Whole-genome analysis of Malawian *Plasmodium falciparum* isolates identifies potential targets of allele-specific immunity to clinical malaria

Zalak Shah¹, Myo T. Naung²,³,⁴, Kara A. Moser⁵, Matthew Adams¹, Andrea G. Buchwald¹, Ankit Dwivedi⁵, Amed Ouattara¹, Karl B Seydel⁶,⁷, Don P. Mathanga⁸, Alyssa E. Barry²,³,⁴,⁹, David Serre⁵, Miriam K. Laufer¹, Joana C. Silva⁸, and Shannon Takala-Harrison¹

¹Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, MD, 21201, USA.  
²Population Health and Immunity Division, Walter Eliza Hall of Medical Institute for Medical Research, Parkville, Victoria, Australia  
³Department of Medical Biology, University of Melbourne, Carlton, Victoria, Australia  
⁴School of Medicine, Deakin University, Geelong, Victoria, Australia  
⁵Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, 21201, USA.  
⁶Department of Osteopathic Medical Specialties, College of Osteopathic Medicine, Michigan State University, East Lansing, Michigan, 48824, USA.  
⁷Blantyre Malaria Project, University of Malawi College of Medicine, Blantyre, Malawi, Africa.  
⁸University of Malawi College of Medicine, Malaria Alert Centre, Blantyre, Malawi.  
⁹Disease Elimination and Maternal and Child Health, Burnet Institute, Melbourne, Victoria, Australia

Corresponding author: Shannon Takala-Harrison stakala@som.umaryland.edu
Abstract

Individuals acquire immunity to clinical malaria after repeated *Plasmodium falciparum* infections. This immunity to disease is thought to reflect the acquisition of a repertoire of responses to multiple alleles in diverse parasite antigens. In previous studies, we identified polymorphic sites within individual antigens that are associated with parasite immune evasion by examining antigen allele dynamics in individuals followed longitudinally. Here we expand this approach by analyzing genome-wide polymorphisms using whole genome sequence data from 140 parasite isolates representing malaria cases from a longitudinal study in Malawi and identify 25 genes that encode likely targets of naturally acquired immunity and that should be further characterized for their potential as vaccine candidates.
INTRODUCTION

Despite recent progress in reducing the burden of malaria, this disease remains a leading cause of mortality worldwide, resulting in an estimated 405,000 deaths in 2018 (WHO | World Malaria Report 2018, n.d.). In areas with high transmission of Plasmodium falciparum, individuals develop immunity to malaria (Marsh & Kinyanjui, 2006). This immunity does not provide sterile protection against all infections, but decreases the risk of clinical disease, and increases with age as individuals are repeatedly exposed to the parasite (Cowman et al., 2016; Ryg-Cornejo et al., 2016). This age-related pattern of immunity to disease is thought to reflect the need for a repertoire of immune responses to multiple alleles in diverse parasite antigens (Buchwald, Sorkin, et al., 2019; Cowman et al., 2016; Marsh & Kinyanjui, 2006; Portugal et al., 2013). The extensive genetic diversity in P. falciparum surface antigens is thought to have evolved over millennia as a means of parasite immune evasion (Weedall & Conway, 2010). Allele-specific immune responses have been demonstrated for several parasite antigens (Barry et al., 2011; Cortés et al., 2005; Crompton et al., 2010; Dutta et al., 2007; Early et al., 2018; Osier et al., 2008; Polley et al., 2007; Tran et al., 2014). In previous work, we examined parasite alleles in repeated infections occurring in individuals followed longitudinally and identified specific polymorphic sites within parasite surface antigens (i.e. AMA1 and MSP1) where amino acid changes were associated with immune escape and increased risk of disease, consistent with allele-specific acquisition of immunity to these antigens (Takala et al., 2007, 2009). Furthermore, malaria subunit vaccines based on a single antigen allele have displayed greater efficacy against parasites with alleles matching the vaccine strain compared to the diverse alleles observed in natural parasite populations (Ouattara et al., 2013, 2015; Takala et al., 2009; Thera et al., 2011). Such allele-specific vaccine efficacy could lead to poor overall vaccine efficacy when the vaccine target allele is at low frequency in the parasite population and could result in selection of non-vaccine alleles capable of vaccine escape (Ouattara et al., 2015). Overcoming this scenario may require the design of a multivalent
malaria vaccine (Ouattara et al., 2015). However, the design of such a vaccine is hampered by an incomplete knowledge of which parasite proteins are targets of acquired natural immunity.

