A paradoxical population structure of *var* DBLα types in Africa

Mun Hua Tan[1†], Kathryn E. Tiedje[1†], Qian Feng[2], Qi Zhan[3], Mercedes Pascual[3], Heejung Shim[2], Yao-ban Chan[2], and Karen P. Day[1] *

[1] Department of Microbiology and Immunology, The University of Melbourne, Bio21 Institute and Peter Doherty Institute, Melbourne, AU
[2] School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne, Melbourne, Australia
[3] Department of Ecology and Evolution, University of Chicago; Chicago, Illinois, USA

[†]Co-first authors

ABSTRACT

The *var* multigene family encodes the *P. falciparum* erythrocyte membrane protein 1 (PfEMP1), which is important in host-parasite interaction as a virulence factor and major surface antigen of the blood stages of the parasite, responsible for maintaining chronic infection. Whilst important in the biology of *P. falciparum*, these genes (50 to 60 genes per parasite genome) are routinely excluded from whole genome analyses due to their hyper-diversity, achieved primarily through recombination. The PfEMP1 head structure almost always consists of a DBLα-CIDR tandem. Categorised into different groups (upsA, upsB, upsC), different head structures have been associated with different ligand-binding affinities and disease severities. We study how conserved individual DBLα types are at the country, regional, and local scales in Sub-Saharan Africa. Using publicly-available sequence datasets and a novel ups classification algorithm, *cUps*, we performed an *in silico* exploration of DBLα conservation through time and space in Africa. In all three ups groups, the population structure of DBLα types in Africa consists of variants occurring at rare, low, moderate, and high frequencies. Non-rare variants were found to be temporally stable in a local area in endemic Ghana. When inspected across different geographical scales, we report different

levels of conservation; while some DBLα types were consistently found in high frequencies in multiple African countries, others were conserved only locally, signifying local preservation of specific types. Underlying this population pattern is the composition of DBLα types within each isolate DBLα repertoire, revealed to also consist of a mix of types found at rare, low, moderate, and high frequencies in the population. We further discuss the adaptive forces and balancing selection, including host genetic factors, potentially shaping the evolution and diversity of DBLα types in Africa.

1. INTRODUCTION

Malaria parasites in endemic areas with high transmission undergo frequent outcrossing in the vector (1,2). The diversification of parasites through meiotic, mitotic, and ectopic recombination results in very high levels of genetic diversity in the *Plasmodium falciparum* parasite population (3–7), particularly of the major variant surface antigens such as the *var* multigene family that encode the *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1). PfEMP1 proteins are expressed on the surface of infected erythrocytes and can bind to host receptors to mediate cytoadhesion and sequestration of infected cells (8–10). Through clonal antigenic variation of *var* genes, whereby different genes are expressed sequentially and exclusively during the blood stage, parasites are also able to effectively evade immune detection, promoting chronic infection within a host (11,12). This high diversity has served as the basis for *var* surveillance (13,14), population genetics (15–17), and estimation of infection complexity (15). However, less work has been done to characterise this diversity and the population structure of antigenic factors of and within these genes.

There are approximately 40 to 60 different *var* genes found across all 14 chromosomes of a *P. falciparum* genome (18). Based on their upstream promoter sequences, *var* genes can be divided into groups of A, B, C, and E, with a minority of genes grouped in two intermediate groups of B/A or B/C (19). The three major 'ups' groups of upsA, upsB, and upsC are associated with different chromosomal locations, transcriptional directions, and sequences (18–21). Genes in the upsA and upsB groups are generally located at the subtelomeric regions whereas genes in the upsC group are found in the central regions of specific chromosomes. UpsA genes have been shown to be transcribed towards centromeres and conversely, upsB and upsC genes are often found to be transcribed towards telomeres.

The extracellular N-terminal PfEMP1 head structure almost always consists of a Duffy-binding-like alpha domain (DBLα) and a cysteine-rich interdomain region (CIDR) (i.e., a DBLα-CIDR tandem). This head structure exists in different configurations of these DBLα and CIDR domain subclasses (e.g., DBLα0-2 with CIDRα1-6, CIDRβ1-7, CIDRδ1-12, CIDRγ1,2) and can influence ligand binding and disease pathogenicity. The prevailing understanding is that *var* genes and DBLα variants (i.e., DBLα types) in the upsA group are generally more conserved compared to those in upsB or upsC groups (i.e., non-upsA) (15,16,19,22). This has largely been attributed to the association of upsA genes to severe malaria, including cerebral malaria, due to their host receptor binding phenotypes (23). Expression of upsA *var* genes encoding the DBLα+CIDRα1 head structure mediate endothelial protein C receptor (EPCR)-binding and/or intercellular adhesion molecule-1 (ICAM-1)-binding has been associated with severe malaria and/or cerebral malaria (24–26). In addition, PfEMP1 in the upsA group containing the DBLα+CIDRβ/δ/γ head structure has been associated with rosetting with uninfected erythrocytes (24). On the other hand, expression of upsB and upsC *var* genes (e.g.,

77 DBLα+CIDRα2-6) have been associated with uncomplicated malaria, commonly mediated by
78 adhesion to the cluster differentiation 36 (CD36) receptor (27,28), and may be more active in
79 establishing chronic infections (29).
80
81 With the exception of one specific *var* gene involved in pregnancy-associated malaria (i.e.,
82 *var2csa*), all other *var* genes encode for a DBLα domain, which is one of the most diverse
83 domains and has been shown to be immunogenic to variant-specific epitopes, recognised
84 serologically in an age-dependent manner (30). Multiple studies have noted that there exists
85 a minority of DBLα types that are highly conserved over various spatial scales (16,31–35).
86 These studies typically focussed on DBLα types or *var* genes found within the highest
87 percentiles (e.g., (35)), at very high frequencies (e.g., (16,31)), or those found globally
88 conserved and prevalent (e.g., (34,35)). Understandably, looking for the most common DBLα
89 or *var* is instinctive in the search for the elusive vaccine candidate targeting the most
90 important group of variant surface antigens of the parasite. However, *var* gene sharing among
91 isolates, particularly those living in high transmission, has been shown to be minimal (33).
92 Given that multiple variants of *var* genes with the same binding phenotype exist, this pre-
93 occupation over the most globally-common types or genes risks overlooking the genetic
94 patterns underlying a local population.
95
96 In the same way that 'severe malaria' must defined by different malaria epidemiologies,
97 patterns of conservation must also be interpreted within the context of an area's local
98 epidemiology. In low-transmission areas in South America and Asia, conservation found
99 across countries and continents likely relates to small population sizes due to founder effects,
100 in which *var* genes have not yet diversified (36–38). In moderate transmission, profiles may

exhibit bias toward more moderate frequency classes, in conjunction with greater diversity within the area. Conservation in high-transmission areas would be most interesting, as these areas possess the epidemiological and genetic characteristics to generate vast diversity. In such a system of high prevalence/incidence, high outcrossing rates, and high genetic diversity of parasites, finding conservation will provide insights into factors constraining and shaping diversity in a highly dynamic system.

