)
rs7879049_A	153829693	0.368	0.385	1.1e-01	1.06 (0.99 - 1.13)	9.0e-01	0.99 (0.88 - 1.1)	1.4e-02	1.13 (1.03 - 1.24)
rs5986875_A	153762392	0.031	0.038	1.3e-01	1.16 (0.96 - 1.39)	1.6e-01	1.26 (0.91 - 1.7)	6.7e-01	0.93 (0.68 - 1.28)
rs2230037_A	153760654	0.352	0.382	1.3e-01	1.05 (0.98 - 1.12)	8.4e-01	0.99 (0.87 - 1.1)	1.7e-01	1.07 (0.97 - 1.18)
rs5986992_A	153776107	0.032	0.036	1.4e-01	1.15 (0.95 - 1.38)	1.1e-01	1.29 (0.94 - 1.8)	7.7e-01	0.95 (0.71 - 1.29)
rs7053878_A	153834100	0.035	0.041	1.4e-01	1.14 (0.96 - 1.35)	9.4e-01	0.99 (0.70 - 1.4)	3.1e-01	1.14 (0.88 - 1.47)
G6PD_X_b36_153413623_A	153760429	0.038	0.035	1.7e-01	0.89 (0.75 - 1.05)	7.1e-01	1.06 (0.79 - 1.4)	8.8e-01	0.98 (0.78 - 1.24)
rs762516_A	153764663	0.061	0.050	2.1e-01	0.91 (0.79 - 1.05)	5.0e-01	1.09 (0.85 - 1.4)	5.1e-02	0.79 (0.62 - 1.00)
rs1734792	153341060	0.177	0.161	2.3e-01	0.95 (0.87 - 1.03)	9.0e-01	0.99 (0.85 - 1.2)	2.2e-01	0.92 (0.81 - 1.05)
hG6PD_plus202	153764217	0.033	0.023	2.3e-01	0.89 (0.73 - 1.08)	4.5e-01	1.13 (0.82 - 1.6)	2.3e-01	0.82 (0.60 - 1.13)
G6PD_X_b36_153414758_A	153761564	0.138	0.139	2.7e-01	1.06 (0.96 - 1.16)	3.1e-01	0.90 (0.75 - 1.1)	2.9e-02	1.16 (1.02 - 1.33)
rs5986877_A	153828269	0.938	0.945	2.7e-01	0.92 (0.80 - 1.06)	2.3e-01	1.21 (0.89 - 1.7)	8.5e-01	1.02 (0.82 - 1.28)
rs60030796_A	153836171	0.089	0.083	2.7e-01	0.94 (0.84 - 1.05)	5.2e-01	0.93 (0.75 - 1.2)	4.7e-01	0.94 (0.79 - 1.11)
rs4898389_A	153827637	0.904	0.910	2.8e-01	0.94 (0.84 - 1.05)	7.8e-01	1.03 (0.83 - 1.3)	3.5e-01	1.09 (0.91 - 1.32)
rs2515905_A	153762075	0.061	0.047	3.0e-01	0.93 (0.80 - 1.07)	4.1e-01	1.11 (0.87 - 1.4)	9.3e-02	0.81 (0.64 - 1.03)
rs2071429_A	153760508	0.936	0.942	3.7e-01	0.94 (0.82 - 1.08)	4.3e-01	1.12 (0.84 - 1.5)	5.5e-01	1.07 (0.85 - 1.35)
G6PD_X_b36_153428979_A	153775785	0.521	0.546	4.5e-01	1.03 (0.96 - 1.09)	5.5e-01	0.96 (0.85 - 1.1)	7.1e-01	1.02 (0.92 - 1.12)
rs763737	153278307	0.430	0.433	4.7e-01	1.02 (0.96 - 1.09)	6.5e-01	1.03 (0.91 - 1.2)	6.2e-01	1.02 (0.93 - 1.13)
rs2230036_A	153760953	0.134	0.138	4.8e-01	1.03 (0.94 - 1.13)	1.9e-01	0.88 (0.73 - 1.1)	8.6e-02	1.12 (0.98 - 1.28)
rs762513	153675171	0.186	0.171	5.5e-01	0.97 (0.90 - 1.06)	2.3e-01	0.90 (0.77 - 1.1)	9.2e-01	0.99 (0.88 - 1.13)
G6PD_X_b36_153426256_A	153773062	0.118	0.119	6.4e-01	1.02 (0.93 - 1.13)	8.3e-02	0.83 (0.68 - 1.0)	6.9e-02	1.14 (0.99 - 1.31)
rs766420	153554404	0.474	0.466	7.6e-01	0.99 (0.93 - 1.06)	7.9e-01	0.98 (0.87 - 1.1)	4.4e-01	0.96 (0.87 - 1.06)
G6PD_X_b36_153426354_A	153773160	0.084	0.085	8.8e-01	1.01 (0.90 - 1.13)	8.5e-01	0.98 (0.78 - 1.2)	5.3e-01	1.06 (0.89 - 1.25)
rs61042368_A	153755336	0.140	0.140	9.1e-01	1.01 (0.92 - 1.10)	8.0e-02	0.85 (0.71 - 1.0)	1.8e-01	1.09 (0.96 - 1.24)

Table S7.7: Logistic tests of association – 1 df dominant. Shown here are the results of the 1 d.f. logistic regression association tests under an dominant model. For each SNP is listed: genomic position (hg19/GRCh37), case and control DAF, as well as P values and OR (with 95% CI) for association tests for all severe malaria (ALL), or either SMA or CM subphenotype.

SNP	Position	Control	Case	P - ALL	Estimated OR	P - SMA	Estimated OR	P - CM	Estimated OR
rs12389569_A	153757734	0.179	0.167	7.5e-04	0.79 (0.70 - 0.91)	9.3e-01	1.01 (0.79 - 1.3)	9.0e-03	0.76 (0.62 - 0.93)
rs915941	153626649	0.434	0.395	1.1e-03	0.83 (0.74 - 0.93)	2.5e-01	0.88 (0.71 - 1.1)	8.7e-02	0.86 (0.73 - 1.02)
rs762515_A	153764528	0.343	0.311	3.9e-03	0.84 (0.75 - 0.95)	6.9e-01	1.04 (0.84 - 1.3)	5.5e-03	0.78 (0.66 - 0.93)
rs12393550_A	153758660	0.308	0.277	5.1e-03	0.84 (0.75 - 0.95)	7.3e-01	1.04 (0.83 - 1.3)	6.0e-03	0.77 (0.64 - 0.93)
hG6PD_968_A	153761240	0.066	0.054	9.0e-03	0.77 (0.63 - 0.94)	3.6e-01	1.17 (0.84 - 1.6)	2.1e-03	0.60 (0.43 - 0.83)
rs28470352_A	153753490	0.310	0.280	1.1e-02	0.86 (0.76 - 0.97)	5.5e-01	1.07 (0.86 - 1.3)	2.4e-02	0.82 (0.69 - 0.97)
hG6PD_542_B	153762655	0.012	0.007	1.2e-02	0.54 (0.33 - 0.87)	5.2e-01	1.25 (0.63 - 2.5)	3.2e-02	0.36 (0.15 - 0.92)
rs915942	153626738	0.441	0.486	2.4e-02	1.14 (1.02 - 1.29)	2.0e-01	1.15 (0.93 - 1.4)	1.9e-01	1.12 (0.94 - 1.34)
rs5986990_A	153761628	0.312	0.280	3.1e-02	0.88 (0.78 - 0.99)	3.3e-01	1.11 (0.90 - 1.4)	4.3e-02	0.83 (0.70 - 0.99)
rs2230037_A	153760654	0.352	0.382	5.4e-02	1.12 (1.00 - 1.25)	9.1e-01	0.99 (0.80 - 1.2)	4.2e-02	1.19 (1.01 - 1.41)
rs2515904_A	153762771	0.062	0.048	6.0e-02	0.82 (0.66 - 1.01)	5.8e-01	1.11 (0.77 - 1.6)	8.3e-03	0.62 (0.43 - 0.88)
hG6PD_plus376	153763492	0.320	0.291	7.6e-02	0.90 (0.80 - 1.01)	2.3e-01	1.14 (0.92 - 1.4)	4.6e-02	0.84 (0.70 - 1.00)
hG6PD_plus202	153764217	0.033	0.023	1.0e-01	0.78 (0.58 - 1.05)	6.9e-01	1.11 (0.67 - 1.8)	9.9e-02	0.67 (0.41 - 1.08)
rs4898389_A	153827637	0.904	0.910	1.1e-01	0.81 (0.62 - 1.05)	7.8e-01	1.08 (0.64 - 1.8)	8.8e-01	1.03 (0.67 - 1.59)
rs1734792	153341060	0.177	0.161	1.3e-01	0.90 (0.79 - 1.03)	8.7e-01	1.02 (0.80 - 1.3)	8.9e-02	0.84 (0.69 - 1.03)
rs7879049_A	153829693	0.368	0.385	1.7e-01	1.08 (0.97 - 1.21)	9.5e-01	1.01 (0.81 - 1.2)	1.8e-02	1.23 (1.04 - 1.45)
rs762516_A	153764663	0.061	0.050	1.8e-01	0.87 (0.70 - 1.07)	4.0e-01	1.17 (0.81 - 1.7)	2.3e-02	0.67 (0.47 - 0.95)
G6PD_X_b36_153413623_A	153760429	0.038	0.035	1.9e-01	0.85 (0.66 - 1.08)	4.6e-01	1.17 (0.77 - 1.8)	5.9e-01	0.91 (0.63 - 1.30)
rs7053878_A	153834100	0.035	0.041	1.9e-01	1.18 (0.92 - 1.51)	6.8e-01	1.10 (0.69 - 1.8)	3.1e-01	1.20 (0.84 - 1.72)
rs60030796_A	153836171	0.089	0.083	1.9e-01	0.89 (0.75 - 1.06)	3.7e-01	0.86 (0.62 - 1.2)	7.6e-01	0.96 (0.75 - 1.23)
rs2515905_A	153762075	0.061	0.047	2.1e-01	0.87 (0.70 - 1.08)	4.0e-01	1.17 (0.81 - 1.7)	3.6e-02	0.69 (0.48 - 0.98)
rs2071429_A	153760508	0.936	0.942	2.6e-01	0.83 (0.59 - 1.15)	1.8e-01	1.73 (0.78 - 3.8)	8.7e-01	1.05 (0.61 - 1.80)
rs5986875_A	153762392	0.031	0.038	2.6e-01	1.17 (0.89 - 1.52)	2.6e-01	1.32 (0.82 - 2.1)	3.9e-01	0.82 (0.53 - 1.29)
rs5986877_A	153828269	0.938	0.945	2.8e-01	0.83 (0.60 - 1.16)	6.3e-02	2.39 (0.95 - 6.0)	9.3e-01	0.98 (0.58 - 1.66)
rs762513	153675171	0.186	0.171	3.0e-01	0.93 (0.82 - 1.06)	2.7e-01	0.87 (0.67 - 1.1)	3.4e-01	0.91 (0.75 - 1.11)
rs5986992_A	153776107	0.032	0.036	3.1e-01	1.14 (0.88 - 1.49)	3.4e-01	1.26 (0.79 - 2.0)	5.4e-01	0.88 (0.57 - 1.34)
G6PD_X_b36_153428979_A	153775785	0.521	0.546	4.5e-01	1.05 (0.93 - 1.19)	5.5e-01	0.93 (0.74 - 1.2)	4.3e-01	1.08 (0.90 - 1.30)
G6PD_X_b36_153414758_A	153761564	0.138	0.139	5.4e-01	1.05 (0.90 - 1.21)	2.9e-01	0.86 (0.64 - 1.1)	7.9e-02	1.21 (0.98 - 1.49)
rs2230036_A	153760953	0.134	0.138	6.5e-01	1.03 (0.90 - 1.19)	2.6e-01	0.85 (0.64 - 1.1)	1.6e-01	1.16 (0.94 - 1.42)
G6PD_X_b36_153426256_A	153773062	0.118	0.119	7.0e-01	1.03 (0.89 - 1.20)	9.5e-02	0.77 (0.57 - 1.0)	6.9e-02	1.22 (0.98 - 1.51)
rs766420	153554404	0.474	0.466	8.6e-01	1.01 (0.90 - 1.14)	9.9e-01	1.00 (0.80 - 1.2)	4.7e-01	0.94 (0.79 - 1.12)
G6PD_X_b36_153426354_A	153773160	0.084	0.085	8.7e-01	1.01 (0.85 - 1.20)	9.2e-01	0.98 (0.71 - 1.4)	4.3e-01	1.10 (0.86 - 1.42)
rs61042368_A	153755336	0.140	0.140	8.8e-01	0.99 (0.86 - 1.14)	1.3e-01	0.81 (0.61 - 1.1)	3.0e-01	1.11 (0.91 - 1.36)
rs763737	153278307	0.430	0.433	9.1e-01	0.99 (0.88 - 1.12)	9.2e-01	0.99 (0.80 - 1.2)	7.4e-01	0.97 (0.82 - 1.15)

Table S7.8: Logistic tests of association – 1 df recessive. Shown here are the results of the 1 d.f. logistic regression association tests under an recessive model. For each SNP is listed: genomic position (hg19/GRCh37), case and control DAF, as well as P values and OR (with 95% CI) for association tests for all severe malaria (ALL), or either SMA or CM subphenotype.

SNP	Position	Control	Case	P - ALL	Estimated OR	P - SMA	Estimated OR	P - CM	Estimated OR
rs915942	153626738	0.441	0.486	1.6e-04	1.26 (1.12 - 1.43)	1.2e-01	1.20 (0.95 - 1.5)	7.5e-03	1.28 (1.068 - 1.53)
rs762515_A	153764528	0.343	0.311	1.4e-03	0.80 (0.70 - 0.92)	9.3e-01	1.01 (0.79 - 1.3)	2.2e-04	0.66 (0.532 - 0.82)
rs12393550_A	153758660	0.308	0.277	6.3e-03	0.81 (0.69 - 0.94)	7.6e-01	0.96 (0.72 - 1.3)	3.2e-03	0.69 (0.543 - 0.88)
rs28470352_A	153753490	0.310	0.280	2.2e-02	0.84 (0.73 - 0.98)	7.9e-01	1.04 (0.80 - 1.3)	1.5e-02	0.75 (0.601 - 0.95)
hG6PD_968_A	153761240	0.066	0.054	2.6e-02	0.70 (0.51 - 0.96)	5.3e-01	1.17 (0.71 - 1.9)	4.6e-02	0.58 (0.338 - 0.99)
hG6PD_plus376	153763492	0.320	0.291	3.4e-02	0.86 (0.74 - 0.99)	6.7e-01	1.06 (0.81 - 1.4)	5.5e-03	0.72 (0.576 - 0.91)
rs915941	153626649	0.434	0.395	6.2e-02	0.89 (0.78 - 1.01)	6.1e-01	0.94 (0.74 - 1.2)	3.4e-01	0.91 (0.753 - 1.10)
rs5986992_A	153776107	0.032	0.036	6.7e-02	1.53 (0.97 - 2.41)	2.7e-02	2.21 (1.10 - 4.4)	7.2e-01	1.14 (0.555 - 2.34)
rs5986875_A	153762392	0.031	0.038	7.7e-02	1.51 (0.96 - 2.38)	1.4e-01	1.77 (0.83 - 3.8)	6.5e-01	1.18 (0.573 - 2.42)
rs7879049_A	153829693	0.368	0.385	1.0e-01	1.11 (0.98 - 1.27)	7.4e-01	0.96 (0.75 - 1.2)	2.6e-02	1.24 (1.026 - 1.50)
rs5986990_A	153761628	0.312	0.280	1.0e-01	0.88 (0.76 - 1.02)	5.9e-01	1.08 (0.82 - 1.4)	3.5e-02	0.78 (0.618 - 0.98)
G6PD_X_b36_153414758_A	153761564	0.138	0.139	1.0e-01	1.20 (0.96 - 1.49)	4.5e-01	0.85 (0.54 - 1.3)	2.0e-02	1.44 (1.058 - 1.95)
rs12389569_A	153757734	0.179	0.167	1.2e-01	0.86 (0.72 - 1.04)	6.9e-01	1.07 (0.77 - 1.5)	1.4e-01	0.81 (0.604 - 1.07)
rs763737	153278307	0.430	0.433	1.3e-01	1.10 (0.97 - 1.25)	3.2e-01	1.12 (0.89 - 1.4)	1.8e-01	1.13 (0.943 - 1.36)
rs7053878_A	153834100	0.035	0.041	1.8e-01	1.33 (0.87 - 2.04)	3.6e-01	0.62 (0.22 - 1.7)	4.9e-01	1.25 (0.662 - 2.36)
rs2515904_A	153762771	0.062	0.048	2.4e-01	0.82 (0.58 - 1.14)	8.2e-01	1.07 (0.61 - 1.9)	2.4e-01	0.72 (0.416 - 1.24)
G6PD_X_b36_153413623_A	153760429	0.038	0.035	2.6e-01	0.80 (0.54 - 1.18)	7.7e-01	0.90 (0.44 - 1.8)	6.2e-01	1.14 (0.674 - 1.93)
rs5986877_A	153828269	0.938	0.945	3.5e-01	0.90 (0.73 - 1.12)	5.5e-01	1.14 (0.74 - 1.7)	7.4e-01	1.06 (0.762 - 1.47)
rs2230036_A	153760953	0.134	0.138	3.5e-01	1.10 (0.90 - 1.36)	1.8e-01	0.74 (0.48 - 1.1)	6.6e-02	1.32 (0.982 - 1.77)
rs762516_A	153764663	0.061	0.050	3.9e-01	0.87 (0.62 - 1.20)	7.9e-01	1.08 (0.61 - 1.9)	3.4e-01	0.77 (0.452 - 1.31)
rs2230037_A	153760654	0.352	0.382	4.2e-01	1.06 (0.92 - 1.21)	7.8e-01	0.97 (0.75 - 1.2)	7.1e-01	1.04 (0.852 - 1.27)
rs766420	153554404	0.474	0.466	4.5e-01	0.95 (0.85 - 1.08)	6.1e-01	0.94 (0.75 - 1.2)	4.8e-01	0.94 (0.781 - 1.12)
G6PD_X_b36_153428979_A	153775785	0.521	0.546	5.2e-01	1.04 (0.92 - 1.17)	6.0e-01	0.94 (0.75 - 1.2)	9.3e-01	0.99 (0.829 - 1.19)
rs2071429_A	153760508	0.936	0.942	5.5e-01	0.94 (0.77 - 1.15)	7.5e-01	1.07 (0.72 - 1.6)	4.5e-01	1.13 (0.820 - 1.57)
hG6PD_542_B	153762655	0.012	0.007	5.6e-01	0.80 (0.37 - 1.71)	7.0e-02	2.39 (0.93 - 6.1)	1.8e-01	0.25 (0.033 - 1.89)
rs4898389_A	153827637	0.904	0.910	5.6e-01	0.95 (0.80 - 1.13)	8.1e-01	1.04 (0.75 - 1.4)	2.1e-01	1.19 (0.908 - 1.56)
rs1734792	153341060	0.177	0.161	5.8e-01	0.95 (0.78 - 1.15)	6.2e-01	0.91 (0.64 - 1.3)	7.7e-01	0.96 (0.719 - 1.28)
rs60030796_A	153836171	0.089	0.083	5.9e-01	0.93 (0.72 - 1.21)	8.8e-01	0.96 (0.60 - 1.5)	2.3e-01	0.77 (0.510 - 1.17)
G6PD_X_b36_153426256_A	153773062	0.118	0.119	6.2e-01	1.06 (0.85 - 1.32)	1.4e-01	0.70 (0.44 - 1.1)	1.5e-01	1.26 (0.918 - 1.73)
rs61042368_A	153755336	0.140	0.140	6.4e-01	1.05 (0.86 - 1.28)	7.9e-02	0.68 (0.44 - 1.0)	1.3e-01	1.25 (0.938 - 1.67)
rs2515905_A	153762075	0.061	0.047	6.4e-01	0.92 (0.66 - 1.30)	5.3e-01	1.20 (0.68 - 2.1)	5.5e-01	0.85 (0.496 - 1.45)
hG6PD_plus202	153764217	0.033	0.023	8.2e-01	0.95 (0.60 - 1.50)	2.6e-01	1.50 (0.74 - 3.0)	8.8e-01	0.95 (0.470 - 1.90)
rs762513	153675171	0.186	0.171	9.0e-01	1.01 (0.84 - 1.22)	2.6e-01	0.81 (0.56 - 1.2)	2.6e-01	1.17 (0.893 - 1.52)
G6PD_X_b36_153426354_A	153773160	0.084	0.085	9.3e-01	1.01 (0.77 - 1.33)	7.8e-01	0.93 (0.55 - 1.6)	8.2e-01	1.05 (0.699 - 1.57)