The advent of technologies allowing whole genome sequencing at epidemiological scales has allowed investigators to transition from investigation of single antigens to performing genome-wide screens to identify loci likely to be involved in the acquisition of protective immunity to malaria, including uncharacterized genes encoding products of unknown function (Gardner et al., 2002; PlasmoDB: The Plasmodium Genomics Resource, n.d.). Although there have been previous genome-wide studies in *P. falciparum* to identify genomic signatures of balancing or diversifying selection at a population level (Amambua-Ngwa et al., 2012; Mobegi et al., 2014; Mu et al., 2007), these studies do not associate identified signatures with clinical outcomes at an individual level, making it difficult to directly link such signatures to immune selection. Expanding on our previous approach where we examined the dynamics of vaccine antigen alleles in individuals’ repeated infections over time in relation to the development of symptoms (Takala et al., 2007, 2009), we compared whole genome sequence data generated from *P. falciparum* infections collected from participants in a longitudinal cohort study conducted in Malawi to identify targets of allele-specific immunity to malaria. Specifically, we compared the frequency of parasite alleles in symptomatic infections occurring in individuals with different levels of malaria immunity to identify significantly differentiated sites, and also compared alleles in repeated infections within an individual versus between individuals to identify polymorphic sites that vary most within individuals. Genes identified using both approaches were considered likely immune targets and were further examined for their potential as vaccine candidates. As a proof of concept of the utility of our approach in identifying targets of allele-specific immune responses, we compared the frequency of alleles in one of the identified antigens (previously considered as a potential vaccine candidate) in individuals with different levels of malaria immunity to test the hypothesis that individuals who are more immune become ill when infected with a parasite having rarer alleles to which they have not yet developed immunity.
RESULTS

Participant/infection characteristics and definition of immune status

To identify targets of allele-specific immunity to malaria, we generated whole-genome sequence data from 140 parasite isolates collected from symptomatic infections occurring in participants in a longitudinal cohort study in Malawi (Buchwald, Sixpence, et al., 2019).

Although age is often used as a proxy for immune status in high transmission areas (Egan et al., 1996; Metzger et al., 2003; Perraut et al., 2017; Tran et al., 2014), this metric does not account for heterogeneous exposure to infectious mosquito bites, which has been observed in endemic areas (Bejon et al., 2010; Bousema et al., 2010; Clark et al., 2008; Elissa et al., 2003; Gaudart et al., 2006; Kang et al., 2018). To better account for heterogeneous exposure at an individual level, we used the proportion of total infections that were symptomatic over the two-year study period to categorize immune status, using the median as a cutoff to define high and low immunity groups. Although these groups consisted of individuals with a range of ages, the median age (13.18 years) of individuals in the group with high immunity was significantly greater

![Figure 1](https://doi.org/10.1101/2020.09.16.20196253) This version posted September 18, 2020. The copyright holder for this [preprint](https://doi.org/10.1101/2020.09.16.20196253) (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.
than the median age (7.27 years) of individuals in the low immunity group (p-value = 0.0001, Wilcoxon rank sum test) (Figure 1, Table 1), as would be expected in a high malaria transmission setting such as Malawi.

Only one infection from each individual was included in comparisons between the high and low immunity groups, with samples selected in a manner to reduce temporal variability between infections (see Methods). DEploid-IBD (Zhu et al., 2017) was used to estimate the proportion of each clone within an infection. Infections without a predominant clone (i.e., where the majority clone had a frequency <60% within the infection) were defined as complex infections. Although the median frequency of the majority clone was not significantly different between infections in the two immunity groups (Supplementary Fig 1, p-value = 0.3372, Wilcoxon rank sum test), the high immunity group had a greater number of complex infections (n=7) than the low immunity group (n=2). These nine complex infections were excluded from further analysis to avoid confounding by infection complexity and misclassification of alleles likely contributing to clinical illness.

Table 1. Participant/infection characteristics in high and low immunity groups*

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>High Immunity (n=28)</th>
<th>Low Immunity (n=33)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median age (years)</td>
<td>13.18</td>
<td>7.27</td>
<td>0.0002*</td>
</tr>
<tr>
<td>Male (%)</td>
<td>36</td>
<td>52</td>
<td>0.33*</td>
</tr>
<tr>
<td>Median parasitemia (parasites/uL)</td>
<td>4710</td>
<td>48200</td>
<td>7.74e-06*</td>
</tr>
<tr>
<td>Median genome coverage 20x (%)</td>
<td>88.73</td>
<td>90.14</td>
<td>0.2673*</td>
</tr>
<tr>
<td>Median depth of coverage</td>
<td>130x</td>
<td>156x</td>
<td>0.2485*</td>
</tr>
</tbody>
</table>

*High complexity infections lacking a predominant parasite clone were excluded
*P-value determined using z-score test for difference in proportions
*P-value determined using Wilcoxon rank sum test
The median parasite density of infections in the high immunity group was significantly lower than in the low immunity group (Table 1, p-value = \(7.737 \times 10^{-6}\), Wilcoxon rank sum test).

However, there was no significant difference in the percentage of the parasite genome with at least 20-fold coverage (Table 1, p-value = 0.2673, Wilcoxon rank sum test), or in the median average depth of coverage (Table 1, p-value = 0.2485, Wilcoxon rank sum test) between whole genome sequence data generated from infections in the two groups.

**Differentiated loci between groups with different levels of immunity to clinical malaria**

We hypothesized that, because of allele-specific immune responses, individuals with greater protective immunity to malaria would experience disease when infected with parasite antigen alleles that are rarer in the parasite population, having already developed immunity to more common alleles circulating in the population. Thus, we would expect significant genetic differentiation between antigen alleles in individuals with high versus low immunity at loci that are targets of allele-specific immunity.