Equipped with publicly-available DBLα data from several populations in Africa and a novel ups classification algorithm (***cUps***) introduced in this paper, this study sought to identify DBLα type conservation beyond those reported in the highest frequencies within and among populations globally. By categorising DBLα types into broad frequency classes of rare to highly-frequent types, we identified two kinds of conservation within all ups groups in our study area in Bongo, Ghana: (1) Conservation of specific types across isolates in a time point and through time (i.e., survey), (2) Conservation of type frequencies (i.e., types were found at relatively stable frequencies through time). We show that these patterns are maintained through the composition of DBLα types within each isolate repertoire, revealed to consist of a mix of types found at rare, low, moderate, and high frequencies in the population. In a spatial analysis, in addition to identifying DBLα types conserved at the continent level, we noted that there are DBLα types conserved at the local and/or regional levels but not necessarily prevalent across wider geographical scales, prompting a discussion on the adaptive forces potentially driving balancing selection and shaping the population structure of DBLα types in Africa.

2. RESULTS

2.1 Description of time-series cross-sectional surveys in Bongo, Ghana

This study analysed publicly-available DBLα tag sequence data from an interrupted time-series study design (i.e., Malaria Reservoir Study (MRS)) (Figure I in Data S1) (13,15,39). This MRS dataset consists of sampling at seven time points from 2012 to 2017 at the end of wet or dry seasons. Each time point represented an age-stratified cross-sectional survey of approximately 2,000 asymptomatic participants per survey (ages from 1 to 97 years old) from two proximal catchment areas (Vea/Gowrie and Soe, with a sampling area ~60 km^2) in Bongo District in Northern Ghana. Surveyed participants (i.e., isolates) represented approximately 15% of the total population that reside in the two catchment areas in Bongo District at a time (Table I in Data S1). This area is characterised by high, seasonal malaria transmission and has undergone several types of interventions, including long-lasting insecticide-treated nets (LLINs) and indoor residual spraying (IRS) that reduced transmission (13,15), and seasonal malaria chemoprevention (SMC) that reduced the burden of infection in children younger than 5 years old (15). Clustering of DBLα tag sequences from seven surveys (S1 to S7) and further post-processing of the dataset (see Methods) resulted in 62,168 representative DBLα types found in 3,166 isolates for this study of DBLα conservation in Bongo (Table I in Data S1).

In a high-transmission setting, the asymptomatic "population" typically consists of "isolates" infected by one or more unique parasite "genomes". This complexity of infections is indicated by multiplicity of infection (MOI), where an isolate with MOI = 1 would represent a single unique parasite genome. Hence, at MOI = 1, an isolate's DBLα repertoire is synonymous to a parasite's DBLα repertoire (i.e., the DBLα repertoire in a single parasite genome). Conversely, at MOI > 1, an isolate's DBLα repertoire would encompass > 1 parasites' DBLα repertoire.

149 **2.2 A novel ups classification algorithm based on *var* DBLα tags**

150 This study introduces ***cUps***, a novel algorithm for classifying DBLα types into the different
151 groups of upsA, upsB, and upsC (Data S2, Figure 1A). At the isolate level, the average isolate
152 repertoire consists of 20.9%, 48.6%, and 30.5% of upsA, upsB, and upsC DBLα types,
153 respectively (Figure II in Data S1). These proportions differ from those reported in (22) that
154 estimated higher proportions of upsB and lower proportions of upsC in isolate repertoires,
155 based on the average of seven genomes. The algorithm shows a tendency to classify more
156 upsB types as upsC types. This is in line with validation results on the algorithm's specificity
157 and sensitivity presented Data S2. A reduced analysis that involved DBLα types with higher
158 confidence in classification (threshold = 8, see Data S2) yielded similar trends and patterns of
159 observation. Genetic similarity by pairwise type sharing (PTS) remains extremely low for all
160 ups groups (median PTS: 4.55% (upsA), 1.00% (upsB), and 2.15% (upsC)) (Figure 1B). The
161 62,168 representative DBLα types from the seven combined MRS surveys were classified into
162 upsA (5.4%), upsB (56.6%), and upsC (37.9%) groups (Table I in Data S1). This points to a
163 highest DBLα richness for the upsB group (35,215 types), followed by the upsC group (23,583
164 types) and upsA group (3,370 types) in combined surveys, and this hierarchy of richness is
165 similarly observed for individual surveys (Figure 1A). The differences in proportions of ups
166 groups at the isolate versus population levels is attributed to the negative relationship
167 between PTS and richness, as a higher level of upsA type sharing will result in lower proportion
168 of unique representative DBLα types in the population.
169
170 **2.3 *Var* DBLα types are conserved in the local Bongo population and through time**
171 The frequency of a DBLα type is defined as the proportion of the population a DBLα type is
172 found in and is calculated in the context of individual surveys (i.e., "survey-specific

frequency"), or of the averaged frequencies of the seven surveys (i.e., "survey-averaged frequency", see Methods for details on frequency calculation and normalisation). To explore the conservation of DBLα types, we further categorised these frequencies into four classes of ***low*** (0%, 1%), ***moderate*** [1%, 5%), ***high*** [5%, 10%), and ***very high*** [10%, 100%] frequencies. While it is well reported that DBLα types in the upsA group are generally more conserved relative to types in the upsB and upsC groups, this study identified conservation of DBLα types in all three ups groups. We describe patterns of DBLα conservation observed in Bongo at the population level in the following subsections.

2.3.1. Conservation of var DBLα types in a surveyed time point

DBLα types are considered conserved in a surveyed time point when found in multiple isolates sampled in a same survey (i.e., moderate-to-high survey-specific frequencies). Overall, distributions of survey-averaged and survey-specific frequencies were found to be strongly and positively skewed, with most DBLα types occurring at low frequencies in <1% of isolates. For all three ups groups, when categorised into frequency classes, the second largest subsets of DBLα types are shown to occur at moderate frequencies between 1% to 5% in each survey, followed by smaller subsets of DBLα types found at higher frequencies exceeding 5% and/or 10% frequencies (Figure 2A). The most frequent DBLα types in the upsA, upsB, and upsC groups were detected at survey-specific frequencies of 61.1%, 42.9%, and 62.0%, respectively. In the different surveys, this study identified hundreds to thousands of moderate-to-highly conserved DBLα types (i.e., ≥1% survey-specific frequencies) in all three ups groups (upsA: 525 to 790 types per survey, upsB: 539 to 1,917 types per survey, upsC: 365 to 1,121 types per survey). This translates into different proportions of DBLα types in each ups group, owing to the higher DBLα type richness of upsB and upsC groups (upsA: 29.2% to

43.9% per survey, upsB: 5.2% to 15.5% per survey, upsC: 5.6% to 14.3% per survey) (Table II in Data S1).