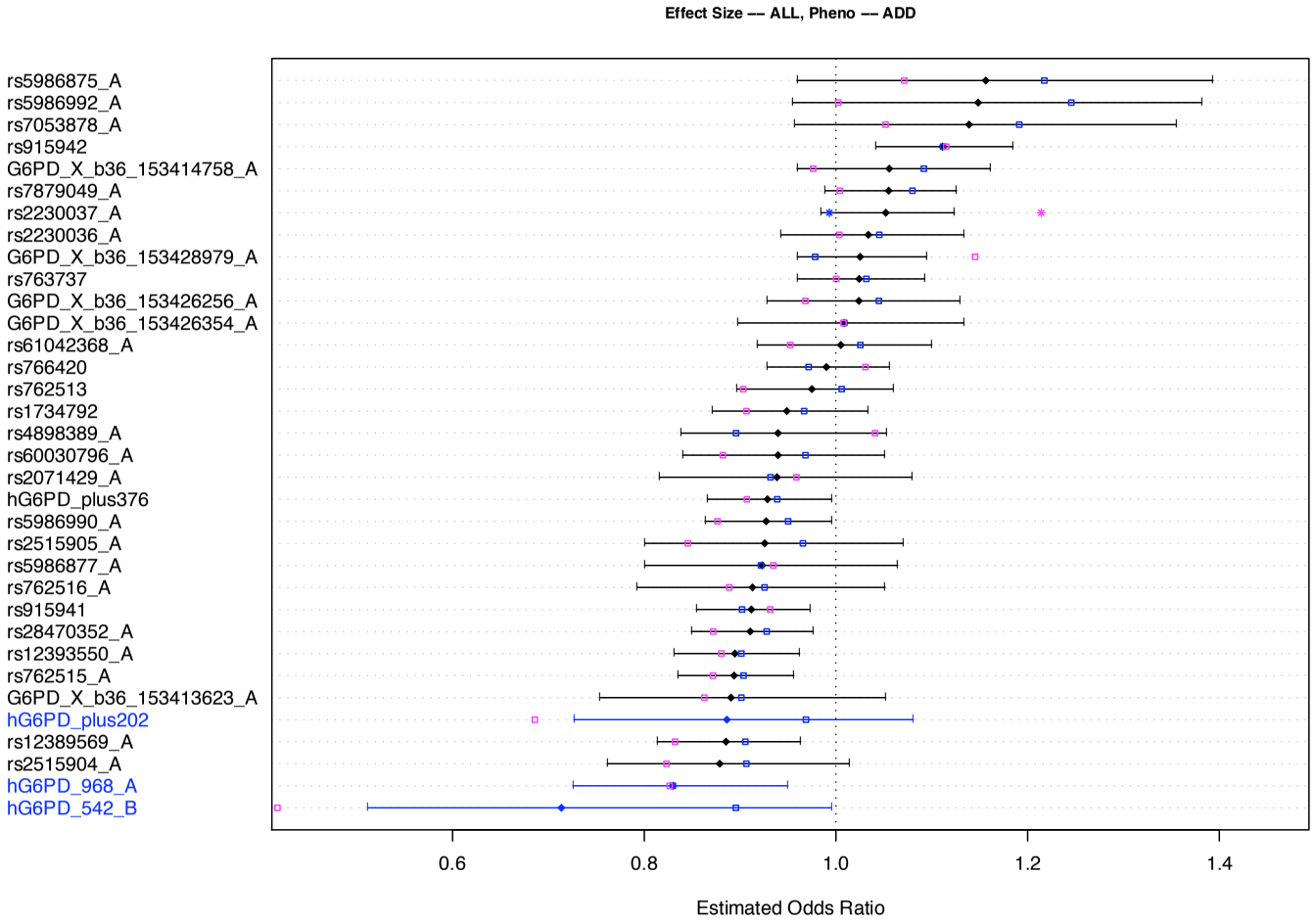


Figure S7.1: Logistic regression association tests – estimated odds ratios (phenotype: ALL, model: additive) The estimated odds ratios for single SNP association tests (ALL) are shown here with 95% CI, as well as estimated OR from analyses using either only males (blue) or females (pink). Significant differences ($P < 0.01$) between male and female estimates are denoted with asterisks. Nonsynonymous SNPs associated with enzyme deficiency are in blue.

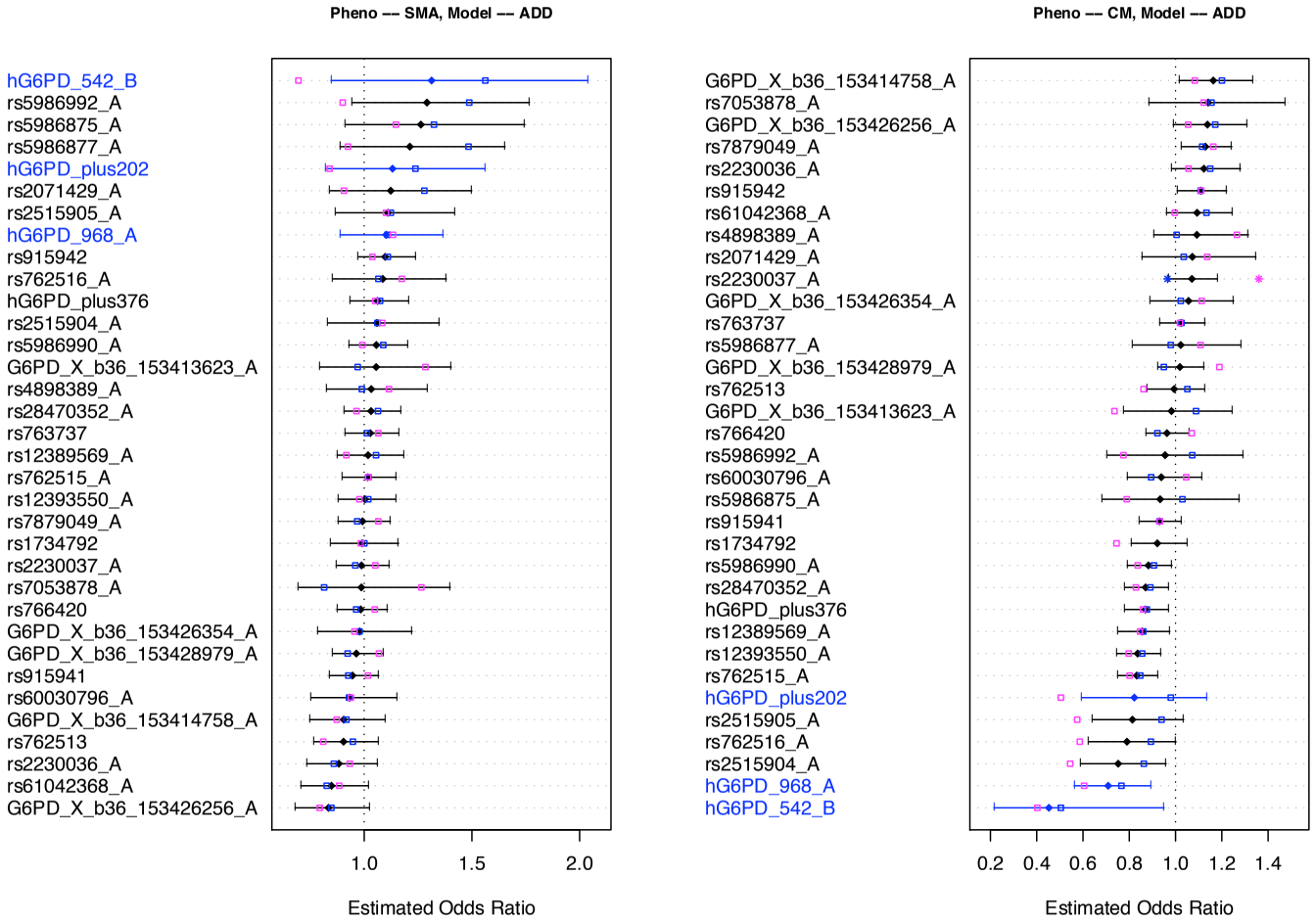


Figure S7.2: Logistic regression association tests – estimated odds ratios (phenotype: SMA, CM, model: additive) The estimated odds ratios for single SNP association tests (SMA, CM) are shown here with 95% CI, as well as for estimated OR from analyses using either only males (blue) or females (pink). Significant differences ($P < 0.01$) between male and female estimates are denoted with asterisks. Nonsynonymous SNPs associated with enzyme deficiency are in blue.

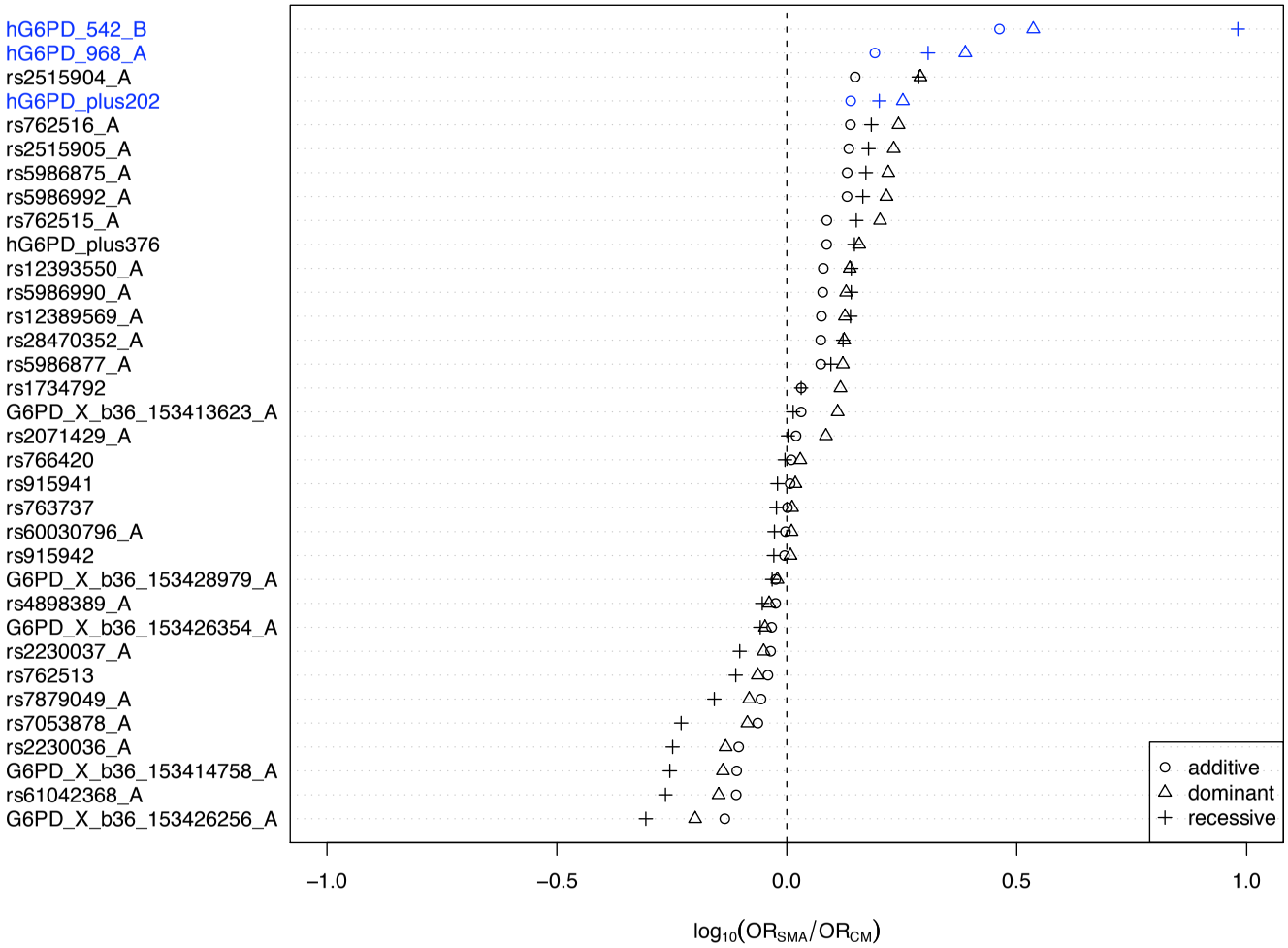


Figure S7.3: Logistic regression association tests – ratio of subphenotype odds ratios. Ratio of estimated OR for SMA versus CM (log10-transformed) under three different genetic models. Nonsynonymous SNPs associated with enzyme deficiency are in blue.

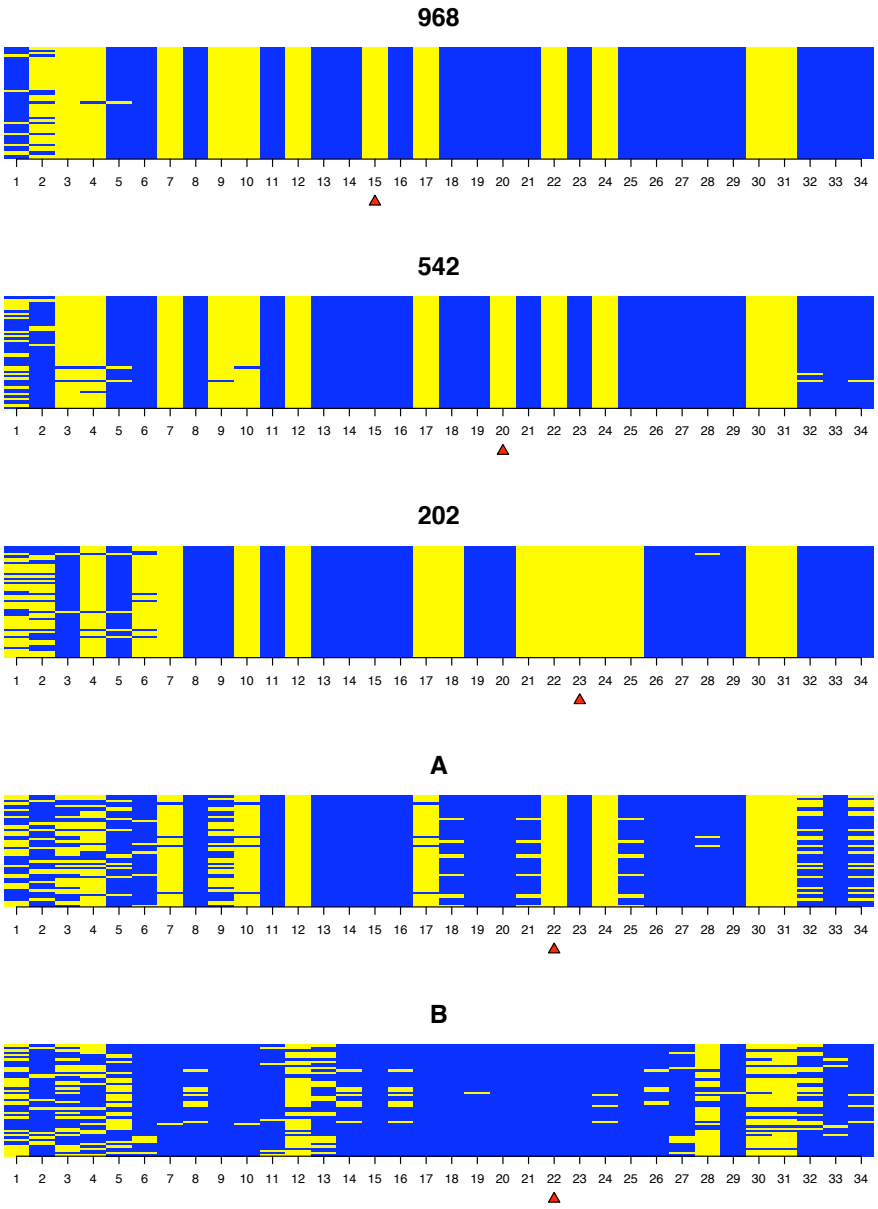


Figure S7.4: Canonical haplotypes at G6PD. A random sampling of fifty haplotypes are shown for the three deficiency (968, 542, 202) and two nondeficiency (A, B) haplotypes to illustrate haplotype structure and phylogeny. Sites are shown across the x-axis and individual haplotypes piled along the y-axis, with ancestral (blue) and derived (yellow) alleles differentiated by color. Also indicated are sites that define a particular haplotype (red arrowhead).