To test this hypothesis, Wright’s fixation index \(F_{ST}\), a measure of genetic differentiation between two populations (Bhatia et al., 2013), was estimated per non-synonymous single nucleotide polymorphism (SNP) to identify genetically differentiated sites between parasites from the high and low immunity groups with the significance threshold based on 10,000 permutations. We identified 160 sites (in 145 genes) in the parasite genome that were significantly differentiated between the two immunity groups \((p\text{-value} \leq 0.0095, \text{Figure 2a, Supplementary Table 1})\). Fifty-five of the genes containing significantly differentiated sites (38%) encode proteins of unknown function that are not associated with any computed or curated molecular function or biological process based on Gene Ontology (Supplementary Table 1). These 145 gene products included some proteins previously identified as potential vaccine candidates, including AMA1 (apical membrane antigen 1) (Takala et al., 2009; Thera et al., 2011, p. 20), ASP (apical sushi protein,
PF3D7_0405900 (Vanegas et al., 2014), CLAG8 (cytoadherence linked asexual protein 8, PF3D7_0831600) (Gupta et al., 2015; Iriko et al., 2008), SLARP (sporozoite and liver asparagine-rich protein, PF3D7_1147000) (van Schaijk et al., 2014) and a conserved protein of unknown function (PF3D7_1359000) (Krzyczmonik et al., 2012).

Figure 2. Genetic differentiation between parasites from high immunity vs. low immunity groups. a) Genome-wide genetic differentiation ($F_{ST}$) between parasites from individuals with higher immunity vs. lower immunity. Each point represents a variable, non-synonymous site. Results are plotted as $-\log_{10}$ $p$-values on the y-axis. The color of each point represents the $F_{ST}$ value, with darker points indicating higher $F_{ST}$ values. The dashed line denotes statistical significance ($p$-value = 0.0095), with $p$-value determined by permutation. b) Nucleotide diversity for significantly differentiated SNPs in parasites from individuals with higher immunity and lower immunity. c) Box-plot of mean allele frequency per individual based on SNPs in the whole genome sequences, and which are significantly differentiated SNPs from (a). Red indicates the high immunity group and blue color indicates the low immunity group.

To further test the hypothesis that individuals who are more immune become symptomatic when infected with parasites having antigen alleles that are rarer in the parasite population, we estimated nucleotide diversity at significantly differentiated non-synonymous sites in parasites from the two immunity groups and observed a significantly greater median nucleotide diversity in
parasites from the high immunity group compared to parasites from the low immunity group (Figure 2b, \( p\text{-value} = 4.74 \times 10^{-05} \), Wilcoxon rank sum test). In addition, we estimated the average frequency of alleles in each infection at both significantly differentiated sites and genome-wide variable sites. The median frequency of alleles at genome-wide variable sites was not significantly different between immunity groups (Figure 2c, \( p\text{-value} = 0.8801 \), Wilcoxon rank sum test); in contrast, at the differentiated sites, the median frequency of alleles was significantly lower in the high immunity group compared to the low immunity group (Figure 2c, \( p\text{-value} = 0.0072 \), Wilcoxon rank sum test). These results are consistent with the scenario that individuals in the high immunity group are infected with parasites having different lower-frequency alleles compared to individuals in the low immunity group who are infected with parasites sharing more common alleles.

Within 23 polyclonal infections, we also compared the proportion of mismatched alleles between the predominant and minor clones at both significantly differentiated sites and genome-wide variable sites in order to assess whether these clones differ at sites thought to be relevant for immunity. We observed a significantly greater median proportion of mismatches between the major and minor clones within an infection at the differentiated sites compared to genome-wide variable sites (Supplementary Figure 2, \( p\text{-value} = 8 \times 10^{-05} \), Wilcoxon rank sum test), consistent with the hypothesis that the predominant clone represents a breakthrough infection that has escaped allele-specific immune responses that maintain minor clones at a subclinical level.

**Loci that differ more within individuals than between individuals**

We expected that allele-specific immune responses would result in a greater proportion of genetic differences in parasites causing multiple symptomatic infections within an individual compared to parasites causing infection in different individuals, at antigenic loci that are targets of immunity. To identify regions of the parasite genome that are most different in infections occurring within an individual *versus* between individuals, we compared the allelic states at each
variable non-synonymous site across the genome between pairs of isolates sampled within an individual and between individuals and estimated the proportion of mismatches per site in each group (Figure 3a). The distribution of the difference in the proportion of mismatches across all

**Figure 3.** Analysis of mismatches in paired samples within and between individuals. a) Illustration of analysis to identify regions of the genome that vary more in parasites causing illness within the same individual over time (within individuals) compared to random pairs of parasites in the population (between individuals). b) Distribution of differences in the proportion of mismatched alleles in the within group and the between group. The difference was calculated as the proportion of mismatches at each non-synonymous SNP in the within group minus the proportion of mismatches at each non-synonymous SNP in the between group. The dashed black line indicates the threshold for the top 1% most different SNPs in the within group compared to the between group.
non-synonymous variable sites between the two groups (within minus between) is shown in Figure 3b (see also Supplementary Fig 3a). The difference in the proportion of mismatches at each site had a median and mode equal to zero, indicating that the proportion of mismatches was not different within and between individuals at most sites. There was no significant correlation between the number of days between infections in a pair and the proportion of mismatches (Supplementary Fig 2b, Pearson's correlation $r = 0.065$, $p$-value $= 0.3043$), suggesting that time was not a significant confounding factor in the analysis. We further examined the top 1% of sites that differed most within individuals compared to between individuals, which included 223 SNPs, located in 173 genes (Supplementary Table 2). Sixty-eight (39%) of these 173 genes encode proteins of unknown function that are not associated with any computed or curated molecular function or biological process based on Gene Ontology, and at least 15 (8.7%) encode for proteins that have been previously identified as potential vaccine candidates.