2.3.2. Conservation of var DBLα type frequencies across multiple surveys and through time

Frequencies of DBLα types are considered conserved if a same DBLα type occurs at stable or similar frequencies in multiple surveys. In all ups groups, a strong positive correlation between survey-specific frequencies and survey-averaged frequencies of DBLα types is shown, indicating that DBLα types occurring at high survey-averaged frequencies were also generally found at high frequencies in specific surveys and likely in multiple surveys (Figure 2C, Figure III in Data S1). Likewise, most DBLα types found at moderate frequencies also showed consistent maintenance of frequencies across multiple surveys. Furthermore, DBLα types are considered conserved through time when found in multiple surveys. This study shows a correlation between DBLα type frequencies (per-survey and survey-averaged) and the number of surveys the types are found in, with most of the relatively conserved DBLα types in all three ups groups also persisting through time (Figure 2B, Figure III and IV in Data S1). All DBLα types occurring at high or very high survey-averaged and/or survey-specific frequencies were seen in all seven surveys. The majority of those with moderate survey-averaged or survey-specific frequencies were found in all seven surveys, with a smaller subset found in four to six surveys. On the other hand, the remainder of DBLα types found at low frequencies (< 1%) can be found in the range of one to seven surveys. This was observed in all three ups groups.

2.4 Isolate repertoires consist of a mix of conserved and rare *var* DBLα types

While isolate DBLα repertoires in high-transmission populations have been reported to be unrelated and largely non-overlapping (15,17), there has not been a detailed exploration of the composition of DBLα types and their respective frequency classes at the isolate repertoire level (i.e., 'per-isolate frequency profiles'). This study showed that all isolate repertoires consist of DBLα types from all three ups groups. More interestingly, these per-isolate frequency profiles show consistent proportions of rare, moderately-frequent, and highly-frequent DBLα types in every isolate, based on survey-specific frequencies (Figure 3, Figure V in Data S1). In all ups groups, these frequency profiles were consistent across isolates within the same survey, regardless of isolates' infection complexities (i.e., MOI). Importantly, the observation of these frequency patterns in MOI = 1 isolates (i.e., isolate non-upsA repertoire size of approximately ≤ 45 in Figure 3) indicates that these per-isolate frequency profiles are a consequence of similar repertoire composition within actual parasite genomes.

Of the three groups, isolates' upsA frequency profiles exhibit largest proportions of DBLα types found at moderate survey-specific frequencies and smaller proportions of DBLα types in the lowest survey-specific frequency class. In contrast, isolates' upsB and upsC frequency profiles both consist of largest proportions of DBLα types in the lowest frequency class (i.e., (0%, 1%)), followed by those in the moderate frequency class. The introduction of interventions did not perturb these frequency profiles, which maintained the composition of different frequency classes, albeit in different proportions. In surveys of the population affected by interventions (e.g., S4 and S5 during and after indoor residual spraying), frequency profiles trended toward a generally larger proportion of low-frequency DBLα types and smaller proportions of higher frequency DBLα types within each isolate (Figure 3). We confirmed this observation of per-isolate frequency profiles using an independent DBLα

sequence dataset (14,33,40), extracted from *var* genes of isolates sampled from Navrongo in Ghana, situated ~30 km adjacent to Bongo district (Figure VI in Data S1).

As expected, genetic similarity between pairwise isolate repertoires (i.e., PTS values) increases when DBLα types of lower frequencies are excluded (Figure VII in Data S1). PTS values range from 0 to 1 representing unrelated to identical isolate repertoires, respectively. Interestingly, even when considering only DBLα types found at very high frequencies, median PTS values remained generally low (median PTS of 0.02, 0.05, 0.13, and 0.19 when considering DBLα types at >0%, ≥1%, ≥5%, ≥10% survey-averaged frequencies, respectively, across all surveys). This indicates that, while every isolate repertoire contains sets of types that are conserved in the population, identical sets of conserved DBLα types are rare. Shifts in PTS distributions were more substantial when exclusively evaluating DBLα types in the upsA or upsC groups relative to the upsB group, consistent with the lower DBLα richness in the two former groups (i.e., the less variants there are in the population, the higher the probability of overlaps).

2.5 Global and local preservation of *var* DBLα types in Africa

Further, a separate spatial study of DBLα conservation in multiple African countries (i.e., "locations") representing West Africa (Ghana, Gabon), Central Africa (Malawi) and East Africa (Uganda) was conducted based on 82,027 DBLα types found in 4,783 isolates (Table I in Data S3, Figure I in Data S3). Similarly, this spatial study showed that the majority of DBLα types were found at low frequencies with smaller proportions seen at higher frequencies. Comparison of DBLα types and frequencies in the four locations showed conservation of the same upsA DBLα types at moderate-to-high frequencies in all locations (i.e., a highly-frequent

DBLα type in Ghana was also found at moderate-to-high frequencies in other analysed locations) (Figure 4). While this was also observed for some highly-frequent DBLα types in the upsB and upsC groups (i.e., non-upsA groups), this study additionally identified some highly-frequent DBLα types in these groups that were present predominantly in a single location, suggesting local selection and preservation of DBLα types in these ups groups (Figure 4, Figure II and III in Data S3). Additionally, as was also observed for isolates in the MRS study, per-isolate frequency profiles in these different locations also consisted of a mix of rare to conserved DBLα types (Figure IV in Data S3). It is worth reminding that these conserved DBLα types make up the minority of all DBLα types in every ups group. An exploration of the relationship between DBLα types and *var* exon 1 sequences revealed that these conserved DBLα types are associated with multiple different *var* exon 1 sequences (Figure V in Data S3), indicating that other parts of the gene were still diversifying even though the DBLα types were maintained in the population. For some of these DBLα types with 1-to-many DBLα-*var* relationships, pairwise nucleotide identity between *var* exon 1 sharing the same DBLα type suggest that some of these *var* sequences could be alleles of a same gene (14). However, the majority of these *var* exon 1 sequences exhibit low shared identity and therefore appear to represent actual different genes (Figure V in Data S3). In a highly-dynamic system where the DBLα domain has been shown *in vitro* to exhibit the highest recombination rate (6), the maintenance of specific DBLα types at high frequencies and through extensive durations could suggest that selection for adaptive advantages.