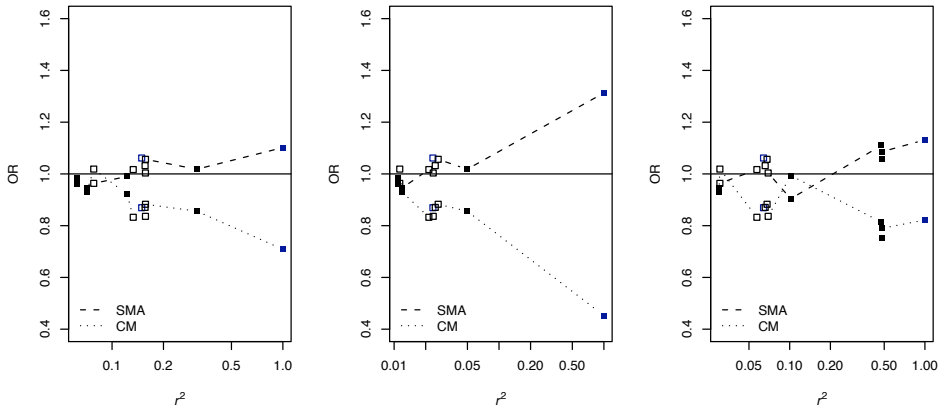


Figure S7.5: Relationship between OR and pairwise r^2 with deficiency SNP. We plot pairwise r^2 (with deficiency alleles) vs. estimated OR for all sites specific to a deficiency allele haplotype (DAH). Sites common to all three DAH (open squares) are shown along with sites unique to each DAH (filled squares), with nonsynonymous variants associated with enzyme deficiency shown in blue (the point furthest to the right, at $r^2 = 1$, is the deficiency allele itself). LD metrics are reviewed in Chapter 2: Supplemental Information.

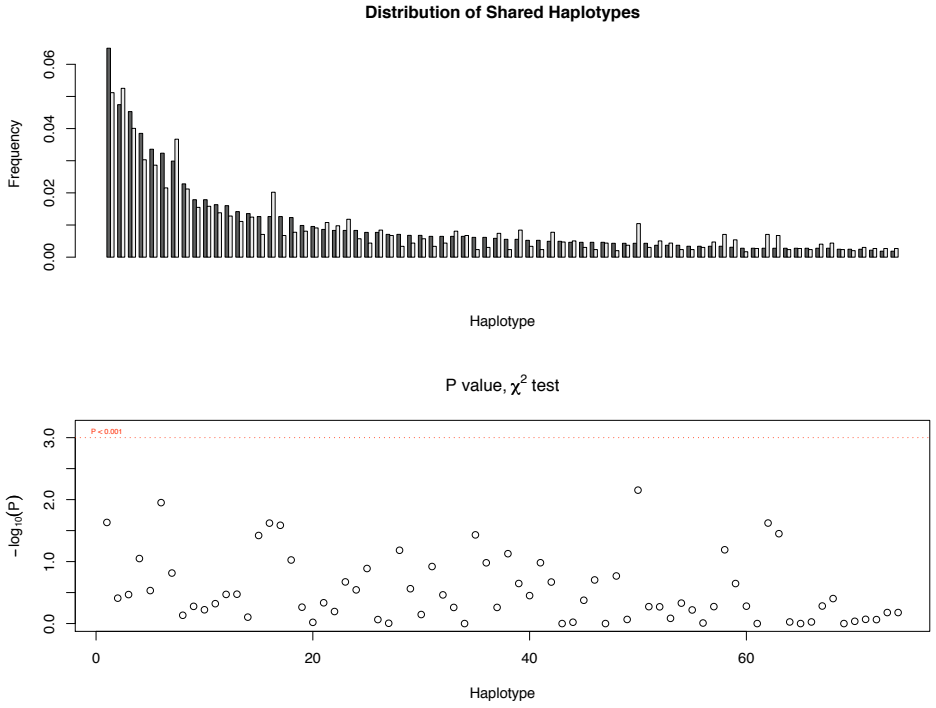


Figure S7.6: Haplotype frequency distribution in males versus females. Given that male haplotypes are known and female haplotypes inferred, this is acts as a check on phasing accuracy by fastPHASE. The frequency distribution of all shared haplotypes between males and females (top) and the P values for corresponding χ^2 tests (bottom) are shown.

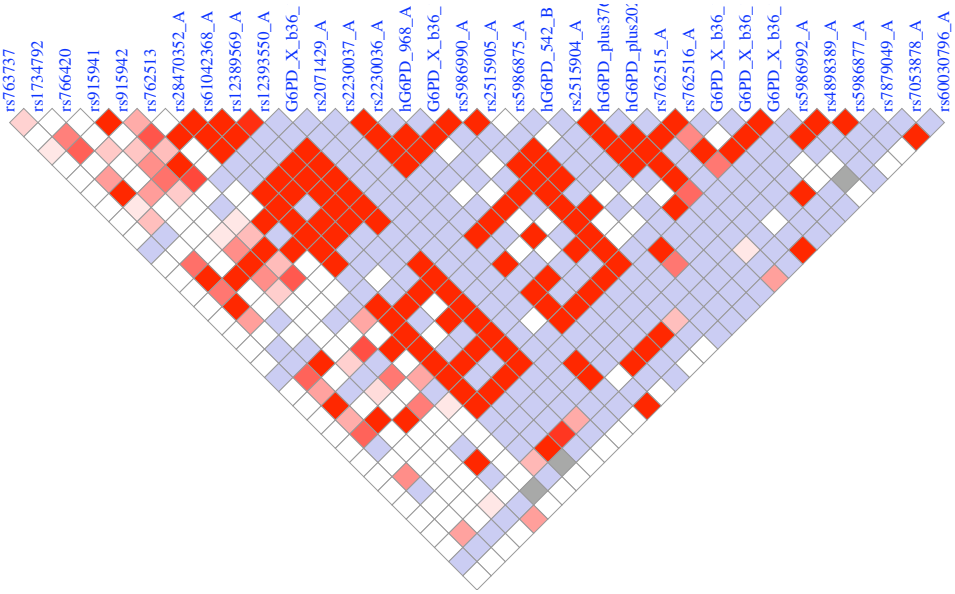


Figure S7.7: LD plot (full): FULA (LD statistic: $\frac{D'}{LOD}$). Pairwise LD matrix is shown in colorimetric form, using the D' and LOD metrics (reviewed in Chapter 2: Supplemental Information). Colors indicate: D' = 1 and LOD $\geq$ 2 (red), D' = 1 and LOD $\leq$ 2 (blue), D' < 1 and LOD $\geq$ 2 (pink), D' < 1 and LOD $\leq$ 2 (white).

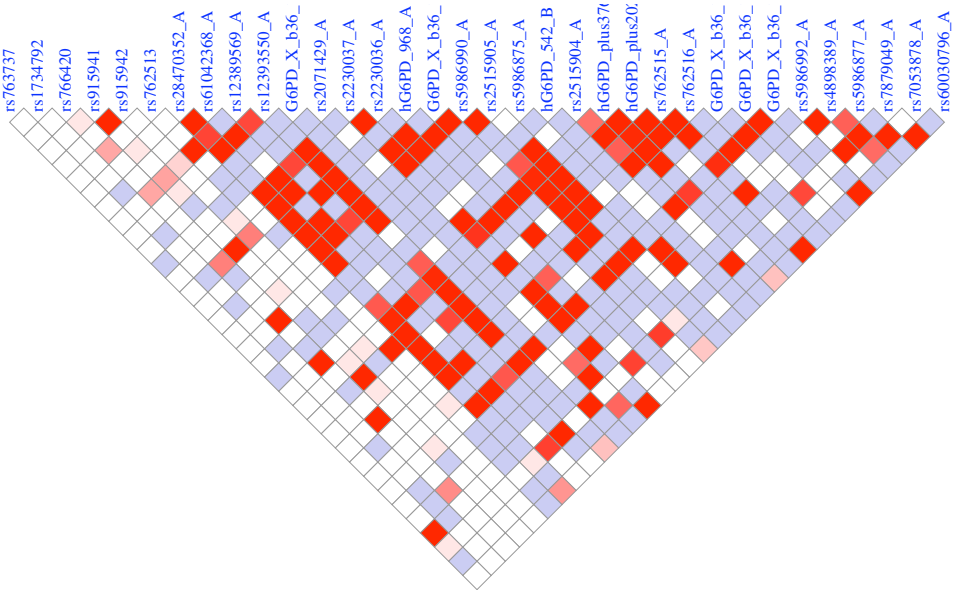


Figure S7.8: LD plot (full): JOLA (LD statistic: $\frac{D'}{LOD}$). Pairwise LD matrix is shown in colorimetric form, using the D' and LOD metrics (reviewed in Chapter 2: Supplemental Information). Colors indicate: D' = 1 and LOD $\geq$ 2 (red), D' = 1 and LOD $\leq$ 2 (blue), D' < 1 and LOD $\geq$ 2 (pink), D' < 1 and LOD $\leq$ 2 (white).

Figure S7.9: LD plot (full): MANDINKA (LD statistic: $\frac{D'}{LOD}$). Pairwise LD matrix is shown in colorimetric form, using the D' and LOD metrics (reviewed in Chapter 2: Supplemental Information). Colors indicate: D' = 1 and LOD $\geq$ 2 (red), D' = 1 and LOD $\leq$ 2 (blue), D' < 1 and LOD $\geq$ 2 (pink), D' < 1 and LOD $\leq$ 2 (white).

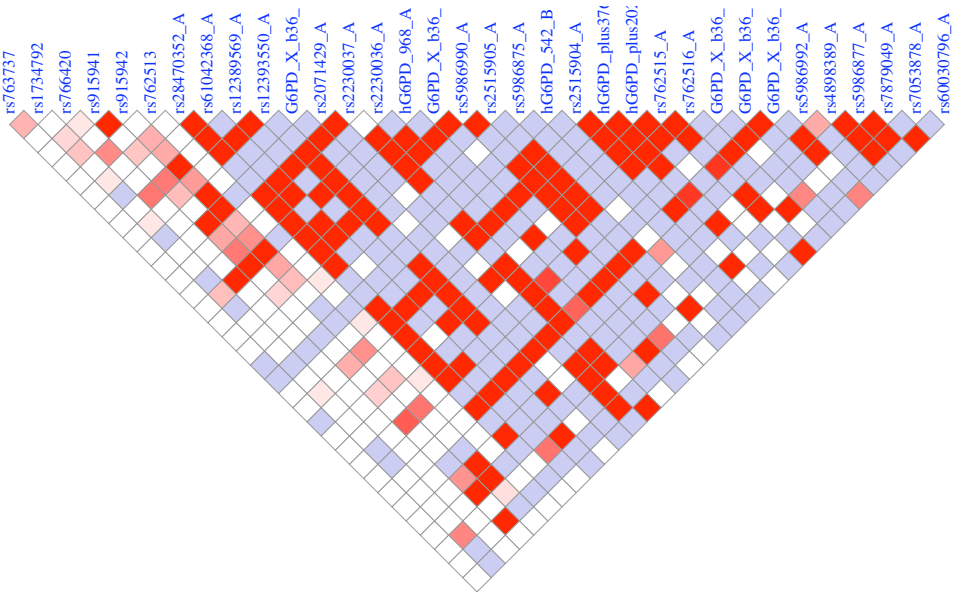


Figure S7.10: LD plot (full): WOLLOF (LD statistic: $\frac{D'}{LOD}$). Pairwise LD matrix is shown in colorimetric form, using the D' and LOD metrics (reviewed in Chapter 2: Supplemental Information). Colors indicate: D' = 1 and LOD $\geq$ 2 (red), D' = 1 and LOD $\leq$ 2 (blue), D' < 1 and LOD $\geq$ 2 (pink), D' < 1 and LOD $\leq$ 2 (white).

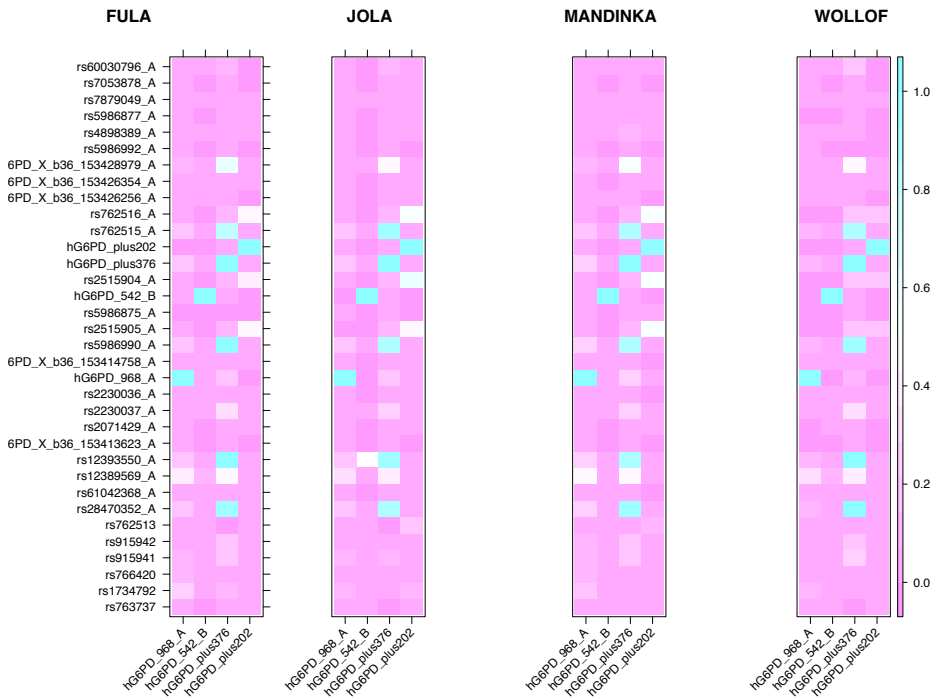


Figure S7.11: LD plot (functional SNPs) – between deficiency-associated SNPs and all other SNPs. Pairwise LD matrices, stratified by ethnicity, show linkage disequilibrium between each deficiency-associated SNP and all other SNPs in colorimetric form (key at right side of plot), using the r^2 metric.

Chapter 8

Conclusion

8.1 Examining the relationship between genetic variation at *G6PD* and malaria – new insights and new tools

The questions addressed in this dissertation all center around the idea of understanding how genetic variation at *G6PD* relates to variation in important molecular and clinical phenotypes, specifically with respect to malaria. While many others have investigated these questions over the past sixty years, my approach differs in the use of new genetic methodologies that enable consideration of questions that were previously intractable. The work presented here reveals important new insights into the genetics and biology of G6PD deficiency and how it relates to malaria, but also contributes new tools that may prove useful for future research.

8.1.1 Defining global genetic variation at G6PD

I began by first considering the characterization of genetic variation at the *G6PD* locus, a prerequisite for all downstream fine-mapping association work. In Chapter 3, new information on population genetic variation at *G6PD* was presented in the form of targeted resequencing and dense genotyping performed at the locus in twelve populations in Africa and southeast Asia, as was a locus-specific database for organizing that data and existing data from the literature. Chapter 4 presented a new methodology for using massively parallel sequencing of pooled, untagged PCR amplicons tiling a locus of interest (e.g. *G6PD* or a GWAS candidate region) to obtain a roughly quantitative estimate of genetic variation in a population. It is hoped that the data and tools presented in these chapters will be of utility to researchers interested in understanding human genetic diversity generally, but also at the *G6PD* locus in particular, and will inform design of both population genetic and association studies at the locus.

8.1.2 Understanding variation in G6PD molecular phenotypes

In the next two chapters, I examined the relationship between genetic variation at *G6PD* and variation in the molecular phenotype of G6PD enzyme activity. In chapter 5, I built upon the genetic variation data collected in Chapter 3 to conduct a fine-mapping association study of G6PD activity in Kenya, confirming a strong effect for one well-described variant (202/A-), and demonstrating a subtle, but still

important effect for other common variants in modulating G6PD activity independently of the 202/A- variant, suggesting that multiple alleles at the locus may influence risk of G6PD deficiency state. Chapter 6 was largely methodological in nature, and concerned the development of a new cytofluorometric tool for typing G6PD deficiency at the level of an individual RBC, one that may be useful as a screening tool for G6PD deficiency in resource-poor settings, especially regions exhibiting allelic heterogeneity. The picture that emerged from these chapters was that a single deficiency variant, here the 202/A- African variant, could exhibit variable biochemical penetrance in different individuals, and that studying the relationship between genetics and molecular phenotypes can reveal useful insights for the design and interpretation of clinical association studies.

8.1.3 Fine-mapping association study of severe malaria in the Gambia

In the final chapter, I considered the association between genetic variation at *G6PD* and susceptibility to severe malaria. Specifically, I asked whether heterogeneity at the level of either functional genetics or differing severe malaria syndromes could explain past conflicting results in the literature. Upon conducting a fine-mapping case-control association study of severe malaria and its constituent clinical subphenotypes in the Gambia, I found that three known deficiency alleles exhibited differing direction of association with respect to two important clinical syndromes, trending towards risk conferred to severe malarial anemia (SMA), but protection with respect to cerebral malaria (CM). The importance of considering allelic heterogeneity was also demonstrated, as pooling of all three deficiency alleles together to conduct association tests of G6PD deficiency state as a whole was able to demonstrate the same trend with respect to CM and SMA, but with stronger statistical significance. These results emphasize the importance of accounting for allelic heterogeneity and differences in clinical presentation in the analysis and interpretation of association studies of severe malaria.

8.2 Dissecting heterogeneity in molecular and clinical phenotypes associated with G6PD deficiency

In seeking to understand the association between G6PD deficiency and certain important clinical phenotypes, the key insight gleaned centers around latent heterogeneity, both at the genetic level, and at the level of individual phenotypes. I discuss this heterogeneity and its implications for the main findings presented in this thesis, as well as for future work, in the sections below.

8.2.1 Exploring variation in G6PD molecular phenotypes

The work presented in Chapter 5 represents the first fine-mapping genetic association study of G6PD enzyme activity. It was a study uniquely powered to address the possibility that other genetic modifiers of G6PD activity existed at the locus, and that such genetic heterogeneity might explain why association studies of severe malaria involving the common African variant 202/A- have yielded different results in different populations [138, 39, 143].

The agnostic approach I took differs from that of previous investigators, whose assumptions about the functional determinism of coding sequence variation implicitly discounted the role that noncoding variation might play. Such work generally involved population screening of biochemical G6PD deficiency, followed by cDNA sequencing of affected individuals [205, 100]. Apart from missing information on noncoding variation, this approach was also insensitive to the impact of subtler functional variants. Such polymorphisms include the SNPs identified in the present study that tagged the H+ and H- haplotypes and doubled the risk to G6PD deficiency state in 202CT heterozygote females.

At least two other large-scale studies have examined the relationship between genotype at 202/A- and G6PD enzyme activity. The first study, conducted by May et al. in Nigeria [242], genotyped both the 202/A- and the 376/A+ polymorphisms, and obtained enzyme activity measurements in 606 individuals (421 children, 185 adults). The second study, conducted by Johnson et al. in Uganda [243], obtained genotypes at 202/A- only, along with G6PD activity data for a total of 600 children. In both studies, there was a strong association between genotype at 202/A- and G6PD activity, but additionally, both studies noted the difficulty of categorizing female heterozygotes, as their enzyme activity levels overlapped significantly with wild-type. This observation has important public health relevance with respect to assessing the hemolytic risk of heterozygote females and their offspring, and was a key motivator for the work undertaken here.