Loci identified as likely targets of immunity by both analytical approaches

Twenty-five genes were identified by both analytical approaches, of which 11 (44%) encode proteins of unknown function (Table 2). Based on publicly available data, 20 of the 25 genes have a moderate to high level of expression in the erythrocytic stage of the parasite life cycle (Toenhake et al., 2018) (Supplementary Table 3). Eight genes have a mutagenesis index score near zero, suggesting that they are likely essential (Zhang et al., 2018), and at least 12 genes have either a known or a predicted transmembrane domain or a signal peptide (PlasmoDB: A Functional Genomic Database for Malaria Parasites. - PubMed - NCBI, n.d.). Among these 25 gene products are proteins whose putative function make them biologically plausible vaccine candidates, including SURFIN4.2 (thought to be involved in formation of the moving junction during erythrocyte invasion) (Quintana et al., 2018), as well as members of the CLAG and PHIST multigene families (both thought to have a role in parasite cytoadherence) (Holt et al., 1999; Proellocks et al., 2014).
Table 2. Gene products identified as likely targets of allele-specific immunity to malaria based on two analytical approaches.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Annotation</th>
<th>GO Function</th>
<th>GO Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF3D7_0311900</td>
<td>heptatricopeptide repeat-containing protein, putative</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_0312500</td>
<td>major facilitator superfamily-related transporter, putative</td>
<td>null</td>
<td>transmembrane transport</td>
</tr>
<tr>
<td>PF3D7_0318200</td>
<td>DNA-directed RNA polymerase II subunit RPB1</td>
<td>DNA-directed 5'-3' RNA polymerase activity</td>
<td>transcription</td>
</tr>
<tr>
<td>PF3D7_0412300</td>
<td>phosphopantothenoylcysteine synthetase, putative</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_0421700</td>
<td>conserved Plasmodium protein, unknown function</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_0424400</td>
<td>surface-associated interspersed protein 4.2 (SURFIN 4.2)</td>
<td>host cell surface binding</td>
<td>entry into host cell</td>
</tr>
<tr>
<td>PF3D7_0511500</td>
<td>RNA pseudouridylate synthase, putative</td>
<td>RNA binding, pseudouridine synthase activity</td>
<td>RNA modification, pseudouridine synthesis</td>
</tr>
<tr>
<td>PF3D7_0522400</td>
<td>conserved Plasmodium protein, unknown function</td>
<td>null</td>
<td>protein localization/transport</td>
</tr>
<tr>
<td>PF3D7_0526600</td>
<td>conserved Plasmodium protein, unknown function</td>
<td>null</td>
<td>lipid biosynthetic process</td>
</tr>
<tr>
<td>PF3D7_0605600</td>
<td>nucleoside diphosphate kinase, putative</td>
<td>nucleoside diphosphate kinase activity</td>
<td>nucleoside diphosphate phosphorylation</td>
</tr>
<tr>
<td>PF3D7_0619600</td>
<td>conserved Plasmodium protein, unknown function</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_0704600</td>
<td>E3 ubiquitin-protein ligase</td>
<td>ubiquitin-protein transferase activity</td>
<td>response to drug</td>
</tr>
<tr>
<td>PF3D7_0710200</td>
<td>conserved Plasmodium protein, unknown function</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_0807700</td>
<td>serine protease DegP</td>
<td>serine-type endopeptidase activity</td>
<td>response to oxidative stress and temperature stimulus</td>
</tr>
<tr>
<td>PF3D7_0831600</td>
<td>cytoadherence linked asexual protein 8</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_0914300</td>
<td>met-10+ like protein, putative</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_1004200</td>
<td>WD repeat-containing protein, putative</td>
<td>protein binding</td>
<td>transport</td>
</tr>
<tr>
<td>PF3D7_1030400</td>
<td>conserved protein, unknown function</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_1033100</td>
<td>S-adenosylmethionine/0rnilthine decarboxylase</td>
<td>adenosyl/methionine decarboxylase activity</td>
<td>spermidine/spermine biosynthetic processes</td>
</tr>
<tr>
<td>PF3D7_1035100</td>
<td>probable protein, unknown function</td>
<td>ATP binding</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_1102500</td>
<td>Plasmodium exported protein (PHISTb), unknown function</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_1149600</td>
<td>DnaJ protein, putative</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>PF3D7_1219100</td>
<td>clathrin heavy chain, putative</td>
<td>clathrin light chain binding, structural molecule activity</td>
<td>clathrin coat assembly, intracellular protein transport, vesicle-mediated transport</td>
</tr>
<tr>
<td>PF3D7_1465800</td>
<td>dynamin beta chain, putative</td>
<td>ATP binding, microtubule motor activity</td>
<td>microtubule-based movement</td>
</tr>
<tr>
<td>PF3D7_1475900</td>
<td>KELT protein</td>
<td>null</td>
<td>null</td>
</tr>
</tbody>
</table>
As a proof of concept of the utility of our approach for identifying targets of allele-specific immunity, we used publicly available sequence data from Pf3K (Pf3k Pilot Data Release 5 / MalariaGEN, n.d.) combined with 156 sequences from Papua New Guinea (PNG) from the MalariaGEN P. falciparum Community Project (MalariaGEN Plasmodium falciparum Community Project, 2016) to examine global diversity in clag8, one of the 25 identified genes that has been considered previously as a potential vaccine candidate, and found high nucleotide diversity, Tajima’s D, and number of segregating sites in the C-terminal region of clag8. This pattern was consistent when comparing data from parasites collected in Malawi to data from parasites from other regions of the world (Figure 4a). A haplotype network generated using clag8 sequences from multiple geographic areas showed no evidence of regional adaptation (Figure 4b), including low $F_{ST}$ values between clag8 sequences from Africa, Asia and PNG (Supplementary Table 5). Further analysis of CLAG8 protein sequences show regions with high protein disorder and B-cell epitope sites, especially in the C-terminal region (Supplementary Fig 4). To test our initial hypothesis that immune individuals become ill when infected with parasite antigen alleles that are rarer in the parasite population, we compared the frequency of clag8 haplotypes based on the C-terminal region of the gene in individuals with different levels of immunity and observed that parasites causing illness in individuals in the high immunity group tended to have clag8 haplotypes that were lower in frequency compared to individuals in the low immunity group, although significance was borderline (Supplementary Fig 5, p-value = 0.099, Wilcoxon rank sum test).
Figure 4. Global diversity of *clag8*. a) Tajima’s D along the *clag8* gene in samples from Thailand, PNG, Malawi, and Ghana. The dotted red lines represent a significant positive Tajima’s D value (≥ 2), suggestive of balancing selection. b) Haplotype network of *clag8* using *P. falciparum* sequence data.
**DISCUSSION**