2.6 Factors driving the conservation of DBLα types remain unknown

The spatial study considered a few possible factors to explain these conserved DBLα types, focusing specifically on 51 and 17 DBLα types in the high and very high location-averaged

292 frequency classes, respectively (i.e., a total of 68 DBLα types with location-averaged
293 frequencies of ≥5%) (Figure 5). Firstly, conservation of *var* genes on specific *P. falciparum*
294 chromosomes 4, 6, 7, and 8 has been previously reported and potentially attributed to
295 selective sweep events associated with antimalarial drug resistance (33). While positional
296 information is unavailable for the DBLα types analysed in this study, sequence alignments
297 show that only six of the 68 highly-frequent DBLα types are homologous to *var* genes on these
298 chromosomes. Furthermore, some of these highly-frequent DBLα types were identified as
299 homologs to the DBLα tags of five *var* genes in *P. praefalciparum*, a Plasmodium species that
300 naturally infects gorillas and is the closest living sister species of *P. falciparum* (41,42).
301 Homologs to the DBLα tag of one *var* gene of another ancestral Laverania species, *P.*
302 *reichenowi*, was identified but occurring at relatively low-to-moderate frequencies
303 (frequencies range from 0.57% to 2.34%). No homologs to the DBLα tags of *var* genes of *P.*
304 *gaboni* were identified. Hence, it is clear that while some of these factors can explain the
305 reason a few of these sequences are conserved, the majority of these conserved DBLα types
306 are still unaccounted for.
307
308 Homologs to other published globally-conserved DBLα types and *var* gene (PF3D7_0617400)
309 are shown in Figure 5. Tonkin-Hill et al. (35) reported a set of 100 most frequent DBLα types
310 conserved in their analysis of ten countries across diverse global regions in Africa,
311 Asia/Oceania, and South America. Homologs to 84 of these globally-conserved DBLα types
312 were identified in this study, with 26 and 11 of these types found in the two highest location-
313 averaged frequency classes. Furthermore, in the context of general prevalence in the
314 analysed African locations, 30 of these 37 DBLα types were found in all four locations, six in
315 three locations, and only one was found in a single location.

316

317 A conserved *P. falciparum var* gene (PF3D7_0617400) was also recently reported in a
318 Gabonese parasite isolate and characterised (34). The homolog to the DBLα tag of this *var*
319 gene was found to occur at high frequencies (ranging from 7.0% to 20.2% in different
320 locations) and present in all locations except, strangely, in Gabon itself. Notably, this
321 conserved PF3D7_0617400 *var* gene is located on chromosome 6, coinciding with the
322 previous reports of haplotypes in linkage disequilibrium on the same chromosome (43,44),
323 though this *var* gene is located outside of this region's cluster. A possible explanation for the
324 absence of this homolog in the Gabon dataset used in this study may be that the isolates were
325 sampled relatively early in the timeline (year 2000), which precedes the switch to artemisinin
326 (ART)-based combination therapies (ACT) in Africa (45,46), suggesting that the selection for
327 this specific type may have still been in progress and may not yet have risen in frequency to
328 result in observed fixation in the population at the time. Additionally, we also checked for
329 possible conservation of a *P. falciparum var* gene (PF3D7_0809100) that has been reported
330 to be expressed in sporozoites and potentially play a role in hepatocyte infection (47), but
331 homologs to the DBLα tag of this gene were found at only <1% frequency in three locations.

332

333 **3. DISCUSSION**

334

335 Extensive DBLα type diversity is reported in areas with high malaria transmission, generated
336 by meiotic and mitotic recombination (3–7) with DBLα repertoire diversity driven by frequent
337 outcrossing in the mosquito vector (1,2), such that we would not expect conservation of
338 types. However, a closer inspection of the population structure of DBLα types reveals
339 conservation of DBLα types ***beyond*** sequences found at very high frequencies or within the

340 highest percentiles. Instead, conservation also encompasses types that are seen stable
341 through time and can be found in a population at various frequencies, be it low, moderate,
342 or high. This study observed the conservation of DBLα types in the three major ups groups
343 within a large natural parasite population in a local area in Bongo. Frequencies of these DBLα
344 types at the population level were shown to be temporally stable over at least five years and
345 through wet and dry seasons. In addition to conservation at the continent level, spatial
346 analysis observed local conservation of specific high-frequency upsB and upsC types in
347 individual countries in Africa, despite the high genetic diversity typically reported for these
348 groups. Global analyses such as (35) and (34) would have uncovered the most conserved types
349 and genes globally but could have missed out on much of local signatures of conservation,
350 which this study has shown to exist within different frequency classes.
351
352 The key result of this analysis is that the frequency pattern of DBLα types that make up every
353 isolate repertoire not only underlies these local population structures but will maintain them.
354 Looking at individual DBLα types found in every isolate repertoire and the corresponding
355 frequency at which each type occurs in the population, per-isolate frequency profiles revealed
356 that every isolate repertoire consists of a mix of low-, moderate-, and high-frequency types,
357 in proportions consistent across all isolates. This presents a paradox in the population
358 structure of DBLα types, where there is a very high diversity of DBLα types found in a
359 population, but each isolate still maintains a combination of low- to moderate- to high-
360 frequency types in its repertoire. Even more interestingly, this paradoxical structure, both at
361 the level of population and isolate, was observed for types in all three groups of upsA, upsB,
362 and upsC and is maintained despite the expectation of frequent outcrossing in these endemic
363 areas.

In high transmission, despite the high rates of outcrossing and recombination, the consistency in these per-isolate frequency profiles suggest a level of constraint on the modularity of each isolate's repertoire, i.e., each isolate repertoire must have a combination of common and rare types while maintaining limited overlaps with other isolates in the population overall. This pattern was observed with both DBLα field data and DBLα encoding sequences identified in assembled *var* exon 1 sequences, thereby excluding any biases from genotyping methods. DBLα-*var* relationships revealed that high-frequency DBLα types are likely to be associated with multiple distinct *var* exon 1 sequences (i.e., 1-to-many), though high sequence similarities were estimated for some pairwise *var* sequences, suggesting that a proportion of these are alleles of a same gene. We make clear that this study describes the conservation of DBLα types, not necessarily the conservation of *var* genes. Conservation of *var* genes can be better studied if we can properly define alleles.

Assuming that there are biological advantages conferred by these conserved types, why do we observe stable presence at different frequencies but not fixation of most of these types in the population? Furthermore, why are these per-isolate frequency profiles maintained? One hypothesis relates to balancing selection as a result of co-evolution between the parasite and the human host population it is infecting, which can occur at the local, regional, and continent levels. The role of PfEMP1 in evading recognition by the host innate immune system would select for its variation, and the DBLα domain has been shown to be immunogenic to variant-specific epitopes and serologically recognised in an age-dependent manner (30). On the other hand, its role in virulence, such as the need to bind to specific host endothelial cells for cytoadhesion or blood cell receptors for rosetting, could select for some level of conservation.