The availability of dense genotyping data at thirty polymorphic sites, coupled with three-fold greater sample size allowed the present study to uncover cryptic heterogeneity in the female heterozygote population and identify a subgroup of 202CT heterozygotes that were at increased risk for G6PD deficiency state by virtue of possessing a particular haplotype signature. Interestingly, the 376/A+ SNP that tags the H- haplotype which confers this added risk was also found to confer lower enzyme activity to 202CT heterozygotes in the May et al. study, though the difference was not statistically significant.

In genotyping thirty variants across the locus, the present study was able to not only identify a subgroup of female heterozygotes at increased hemolytic risk, but also implicate a role for common, and potentially noncoding, variation in modulating the phenotype of what was thought to be a Mendelian trait. Multi-site haplotype analyses proved critical in elucidating the role of these variants, demonstrating that two common haplotype motifs were associated with changes in G6PD enzyme activity independent of 202/A-. Additionally, the absence of any uniquely strong signal of association (c.f. 202/A-) across the twelve sites that defined the H+ and H- haplotypes suggested that the causal variant was not among these SNPs, and that it might not even be at the *G6PD* locus itself, as two of these tagging SNPs were hundreds of kilobases away from the gene. One possibility is that there could be multiple variants acting in concert to affect protein expression and/or function, for example the 376/A+ nonsynonymous variant

and the G6PD_X_b36_153428979 5' UTR splice site variant. Another possibility is that interactions between RBC may affect gross activity in mosaic females. Thus, in a 202CT female, whose RBC consist of a mixture of 202C (wild-type) and 202T (deficient) RBC, the oxidant species generated in 202T cells could synergistically impact 202C cells co-expressing the H- haplotype by exacerbating the mild deficiency already present in those cells.

Finally, the simpler question relevant to interpreting past association studies of severe malaria centers around whether there is evidence of a causal role for 202/A- with respect to both molecular and clinical phenotypes. My work demonstrated a strong association, that while not unexpected, was nonetheless strongly suggestive of a causal role with respect to enzyme deficiency, as the 202/A- allele, among SNPs covering several hundred kilobases, exhibited a strength of association with G6PD activity that other SNPs only mirror by virtue of strong linkage disequilibrium with it. Furthermore, the same 202T allele has been reported to be sufficient to cause class 3 enzyme deficiency on its own in a Japanese family where the mutation occurred by chance on a different, non-African genetic background [224]. Altogether, then, compelling evidence exists to support a causal role for the 202/A- allele with respect to G6PD enzyme deficiency, though it is still possible that the trait itself is only indirectly associated with the variant(s) that are causal with respect to clinical phenotypes.

8.2.2 Decomposing variation in G6PD clinical phenotypes

The results presented in Chapter 7 were likewise reliant on dense genotyping at the *G6PD* locus and large sample size. These methodological advantages enabled detection of previously unrecognized heterogeneity at the level of clinical phenotypes, and reinforced the importance of accounting for allelic heterogeneity at *G6PD* in west Africa, as effects that were subtle when considering any single deficiency allele alone were made statistically robust when pooling all three SNPs together and considering G6PD deficiency as a trait. Additionally, as with the biochemical phenotype work discussed above, the utility of considering haplotype structure was demonstrated, as each deficiency allele was found to exhibit more extreme effect sizes for association with both clinical subphenotypes relative to other alleles specific to deficiency allele haplotypes.

8.2.2.1 Association signal resolution and the malaria protection hypothesis

The purpose of the Gambian fine-mapping study was to critically evaluate the malaria protection hypothesis and determine if the locus *G6PD* locus generally, or deficiency alleles specifically, were associated with protection from severe malaria. In contrast to the biochemical phenotype study in Chapter 5, or fine-mapping work done for the HbS polymorphism at the β-globin locus [174], I was unable to identify any single variant that stood out with a uniquely strong association signal. Indeed, variants several hundred kilobases away from the gene exhibited modest signals of association of similar magnitude to those found at the gene itself, and no signal met the genome-wide association study (GWAS) threshold used in the same Gambia cohort by Jallow et al. of $P < 10^{-4}$ [174]. Several possible explanations for

this phenomenon exist, and these are discussed below.

One possible explanation that has been previously proffered by Jallow et al. [174] and Teo et al. [173] concerns the structure of linkage disequilibrium (LD) in the Gambia. Here, as in other African populations, the strength of LD between neighboring loci falls off with distance more rapidly than in non-African populations, as was observed by the HapMap project [161]. The practical implication of this population genetic phenomenon is that association studies relying on a tag-SNP approach (e.g. GWAS), where the causal variant is unlikely to be typed directly, and is instead indirectly inferred by the presence of an association signal at a neigboring SNP with which it is correlated, will require greater SNP coverage density to achieve similar study power in African populations. While there is strong evidence implicating the functional role of G6PD deficiency alleles with respect to biochemical phenotypes, it is still nonetheless possible that these alleles are not directly responsible for clinical protection. As G6PD is located on a gene-dense region of chromosome X (Xq28), it is plausible that some other gene might actually harbor the true causal variant that mediates protection from malaria. Indeed, the ancestral haplotype analysis presented here suggested a common genetic background for all three deficiency alleles, so there may be an older and physically distant causal mutation that exhibits differential LD with respect to the deficiency alleles. If this pattern of LD varied between different African populations, this might also explain the geographic heterogeneity of past association studies.

Another possible explanation is that a combination of functional allelic heterogeneity and small effect sizes limits the signal resolution in a fine-mapping experiment at the locus. In other words, even if the deficiency SNPs represented true causal variants of modest effect, their shared ancestral history creates overlapping patterns of LD with neighboring SNPs such that no single variant will uniquely stand out. Another consequence of allelic heterogeneity is that the single SNP association tests conducted lose some power, in that the control population consists of a proportion of individuals with the other deficiency alleles not being considered in a specific test. In order to address this latter issue, and achieve greater statistical power, association tests were conducted that considered G6PD deficiency as a trait, and thus more directly addressed the malaria protection hypothesis.

8.2.2.2 Rationalizing differential association with clinical subphenotypes

In examining association with major clinical subphenotypes of severe malaria, a differential direction of association was observed, with protection conferred against CM but risk with respect to SMA. This paradoxical result, where a single gene was involved both in protection from and susceptibility to different manifestations of the same disease, contrasts with sickle trait, which was found to be strongly protective against both SMA and CM. These differences may reflect pathophysiological distinctions in how each trait protects against malaria.

Protection conferred by sickle trait The relationship between sickle trait (HbAS) and malaria has been well-studied for over a half century, and in contrast to G6PD deficiency, the evidence for protection obtained by clinical epidemiology studies has been unequivocal, with protection seemingly conferred against all forms of disease, mild and severe, SMA and CM [132, 133, 134, 138, 139, 18, 140]. Additionally, there is strong experimental and clinical evidence suggesting that the AS red cell is inhospitable for parasite growth [145, 146] and more likely to be cleared at a higher rate by splenic macrophages [147]. Later work has added observations that might further explain sickle protection from severe malaria, including the finding of reduced rosetting [282], impaired cytoadherence [155], and greater tolerance to cell-free hemoglobin generated during red cell lysis [87], but in contrast to G6PD deficiency, where both *in vivo* parasite density studies [135, 136, 19] and *in vitro* experimental work [148, 149, 150] have yielded conflicting results, the sum of all evidence for HbAS suggests that the level at which protection is conferred involves an earlier stage in the progression of disease pathogenesis. It is because sickle trait exerts protection early on that we see an even stronger depletion of HbAS alleles in those patients who progress to the most severe manifestations of the disease, which is reflected in the extremely strong protective effect observed for all forms of severe malaria in association studies.

Risk conferred by G6PD deficiency to SMA As G6PD deficiency is known to predispose individuals to episodes of acute hemolysis in the context of certain infections, one simple explanation for risk conferred to SMA is that the oxidant stress generated during the immune response to malaria infection is especially injurious to G6PD-deficient (and thus glutathione-deficient) RBC. Given that severe malarial anemia is thought to involve oxidative damage to and more rapid senescence of uninfected RBC, and that the 202/A- enzyme variant is thought to degrade rapidly as RBC age, this explanation is quite attractive. In fact, variant enzyme stability may be related to RBC survival via a synergistic feedforward mechanism where G6PD deficiency predisposes to oxidant damage in RBC, which then leads to faster degradation of any residual active G6PD protein, and thus hastens senescence and destruction by splenic macrophages. Consistent with this idea is the observation that hallmarks of senescence, including enhanced binding of autologous IgG and complement C3, as well as increased extracellular exposure of phosphatidylserine groups on the cell membrane, are increased in oxidatively-stressed G6PD deficient RBC [150, 75].

An alternate, and even simpler explanation invokes SMA case classification. The hemoglobin concentration (Hb) that defines SMA according to the WHO is 5 g/dL, a threshold that might be easier for a G6PD-deficient child with malaria to meet if infection is superimposed upon a mild chronic anemia, which itself could be the result of not only oxidative stress in the red cell, but also differences in other factors that predispose to anemia, e.g. iron metabolism or hookworm susceptibility [28].

Finally, it is also possible that erythropoiesis may be diminished in G6PD deficient individuals, perhaps by diminishing production of erythroid precursor responsiveness to erythropoietin, though one might

intuitively expect the opposite to be true as a compensatory mechanism to mitigate chronic subclinical hemolysis in such individuals. Interestingly, experiments done using mouse embryonic stem cells wholly-deficient in G6PD activity demonstrates that G6PD activity is indispensable in order for erythroid precursors to avoid apoptosis secondary to oxidative damage generated by newly formed hemoglobin [283].

Protection conferred by G6PD deficiency against CM The protection conferred against CM is somewhat perplexing when one considers the biological function of G6PD throughout the body. In fact, there are many reasons why one might assume the opposite to be true, namely that G6PD deficiency should confer risk to CM. For example, G6PD activity and NADPH are necessary for the activities of nitric oxide synthase (NOS) and heme oxygenase, two enzymes whose activities are thought to be important in limiting the severity of malarial illness. For example, heme oxygenase activity is known to protect against CM in mice via production of carbon monoxide which limits the oxidation of cell-free hemoglobin and the release of the pro-oxidant heme moiety [87]. Likewise, low nitric oxide bioavailability has been linked to CM pathogenesis in rodents and humans [284, 285], and increased plasma levels of asymmetric dimethylarginine (ADMA), an endogenous inhibitor of NOS, were recently found to be associated with severe malaria in Indonesian adults [286].

In spite of the aforementioned reasons why G6PD deficiency might serve to exacerbate CM, there has been a consistent protective effect observed in all three association studies conducted [138, 39, 19], in three different populations, using three different G6PD typing methods (see Chapter 7 for a full discussion). While several mechanisms could plausibly mediate this effect given the role of G6PD in fundamental physiological processes, I consider two strong possibilities below, one at the level of the deficient red cell, and the other in the deficient phagocyte.

In the G6PD deficient red cell, limited bioavailability of glutathione might impact the ability of the parasite to properly export PfEMP1 to the cell surface, which would not only impact the ability of the parasite to evade the host immune system via sequestration, but also limit the ability of the parasite to adhere to cerebral microvascular endothelium and form rosettes with uninfected RBC. A more rapid immune response would limit parasite density and the accumulated pathophysiological effects of a persistent infection [10], while weaker adhesion properties for infected RBC (iRBC) would serve to diminish the degree of endothelial and coagulation pathway activation and the immunopathogenesis thought to be contributory to CM [154].

In the same vein, the immunopathogenesis of CM might also be limited by impaired G6PD function in phagocytic cells, principally macrophages. In order to lyse iRBC and free merozoites, macrophages generate reactive oxygen species (ROS) such as superoxide and peroxynitrite via NADPH oxidase and

inducible NOS (iNOS), respectively. With less NADPH available, less ROS will be produced, and fewer off-target effects that contribute to systemic immune activation will result. While a severely impaired respiratory burst is detrimental to clearing infection [84], and poor systemic bioavailability of NO limits vascular reactivity [80], it is possible that a tempered respiratory burst might actually be beneficial in limiting endothelial activation, platelet and white cell recruitment, fibrin deposition, etc. Indeed, evidence from cardiovascular research suggests that deficient G6PD activity and limited NADPH oxidase activity in macrophages might be linked to decreased atherosclerotic plaque formation and decreased risk for cardiovascular disease [85, 107, 108].

8.2.2.3 The absence of sex-specific differences in association

While previous association studies of severe malaria have claimed sex-specific protective effects, the study presented here, which is the first to conduct formal hypothesis tests for differences in allelic odds ratios between males and females, found no significant differences for any of the deficiency alleles tested. Additionally, as detailed in Chapter 7, the claims of previous investigators such as Bienzle et al. [136] and Guindo et al. [19] regarding female- and male-specific protection, respectively, must be viewed cautiously in light of specific methodological issues and incongruities found in each study. Furthermore, undue focus on sex specificity leads to gender-stratified association testing which is inherently half as powerful in addressing the main question relating to the malaria protection hypothesis, namely whether an association between G6PD deficiency and severe malaria exists. Indeed, Clayton recommends that sex-stratified testing be avoided on the X chromosome, provided that males and females do not differ in mean allele frequency and that a genetic model where female heterozygotes are treated as intermediate between both genotypic extremes is plausible [189]. Here, both conditions are met, as male and female allele frequencies are similar for all deficiency SNPs, and the biochemical phenotype data presented in Chapter 5 validates the assumptions of our genetic model, where female heterozygotes demonstrate G6PD enzyme activity intermediate between wild-type and deficient individuals. Finally, it is worth mentioning that the false positive rate of sex-specificity claims in association studies is extraordinarily high as documented in a meta-analysis by Patsopoulos et al. [144].

8.3 Future directions – toward a better understanding of the functional genetics of G6PD deficiency

In the next ten years, there will be a considerable increase in the amount of genetic variation data available throughout the genome, and as the cost of sequencing continues to drop, it will likely become cost-effective to obtain whole genome sequence for all individuals in a genome-wide association study. Similarly, more and more collaborative work will be done in the form of disease- and task-specific consortia, such as MalariaGEN and the Worldwide Antimalarial Resistance Network (WWARN). Taken together, these developments will generate massive amounts of data, and have the potential to solve outstanding scientific puzzles, such as the one considered in this dissertation. Perhaps with the combination

of whole-genome sequence data and the large sample sizes possible in consortia, it could even become trivial to identify novel loci associated with protection from disease.

Even assuming such an optimistic scenario, what would remain outstanding is the missing link between genetics and clinical outcomes, namely the molecular mechanisms that underlie disease association, and how such knowledge can be translated into clinical application. Sickle cell trait provides an informative case-in-point. While the malaria-protective effect was identified nearly sixty years ago, and studies using a variety of experimental and natural systems suggest that it is some combination of decreased parasite viability and virulence, as well as host adaptation that is responsible for protection, it remains unclear today what the relative contribution of these different molecular mechanisms are, and none have been exploited to inform antimalarial drug design or the development of supportive therapy.

For G6PD deficiency, the problem may be even more challenging, for all the reasons that have been discussed thus far, including allelic heterogeneity, differential biochemical penetrance, as well as conflicting results from both clinical association studies and *in vitro* mechanistic work.

8.3.1 New avenues in exploring molecular phenotypes

Studying G6PD deficiency is of interest to malaria researchers not only because of the malaria protection hypothesis, but also because of important drug sensitivity reactions that occur with administration of certain antimalarials, most notably primaquine. As mass primaquine administration has been discussed recently as a means to concurrently reduce *P. vivax* and *P. falciparum* transmission as part of malaria elimination campaigns, the high prevalence of G6PD deficiency in many malaria-endemic regions becomes an important issue of global public health. As much remains unclear about who is susceptible to primaquine-induced hemolysis, further research will be necessary in different populations to determine the risk conferred by varying degrees of enzyme deficiency, and whether any informative genetic modifiers of that risk can be identified that would increase the specificity of genetic testing. It would also be worth testing the positive predictive value of various biochemical assays as they may be more cost-effective in a resource-poor setting where multiple deficiency alleles coexist.

Repeating the experiment A necessary first step in understanding the work presented in Chapter 5 will be to simply repeat the experiment, in a different part of Africa, with different environmental exposures and genetic backgrounds, and importantly, one where other G6PD deficiency alleles are present. An ideal first location would be the Gambia, where three deficiency alleles of varying biochemical penetrance are present at greater than 1% frequency, and where three of the large case-control studies of severe malaria at *G6PD* have now been conducted. Together with the Kenyan data, a pooled meta-analysis would be even better powered to detect subtle effects that impact G6PD activity. It would be especially intriguing to repeat this study in southeast Asia or Oceania, regions with even greater known

functional allelic heterogeneity, to see if G6PD activity there is similarly mediated by the independent effects of multiple variants, both coding and noncoding, at the locus.

Refining our understanding of functional genetics In the Kenyan study, it is likely that the causal alleles underlying the minor determinants of G6PD were not part of the present study, as described earlier. Additionally, in our work, as well as in other studies conducted in east Africa [243], some phenotypically G6PD deficient individuals were wild-type at the 202/A- SNP. To identify sources of outstanding unexplained variation in our data, one useful future experiment will be simply to add more complete genetic data to the Kenyan enzyme activity data set. One simple, cost-effective way to better resolve signals near the locus itself would be via imputation of untyped SNPs using whole genome sequence data for the HapMap Luhya as a reference panel. Whole-genome genotyping or sequencing data would be even more useful, as G6PD activity is likely to be determined by the independent and additive effects of multiple loci given the number of biochemical pathways (e.g. nitric oxide generation, reduction of glutathione) and known endogenous regulators (e.g. p53 [76], DHEA [75]) with which it interacts. Also, assuming that G6PD enzyme activity is related to thermal instability in the aging erythrocyte, it may be that factors that affect RBC lifespan, such as oxidant stress or reticuloendothelial system function, would impact gross activity, and these could only be fully elucidated via a whole-genome approach.

It will also be useful to conduct future work that considers other molecular phenotypes related to G6PD deficiency. For example, the cytofluorometric assay presented in chapter 6 decomposes gross, population-level G6PD activity in hemolysate to that present in individual RBC, a gain in information content that might provide valuable insight relevant to RBC senescence and variable enzyme activity in female heterozygotes. Other metabolites will also be important to examine, especially glutathione, given that it is a critical driver of demand for G6PD activity in the RBC and is easily assayed using dichlorofluorescein [287]. Additionally, studying G6PD activity in nucleated cells will be useful, especially hepatocytes and macrophages, given their roles in neonatal jaundice and malaria. While it has been reported that the G6PD activity of nucleated cells in a 202/A- individual is roughly similar to that of wild-type [97, 288], these studies have generally been conducted at steady state (i.e. basal activity state) and in only a few individuals. Larger studies examining G6PD activity in hepatocytes and macrophages under conditions of stress, e.g. conditions that mimic the bilirubin conjugation load in a newborn, or the respiratory burst stimulated by IFN-γ, will be useful in order to understand the impact of G6PD enzyme deficiency in important nucleated cells that contribute to common pathophysiology.