Previous studies have shown evidence of allele-specific acquisition of immunity to *P. falciparum* in single genes or proteins (Cortés et al., 2005; Early et al., 2018; Osier et al., 2008; Polley et al., 2007; Takala et al., 2007, 2009) that have been identified as potential vaccine antigens based on traditional vaccinology approaches that empirically identify immunogenic proteins. Other studies have performed genome-wide screens to identify genomic signatures of balancing selection but lack individual-level associations with clinical outcomes that might link specific signatures with protective immune responses (Amambua-Ngwa et al., 2012; Mobegi et al., 2014; Mu et al., 2007).

In this study, we conducted a genome-wide, individual-level analysis to identify targets of allele-specific immunity to clinical malaria using *P. falciparum* whole genome sequencing data by identifying parasite genes that are genetically differentiated between individuals with different levels of immunity to malaria and genomic regions that are most different in parasites causing illness within the same individual versus between individuals. Twenty-five genes were identified using both analytical approaches and encode likely targets of allele-specific acquired immunity to clinical malaria, including genes thought to be involved in erythrocyte invasion and cytoadherence, among other functions. Examination of global diversity in clag8, a gene whose product has previously been considered as a vaccine candidate, provided evidence of immune selection in the C-terminal region of the gene and did not indicate geographical differences that might be indicative of local adaptation or genetic drift. Consistent with the hypothesis of allele-specific acquisition of immunity, clag8 haplotypes of parasites causing illness in individuals with greater protective immunity were lower in frequency compared to those causing illness in less immune individuals. These findings support the utility of our approach for identification of targets of allele-specific immunity and the further investigation of these 25 loci as potential vaccine candidates.

In this study, we estimated the genetic complexity of infections and observed that individuals in the group with higher immunity generally had more complex infections than...
individuals with lower immunity. Although not statistically significant, this pattern is broadly consistent with results from other studies that have reported associations between infection complexity and either age and/or risk of disease (Färnert et al., 1999; Takala et al., 2007). The greater complexity of infections in individuals with higher immunity is consistent with the idea that immune individuals are capable of maintaining some parasite clones at a subclinical level and could suggest a possible role of polyclonal infections in maintaining immunity through continuous exposure to different strains (Bereczky et al., 2007; Sondén et al., 2015). In polyclonal infections from our data set, we observed that the predominant clone was more likely to have different alleles than the minor clone at sites identified as potential targets of immunity in our analysis. This finding is consistent with the hypothesis that minority clones are maintained at a subclinical level by acquired immunity, while the predominant clone is able to escape the immune response (because it has unrecognized alleles), resulting in a symptomatic infection. Prior to downstream analysis, infections lacking a predominant clone were excluded to avoid confounding by infection complexity and misclassification of clones likely responsible for disease symptoms. The exclusion of infections lacking a predominant clone led to the removal of a greater proportion of infections from the high immunity group than the low immunity group, which may have resulted in an underestimation of parasite diversity in the high immunity group. However, we do not believe this underestimation impacted our conclusions, as infections within the high immunity group were significantly more diverse at differentiated loci even after exclusion of these highly complex infections.