388 This tension between the dual different roles, both of which relate to local host genetics,
389 could be creating the observed paradoxical pattern.
390
391 Stochastic simulations and network analyses have provided clear evidence for a role of
392 immune selection or negative frequency-dependent selection resulting from specific immune
393 memory, which is a form of balancing selection, in shaping antigenic diversity within natural
394 populations. As antibody-mediated immunity plays a significant role in recognition of PfEMP1
395 variants, we hypothesise that another possible driver of balancing selection is the arm's race
396 between the parasite PfEMP1 variants and host HLA class II haplotypes (48–51). Similar to our
397 finding of local signatures of DBLα type conservation against a highly-diverse background,
398 there are also geographic differences in HLA class II alleles across the African continent (52–
399 54). Immune evasion of common local HLA class II alleles could drive DBLα types or *var* genes
400 containing these types to persist through time at stable low-to-moderate frequencies.
401 Alternatively, genetic variation in a parasite's DBLα or *var* repertoires may have been shaped
402 by underlying differences in host receptors of varying spatial niches and if not, these types
403 could be in linkage disequilibrium with other proximal domains (e.g., CIDR) or genes vital to
404 these roles.
405
406 The consistency of these per-isolate frequency profile patterns is striking and suggests that
407 maintaining such frequency profiles within a parasite repertoire is advantageous to the
408 parasite. Having a range of rare to common types may allow malaria parasites to adapt to
409 host factors in order to persist through dynamics and competition within and between hosts.
410 The translational implication of this work suggests that breaking this pattern to what is seen

in low transmission i.e., high relatedness of *var* repertoires and clonality could be a target of elimination efforts.

4. MATERIALS AND METHODS

4.1 Data sources and types

Conservation analyses were performed on a small ~450bp region of a *var* gene that encodes a portion of the DBLα domain of PfEMP1 (i.e., DBLα tags) (55,56). DBLα tag sequences included in this study were either generated from targeted amplicon sequencing (15–17) or extracted from assembled *var* gene sequences (14,33,40). This made available DBLα tag datasets of varying sizes from Africa and Asia, which were clustered to generate representative DBLα types (see Section 4.2). However, the scope of this study on DBLα conservation was limited to African locations only, with higher transmission, because lower transmission areas may present a different context underlying conservation (e.g., clonality or smaller population sizes). Data in Africa were available from West Africa (Senegal, Gambia, Guinea, Mali, Ghana, Gabon), Central Africa (Congo, Malawi) and East Africa (Uganda, Kenya) (Table I in Data S1, Table I in Data S3). However, most of these African countries were excluded due to limited dataset sizes (number of isolates < 100), resulting in a final analysis from four locations in Africa (i.e., Ghana, Gabon, Malawi, Uganda). Sources and methods that the different studies used to generate these DBLα tag datasets are described in the following subsections.

4.1.1 DBLα tags from targeted amplicon sequencing data

Published DBLα tag datasets from three locations were generated from targeted amplicon sequencing (Table I in Data S1, Table I in Data S3). Amplicon sequencing of DBLα tag

435 sequences involves PCR amplification of a small sequence region encoding the DBLα domain
436 of PfEMP1 with degenerate primers (55,56), followed by high throughput sequencing on
437 either the Illumina MiSeq platform (Ghana, (13,15,39)) or on the 454 sequencing platform
438 (Gabon, (17)), Uganda, (16)). These include sequences from:

439 (I) One area (Bongo) in Ghana: dataset spans seven time points (surveys) from 2012
440 to 2017 through sampling of asymptomatic individuals through multiple dry and
441 wet seasons.

442 (II) Six areas (Apac, Arua, Jinja, Kanungu, Kyenjojo, Tororo) in Uganda: dataset
443 included sampling of clinical isolates over two years.

444 (III) One area (Bakoumba) in Gabon: dataset included sampling of asymptomatic
445 children in one year.

446

447 **4.1.2 DBLα tags from assembled *var* gene sequences**

448 Published *var* gene sequences (from isolates in Africa and Asia) were downloaded from the
449 'Full Dataset' published by (33). DBLα tag sequences were identified and extracted from *var*
450 gene sequences (regardless of *var* gene completeness) as described in (14). Briefly, domain
451 annotations provided by (33) were used to extract nucleotide sequences encoding the DBLα
452 domain. These extracted sequences were further translated into the best reading frames and,
453 using *hmmsearch* (57), the resulting amino acid sequences were further searched against
454 positions 189 to 430 of the PFAM profile alignment (PF05424_seed.txt) to identify the 'tag'
455 region (domain score cut-off of 60 and ≥100 aligned positions) and to ultimately extract the
456 DBLα tag sequence that would have been amplified with degenerate primers (55,56).

457

458 **4.2 Clustering of DBLα tags into DBLα types**

459 DBLα tags (Africa and Asia) were translated into amino acid sequences and any untranslatable
460 sequences (i.e., stop codons in reading frame) were excluded. The remaining DBLα tags were
461 combined and clustered with *clusterDBLa* (58), using a 96% nucleotide identity threshold (31)
462 to produce representative DBLα types. This also generated a binary matrix detailing the
463 presence/absence matrix of each DBLα type in each isolate.
464
465 **4.3 Classification of DBLα types into domain classes and ups groups**
466 The *classifyDBLα* pipeline (16) was used to classify DBLα types into DBLα domain classes of
467 DBLα0, DBLα1, or DBLα2, in order to confirm that sequences were indeed those encoding the
468 DBLα domain of PfEMP1. In addition, a novel algorithm (***cUps***) described in this study was
469 used to classify DBLα types into the most probable ups group (i.e., upsA, upsB, or upsC),
470 accompanied by assignment probability values. For each DBLα type, ups groups were assigned
471 according to the prediction with the highest assignment probability. We describe this novel
472 classification algorithm below as well as in Data S2). An implementation of the algorithm is
473 available at https://github.com/qianfeng2/cUps.
474
475 Through the alignment and clustering of 2kb sequences upstream of *var* genes, followed by
476 the classification *var* genes into ups groups by Neighbour-joining (NJ) and Markov clustering
477 (MCL) methods (trees available in Data S2), a reference dataset of DBLα tag sequences was
478 generated from 846 *var* genes from 16 P. falciparum genomes ((59) and NCBI). We begin with
479 this reference database of DBLα tag sequences with ups groups and DBLα subclasses known.
480 For each category (ups group/DBLα subclass combination), we align the reference sequences
481 in the category using Clustal Omega v1.2.4 (60), then fit a profile hidden Markov model (61)
482 using HMMER v3.2.1 (57) with default settings.

For a given query sequence (representing a DBLα type), we calculate the likelihood of the query sequence being drawn from the profile HMM of each category, using the forward algorithm. The posterior probability for each category is then calculated using Bayes' Theorem, with the prior probabilities of each category calculated from the reference database. Summing over DBLα subclasses gives the posterior probability for each ups group (i.e., assignment probability). The query sequence can be classified to the ups group with the highest assignment probability. Although we do not do so in this paper, a threshold may optionally be applied, so that sequences with highest assignment probability below the threshold are categorised as `unclassified'. Alternatively, a summary statistic may weight each ups group by the assignment probability. This method is described in much more detail, with verification (Feng, submitted).