Predicting risk of hemolysis The practical importance to modeling the genetic basis for variation in molecular phenotypes such as G6PD activity has to do with assessing individual and population risk of hemolytic events given environmental and infectious agent pressure. For example, if primaquine is to be mass administered in a local malaria elimination campaign, it is important to be able to assess and closely monitor those with the greatest likelihood of developing drug-related complications. One

simple genetic approach would be to conduct a prospective cohort study looking at the ability of one or more SNPs to predict the likelihood of hemolysis. An ideal genetic assay would need to be sensitive enough to predict most hemolytic events, but specific enough to avoid undue burden on health systems infrastructure. A complementary laboratory approach would be to develop better, more productive in vitro assays of primaquine sensitivity in variant RBC, one that identifies and assays the key biomolecular harbingers of hemolysis. Finally, animal models using a humanized liver might also be informative as they would better mimic the temporal course of oxidant insult from certain drug metabolites in a living host.

8.3.2 Further investigations into clinical phenotypes

Finally, solving the question of the molecular mechanism of protection may require new technologies and new ways of thinking about G6PD. As the enzyme is found throughout the body, and its activity is important for numerous metabolic pathways, it will be important for researchers to consider mechanisms of protection that do not involve the red blood cell. While the finding that G6PD deficiency is differentially associated with SMA and CM suggests that the molecular mechanism(s) underlying protection conferred by G6PD deficiency might differ from that of sickle cell trait, 'SMA' and 'CM' are merely useful clinical categories that, while important, belie the clinical heterogeneity and pathophysiological complexity of severe malaria.

Replication studies As with the molecular phenotype work in Chapter 5, the first important step in contextualizing the Gambian association study will be to conduct similar studies in other populations, to see if any consistent patterns emerge. While we know from examination of past studies of severe malaria that protection from CM has been consistent in association studies of G6PD deficiency, it is nonetheless critical to replicate this observation using a standardized study design including, when possible, cord blood controls, and precise case definitions that vary minimally across study sites. A large multi-centre replication study also allows for the statistical power to address risk conferred to SMA. As differences in environmental exposures, especially malaria endemicity, are known to impact the presentation of severe malaria, it is possible that there will be between-study heterogeneity in effect size and direction, though one would expect a strong and authentic effect to be relatively consistent for a given clinical subphenotype. As a protective effect has yet to be demonstrated in east Africa, that would be an ideal location in which to do the first replication study. Finally, a gold standard study to causally implicate G6PD deficiency in protection from severe malaria would involve assessment of the effect of non-African deficiency alleles on a non-African genetic background, in an area with a distinct set of environmental exposures (e.g. southeast Asia).

Unraveling mechanism In seeking to understand the mechanistic aspects of protection, genetic variation data of more depth and breadth will be necessary, and whole genome sequencing data will likely be best suited for this, as whole genome genotyping chips used in GWAS have insufficiently dense coverage

near *G6PD* to efficiently tag any putative causal variant [174], especially given the low levels of LD found in African populations. In my study in the Gambia, a definitive causal role for G6PD deficiency alleles was not firmly established. Only with more dense genotyping data at and around the locus will the association signals present be clarified. An additional benefit to whole genome variation data will be the opportunity to investigate epistatic interactions between G6PD and other essential genes with which it is known to interact, for example cell cycle regulator p53 [76] or nitric oxide synthase [83].

Building upon the work presented in Chapter 5, it would also be informative to couple G6PD enzyme activity data with genetic data to predict clinical outcomes. Additionally, markers of poor clinical prognosis such as lactate or bilirubin, or of nitric oxide bioavailability such as ADMA [286] or nitrite [285], would add value to understanding disease pathogenesis of severe malaria and CM as it relates to G6PD. As discussed earlier, CM might also be associated with an overactive respiratory burst, and it would be useful and facile to assess macrophage respiratory burst function in a G6PD deficient subject using a standard clinical laboratory test such as the nitro-blue tetrazolium test [84]. With respect to SMA, reticulocyte measurements made during and after an infection could be compared between G6PD deficient and normal individuals to ascertain the efficiency of erythropoietic drive, and markers of activated macrophage function such as neopterin [24] could be useful in understanding the impact of hypersplenism to SMA in G6PD deficiency subjects. What will also be important in the future is merging detailed clinical data from quantitative biomarkers with proteomics or metabolomics data to provide a systems approach to understanding disease pathogenesis. If indeed the ubiquity and housekeeping functions of G6PD make it relevant to many pathways involved in severe malaria, then such systems approaches may prove to be the most informative of all.

Experimental work that would be most useful in understanding the altered pathogenesis of severe malaria in G6PD deficient individuals would include using a mouse model of experimental CM (ECM) using *P. chabaudi chabaudi*, which shares many common features with human CM. Although a G6PD-null mouse model is embryonic lethal, an allelic exchange strategy where a hypomorphic human G6PD allele (e.g. 202/A-) replaces the native mouse gene would be possible, and has already demonstrated utility in a mouse ECM model expressing human sickle cell trait [87]. A rodent model would allow for invasive experimentation that cannot ethically be performed in humans, for example biopsies of liver to assess hepatocyte function and sporozoite replication efficiency, spleen to assess macrophage function, and bone marrow to assess hematopoiesis. Another important line of experimental work would involve assessing the PfEMP1-mediated virulence phenotypes of cytoadherence and rosetting in vitro using parasite field isolates. As such adhesive phenotypes have been shown to be diminished in other oxidatively-stressed RBC, e.g. HbS [155] and HbC [154], it will be useful to see if similar observations hold true for G6PD deficient RBC.

8.3.3 Final thoughts

Investigations of natural variation in glucose-6-phosphate dehydrogenase will likely continue to yield interesting insights into physiologic and disease states. What is clear from the data presented in this thesis is that no simple explanations exist for the link between genetics, molecular function, and clinical sequelae. Further insights must be gained, from more detailed clinical and laboratory studies conducted in different parts of the world, in order to better understand and predict the benefits and risks associated with the world's most common enzyme disorder.

Bibliography

[1] Hay, S. I., Okiro, E. A., Gething, P. W., Patil, A. P., Tatem, A. J., Guerra, C. A., and Snow, R. W. Estimating the global clinical burden of Plasmodium falciparum malaria in 2007. *PLoS Med*, 7(6):e1000290, Jun 2010. doi:10.1371/journal.pmed.1000290.

[2] Greenwood, B. M., Bojang, K., Whitty, C. J. M., and Targett, G. A. T. Malaria. *Lancet*, 365(9469):1487–98, Jan 2005. doi:10.1016/S0140-6736(05)66420-3.

[3] Sternberg, G. M. The malarial germ of Laveran. *Science*, 7(165):297–9, Apr 1886. doi: 10.1126/science.ns-7.165.297.

[4] Ross, R. Observations on a Condition Necessary to the Transformation of the Malaria Crescent. *Br Med J*, 1(1883):251–5, Jan 1897.

[5] Grassi, B., Bignami, A., and Bastianelli, G. Ulteriore ricerche sul ciclo dei parassiti malarici umani sul corpo del zanzarone. *Atti Reale Accad Lincei*, 8:21–28, 1899.

[6] Garnham, P. Malaria parasites and other Haemosporidia. *Malaria Parasites and Other Haemosporidia.*, Jan 1966.

[7] Liu, W., Li, Y., Learn, G. H., Rudicell, R. S., Robertson, J. D., Keele, B. F., Ndjango, J.-B. N., Sanz, C. M., Morgan, D. B., Locatelli, S., et al. Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature*, 467(7314):420–5, Sep 2010. doi:10.1038/nature09442.

[8] Singh, A. P., Buscaglia, C. A., Wang, Q., Levay, A., Nussenzweig, D. R., Walker, J. R., Winzeler, E. A., Fujii, H., Fontoura, B. M. A., and Nussenzweig, V. Plasmodium circumsporozoite protein promotes the development of the liver stages of the parasite. *Cell*, 131(3):492–504, Nov 2007. doi:10.1016/j.cell.2007.09.013.

[9] Krotoski, W. A., Collins, W. E., Bray, R. S., Garnham, P. C., Cogswell, F. B., Gwadz, R. W., Killick-Kendrick, R., Wolf, R., Sinden, R., Koontz, L. C., et al. Demonstration of hypnozoites in sporozoite-transmitted Plasmodium vivax infection. *Am J Trop Med Hyg*, 31(6):1291–3, Nov 1982.

[10] Miller, L. H., Baruch, D. I., Marsh, K., and Doumbo, O. K. The pathogenic basis of malaria. *Nature*, 415(6872):673–9, Feb 2002. doi:10.1038/415673a.

[11] Miller, L. H., Mason, S. J., Dvorak, J. A., McGinniss, M. H., and Rothman, I. K. Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants. *Science*, 189(4202):561–3, Aug 1975.

[12] Nguitragool, W., Bokhari, A. A. B., Pillai, A. D., Rayavara, K., Sharma, P., Turpin, B., Aravind, L., and Desai, S. A. Malaria parasite clag3 genes determine channel-mediated nutrient uptake by infected red blood cells. *Cell*, 145(5):665–77, May 2011. doi:10.1016/j.cell.2011.05.002.

[13] Atamna, H., Pascarmona, G., and Ginsburg, H. Hexose-monophosphate shunt activity in intact Plasmodium falciparum-infected erythrocytes and in free parasites. *Mol Biochem Parasitol*, 67(1):79–89, Aug 1994.

[14] Eckman, J. R., Modler, S., Eaton, J. W., Berger, E., and Engel, R. R. Host heme catabolism in drug-sensitive and drug-resistant malaria. *J Lab Clin Med*, 90(4):767–70, Oct 1977.

[15] Ralph, S. A., van Dooren, G. G., Waller, R. F., Crawford, M. J., Fraunholz, M. J., Foth, B. J., Tonkin, C. J., Roos, D. S., and McFadden, G. I. Tropical infectious diseases: metabolic maps and functions of the Plasmodium falciparum apicoplast. *Nat Rev Microbiol*, 2(3):203–16, Mar 2004. doi:10.1038/nrmicro843.

[16] Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., and Hay, S. I. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, 434(7030):214–7, Mar 2005. doi:10.1038/nature03342.

[17] Brewster, D. R., Kwiatkowski, D., and White, N. J. Neurological sequelae of cerebral malaria in children. *Lancet*, 336(8722):1039–43, Oct 1990.

[18] Agarwal, A., Guindo, A., Cissoko, Y., Taylor, J. G., Coulibaly, D., Koné, A., Kayentao, K., Djimde, A., Plowe, C. V., Doumbo, O., et al. Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood*, 96(7):2358–63, Oct 2000.

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

[20] White, N. J. The treatment of malaria. *N Engl J Med*, 335(11):800–6, Sep 1996. doi:10.1056/NEJM199609123351107.

[21] Marsh, K., Forster, D., Waruiru, C., Mwangi, I., Winstanley, M., Marsh, V., Newton, C., Winstanley, P., Warn, P., and Peshu, N. Indicators of life-threatening malaria in African children. *N Engl J Med*, 332(21):1399–404, May 1995. doi:10.1056/NEJM199505253322102.

[22] WHO. Malaria case management: operations manual. pages 1–122, Feb 2010.

[23] Looareesuwan, S., Davis, T. M., Pukrittayakamee, S., Supanaranond, W., Desakorn, V., Silamut, K., Krishna, S., Boonamrung, S., and White, N. J. Erythrocyte survival in severe falciparum malaria. *Acta Trop*, 48(4):263–70, Feb 1991.

[24] Schofield, L. and Grau, G. E. Immunological processes in malaria pathogenesis. *Nat Rev Immunol*, 5(9):722–35, Sep 2005. doi:10.1038/nri1686.

[25] Jakeman, G. N., Saul, A., Hogarth, W. L., and Collins, W. E. Anaemia of acute malaria infections in non-immune patients primarily results from destruction of uninfected erythrocytes. *Parasitology*, 119 (Pt 2):127–33, Aug 1999.

[26] Kurtzhals, J. A., Rodrigues, O., Addae, M., Commey, J. O., Nkrumah, F. K., and Hviid, L. Reversible suppression of bone marrow response to erythropoietin in Plasmodium falciparum malaria. *Br J Haematol*, 97(1):169–74, Apr 1997.

[27] Chang, K.-H., Tam, M., and Stevenson, M. M. Inappropriately low reticulocytosis in severe malarial anemia correlates with suppression in the development of late erythroid precursors. *Blood*, 103(10):3727–35, May 2004. doi:10.1182/blood-2003-08-2887.

[28] Calis, J. C. J., Phiri, K. S., Faragher, E. B., Brabin, B. J., Bates, I., Cuevas, L. E., de Haan, R. J., Phiri, A. I., Malange, P., Khoka, M., et al. Severe anemia in Malawian children. *N Engl J Med*, 358(9):888–99, Feb 2008. doi:10.1056/NEJMoa072727.

[29] Marchiafava, E. and Bignami, A. On summer-autumn malaria fevers (translated from the Italian version by J.H. Thompson). 1894.

[30] Turner, G. D., Morrison, H., Jones, M., Davis, T. M., Looareesuwan, S., Buley, I. D., Gatter, K. C., Newbold, C. I., Pukritayakamee, S., and Nagachinta, B. An immunohistochemical study of the pathology of fatal malaria. Evidence for widespread endothelial activation and a potential role for intercellular adhesion molecule-1 in cerebral sequestration. *Am J Pathol*, 145(5):1057–69, Nov 1994.

[31] Dorovini-Zis, K., Schmidt, K., Huynh, H., Fu, W., Whitten, R. O., Milner, D., Kamiza, S., Molyneux, M., and Taylor, T. E. The neuropathology of fatal cerebral malaria in Malawian children. *Am J Pathol*, 178(5):2146–58, May 2011. doi:10.1016/j.ajpath.2011.01.016.

[32] Taylor, T. E., Fu, W. J., Carr, R. A., Whitten, R. O., Mueller, J. S., Fosiko, N. G., Lewallen, S., Liomba, N. G., Molyneux, M. E., and Mueller, J. G. Differentiating the pathologies of cerebral malaria by postmortem parasite counts. *Nat Med*, 10(2):143–5, Feb 2004. doi:10.1038/nm986.

[33] Amani, V., Vigário, A. M., Belnoue, E., Marussig, M., Fonseca, L., Mazier, D., and Rénia, L. Involvement of IFN-gamma receptor-medicated signaling in pathology and anti-malarial immunity induced by Plasmodium berghei infection. *Eur J Immunol*, 30(6):1646–55, Jun 2000. doi:10.1002/1521-4141(200006)30:6<1646::AID-IMMU1646>3.0.CO;2-0.

[34] Stevenson, M. M. and Riley, E. M. Innate immunity to malaria. *Nat Rev Immunol*, 4(3):169–80, Mar 2004. doi:10.1038/nri1311.

[35] Mishra, S. K. and Newton, C. R. J. C. Diagnosis and management of the neurological complications of falciparum malaria. *Nat Rev Neurol*, 5(4):189–98, Apr 2009. doi:10.1038/nrneurol.2009.23.

[36] Awandare, G. A., Goka, B., Boeuf, P., Tetteh, J. K. A., Kurtzhals, J. A. L., Behr, C., and Akanmori, B. D. Increased levels of inflammatory mediators in children with severe Plasmodium falciparum malaria with respiratory distress. *J Infect Dis*, 194(10):1438–46, Nov 2006. doi:10.1086/508547.

[37] Mbogo, C. N., Kabiru, E. W., Glass, G. E., Forster, D., Snow, R. W., Khamala, C. P., Ouma, J. H., Githure, J. I., Marsh, K., and Beier, J. C. Vector-related case-control study of severe malaria in Kilifi District, Kenya. *Am J Trop Med Hyg*, 60(5):781–5, May 1999.

[38] Snow, R. W., Omumbo, J. A., Lowe, B., Molyneux, C. S., Obiero, J. O., Palmer, A., Weber, M. W., Pinder, M., Nahlen, B., Obonyo, C., et al. Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *Lancet*, 349(9066):1650–4, Jun 1997. doi:10.1016/S0140-6736(97)02038-2.

[39] Ruwende, C., Khoo, S. C., Snow, R. W., Yates, S. N., Kwiatkowski, D., Gupta, S., Warn, P., Allsopp, C. E., Gilbert, S. C., and Peschu, N. Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature*, 376(6537):246–9, Jul 1995. doi:10.1038/376246a0.

[40] Gupta, S., Snow, R. W., Donnelly, C. A., Marsh, K., and Newbold, C. Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nat Med*, 5(3):340–3, Mar 1999. doi:10.1038/6560.

[41] Nardin, E. H. and Nussenzweig, R. S. T cell responses to pre-erythrocytic stages of malaria: role in protection and vaccine development against pre-erythrocytic stages. *Annu Rev Immunol*, 11:687–727, Jan 1993. doi:10.1146/annurev.iy.11.040193.003351.

[42] Herrington, D., Davis, J., Nardin, E., Beier, M., Cortese, J., Eddy, H., Losonsky, G., Hollingdale, M., Sztein, M., and Levine, M. Successful immunization of humans with irradiated malaria sporozoites: humoral and cellular responses of the protected individuals. *Am J Trop Med Hyg*, 45(5):539–47, Nov 1991.

[43] Alonso, P. L., Sacarlal, J., Aponte, J. J., Leach, A., Macete, E., Aide, P., Sigauque, B., Milman, J., Mandomando, I., Bassat, Q., et al. Duration of protection with RTS,S/AS02A malaria vaccine in prevention of Plasmodium falciparum disease in Mozambican children: single-blind extended follow-up of a randomised controlled trial. *Lancet*, 366(9502):2012–8, Dec 2005. doi:10.1016/S0140-6736(05)67669-6.

[44] Plebanski, M. and Hill, A. V. The immunology of malaria infection. *Curr Opin Immunol*, 12(4):437–41, Aug 2000.

[45] Meding, S. J. and Langhorne, J. CD4+ T cells and B cells are necessary for the transfer of protective immunity to Plasmodium chabaudi chabaudi. *Eur J Immunol*, 21(6):1433–8, Jun 1991. doi:10.1002/eji.1830210616.

[46] van der Heyde, H. C., Pepper, B., Batchelder, J., Cigel, F., and Weidanz, W. P. The time course of selected malarial infections in cytokine-deficient mice. *Exp Parasitol*, 85(2):206–13, Feb 1997. doi:10.1006/expr.1996.4132.