We hypothesized that individuals who have a higher degree of immunity to malaria would have a greater risk of malaria symptoms when infected with parasites having protein variants that are of lower frequency in the local parasite population, owing to their having already acquired immunity to variants encoded by more common alleles. When we estimated nucleotide diversity at differentiated loci in both immunity groups, we found significantly greater diversity in the group with higher immunity compared to the group with lower immunity. This difference in diversity may
reflect the fact that individuals with higher immunity have symptomatic infections with parasite alleles that are rarer in the population and therefore more likely to be different from one another at differentiated loci. This scenario is also supported by our finding that the median frequency of the infecting allele was significantly lower in more immune individuals compared to less immune individuals when examining differentiated loci. These results also demonstrate the importance of accounting for rare alleles in vaccine design, as they may also lead to escape from vaccine-induced immunity.

Genome-wide screens for signatures of balancing selection have previously identified some of the 25 genes that we identified in this study, including PF3D7_0710200, clag8 and other genes from the surfin and phist multigene families (Amambua-Ngwa et al., 2012; Mobegi et al., 2014; Mu et al., 2007). PF3D7_0710200, a gene encoding a conserved protein of unknown function, was identified by three separate studies as a potential immune target (Amambua-Ngwa et al., 2012; Mobegi et al., 2014; Mu et al., 2007); however, little is known about the function of this protein. Indeed, 11 of the 25 genes (44%) identified in this study encode for proteins of unknown function, highlighting the importance of further genetic screens to determine the role of such genes, which make up ~35% of the parasite genome (Sexton et al., 2019). A recent study has suggested that most genes involved in host-parasite interactions, including many known antigens, are non-essential (Zhang et al., 2018); however, eight of the 25 genes identified in this study would be considered essential genes, based on their low mutagenesis index score (Zhang et al., 2018) (Supplementary Table 4). Such essential genes may be attractive vaccine targets as there may be less redundancy in function that would allow the parasite to adapt and escape vaccine-induced inhibition.

As a proof of concept of the utility of our approach for identifying targets of allele-specific immunity, we further examined the global diversity of one of the 25 identified genes, clag8, which has previously been considered as a malaria vaccine candidate. clag8 belongs to the clag multigene family and is one of the least studied genes in the clag family. clag8 is highly expressed
during the first few hours following erythrocyte invasion (early ring stage), displays decreased expression from 5 hours to 35 hours post invasion, and then increased expression during the schizont stage (Toenhake et al., 2018). It is thought to be part of the RhopH complex, formed by the members of rhoph1/clag gene families. The RhopH complex is an erythrocyte-binding protein complex inside the rhoptry and has been suggested to play an important role in establishment of the parasitophorous vacuole (Iriko et al., 2008; Kaneko et al., 2005; Sam-Yellowe & Perkins, 1991). Previous studies have reported evidence of positive diversifying selection in this gene, with high nucleotide diversity and a high proportion of non-synonymous substitutions per site (dN) (Iriko et al., 2008). Our global analysis of clag8 diversity displayed high nucleotide diversity and Tajima’s D values in the C-terminal region of the gene, as well as evidence of high protein disorder and predicted B-cell epitopes in the C-terminal region of the protein, suggesting that it is likely to be immunogenic (Guy et al., 2015, p. 20). These results were consistent in all the countries included in the dataset, which represented parasite isolates from three major malaria endemic regions (Pf3k Pilot Data Release 5 / MalariaGEN, n.d.). Additionally, clag8 haplotypes displayed no evidence of geographical adaptation, with major haplotypes being observed in all geographic areas, and seemed to cluster into three main groups. Further studies to identify functional epitopes within the protein and potential cross-reactivity are necessary to determine whether related haplotypes can be grouped into serotypes for the purpose of designing a broadly protective vaccine. At an individual level, and in support of our overarching hypothesis, individuals with higher immunity to malaria were infected with clag8 haplotypes that were less frequent compared to the clag8 haplotypes infecting individuals with lower immunity.

It is noteworthy that our combined list of genes based on both analytical approaches did not include leading blood stage vaccine candidates such as AMA1 and the MSPs, although some of these genes were identified in a single approach. Using a protein microarray, Crompton et al. found that antibody responses to leading vaccine candidates, such as AMA1, MSP1, and MSP2, did not distinguish individuals who were protected from clinical infection versus those who were
not in a cohort of individuals from Mali (Crompton et al., 2010), suggesting the possibility that responses to these proteins are not the primary drivers of clinical immunity. Other studies (Akpogheneta et al., 2008; Kinyanjui et al., 2007; Takala et al., 2009) have supported the hypothesis that responses to antigens such as AMA1 contribute to allele-specific clinical immunity but may be short-lived. In our analyses, infection pairs within an individual were not necessarily consecutive infections, with the time between infections ranging from 27 to 696 days. Although we did not see a significant correlation between the proportion of allele mismatches and the number of days between infections in a pair, it is possible that we could have failed to identify antigens involved in allele-specific acquisition of immunity that elicit short-lived immune responses. Studies with larger sample size would likely be required to distinguish antigens that have different antibody kinetics.

In addition, analyses in this study included only the core genome, owing to our use of selective whole genome amplification to enrich for parasite DNA prior to sequencing, which has been shown to result in poor sequencing coverage in the telomeric and centromeric regions (Oyola et al., 2016; Shah et al., 2020). Limiting analysis to the core genome could have prevented us from identifying members of multigene families that may be important for development of immunity to clinical malaria, as many of these genes are located in these low-coverage regions of the genome. However, because multigene families have been implicated in severe malaria, it is possible that immunity to these diverse antigens may be more relevant to preventing severe rather than uncomplicated malaria.