4.4 Exclusion of DBLα types, isolates, and populations from the final DBLα type dataset

Only the DBLα types that were successfully classified into a DBLα domain class (i.e., DBLα0, DBLα1, or DBLα2) were retained in the final dataset. Subsequently, isolates with < 20 DBLα types were also removed from dataset to ensure robust analyses downstream (Table I in Data S1, Table I in Data S3). Specifically for the time-series dataset from the Malaria Reservoir Study (MRS) in Bongo, Ghana (13,15,39), submicroscopic or symptomatic isolates were additionally excluded from the dataset. Further, using *blastn* (≥96% nucleotide identity, ≥95% query coverage) (62), DBLα types with homology to isolate-transcendent *var1*, *var2csa*, and *var3* sequences (sequences from (22,33)) were excluded to remove putative DBLα types previously reported as isolate-transcendent (20,21). Finally, given that frequency classes and profiles were calculated based on proportional frequencies, only locations with datasets of ≥100

507 isolates were retained. This resulted in the exclusion of six African countries from this study
508 ("*" in Table I in Data S3).
509
510 **4.5 Genetic similarity between pairwise isolate repertoires**
511 The pairwise type sharing metric (PTS) (31) is used to estimate the overlap between pairwise
512 isolate repertoires (e.g., isolates *i* & *j*). Specifically:
513
514 $$PTS = \frac{2 * shared_{ij}}{Size_{ij} + Size_{ij}}$$
515
516 where $shared_{ij}$ is the number of shared DBLα types between repertoires of isolates *i* and *j*,
517 and $Size_i$ and $Size_j$ are the total number of DBLα types (i.e., repertoire sizes) of isolates *i* and *j*,
518 respectively. A value of 0 indicates the absence of sharing between two isolates whereas a
519 value of 1 indicates completely identical isolate repertoires.
520
521 **4.6 Calculation of DBLα type frequencies and assignments into frequency classes**
522 Depending on the analysis, a population can be the collection of isolates sampled at a specific
523 survey or time point in the time-series analyses (i.e., by year or survey in the MRS data) or the
524 collection of isolates sampled from a specific region or location/country in the spatial
525 analyses. Raw frequencies of DBLα types were defined at the survey or location level in counts
526 (i.e., number of isolates with a particular DBLα type in each survey or location). Raw (count)
527 frequencies were converted into proportional frequencies through division of count
528 frequencies by the total number of isolates at a corresponding time point or location, leading
529 to "survey-specific frequencies" or "location-specific frequencies". Subsequently, these

530 frequencies were further categorised into frequency classes of 0%, ***low*** (0%, 1%), ***moderate***
531 [1%, 5%), ***high*** [5%, 10%), and ***very high*** [10%, 100%].
532
533 Given the substantial differences in dataset sizes across surveys or locations (e.g., 499 isolates
534 for Uganda *versus* 176 isolates for Gabon), simply summing isolates across datasets of
535 multiple surveys or locations would bias total frequencies to reflect those of larger datasets.
536 Hence, averaged frequencies were used instead as a means to normalise total frequencies by
537 isolate counts in each survey or location. For example, a DBLα type found in 10 out of 100
538 isolates for location A and 10 out of 500 isolates for location B would be reported to have 10%
539 and 2% frequencies for locations A and B, respectively. This would yield a crude total
540 frequency of 3.33% (20 of 600 isolates), which is more reflective of the frequency observed
541 in location B even though the DBLα type was found at relatively high frequency at location A.
542 In this instance, with normalisation, an averaged frequency of 6% would be estimated (12 of
543 200 isolates), reducing the bias towards larger dataset sizes. Given the focus of this study on
544 conserved DBLα types, this normalisation method provides a less biased approach in
545 identifying DBLα types that are found at high frequencies but not necessarily uniformly across
546 all datasets.
547
548 **4.7 Determination of DBLα-*var* relationships**
549 For two locations (Ghana and Malawi), *var* gene sequences were available from assemblies
550 performed by (33). DBLα-*var* relationships were determined using complete *var* exon 1
551 sequences that are bounded by an N-terminal segment (NTS) and a transmembrane region
552 (TM) on the 5' and 3' ends of exon 1, respectively (14). Briefly, using *vsearch* (63), DBLα types
553 were globally aligned to *var* exon 1 sequences from the same location (e.g., Malawi DBLα

554 types to Malawi *var* exon 1). Given that DBLα types were generated from clustering at a 96%
555 nucleotide identity threshold, these types were aligned to *var* exon 1 sequences at the same
556 threshold of 96% identity, calculated over the alignment length and excluding terminal gaps
557 *(--iddef 2*). The relationship between a DBLα type and distinct *var* exon 1 was determined
558 based on the number of unique *var* exon 1 sequences sharing a same DBLα type (e.g., a 1-to-
559 *n* DBLα-*var* relationship is defined as a DBLα type found in *n* unique *var* exon 1).
560
561 For each group of *var* exon 1 that share a same DBLα type, an all *vs* all sequence alignment of
562 *var* exon 1 sequences in the group was performed using the *allpairs_global* option within
563 *vsearch* (63) and set to include all pairwise alignments *(--acceptall*). Pairwise nucleotide
564 identities were estimated based on calculations over whole alignment lengths, including
565 terminal gaps (*--iddef 1*), to account for differences in pairs of *var* exon 1 of variable lengths.
566
567 **4.8 Search for homology to other DBLα types or *var* genes**
568 **4.8.1 *Var* genes in association with selective sweeps on select chromosomes**
569 Published work reported conserved *var* genes on chromosomes 4, 6, 7 and 8 associated with
570 selective sweep events, potentially due to drug resistance or other factors. Accession
571 numbers of these conserved genes were obtained from the author (33,59) and used as
572 reference. Using *blastn* (62), DBLα types were searched against these reference sequences
573 and hits from alignments were reported (≥96% nucleotide identity, ≥95% query coverage).
574
575 **4.8.2 *Var* genes in primate *Plasmodium* species**
576 *Var* genes from three *Plasmodium* species, *P. praefalciparum*, *P. reichenowi* and *P. gaboni*,
577 were downloaded from PlasmoDB (41) and used as reference. Using *blastn* (62), DBLα types

were searched against these reference sequences and hits from alignments were reported (≥96% nucleotide identity, ≥95% query coverage).

4.8.3 Globally-conserved DBLα types or *var* genes

The 100 most frequent DBLα sequences reported in the global analysis by (35) was used as reference. Using *blastn* (62), DBLα types were searched against these reference sequences and hits from alignments were reported (≥96% nucleotide identity, ≥95% query coverage). The same search parameters and thresholds were applied in searching for homologs to *var* gene sequence PF3D7_0617400, a conserved 3D7 *var* gene reported by (34). Homologs to *var* gene sequence PF3D7_0809100, shown by (47) to be expressed at the sporozoite stage, were also searched for.

ACKNOWLEDGMENTS

This publication uses *var* gene data assembled from data generated from the MalariaGEN *Plasmodium falciparum* Community Project. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

REFERENCES

1. Babiker HA, Ranford-Cartwright LC, Currie D, Charlwood JD, Billingsley P, Teuscher T, et al. Random mating in a natural population of the malaria parasite Plasmodium falciparum. Parasitology. 2009/04/06. 1994;109(4):413–21.
2. Paul REL, Packer MJ, Walmsley M, Lagog M, Ranford-Cartwright LC, Paru R, et al. Mating Patterns in Malaria Parasite Populations of Papua New Guinea. Science (80-). 1995;269(5231):1709–11.
3. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature. 2000;407(6807):1018–22.
4. Duffy MF, Byrne TJ, Carret C, Ivens A, Brown G V. Ectopic Recombination of a Malaria var Gene during Mitosis Associated with an Altered var Switch Rate. J Mol Biol.