[47] Mohan, K., Moulin, P., and Stevenson, M. M. Natural killer cell cytokine production, not cytotoxicity, contributes to resistance against blood-stage Plasmodium chabaudi AS infection. *J Immunol*, 159(10):4990–8, Nov 1997.

[48] Schofield, L., McConville, M. J., Hansen, D., Campbell, A. S., Fraser-Reid, B., Grusby, M. J., and Tachado, S. D. CD1d-restricted immunoglobulin G formation to GPI-anchored antigens mediated by NKT cells. *Science*, 283(5399):225–9, Jan 1999.

[49] Vijaykumar, M., Naik, R. S., and Gowda, D. C. Plasmodium falciparum glycosylphosphatidylinositol-induced TNF-alpha secretion by macrophages is mediated without membrane insertion or endocytosis. *J Biol Chem*, 276(10):6909–12, Mar 2001. doi:10.1074/jbc.C100007200.

[50] Urban, B. C., Ferguson, D. J., Pain, A., Willcox, N., Plebanski, M., Austyn, J. M., and Roberts, D. J. Plasmodium falciparum-infected erythrocytes modulate the maturation of dendritic cells. *Nature*, 400(6739):73–7, Jul 1999. doi:10.1038/21900.

[51] Kyes, S., Horrocks, P., and Newbold, C. Antigenic variation at the infected red cell surface in malaria. *Annu Rev Microbiol*, 55:673–707, Jan 2001. doi:10.1146/annurev.micro.55.1.673.

[52] Eaton, M. D. THE AGGLUTINATION OF PLASMODIUM KNOWLESI BY IMMUNE SERUM. *J Exp Med*, 67(6):857–70, May 1938.

[53] Brown, K. N. and Brown, I. N. Immunity to malaria: antigenic variation in chronic infections of Plasmodium knowlesi. *Nature*, 208(5017):1286–8, Dec 1965.

[54] Marsh, K. and Howard, R. J. Antigens induced on erythrocytes by P. falciparum: expression of diverse and conserved determinants. *Science*, 231(4734):150–3, Jan 1986.

[55] Leech, J. H., Barnwell, J. W., Miller, L. H., and Howard, R. J. Identification of a strain-specific malarial antigen exposed on the surface of Plasmodium falciparum-infected erythrocytes. *J Exp Med*, 159(6):1567–75, Jun 1984.

[56] Su, X. Z., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., Ravetch, J. A., and Wellems, T. E. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell*, 82(1):89–100, Jul 1995.

[57] Ockenhouse, C. F., Tandon, N. N., Magowan, C., Jamieson, G. A., and Chulay, J. D. Identification of a platelet membrane glycoprotein as a falciparum malaria sequestration receptor. *Science*, 243(4897):1469–71, Mar 1989.

[58] Ruangjirachuporn, W., Afzelius, B. A., Helmby, H., Hill, A. V., Greenwood, B. M., Carlson, J., Berzins, K., Perlmann, P., and Wahlgren, M. Ultrastructural analysis of fresh Plasmodium falciparum-infected erythrocytes and their cytoadherence to human leukocytes. *Am J Trop Med Hyg*, 46(5):511–9, May 1992.

[59] Newbold, C., Warn, P., Black, G., Berendt, A., Craig, A., Snow, B., Msobo, M., Peshu, N., and Marsh, K. Receptor-specific adhesion and clinical disease in Plasmodium falciparum. *Am J Trop Med Hyg*, 57(4):389–98, Oct 1997.

[60] Ochola, L. B., Siddondo, B. R., Ocholla, H., Nkya, S., Kimani, E. N., Williams, T. N., Makale, J. O., Liljander, A., Urban, B. C., Bull, P. C., et al. Specific receptor usage in Plasmodium falciparum cytoadherence is associated with disease outcome. *PLoS ONE*, 6(3):e14741, Jan 2011. doi:10.1371/journal.pone.0014741.

[61] Treutiger, C. J., Hedlund, I., Helmby, H., Carlson, J., Jepson, A., Twumasi, P., Kwiatkowski, D., Greenwood, B. M., and Wahlgren, M. Rosette formation in Plasmodium falciparum isolates and anti-rosette activity of sera from Gambians with cerebral or uncomplicated malaria. *Am J Trop Med Hyg*, 46(5):503–10, May 1992.

[62] Rowe, A., Obeiro, J., Newbold, C. I., and Marsh, K. Plasmodium falciparum rosetting is associated with malaria severity in Kenya. *Infect Immun*, 63(6):2323–6, Jun 1995.

[63] Fried, M. and Duffy, P. E. Adherence of Plasmodium falciparum to chondroitin sulfate A in the human placenta. *Science*, 272(5267):1502–4, Jun 1996.

[64] White, N. J. Artemisinin resistance–the clock is ticking. *Lancet*, 376(9758):2051–2, Dec 2010. doi:10.1016/S0140-6736(10)61963-0.

[65] Alonso, P. L., Brown, G., Arevalo-Herrera, M., Binka, F., Chitnis, C., Collins, F., Doumbo, O. K., Greenwood, B., Hall, B. F., Levine, M. M., et al. A research agenda to underpin malaria eradication. *PLoS Med*, 8(1):e1000406, Jan 2011. doi:10.1371/journal.pmed.1000406.

[66] Moonen, B., Cohen, J. M., Snow, R. W., Slutsker, L., Drakeley, C., Smith, D. L., Abeyasinghe, R. R., Rodriguez, M. H., Maharaj, R., Tanner, M., et al. Operational strategies to achieve and maintain malaria elimination. *Lancet*, 376(9752):1592–603, Nov 2010. doi:10.1016/S0140-6736(10)61269-X.

[67] Carter, R., Mendis, K. N., Miller, L. H., Molineaux, L., and Saul, A. Malaria transmission-blocking vaccines–how can their development be supported? *Nat Med*, 6(3):241–4, Mar 2000. doi:10.1038/73062.

[68] Buchholz, K., Burke, T. A., Williamson, K. C., Wiegand, R. C., Wirth, D. F., and Marti, M. A high-throughput screen targeting malaria transmission stages opens new avenues for drug development. *J Infect Dis*, 203(10):1445–53, May 2011. doi:10.1093/infdis/jir037.

[69] Friedman, J. F., Kwena, A. M., Mirel, L. B., Kariuki, S. K., Terlouw, D. J., Phillips-Howard, P. A., Hawley, W. A., Nahlen, B. L., Shi, Y. P., and ter Kuile, F. O. Malaria and nutritional status among pre-school children: results from cross-sectional surveys in western Kenya. *Am J Trop Med Hyg*, 73(4):698–704, Oct 2005.

[70] Valin, N., Dieng, A. B., Chippaux, J. P., Hesran, J. Y. L., Gaye, O., and Garcia, A. Low malaria morbidity in a cohort of Senegalese children with free access to health structures. *Parasite*, 13(1):79–81, Mar 2006.

[71] Greenwood, B. Can malaria be eliminated? *Trans R Soc Trop Med Hyg*, 103 Suppl 1:S2–5, Apr 2009. doi:10.1016/j.trstmh.2008.10.027.

[72] Clarke, J. L., Scopes, D. A., Sodeinde, O., and Mason, P. J. Glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase. A novel bifunctional enzyme in malaria parasites. *Eur J Biochem*, 268(7):2013–9, Mar 2001.

[73] Cappellini, M. D. and Fiorelli, G. Glucose-6-phosphate dehydrogenase deficiency. *Lancet*, 371(9606):64–74, Jan 2008. doi:10.1016/S0140-6736(08)60073-2.

[74] Kotaka, M., Gover, S., Vandeputte-Rutten, L., Au, S. W. N., Lam, V. M. S., and Adams, M. J. Structural studies of glucose-6-phosphate and NADP+ binding to human glucose-6-phosphate dehydrogenase. *Acta Crystallogr D Biol Crystallogr*, 61(Pt 5):495–504, May 2005. doi:10.1107/S0907444905002350.

[75] Ayi, K., Giribaldi, G., Skorokhod, A., Schwarzer, E., Prendergast, P. T., and Arese, P. 16alpha-bromoepiandrosterone, an antimalarial analogue of the hormone dehydroepiandrosterone, enhances phagocytosis of ring stage parasitized erythrocytes: a novel mechanism for antimalarial activity. *Antimicrob Agents Chemother*, 46(10):3180–4, Sep 2002.

[76] Jiang, P., Du, W., Wang, X., Mancuso, A., Gao, X., Wu, M., and Yang, X. p53 regulates biosynthesis through direct inactivation of glucose-6-phosphate dehydrogenase. *Nat Cell Biol*, 13(3):310–6, Mar 2011. doi:10.1038/ncb2172.

[77] Gaetani, G. D., Parker, J. C., and Kirkman, H. N. Intracellular restraint: a new basis for the limitation in response to oxidative stress in human erythrocytes containing low-activity variants of glucose-6-phosphate dehydrogenase. *Proc Natl Acad Sci USA*, 71(9):3584–7, Sep 1974.

[78] Franzè, A., Ferrante, M. I., Fusco, F., Santoro, A., Sanzari, E., Martini, G., and Ursini, M. V. Molecular anatomy of the human glucose 6-phosphate dehydrogenase core promoter. *FEBS Lett*, 437(3):313–8, Oct 1998.

[79] Philippe, M., Larondelle, Y., Lemaigre, F., Mariamé, B., Delhez, H., Mason, P., Luzzatto, L., and Rousseau, G. G. Promoter function of the human glucose-6-phosphate dehydrogenase gene depends on two GC boxes that are cell specifically controlled. *Eur J Biochem*, 226(2):377–84, Nov 1994.

[80] Leopold, J. A., Dam, A., Maron, B. A., Scribner, A. W., Liao, R., Handy, D. E., Stanton, R. C., Pitt, B., and Loscalzo, J. Aldosterone impairs vascular reactivity by decreasing glucose-6-phosphate dehydrogenase activity. *Nat Med*, 13(2):189–97, Feb 2007. doi:10.1038/nm1545.

[81] Johnston, C. M., Lovell, F. L., Leongamornlert, D. A., Stranger, B. E., Dermitzakis, E. T., and Ross, M. T. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet*, 4(1):e9, Jan 2008. doi:10.1371/journal.pgen.0040009.

[82] Au, W.-Y., Lam, V., Pang, A., Lee, W.-M., Chan, J. L. C., Song, Y.-Q., Ma, E. S., and Kwong, Y.-L. Glucose-6-phosphate dehydrogenase deficiency in female octogenarians, nanogenarians, and centenarians. *J Gerontol A Biol Sci Med Sci*, 61(10):1086–9, Oct 2006.

[83] Leopold, J. A., Walker, J., Scribner, A. W., Voetsch, B., Zhang, Y.-Y., Loscalzo, A. J., Stanton, R. C., and Loscalzo, J. Glucose-6-phosphate dehydrogenase modulates vascular endothelial growth factor-mediated angiogenesis. *J Biol Chem*, 278(34):32100–6, Aug 2003. doi:10.1074/jbc.M301293200.

[84] Agudelo-Flórez, P., Costa-Carvalho, B. T., López, J. A., Redher, J., Newburger, P. E., Olalla-Saad, S. T., and Condino-Neto, A. Association of glucose-6-phosphate dehydrogenase deficiency and X-linked chronic granulomatous disease in a child with anemia and recurrent infections. *Am J Hematol*, 75(3):151–6, Mar 2004. doi:10.1002/ajh.10477.

[85] Matsui, R., Xu, S., Maitland, K. A., Mastroianni, R., Leopold, J. A., Handy, D. E., Loscalzo, J., and Cohen, R. A. Glucose-6-phosphate dehydrogenase deficiency decreases vascular superoxide and atherosclerotic lesions in apolipoprotein E(-/-) mice. *Arterioscler Thromb Vasc Biol*, 26(4):910–6, Apr 2006. doi:10.1161/01.ATV.0000205850.49390.3b.

[86] Upperman, J. S., Pillage, G., Siddiqi, M. Q., Zeevi, A., Kelly, N., Ford, H. R., Kammerer, C., and Spolarics, Z. Dominance of high-producing interleukin 6 and low-producing interleukin 10 and interferon gamma alleles in glucose-6-phosphate dehydrogenase-deficient trauma patients. *Shock*, 23(3):197–201, Feb 2005.

[87] Ferreira, A., Marguti, I., Bechmann, I., Jeney, V., Chora, A., Palha, N. R., Rebelo, S., Henri, A., Beuzard, Y., and Soares, M. P. Sickle hemoglobin confers tolerance to Plasmodium infection. *Cell*, 145(3):398–409, Apr 2011. doi:10.1016/j.cell.2011.03.049.

[88] Nkhoma, E. T., Poole, C., Vannappagari, V., Hall, S. A., and Beutler, E. The global prevalence of glucose-6-phosphate dehydrogenase deficiency: a systematic review and meta-analysis. *Blood Cells Mol Dis*, 42(3):267–78, Dec 2009. doi:10.1016/j.bcmd.2008.12.005.

[89] Hockwald, R., Arnold, J., Clayman, C., and Alving, A. Toxicity of primaquine in Negroes. *J Am Med Assoc*, 149(17):1568–70, Aug 1952.

[90] Beutler, E. Glucose-6-phosphate dehydrogenase deficiency: a historical perspective. *Blood*, 111(1):16–24, Dec 2008. doi:10.1182/blood-2007-04-077412.

[91] Dern, R. J., Beutler, E., and Alving, A. The hemolytic effect of primaquine. II. The natural course of the hemolytic anemia and the mechanism of its self-limited character. *J Lab Clin Med*, 44(2):171–6, Aug 1954.

[92] Carson, P. E., Flanagan, C., Ickes, C., and Alving, A. Enzymatic deficiency in primaquine-sensitive erythrocytes. *Science*, 124(3220):484–5, Sep 1956.

[93] Beutler, E., Yeh, M., and Fairbanks, V. The normal human female as a mosaic of X-chromosome activity: studies using the gene for C-6-PD-deficiency as a marker. *Proc Natl Acad Sci USA*, 48:9–16, Jan 1962.

[94] Beutler, E. and Vulliamy, T. J. Hematologically important mutations: glucose-6-phosphate dehydrogenase. *Blood Cells Mol Dis*, 28(2):93–103, Dec 2002.

[95] Piomelli, S., Corash, L. M., Davenport, D. D., Miraglia, J., and Amorosi, E. L. In vivo lability of glucose-6-phosphate dehydrogenase in GdA- and GdMediterranean deficiency. *J Clin Invest*, 47(4):940–8, Mar 1968. doi:10.1172/JCI105786.

[96] Babalola, O., Cancedda, R., and Luzzatto, L. Genetic variants of glucose 6-phosphate dehydrogenase from human erythrocytes: unique properties of the A - variant isolated from "deficient" cells. *Proc Natl Acad Sci USA*, 69(4):946–50, Mar 1972.

[97] MARKS, P. A., GROSS, R. T., and HURWITZ, R. E. Gene action in erythrocyte deficiency of glucose-6-phosphate dehydrogenase: tissue enzyme-levels. *Nature*, 183(4670):1266–7, May 1959.

[98] Yoshida, A., Stamatoyannopoulos, G., and Motulsky, A. G. Negro variant of glucose-6-phosphate dehydrogenase deficiency (A-) in man. *Science*, 155(758):97–9, Jan 1967.

[99] Yoshida, A. Hemolytic anemia and G6PD deficiency. *Science*, 179(73):532–7, Feb 1973.

[100] Beutler, E., Kuhl, W., Vives-Corrons, J. L., and Prchal, J. T. Molecular heterogeneity of glucose-6-phosphate dehydrogenase A-. *Blood*, 74(7):2550–5, Nov 1989.

[101] Hue, N. T., Charlieu, J. P., Chau, T. T. H., Day, N., Farrar, J. J., Hien, T. T., and Dunstan, S. J. Glucose-6-phosphate dehydrogenase (G6PD) mutations and haemoglobinuria syndrome in the Vietnamese population. *Malar J*, 8:152, Dec 2009. doi:10.1186/1475-2875-8-152.

[102] Kaplan, M., Beutler, E., Vreman, H. J., Hammerman, C., Levy-Lahad, E., Renbaum, P., and Stevenson, D. K. Neonatal hyperbilirubinemia in glucose-6-phosphate dehydrogenase-deficient heterozygotes. *Pediatrics*, 104(1 Pt 1):68–74, Jun 1999.

[103] Beutler, E. G6PD: population genetics and clinical manifestations. *Blood Rev*, 10(1):45–52, Feb 1996.

[104] Saha, N. Association of glucose-6-phosphate dehydrogenase deficiency with diabetes mellitus in ethnic groups of Singapore. *J Med Genet*, 16(6):431–4, Dec 1979.

[105] Saeed, T. K., Hamamy, H. A., and Alwan, A. A. Association of glucose-6-phosphate dehydrogenase deficiency with diabetes mellitus. *Diabet Med*, 2(2):110–2, Mar 1985.

[106] Niazi, G. A. Glucose-6-phosphate dehydrogenase deficiency and diabetes mellitus. *Int J Hematol*, 54(4):295–8, Aug 1991.

[107] Meloni, L., Manca, M. R., Loddo, I., Cioglia, G., Cocco, P., Schwartz, A., Muntoni, S., and Muntoni, S. Glucose-6-phosphate dehydrogenase deficiency protects against coronary heart disease. *J Inherit Metab Dis*, 31(3):412–7, Jun 2008. doi:10.1007/s10545-008-0704-5.

[108] Gupte, R. S., Vijay, V., Marks, B., Levine, R. J., Sabbah, H. N., Wolin, M. S., Recchia, F. A., and Gupte, S. A. Upregulation of glucose-6-phosphate dehydrogenase and NAD(P)H oxidase activity increases oxidative stress in failing human heart. *J Card Fail*, 13(6):497–506, Aug 2007. doi: 10.1016/j.cardfail.2007.04.003.

[109] Schutzman, D. L. and Porat, R. Glucose-6-phosphate dehydrogenase deficiency: another risk factor for necrotizing enterocolitis? *J Pediatr*, 151(4):435–7, Oct 2007. doi:10.1016/j.jpeds.2007.06.041.

[110] Haldane, J. The rate of mutations of human genes. *Proceedings of the Eighth International Congress of Genetics*, Dec 1949.

[111] Sørensen, T. I., Nielsen, G. G., Andersen, P. K., and Teasdale, T. W. Genetic and environmental influences on premature death in adult adoptees. *N Engl J Med*, 318(12):727–32, Mar 1988.

[112] Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet*, 77(2):171–92, Jul 2005. doi:10.1086/432519.