Here, we describe a promising genome-wide, individual-level approach to identify potential targets of allele-specific immunity to clinical malaria. Using this approach, we identified 25 genes, many of unknown function, that encode proteins that can be further characterized for their potential as candidates for a multivalent subunit malaria vaccine. Although further immunological validation will be necessary to confirm that the patterns observed in this study result from allele-
specific immune responses, our results support the utility of this genomic epidemiology approach to identify new vaccine candidate antigens.

METHODS

Study design and samples

Parasite isolates were collected from participants in a longitudinal cohort study conducted in a rural area in southern Malawi. Details about the participants and study procedures have been described previously by Buchwald et al (Buchwald, Sixpence, et al., 2019). The data analyzed in this study were generated from red blood cell pellets collected from symptomatic, uncomplicated malaria infections. Samples were collected under protocols approved by the ethics committees at the College of Medicine in Blantyre, Malawi, and the University of Maryland, Baltimore, and with the informed consent of the participants or their guardians. The median parasitemia of the sampled infections as determined by microscopy was 21,960 parasites/μL and ranged from 0 parasites/μL (but positive by a rapid diagnostic test) to 241,260 parasites/μL. All samples were confirmed to be positive for P. falciparum via PCR. To ensure only independent infections were included in the analysis, infections within an individual separated by <14 days were excluded.

DNA from red blood cell pellets was extracted using the method of Zainabadi et al (Zainabadi et al., 2017). Extracted DNA was enriched for parasite DNA using an optimized selective whole genome amplification approach described by Shah et al (Shah et al., 2020).

Whole genome sequencing

Genomic DNA libraries were constructed for sequencing using the KAPA Library Preparation Kit (Kapa Biosystems, Woburn, MA). DNA (≥ 200 ng) was fragmented with the Covaris E210 to ~200 bp. Libraries were prepared using a modified version of the manufacturer’s protocol. The DNA was purified between enzymatic reactions and library size selection was performed with AMPure
XT beads. Libraries were assessed for concentration and fragment size using the DNA High
Sensitivity Assay on the LabChip GX (Perkin Elmer, Waltham, MA). Library concentrations were
also assessed by qPCR using the KAPA Library Quantification Kit. Libraries were pooled and
subsequently sequenced on an Illumina HiSeq 4000 (Illumina, San Diego, CA) to generate 150
bp paired-end reads. Sequencing data are available at NCBI under the project numbers and
accession numbers listed in Supplementary Table 6.

Read mapping and SNP Calling

Sequencing data were analyzed by mapping raw fastq files to the 3D7 reference genome using
Bowtie2 (Langmead & Salzberg, 2012). Binary Alignment Map (BAM) files were processed
following the GATK Best Practices workflow to obtain analysis-ready reads (DePristo et al., 2011;
Van der Auwera et al., 2013). Bedtools (Quinlan & Hall, 2010) was used to generate coverage
and depth estimates from the processed reads, and the GATK Best Practices workflow was
followed for variant calling (DePristo et al., 2011; Van der Auwera et al., 2013). Haplotype Caller
was used to create genomic variant call format (GVCF) files for each sample and joint SNP Calling
was performed (GATK v3.7). Variants were removed if they met the following filtering criteria:
variant confidence/quality by depth (QD) < 2.0, strand bias (FS) > 60.0, root mean square of the
mapping quality (MQ) < 40.0, mapping quality rank sum (MQRankSum) < -12.5, read position
rank sum (ReadPosRankSum) < -8.0, quality (QUAL) < 50. Variant sites with >20% missing
genotypes and samples with >30% missing data were additionally removed using vcftools.
Variants were also removed if the minor allele was not present in at least two samples. Only the
core genome was used for further analysis. The median percentage of the genome covered with
at least 20 reads was 88.87% (Shah et al., 2020).

Definition of immune status
The degree of immunity to clinical malaria was defined based on the proportion of symptomatic infections out of all *P. falciparum* infections experienced by each study participant over the course of the two-year study. To account for exposure, individuals with less than five total infections, including symptomatic and asymptomatic infections, were excluded from the analysis. The median proportion of symptomatic infections was used as the cutoff to categorize individuals into higher and lower immunity groups.

**Complexity of infection and genetic differentiation**

Only one infection from each individual was included in comparisons between high and low immunity groups. Infections were selected based on proximity to the median of the distribution of sampling dates to reduce temporal variability. DEploid-IBD (Zhu et al., 2017) was used to estimate the proportion of each clone within an infection. Infections without a predominant clone (i.e., where the majority clone had a frequency <60% within the infection) were defined as complex infections and were excluded from downstream analysis. For the remaining samples, the major allele was called at heterozygous positions if the allele was supported by ≥70% of reads; otherwise, the genotype was coded as missing. A Wilcoxon rank sum test was used to assess differences in the frequency of the majority clone in infections from the two immunity groups.