2009;389(3):453–69.
5. Bopp SER, Manary MJ, Bright AT, Johnston GL, Dharia N V, Luna FL, et al. Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families. PLOS Genet. 2013;9(2):e1003293.
6. Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, et al. Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. PLOS Genet. 2014 Dec 18;10(12):e1004812.
7. Zhang X, Alexander N, Leonardi I, Mason C, Kirkman LA, Deitsch KW. Rapid antigen diversification through mitotic recombination in the human malaria parasite Plasmodium falciparum. PLOS Biol. 2019;17(5):e3000271.
8. Baruch DI. Adhesive receptors on malaria-parasitized red cells. Best Pract Res Clin Haematol. 1999;12(4):747–61.
9. Newbold C, Craig A, Kyes S, Rowe A, Fernandez-Reyes D, Fagan T. Cytoadherence, pathogenesis and the infected red cell surface in Plasmodium falciparum. Int J Parasitol. 1999;29(6):927–37.
10. Chen Q, Schlichtherle M, Wahlgren M. Molecular Aspects of Severe Malaria. Clin Microbiol Rev. 2000;13(3):439–50.
11. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, et al. Switches in expression of plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell. 1995;82(1):101–10.
12. Su X, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes. Cell. 1995;82(1):89–100.
13. Tiedje KE, Oduro AR, Bangre O, Amenga-Etego L, Dadzie SK, Appawu MA, et al. Indoor residual spraying with a non-pyrethroid insecticide reduces the reservoir of Plasmodium falciparum in a high-transmission area in northern Ghana. PLOS Glob Public Heal. 2022 May 18;2(5):e0000285.
14. Tan MH, Shim H, Chan Y, Day KP. Unravelling var complexity: Relationship between DBLα types and var genes in Plasmodium falciparum. Vol. 1, Frontiers in Parasitology. 2023.
15. Tiedje KE, Zhan Q, Ruybal-Pesántez S, Tonkin-Hill G, He Q, Tan MH, et al. Measuring changes in Plasmodium falciparum census population size in response to sequential malaria control interventions. medRxiv. 2023;
16. Ruybal-Pesántez S, Tiedje KE, Tonkin-Hill G, Rask TS, Kamya MR, Greenhouse B, et al. Population genomics of virulence genes of Plasmodium falciparum in clinical isolates from Uganda. Sci Rep. 2017;7(1):11810.
17. Day KP, Artzy-Randrup Y, Tiedje KE, Rougeron V, Chen DS, Rask TS, et al. Evidence of strain structure in Plasmodium falciparum var gene repertoires in children from Gabon, West Africa. Proc Natl Acad Sci. 2017;114(20):E4103–11.
18. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002;419(6906):498–511.
19. Lavstsen T, Salanti A, Jensen ATR, Arnot DE, Theander TG. Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malar J. 2003;2(1):27.

20. Kyes SA, Kraemer SM, Smith JD. Antigenic Variation in Plasmodium falciparum: Gene Organization and Regulation of the var Multigene Family. Eukaryot Cell. 2007 Sep 1;6(9):1511–20.
21. Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, Christodoulou Z, et al. Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates. BMC Genomics. 2007;8(1):45.
22. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T. Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer. PLOS Comput Biol. 2010 Sep 16;6(9):e1000933.
23. Walker IS, Rogerson SJ. Pathogenicity and virulence of malaria: Sticky problems and tricky solutions. Virulence. 2023 Dec 31;14(1):2150456.
24. Smith JD, Rowe JA, Higgins MK, Lavstsen T. Malaria's deadly grip: cytoadhesion of Plasmodium falciparum-infected erythrocytes. Cell Microbiol. 2013 Dec 1;15(12):1976–83.
25. Lau CKY, Turner L, Jespersen JS, Lowe ED, Petersen B, Wang CW, et al. Structural Conservation Despite Huge Sequence Diversity Allows EPCR Binding by the PfEMP1 Family Implicated in Severe Childhood Malaria. Cell Host Microbe. 2015 Jan 14;17(1):118–29.
26. Turner L, Lavstsen T, Berger SS, Wang CW, Petersen JE V, Avril M, et al. Severe malaria is associated with parasite binding to endothelial protein C receptor. Nature. 2013;498(7455):502–5.
27. Hsieh F-L, Turner L, Bolla JR, Robinson C V, Lavstsen T, Higgins MK. The structural basis for CD36 binding by the malaria parasite. Nat Commun. 2016;7(1):12837.
28. Robinson BA, Welch TL, Smith JD. Widespread functional specialization of Plasmodium falciparum erythrocyte membrane protein 1 family members to bind CD36 analysed across a parasite genome. Mol Microbiol. 2003 Mar 1;47(5):1265–78.
29. Smith JD. The role of PfEMP1 adhesion domain classification in Plasmodium falciparum pathogenesis research. Mol Biochem Parasitol. 2014;195(2):82–7.
30. Barry AE, Trieu A, Fowkes FJI, Pablo J, Kalantari-Dehaghi M, Jasinskas A, et al. The Stability and Complexity of Antibody Responses to the Major Surface Antigen of Plasmodium falciparum Are Associated with Age in a Malaria Endemic Area. Mol Cell Proteomics. 2011;10(11).
31. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, et al. Population Genomics of the Immune Evasion (var) Genes of Plasmodium falciparum. PLOS Pathog. 2007 Mar 16;3(3):e34.
32. Rougeron V, Tiedje KE, Chen DS, Rask TS, Gamboa D, Maestre A, et al. Evolutionary structure of Plasmodium falciparum major variant surface antigen genes in South America: Implications for epidemic transmission and surveillance. Ecol Evol. 2017 Nov 1;7(22):9376–90.
33. Otto TD, Assefa SA, Böhme U, Sanders MJ, Kwiatkowski DP, Null N, et al. Evolutionary analysis of the most polymorphic gene family in falciparum malaria [version 1; peer review: 1 approved, 2 approved with reservations]. Wellcome Open Res. 2019;4(193).
34. Dimonte S, Bruske EI, Enderes C, Otto TD, Turner L, Kremsner P, et al. Identification of a conserved var gene in different Plasmodium falciparum strains. Malar J. 2020;19(1):194.
35. Tonkin-Hill G, Ruybal-Pesántez S, Tiedje KE, Rougeron V, Duffy MF, Zakeri S, et al. Evolutionary analyses of the major variant surface antigen-encoding genes reveal

population structure of Plasmodium falciparum within and between continents. PLOS Genet. 2021 Feb 25;17(2):e1009269.