[113] Sakai, Y., Kobayashi, S., Shibata, H., Furuumi, H., Endo, T., Fucharoen, S., Hamano, S., Acharya, G. P., Kawasaki, T., and Fukumaki, Y. Molecular analysis of alpha-thalassemia in Nepal: correlation with malaria endemicity. *J Hum Genet*, 45(3):127–32, Jan 2000. doi:10.1007/s100380050198.

[114] Flint, J., Hill, A. V., Bowden, D. K., Oppenheimer, S. J., Sill, P. R., Serjeantson, S. W., Bana-Koiri, J., Bhatia, K., Alpers, M. P., and Boyce, A. J. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature*, 321(6072):744–50, Dec 1986. doi:10.1038/321744a0.

[115] Modiano, G., Morpurgo, G., Terrenato, L., Novelletto, A., Rienzo, A. D., Colombo, B., Purpura, M., Mariani, M., Santachiara-Benerecetti, S., and Brega, A. Protection against malaria morbidity: near-fixation of the alpha-thalassemia gene in a Nepalese population. *Am J Hum Genet*, 48(2):390–7, Feb 1991.

[116] Allison, A. C. Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans. *Nature*, 186:531–2, May 1960.

[117] Flatz, G. and Sringam, S. Malaria and glucose-6-phosphate dehydrogenase deficiency in Thailand. *Lancet*, 2(7320):1248–50, Dec 1963.

[118] Kidson, C. and Gorman, J. G. A challenge to the concept of selection by malaria in glucose-6-phosphate dehydrogenase deficiency. *Nature*, 196:49–51, Oct 1962.

[119] Eng, L. Glucose-6-phosphate dehydrogenase deficiency in Indonesia. *Nature*, Jan 1964.

[120] Kidson, C. and Gajdusek, D. Congenital defects of the central nervous system associated with hyperendemic goiter in a neolithic highland society of Netherlands New Guinea II. Glucose-6-phosphate dehydrogenase in the mulia population. *Pediatrics*, 29:364–8, Mar 1962.

[121] Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, 293(5529):455–62, Jul 2001. doi:10.1126/science.1061573.

[122] Saunders, M. A., Hammer, M. F., and Nachman, M. W. Nucleotide variability at G6pd and the signature of malarial selection in humans. *Genetics*, 162(4):1849–61, Nov 2002.

[123] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–7, Oct 2002. doi:10.1038/nature01140.

[124] Saunders, M. A., Slatkin, M., Garner, C., Hammer, M. F., and Nachman, M. W. The extent of linkage disequilibrium caused by selection on G6PD in humans. *Genetics*, 171(3):1219–29, Oct 2005. doi:10.1534/genetics.105.048140.

[125] Louicharoen, C., Patin, E., Paul, R., Nuchprayoon, I., Witoonpanich, B., Peerapittayamongkol, C., Casademont, I., Sura, T., Laird, N. M., Singhasivanon, P., et al. Positively selected G6PD-Mahidol mutation reduces Plasmodium vivax density in Southeast Asians. *Science*, 326(5959):1546–9, Dec 2009. doi:10.1126/science.1178849.

[126] Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol*, 4(3):e72, Feb 2006. doi:10.1371/journal.pbio.0040072.

[127] Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–8, Oct 2007. doi:10.1038/nature06250.

[128] English, M., Ngama, M., Musumba, C., Wamola, B., Bwika, J., Mohammed, S., Ahmed, M., Mwarumba, S., Ouma, B., McHugh, K., et al. Causes and outcome of young infant admissions to a Kenyan district hospital. *Arch Dis Child*, 88(5):438–43, May 2003.

[129] Ogunlesi, T. A., Dedeke, I. O. F., Adekanmbi, A. F., Fetuga, M. B., and Ogunfowora, O. B. The incidence and outcome of bilirubin encephalopathy in Nigeria: a bi-centre study. *Niger J Med*, 16(4):354–9, Jan 2007.

[130] Gordon, A. L., English, M., Dzombo, J. T., Karisa, M., and Newton, C. R. J. C. Neurological and developmental outcome of neonatal jaundice and sepsis in rural Kenya. *Trop Med Int Health*, 10(11):1114–20, Nov 2005. doi:10.1111/j.1365-3156.2005.01496.x.

[131] Motulsky, A. G., Vandepitte, J., and Fraser, G. R. Population genetic studies in the Congo. I. Glucose-6-phosphate dehydrogenase deficiency, hemoglobin S, and malaria. *Am J Hum Genet*, 18(6):514–37, Oct 1966.

[132] Allison, A. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br Med J*, Dec 1954.

[133] Williams, T. N., Mwangi, T. W., Wambua, S., Alexander, N. D., Kortok, M., Snow, R. W., and Marsh, K. Sickle cell trait and the risk of Plasmodium falciparum malaria and other childhood diseases. *J Infect Dis*, 192(1):178–86, Jul 2005. doi:10.1086/430744.

[134] May, J., Evans, J. A., Timmann, C., Ehmen, C., Busch, W., Thye, T., Agbenyega, T., and Horstmann, R. D. Hemoglobin variants and disease manifestations in severe falciparum malaria. *JAMA*, 297(20):2220–6, May 2007. doi:10.1001/jama.297.20.2220.

[135] Allison, A. C. and Clyde, D. F. Malaria in African children with deficient erythrocyte glucose-6-phosphate dehydrogenase. *Br Med J*, 1(5236):1346–9, May 1961.

[136] Bienzle, U., Ayeni, O., Lucas, A., and Luzzatto, L. Glucose-6-phosphate dehydrogenase and malaria. Greater resistance of females heterozygous for enzyme deficiency and of males with non-deficient variant. *Lancet*, 1(7742):107, 1972.

[137] Kruatrachue, M., Charoenlarp, P., Chongsuphajaisiddhi, T., and Harinasuta, C. Erythrocyte glucose-6-phosphate dehydrogenase and malaria in Thailand. *Lancet*, 2(7267):1183–6, Dec 1962.

[138] Gilles, H. M., Fletcher, K. A., Hendrickse, R. G., Lindner, R., Reddy, S., and Allan, N. Glucose-6-phosphate-dehydrogenase deficiency, sickling, and malaria in African children in South Western Nigeria. *Lancet*, 1(7482):138–40, Jan 1967.

[139] Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., and Greenwood, B. M. Common west African HLA antigens are associated with protection from severe malaria. *Nature*, 352(6336):595–600, Aug 1991. doi: 10.1038/352595a0.

[140] Modiano, D., Luoni, G., Sirima, B. S., Simporé, J., Verra, F., Konaté, A., Rastrelli, E., Olivieri, A., Calissano, C., Paganotti, G. M., et al. Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature*, 414(6861):305–8, Nov 2001. doi:10.1038/35104556.

[141] Williams, T. N., Mwangi, T. W., Wambua, S., Peto, T. E. A., Weatherall, D. J., Gupta, S., Recker, M., Penman, B. S., Uyoga, S., Macharia, A., et al. Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. *Nat Genet*, 37(11):1253–7, Nov 2005. doi:10.1038/ng1660.

[142] Luzzatto, L., Usanga, F. A., and Reddy, S. Glucose-6-phosphate dehydrogenase deficient red cells: resistance to infection by malarial parasites. *Science*, 164(881):839–42, May 1969.

[143] Clark, T., Fry, A., Auburn, S., Campino, S., Diakite, M., Green, A., Richardson, A., Teo, Y., Small, K., Wilson, J., et al. Allelic heterogeneity of G6PD deficiency in West Africa and severe malaria susceptibility. *Eur J Hum Genet*, Feb 2009. doi:10.1038/ejhg.2009.8.

[144] Patsopoulos, N. A., Tatsioni, A., and Ioannidis, J. P. A. Claims of sex differences: an empirical assessment in genetic associations. *JAMA*, 298(8):880–93, Aug 2007. doi:10.1001/jama.298.8.880.

[145] Friedman, M. Erythrocytic mechanism of sickle cell resistance to malaria. *Proceedings of the National Academy of Sciences*, Dec 1978.

[146] Pasvol, G., Weatherall, D. J., and Wilson, R. J. Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria. *Nature*, 274(5672):701–3, Aug 1978.

[147] Luzzatto, L., Nwachuku-Jarrett, E. S., and Reddy, S. Increased sickling of parasitised erythrocytes as mechanism of resistance against malaria in the sickle-cell trait. *Lancet*, 1(7642):319–21, Feb 1970.

[148] Roth, E. F., Raventos-Suarez, C., Rinaldi, A., and Nagel, R. L. Glucose-6-phosphate dehydrogenase deficiency inhibits in vitro growth of Plasmodium falciparum. *Proc Natl Acad Sci USA*, 80(1):298–9, Dec 1983.

[149] Friedman, M. J. Oxidant damage mediates variant red cell resistance to malaria. *Nature*, 280(5719):245–7, Jul 1979.

[150] Cappadoro, M., Giribaldi, G., O'Brien, E., Turrini, F., Mannu, F., Ulliers, D., Simula, G., Luzzatto, L., and Arese, P. Early phagocytosis of glucose-6-phosphate dehydrogenase (G6PD)-deficient erythrocytes parasitized by Plasmodium falciparum may explain malaria protection in G6PD deficiency. *Blood*, 92(7):2527–34, Sep 1998.

[151] Usanga, E. A. and Luzzatto, L. Adaptation of Plasmodium falciparum to glucose 6-phosphate dehydrogenase-deficient host red cells by production of parasite-encoded enzyme. *Nature*, 313(6005):793–5, Dec 1985.

[152] Kurdi-Haidar, B. and Luzzatto, L. Expression and characterization of glucose-6-phosphate dehydrogenase of Plasmodium falciparum. *Mol Biochem Parasitol*, 41(1):83–91, May 1990.

[153] Rowe, J. A., Handel, I. G., Thera, M. A., Deans, A.-M., Lyke, K. E., Koné, A., Diallo, D. A., Raza, A., Kai, O., Marsh, K., et al. Blood group O protects against severe Plasmodium falciparum malaria through the mechanism of reduced rosetting. *Proc Natl Acad Sci USA*, 104(44):17471–6, Oct 2007. doi:10.1073/pnas.0705390104.

[154] Fairhurst, R. M., Baruch, D. I., Brittain, N. J., Ostera, G. R., Wallach, J. S., Hoang, H. L., Hayton, K., Guindo, A., Makobongo, M. O., Schwartz, O. M., et al. Abnormal display of PfEMP-1 on erythrocytes carrying haemoglobin C may protect against malaria. *Nature*, 435(7045):1117–21, Jun 2005. doi:10.1038/nature03631.

[155] Cholera, R., Brittain, N. J., Gillrie, M. R., Lopera-Mesa, T. M., Diakité, S. A. S., Arie, T., Krause, M. A., Guindo, A., Tubman, A., Fujioka, H., et al. Impaired cytoadherence of Plasmodium falciparum-infected erythrocytes containing sickle hemoglobin. *Proc Natl Acad Sci USA*, 105(3):991–6, Jan 2008. doi:10.1073/pnas.0711401105.

[156] Williams, T. N., Mwangi, T. W., Roberts, D. J., Alexander, N. D., Weatherall, D. J., Wambua, S., Kortok, M., Snow, R. W., and Marsh, K. An immune basis for malaria protection by the sickle cell trait. *PLoS Med*, 2(5):e128, May 2005. doi:10.1371/journal.pmed.0020128.

[157] Fry, A. E., Griffiths, M. J., Auburn, S., Diakite, M., Forton, J. T., Green, A., Richardson, A., Wilson, J., Jallow, M., Sisay-Joof, F., et al. Common variation in the ABO glycosyltransferase is associated with susceptibility to severe Plasmodium falciparum malaria. *Hum Mol Genet*, 17(4):567–76, Feb 2008. doi:10.1093/hmg/ddm331.

[158] Hartl, D. L. and Clark, A. G. Principles of population genetics. page 652, Jan 2007.

[159] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001. doi:10.1038/35057062.

[160] Consortium, I. H. The International HapMap Project. *Nature*, 426(6968):789–96, Dec 2003. doi:10.1038/nature02168.

[161] Consortium, I. H. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, Oct 2005. doi:10.1038/nature04226.

[162] Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–33, Feb 2001. doi:10.1038/35057149.

[163] Hawkins, R. D., Hon, G. C., and Ren, B. Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–86, Jul 2010. doi:10.1038/nrg2795.

[164] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, Nov 2008. doi:10.1038/nature07517.

[165] Kaiser, J. DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science*, 319(5862):395, Jan 2008. doi:10.1126/science.319.5862.395.

[166] Lander, E. Marshall Nirenberg lecture. 2011.

[167] Wall, J. D. and Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet*, 4(8):587–97, Jul 2003. doi:10.1038/nrg1123.

[168] Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, Sep 1996.

[169] Pennisi, E. Breakthrough of the year. Human genetic variation. *Science*, 318(5858):1842–3, Dec 2007. doi:10.1126/science.318.5858.1842.

[170] Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–48, Nov 2010. doi: 10.1038/ng.686.

[171] Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–8, Oct 2010. doi:10.1038/nature09410.

[172] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, Jul 2007. doi:10.1038/ng2088.

[173] Teo, Y.-Y., Small, K. S., and Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet*, 11(2):149–60, Feb 2010. doi:10.1038/nrg2731.

[174] Jallow, M., Teo, Y., Small, K., Rockett, K., Deloukas, P., Clark, T., Kivinen, K., Bojang, K., Conway, D., Pinder, M., et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet*, May 2009. doi:10.1038/ng.388.

[175] Mardis, E. R. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9:387–402, Dec 2008. doi:10.1146/annurev.genom.9.081307.164359.

[176] Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. A large genome center's improvements to the Illumina sequencing system. *Nat Methods*, 5(12):1005–10, Nov 2008. doi:10.1038/nmeth.1270.

[177] Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, 6(4):291–5, Mar 2009. doi:10.1038/nmeth.1311.

[178] Tang, K., Fu, D. J., Julien, D., Braun, A., Cantor, C. R., and Köster, H. Chip-based genotyping by mass spectrometry. *Proc Natl Acad Sci USA*, 96(18):10016–20, Aug 1999.

[179] Gabriel, S., Ziaugra, L., and Tabbaa, D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet*, Chapter 2:Unit 2.12, Dec 2009. doi: 10.1002/0471142905.hg0212s60.

[180] Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci USA*, 89(13):5847–51, Jul 1992.

[181] RDevelopmentCoreTeam. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Jan 2008.

[182] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, Sep 2007. doi:10.1086/519795.

[183] Clayton, D. and Leung, H.-T. An R package for analysis of whole-genome association studies. *Hum Hered*, 64(1):45–51, Dec 2007. doi:10.1159/000101422.

[184] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, Oct 2002. doi:10.1101/gr.361602.

[185] Scheet, P. and Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–44, Mar 2006. doi:10.1086/502802.

[186] Network, M. G. E. A global network for investigating the genomic epidemiology of malaria. *Nature*, 456(7223):732–7, Dec 2008. doi:10.1038/nature07632.

[187] Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA*, 99(8):5261–6, Apr 2002. doi:10.1073/pnas.082089499.

[188] Li, H., Ruan, J., and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8, Oct 2008. doi:10.1101/gr.078212.108.

[189] Clayton, D. Testing for association on the X chromosome. *Biostatistics (Oxford, England)*, 9(4):593–600, Sep 2008. doi:10.1093/biostatistics/kxn007.

[190] Brewer, G. J., Tarlov, A., and Alving, A. Methaemoglobin reduction test: a new, simple, in vitro test for identifying primaquine-sensitivity. *Bull World Health Organ*, 22:633–40, Jan 1960.

[191] Beutler, E. and Baluda, M. C. Methemoglobin reduction: Studies of the interaction between cell populations and of the role of methylene blue. *Blood*, 22:323–33, Sep 1963.

[192] Slatkin, M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, 9(6):477–85, Jun 2008. doi:10.1038/nrg2361.

[193] Stephens, M. and Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–9, Oct 2003. doi:10.1086/379378.

[194] Stephens, M., Smith, N. J., and Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, Apr 2001. doi:10.1086/319501.

[195] Weir, B. S. Genetic data analysis 2âĂŐ. page 445, Jan 1996.

[196] Holder, M. and Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet*, 4(4):275–84, Mar 2003. doi:10.1038/nrg1044.

[197] Cavalli-Sforza, L. L. and Edwards, A. W. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet*, 19(3 Pt 1):233–57, Apr 1967.

[198] Felsenstein, J. Evolutionary Trees From Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates. *Evolution*, 35(6):1229–1242, Nov 1981.

[199] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, Jun 1901.

[200] Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. The history and geography of human genesâĂŐ. page 1059, Jan 1994.

[201] Novembre, J. and Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*, 40(5):646–9, Apr 2008. doi:10.1038/ng.139.

[202] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, Jul 2006. doi:10.1038/ng1847.

[203] Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Cox, M. P., and Wall, J. D. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nat Genet*, 42(10):830–1, Oct 2010. doi:10.1038/ng.651.

[204] Vulliamy, T., Luzzatto, L., Hirono, A., and Beutler, E. Hematologically important mutations: glucose-6-phosphate dehydrogenase. *Blood Cells Mol Dis*, 23(2):302–13, Jul 1997.

[205] Hirono, A. and Beutler, E. Molecular cloning and nucleotide sequence of cDNA for human glucose-6-phosphate dehydrogenase variant A(-). *Proc Natl Acad Sci USA*, 85(11):3951–4, May 1988.

[206] Vulliamy, T. J., Othman, A., Town, M., Nathwani, A., Falusi, A. G., Mason, P. J., and Luzzatto, L. Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variants A and A-. *Proc Natl Acad Sci USA*, 88(19):8568–71, Sep 1991.

[207] Ganczakowski, M., Town, M., Bowden, D. K., Vulliamy, T. J., Kaneko, A., Clegg, J. B., Weatherall, D. J., and Luzzatto, L. Multiple glucose 6-phosphate dehydrogenase-deficient variants correlate with malaria endemicity in the Vanuatu archipelago (southwestern Pacific). *Am J Hum Genet*, 56(1):294–301, Dec 1995.

[208] Verrelli, B. C., McDonald, J. H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A. N., Loiselet, J., and Tishkoff, S. A. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am J Hum Genet*, 71(5):1112–28, Oct 2002. doi:10.1086/344345.

[209] Kwok, C. J., Martin, A. C. R., Au, S. W. N., and Lam, V. M. S. G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. *Hum Mutat*, 19(3):217–24, Feb 2002. doi:10.1002/humu.10036.

[210] Zweig, A. S., Karolchik, D., Kuhn, R. M., Haussler, D., and Kent, W. J. UCSC genome browser tutorial. *Genomics*, 92(2):75–84, Jul 2008. doi:10.1016/j.ygeno.2008.02.003.

[211] Taylor, J., Schenck, I., Blankenberg, D., and Nekrutenko, A. Using galaxy to perform large-scale interactive data analyses. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]*, Chapter 10:Unit 10.5, Aug 2007. doi:10.1002/0471250953.bi1005s19.