Vcftools (Danecek et al., 2011) was used to estimate Weir and Cockerham $F_{ST}$ in variable non-synonymous, bi-allelic sites. Significance was determined using 10,000 permutations, where the observed population was resampled without replacement. Nucleotide diversity at significantly differentiated sites was estimated using vcftools (Danecek et al., 2011). PlasmoDB (v44) (*PlasmoDB: The Plasmodium Genomics Resource*, n.d.) was used to identify genes containing differentiated SNPs.

In all polyclonal infections, the major and minor clones (defined by clone frequencies obtained from DEploid-IBD (Zhu et al., 2017)) were compared, provided clone frequency was less than 80% and greater than 10% ($n = 23$). At each non-synonymous site, the proportion of samples
with mismatched alleles from major and minor clones was estimated. The proportion of mismatches was then compared between significantly differentiated sites and all remaining variable sites from the genome. The \( p \)-value was estimated by conducting a Wilcoxon rank sum test to determine if there is a significant difference in mismatches between clones at different sites versus remaining genome-wide variable sites.

**Paired infection analysis**

Individuals with parasite whole genome sequence data from at least two symptomatic infections occurring at least 14 days apart were included in the comparison of infections occurring within the same host to infections occurring in different hosts. Multi-allelic sites were included in the analysis of paired infections, in contrast to analyses of genetic differentiation. The ‘within’ group included all pairs of parasites collected at different time points from the same individual. The ‘between group’, included all pairs of parasites from different individuals. A total of 116 samples were included in this study. The within group contained 124 pairs of samples and the between group contained 6546 pairs of samples. For all pairs, the allelic state was compared at each site and the proportion of pairs with non-matching allelic states was estimated by site (illustrated in Figure 3). The difference between the within group and the between group was calculated by subtracting the proportion of pairs with non-matching allelic states for each site. The \( p \)-value was estimated by conducting a one-sided z-test using the difference in proportion of mismatched alleles between the two groups. PlasmoDB (*PlasmoDB: The Plasmodium Genomics Resource*, n.d.) was used to identify genes containing the SNPs of interest.

**Global diversity in clag8**

The MalariaGEN Pf3K project release 5.1 data (*Pf3k Pilot Data Release 5 | MalariaGEN*, n.d.) was used to estimate global diversity in these genes identified in this study. The Pf3K dataset includes whole genome sequencing data from 2,512 samples collected in multiple locations in Asia and
Africa. Data from 156 additional isolates from Papua New Guinea were also included in the analysis. VaxPack ([https://github.com/BarryLab01/vaxpack](https://github.com/BarryLab01/vaxpack)) was used for global population genetic analysis. GATKv4.0 was used for variant calling. Samples containing ambiguous bases were removed. Singleton SNPs were converted back to reference to prevent false positive variants. Nucleotide diversity and Tajima’s D were calculated for all polymorphic sites separately for every country that had a sample size greater than 50. Templeton, Crandall, and Sing (TCS) (Clement et al., 2002) method on PopArt (Popart: Full-feature Software for Haplotype Network Construction - Leigh - 2015 - Methods in Ecology and Evolution - Wiley Online Library, n.d.) was used to construct the haplotype network using non-synonymous SNPs. Protein disorder and B-cell epitopes were predicted using PlasmoSIP (Guy et al., 2015). The haplotype frequencies of the C-terminal region in Malawian isolates from different immunity groups were estimated for non-synonymous sites using DnaSP v6 (Rozas et al., 2017).

Acknowledgements

We thank the participants in the Mfera Cohort Study. We would like to thank Biraj Shrestha and Gillian Mbambo for their assistance with sample processing. We would also like to acknowledge Timothy D. O’Connor, Michael P. Cummings and Alexis Boleda for their valuable input and suggestions, and Terrie Taylor for her role in leading the Malawi International Center of Excellence for Malaria Research. This work was supported by funding from the following awards granted by the National Institutes of Health: R01AI101713, R01AI125579, R01AI141900, U19AI110820, K24AI114996, and the Malawi International Center of Excellence for Malaria Research U19AI089683.

Author Contributions

Conceived and designed the study: ZS, MKL, JCS and ST-H. Collected and provided the samples: AGB, KBS, DPM, and MKL. Sequence processing and variant calling: ZS, KAM, and
JCS. Analyzed the data: ZS, MTN, KAM, AD, DS, AEB, and ST-H. Interpreted the results: ZS, MA, AO, KBS, DS, AEB, MKL, JCS, and ST-H. Wrote the manuscript: ZS and ST-H. All authors have read and had the opportunity to comment on the manuscript.

Competing Interests
The authors declare no competing interests.

References


Biochemical Parasitology, 143(1), 20–28.

https://doi.org/10.1016/j.molbiopara.2005.05.003


https://doi.org/10.12688/gatesopenres.12838.2


Plasmodium falciparum in West African populations of differing infection endemicity. 

_Molecular Biology and Evolution_, 31(6), 1490–1499.

https://doi.org/10.1093/molbev/msu106


https://doi.org/10.1038/ng1924


https://doi.org/10.1093/infdis/jis709


https://doi.org/10.1186/s12936-020-03195-8


https://doi.org/10.1093/infdis/jiv088


https://doi.org/10.1126/scitranslmed.3000257


https://doi.org/10.1371/journal.pmed.0040093