[212] Cockerham, C. C. and Weir, B. S. Estimation of inbreeding parameters in stratified populations. *Ann Hum Genet*, 50(Pt 3):271–81, Jun 1986.

[213] Menozzi, P., Piazza, A., and Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–92, Aug 1978.

[214] Felsenstein, J. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, Jan 1989.

[215] Page, R. D. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci*, 12(4):357–8, Aug 1996.

[216] Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–72, Apr 2010. doi:10.1038/nature08872.

[217] Wei, W., Pelechano, V., Järvelin, A. I., and Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet*, 27(7):267–76, Jul 2011. doi:10.1016/j.tig.2011.04.002.

[218] Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–4, Feb 2008. doi:10.1126/science.1153717.

[219] Xi, T., Jones, I. M., and Mohrenweiser, H. W. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*, 83(6):970–9, Jun 2004. doi:10.1016/j.ygeno.2003.12.016.

[220] Naylor, C. E., Rowland, P., Basak, A. K., Gover, S., Mason, P. J., Bautista, J. M., Vulliamy, T. J., Luzzatto, L., and Adams, M. J. Glucose 6-phosphate dehydrogenase mutations causing enzyme deficiency in a model of the tertiary structure of the human enzyme. *Blood*, 87(7):2974–82, Mar 1996.

[221] Au, S. W., Gover, S., Lam, V. M., and Adams, M. J. Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural NADP(+) molecule and provides insights into enzyme deficiency. *Structure*, 8(3):293–303, Mar 2000.

[222] Bao, L. and Cui, Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, 21(10):2185–90, May 2005. doi:10.1093/bioinformatics/bti365.

[223] Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9, Apr 2010. doi:10.1038/nmeth0410-248.

[224] Hirono, A., Kawate, K., Honda, A., Fujii, H., and Miwa, S. A single mutation 202G>A in the human glucose-6-phosphate dehydrogenase gene (G6PD) can cause acute hemolysis by itself. *Blood*, 99(4):1498, Feb 2002.

[225] Group, N. H. W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., et al. The NIH Human Microbiome Project. *Genome Res*, 19(12):2317–23, Dec 2009. doi:10.1101/gr.096651.109.

[226] Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–7, Apr 2010. doi:10.1038/nature08903.

[227] Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–80, Oct 2009. doi:10.1038/nrg2641.

[228] Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Richmond, T. A., Middle, C. M., Rodesch, M. J., Packard, C. J., et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*, 4(11):903–5, Oct 2007. doi:10.1038/nmeth1111.

[229] Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., and Zwick, M. E. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods*, 4(11):907–9, Oct 2007. doi:10.1038/nmeth1109.

[230] Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., LeProust, E. M., Peck, B. J., Emig, C. J., Dahl, F., et al. Multiplex amplification of large sets of human exons. *Nat Methods*, 4(11):931–6, Oct 2007. doi:10.1038/nmeth1110.

[231] Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. Target-enrichment strategies for next-generation sequencing. *Nat Methods*, 7(2):111–8, Feb 2010. doi:10.1038/nmeth.1419.

[232] Craig, D. W., Pearson, J. V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J. J., Pawlowski, T. L., Laub, T., Nunn, G., Stephan, D. A., et al. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods*, 5(10):887–93, Sep 2008. doi:10.1038/nmeth.1251.

[233] Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., and Mockler, T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res*, 36(19):e122, Nov 2008. doi:10.1093/nar/gkn502.

[234] Bansal, V., Harismendy, O., Tewhey, R., Murray, S. S., Schork, N. J., Topol, E. J., and Frazer, K. A. Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res*, 20(4):537–45, Apr 2010. doi:10.1101/gr.100040.109.

[235] Tassell, C. P. V., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., and Sonstegard, T. S. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods*, 5(3):247–52, Mar 2008. doi:10.1038/nmeth.1185.

[236] Druley, T. E., Vallania, F. L. M., Wegner, D. J., Varley, K. E., Knowles, O. L., Bonds, J. A., Robison, S. W., Doniger, S. W., Hamvas, A., Cole, F. S., et al. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods*, 6(4):263–5, Mar 2009. doi:10.1038/nmeth.1307.

[237] Ingman, M. and Gyllensten, U. SNP frequency estimation using massively parallel sequencing of pooled DNA. *Eur J Hum Genet*, 17(3):383–6, Mar 2009. doi:10.1038/ejhg.2008.182.

[238] Out, A. A., van Minderhout, I. J. H. M., Goeman, J. J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P. E. M., Tops, C. M. J., et al. Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat*, 30(12):1703–12, Dec 2009. doi: 10.1002/humu.21122.

[239] Andreson, R., Puurand, T., and Remm, M. SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. *Nucleic Acids Res*, 34(Web Server issue):W651–5, Jun 2006. doi: 10.1093/nar/gkl125.

[240] Rozen, S. and Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, Dec 2000.

[241] Shekalaghe, S. A., ter Braak, R., Daou, M., Kavishe, R., van den Bijllaardt, W., van den Bosch, S., Koenderink, J. B., Luty, A. J. F., Whitty, C. J. M., Drakeley, C., et al. In Tanzania, hemolysis after a single dose of primaquine coadministered with an artemisinin is not restricted to glucose-6-phosphate dehydrogenase-deficient (G6PD A-) individuals. *Antimicrob Agents Chemother*, 54(5):1762–8, May 2010. doi:10.1128/AAC.01135-09.

[242] May, J., Meyer, C. G., Grossterlinden, L., Ademowo, O. G., Mockenhaupt, F. P., Olumese, P. E., Falusi, A. G., Luzzatto, L., and Bienzle, U. Red cell glucose-6-phosphate dehydrogenase status and pyruvate kinase activity in a Nigerian population. *Trop Med Int Health*, 5(2):119–23, Jan 2000.

[243] Johnson, M. K., Clark, T. D., Njama-Meya, D., Rosenthal, P. J., and Parikh, S. Impact of the method of G6PD deficiency assessment on genetic association studies of malaria susceptibility. *PLoS ONE*, 4(9):e7246, Dec 2009. doi:10.1371/journal.pone.0007246.

[244] White, N. J. The role of anti-malarial drugs in eliminating malaria. *Malar J*, 7 Suppl 1:S8, Jan 2008. doi:10.1186/1475-2875-7-S1-S8.

[245] Plenge, R. M., Hendrich, B. D., Schwartz, C., Arena, J. F., Naumova, A., Sapienza, C., Winter, R. M., and Willard, H. F. A promoter mutation in the XIST gene in two unrelated families with skewed X-chromosome inactivation. *Nat Genet*, 17(3):353–6, Nov 1997. doi:10.1038/ng1197-353.

[246] Boyer, S., Porter, I., and Weilbacher, R. Electrophoretic heterogeneity of glucose-6-phosphate dehydrogenase and its relationship to enzyme deficiency in man. *Proc Natl Acad Sci USA*, 48:1868–76, Oct 1962.

[247] Glock, G. and McLean, P. Glucose-6-phosphate dehydrogenase activity of rat liver. *Nature*, 170(4316):119–20, Jul 1952.

[248] Beutler, E. A series of new screening procedures for pyruvate kinase deficiency, glucose-6-phosphate dehydrogenase deficiency, and glutathione reductase deficiency. *Blood*, 28(4):553–62, Sep 1966.

[249] Dickens, F. and Glock, G. Direct oxidation of glucose-6-phosphate, 6-phosphogluconate and pentose-5-phosphates by enzymes of animal origin. *Biochem J*, 50(1):81–95, Nov 1951.

[250] Guilbault, G. G. Fluorometric determination of dehydrogenase activity using resorufin. *Meth Enzymol*, 41:53–6, Jan 1975.

[251] Hess, R., Scarpelli, D., and Pearse, A. Cytochemical localization of pyridine nucleotide-linked dehydrogenases. *Nature*, 181(4622):1531–2, May 1958.

[252] Jalloh, A., Tantular, I. S., Pusarawati, S., Kawilarang, A. P., Kerong, H., Lin, K., Ferreira, M. U., Matsuoka, H., Arai, M., Kita, K., et al. Rapid epidemiologic assessment of glucose-6-phosphate dehydrogenase deficiency in malaria-endemic areas in Southeast Asia using a novel diagnostic kit. *Trop Med Int Health*, 9(5):615–23, Apr 2004. doi:10.1111/j.1365-3156.2004.01237.x.

[253] Gall, J. C., Brewer, G. J., and Dern, R. J. Studies of Glucose-6-Phosphate Dehydrogenase Activity of Individual Erythrocytes: The Methemoglobin-Elution Test for Identification of Females Heterozygous for G6PD Deficiency. *Am J Hum Genet*, 17(4):359–68, Jul 1965.

[254] van Noorden, C. J., Dolbeare, F., and Aten, J. Flow cytofluorometric analysis of enzyme reactions based on quenching of fluorescence by the final reaction product: detection of glucose-6-phosphate dehydrogenase deficiency in human erythrocytes. *J Histochem Cytochem*, 37(9):1313–8, Aug 1989.

[255] Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., Spidlen, J., Strain, E., and Gentleman, R. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 10:106, Jan 2009. doi:10.1186/1471-2105-10-106.

[256] Rinaldi, A., Filippi, G., and Siniscalco, M. Variability of red cell phenotypes between and within individuals in an unbiased sample of 77 heterozygotes for G6PD deficiency in Sardinia. *Am J Hum Genet*, 28(5):496–505, Sep 1976.

[257] Evelyn, K. A. and Malloy, H. T. Microdetermination of oxyhemoglobin, methemoglobin, and sulfhemoglobin in a single sample of blood. *Journal of Biological Chemistry*, Jan 1938.

[258] Betke, K. Peroxidase activity of the hemoglobin. *Biochem Z*, 321(4):271–91, Jan 1951.

[259] Kleihauer, E. and Betke, K. Elution procedure for the demonstration of methaemoglobin in red cells of human blood smears. *Nature*, 199:1196–7, Sep 1963.

[260] Grinstein, M. and Watson, C. J. Studies of protoporphyrin; a simple and improved method for the preparation of pure protoporphyrin from hemoglobin. *J Biol Chem*, 167(2):515–9, Feb 1947.

[261] Granick, S. and Levere, R. D. The intracellular localization of heme by a fluorescence technique. *J Cell Biol*, 26(1):167–76, Jul 1965.

[262] Schwartz, S., Dahl, J., Ellefson, M., and Ahlquist, D. The "HemoQuant" test: a specific and quantitative determination of heme (hemoglobin) in feces and other materials. *Clin Chem*, 29(12):2061–7, Dec 1983.

[263] Nagababu, E., Chrest, F. J., and Rifkind, J. M. The origin of red cell fluorescence caused by hydrogen peroxide treatment. *Free Radic Biol Med*, 29(7):659–63, Oct 2000.

[264] Flynn, T. P., Allen, D. W., Johnson, G. J., and White, J. G. Oxidant damage of the lipids and proteins of the erythrocyte membranes in unstable hemoglobin disease. Evidence for the role of lipid peroxidation. *J Clin Invest*, 71(5):1215–23, May 1983.

[265] Kikugawa, K., Kosugi, H., and Asakura, T. Effect of malondialdehyde, a product of lipid peroxidation, on the function and stability of hemoglobin. *Arch Biochem Biophys*, 229(1):7–14, Feb 1984.

[266] Nagababu, E., Mohanty, J. G., Bhamidipaty, S., Ostera, G. R., and Rifkind, J. M. Role of the membrane in the formation of heme degradation products in red blood cells. *Life Sci*, 86(3-4):133–8, Jan 2010. doi:10.1016/j.lfs.2009.11.015.

[267] Tinley, K. E., Loughlin, A. M., Jepson, A., and Barnett, E. D. Evaluation of a rapid qualitative enzyme chromatographic test for glucose-6-phosphate dehydrogenase deficiency. *Am J Trop Med Hyg*, 82(2):210–4, Feb 2010. doi:10.4269/ajtmh.2010.09-0416.

[268] Mackinnon, M. J., Mwangi, T. W., Snow, R. W., Marsh, K., and Williams, T. N. Heritability of malaria in Africa. *PLoS Med*, 2(12):e340, Dec 2005. doi:10.1371/journal.pmed.0020340.

[269] Martin, S. K., Miller, L. H., Alling, D., Okoye, V. C., Esan, G. J., Osunkoya, B. O., and Deane, M. Severe malaria and glucose-6-phosphate-dehydrogenase deficiency: a reappraisal of the malaria/G-6-P.D. hypothesis. *Lancet*, 1(8115):524–6, Mar 1979.

[270] Leslie, T., Briceño, M., Mayan, I., Mohammed, N., Klinkenberg, E., Sibley, C. H., Whitty, C. J. M., and Rowland, M. The impact of phenotypic and genotypic G6PD deficiency on risk of plasmodium vivax infection: a case-control study amongst Afghan refugees in Pakistan. *PLoS Med*, 7(5):e1000283, Jan 2010. doi:10.1371/journal.pmed.1000283.

[271] Greenwood, B., Marsh, K., and Snow, R. Why do some African children develop severe malaria? *Parasitol Today (Regul Ed)*, 7(10):277–81, Oct 1991.

[272] Araujo, C. D., Migot-Nabias, F., Guitard, J., Pelleau, S., Vulliamy, T., and Ducrocq, R. The role of the G6PD AEth376G/968C allele in glucose-6-phosphate dehydrogenase deficiency in the seerer population of Senegal. *Haematologica*, 91(2):262–3, Feb 2006.

[273] Beutler, E., Kuhl, W., Sáenz, G. F., and Rodríguez, W. Mutation analysis of glucose-6-phosphate dehydrogenase (G6PD) variants in Costa Rica. *Hum Genet*, 87(4):462–4, Jul 1991.

[274] Mason, P. J., Bautista, J. M., and Gilsanz, F. G6PD deficiency: the genotype-phenotype association. *Blood Rev*, 21(5):267–83, Aug 2007. doi:10.1016/j.blre.2007.05.002.

[275] Segeja, M. D., Mmbando, B. P., Kamugisha, M. L., Akida, J. A., Savaeli, Z. X., Minja, D. T., Msangeni, H. A., and Lemnge, M. M. Prevalence of glucose-6-phosphate dehydrogenase deficiency and haemoglobin S in high and moderate malaria transmission areas of Muheza, north-eastern Tanzania. *Tanzania journal of health research*, 10(1):9–13, Dec 2008.

[276] Brabin, B. J., Prinsen-Geerligs, P. D., Verhoeff, F. H., Fletcher, K. A., Chimsuku, L. H. E., Ngwira, B. M., Leich, O. J., and Broadhead, R. L. Haematological profiles of the people of rural southern Malawi: an overview. *Ann Trop Med Parasitol*, 98(1):71–83, Dec 2004. doi:10.1179/000349804225003055.

[277] Oduro, A. R., Koram, K. A., Rogers, W., Atuguba, F., Ansah, P., Anyorigiya, T., Ansah, A., Anto, F., Mensah, N., Hodgson, A., et al. Severe falciparum malaria in young children of the Kassena-Nankana district of northern Ghana. *Malar J*, 6:96, Dec 2007. doi:10.1186/1475-2875-6-96.

[278] Ackerman, H., Usen, S., Jallow, M., Sisay-Joof, F., Pinder, M., and Kwiatkowski, D. P. A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. *Ann Hum Genet*, 69(Pt 5):559–65, Sep 2005. doi:10.1111/j.1529-8817.2005.00180.x.

[279] Allen, S. J., O'Donnell, A., Alexander, N. D., Alpers, M. P., Peto, T. E., Clegg, J. B., and Weatherall, D. J. alpha+-Thalassemia protects children against disease caused by other infections as well as malaria. *Proc Natl Acad Sci USA*, 94(26):14736–41, Dec 1997.

[280] Williams, T. N., Wambua, S., Uyoga, S., Macharia, A., Mwacharo, J. K., Newton, C. R. J. C., and Maitland, K. Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal Plasmodium falciparum malaria on the coast of Kenya. *Blood*, 106(1):368–71, Jul 2005. doi:10.1182/blood-2005-01-0313.

[281] Beutler, E., Dern, R., and Flanagan, C. Effect of sickle-cell trait on resistance to malaria. *Br Med J*, 1(4923):1189–91, May 1955.

[282] Carlson, J., Nash, G. B., Gabutti, V., al Yaman, F., and Wahlgren, M. Natural protection against severe Plasmodium falciparum malaria due to impaired rosette formation. *Blood*, 84(11):3909–14, Dec 1994.

[283] Paglialunga, F., Fico, A., Iaccarino, I., Notaro, R., Luzzatto, L., Martini, G., and Filosa, S. G6PD is indispensable for erythropoiesis after the embryonic-adult hemoglobin switch. *Blood*, 104(10):3148–52, Nov 2004. doi:10.1182/blood-2004-03-0835.

[284] Morahan, G., Boutlis, C. S., Huang, D., Pain, A., Saunders, J. R., Hobbs, M. R., Granger, D. L., Weinberg, J. B., Peshu, N., Mwaikambo, E. D., et al. A promoter polymorphism in the gene encoding interleukin-12 p40 (IL12B) is associated with mortality from cerebral malaria and with reduced nitric oxide production. *Genes Immun*, 3(7):414–8, Nov 2002. doi:10.1038/sj.gene.6363909.

[285] Gramaglia, I., Sobolewski, P., Meays, D., Contreras, R., Nolan, J. P., Frangos, J. A., Intaglietta, M., and van der Heyde, H. C. Low nitric oxide bioavailability contributes to the genesis of experimental cerebral malaria. *Nat Med*, 12(12):1417–22, Dec 2006. doi:10.1038/nm1499.

[286] Yeo, T. W., Lampah, D. A., Tjitra, E., Gitawati, R., Darcy, C. J., Jones, C., Kenangalem, E., McNeil, Y. R., Granger, D. L., Lopansri, B. K., et al. Increased asymmetric dimethylarginine in severe falciparum malaria: association with impaired nitric oxide bioavailability and fatal outcome. *PLoS Pathog*, 6(4):e1000868, Apr 2010. doi:10.1371/journal.ppat.1000868.

[287] Ho, H.-Y., Cheng, M.-L., Weng, S.-F., Chang, L., Yeh, T.-T., Shih, S.-R., and Chiu, D. T.-Y. Glucose-6-phosphate dehydrogenase deficiency enhances enterovirus 71 infection. *J Gen Virol*, 89(Pt 9):2080–9, Sep 2008. doi:10.1099/vir.0.2008/001404-0.

[288] Mordmüller, B., Turrini, F., Long, H., Kremsner, P. G., and Arese, P. Neutrophils and monocytes from subjects with the Mediterranean G6PD variant: effect of Plasmodium falciparum hemozoin on G6PD activity, oxidative burst and cytokine production. *Eur Cytokine Netw*, 9(3):239–45, Sep 1998.