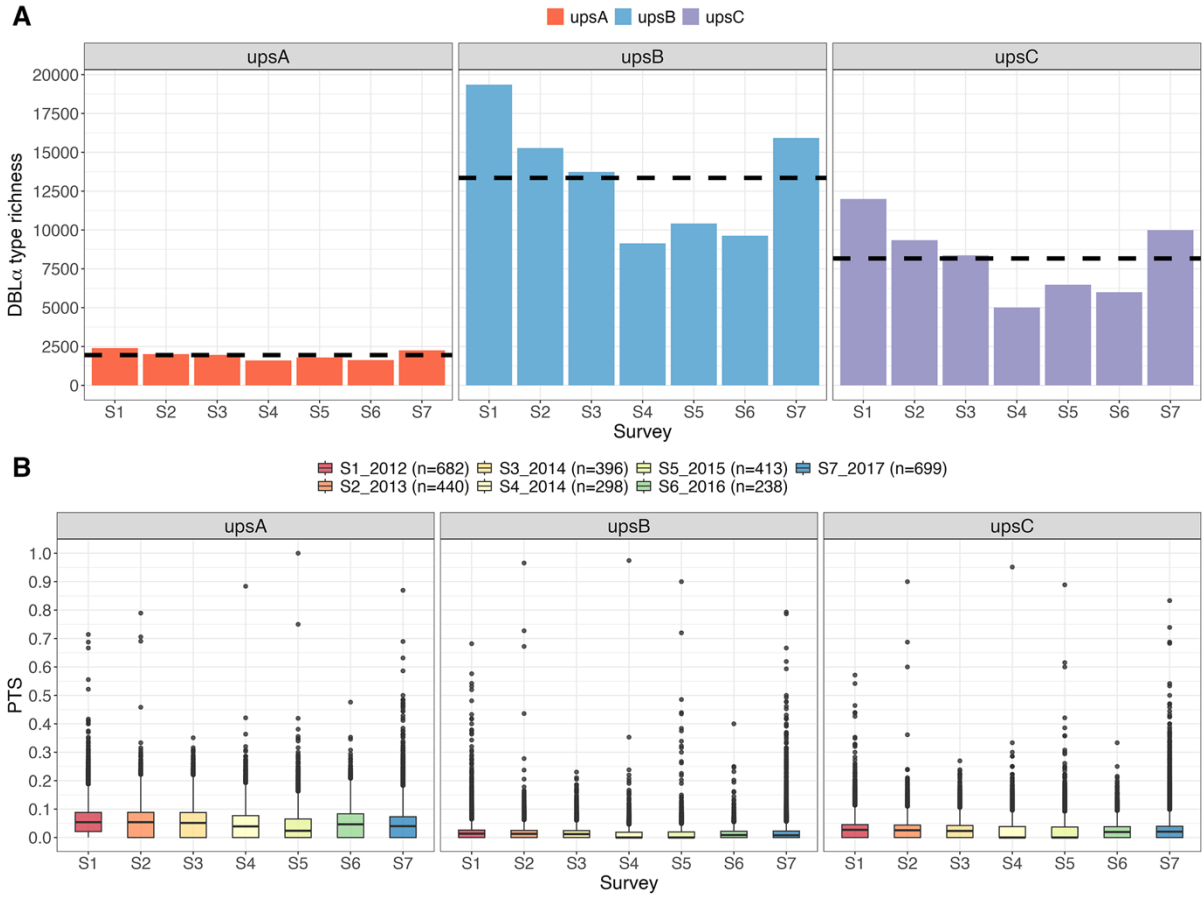
36. Albrecht L, Merino EF, Hoffmann EHE, Ferreira MU, de Mattos Ferreira RG, Osakabe AL, et al. Extense variant gene family repertoire overlap in Western Amazon Plasmodium falciparum isolates. Mol Biochem Parasitol. 2006;150(2):157–65.
37. Albrecht L, Castiñeiras C, Carvalho BO, Ladeia-Andrade S, Santos da Silva N, Hoffmann EHE, et al. The South American Plasmodium falciparum var gene repertoire is limited, highly shared and possibly lacks several antigenic types. Gene. 2010;453(1):37–44.
38. Ruybal-Pesántez S, Sáenz FE, Deed SL, Johnson EK, Larremore DB, Vera-Arias CA, et al. Molecular epidemiology of continued Plasmodium falciparum disease transmission after an outbreak in Ecuador. Vol. 4, Frontiers in Tropical Diseases. 2023.
39. Tiedje KE, Oduro AR, Agongo G, Anyorigiya T, Azongo D, Awine T, et al. Seasonal Variation in the Epidemiology of Asymptomatic Plasmodium falciparum Infections across Two Catchment Areas in Bongo District, Ghana. Am Soc Trop Med Hyg. 2017;97(1):199–212.
40. MalariaGen, Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples [version 2; peer review: 2 approved]. Wellcome Open Res. 2021;6:42.
41. Otto TD, Gilabert A, Crellen T, Böhme U, Arnathau C, Sanders M, et al. Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. Nat Microbiol. 2018;3(6):687–97.
42. Sharp PM, Plenderleith LJ, Hahn BH. Ape Origins of Human Malaria. Annu Rev Microbiol. 2020 Sep 8;74(1):39–63.
43. Amambua-Ngwa A, Park DJ, Volkman SK, Barnes KG, Bei AK, Lukens AK, et al. SNP Genotyping Identifies New Signatures of Selection in a Deep Sample of West African Plasmodium falciparum Malaria Parasites. Mol Biol Evol. 2012;29(11):3249–53.
44. Amambua-Ngwa A, Danso B, Worwui A, Ceesay S, Davies N, Jeffries D, et al. Exceptionally long-range haplotypes in Plasmodium falciparum chromosome 6 maintained in an endemic African population. Malar J. 2016;15(1):515.
45. Eastman RT, Fidock DA. Artemisinin-based combination therapies: a vital tool in efforts to eliminate malaria. Nat Rev Microbiol. 2009;7(12):864–74.
46. Bosman A, Mendis. KN. A Major Transition in Malaria Treatment: The Adoption and Deployment of Artemisinin-Based Combination Therapies. In: Breman JG, Alilio MS, White NJ, editors. Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume 77(6) of American Journal of Tropical Medicine and Hygiene. Northbrook (IL): American Society of Tropical Medicine and Hygiene; 2007.
47. Zanghì G, Vembar SS, Baumgarten S, Ding S, Guizetti J, Bryant JM, et al. A Specific PfEMP1 Is Expressed in P. falciparum Sporozoites and Plays a Role in Hepatocyte Infection. Cell Rep. 2018 Mar 13;22(11):2951–63.
48. Hill AVS, Allsopp CEM, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, et al. Common West African HLA antigens are associated with protection from severe malaria. Nature. 1991;352(6336):595–600.
49. Hill AVS, Yates SNR, Allsopp CEM, Gupta S, Gilbert SC, Lalvani A, et al. Human Leukocyte Antigens and Natural Selection by Malaria. Philos Trans Biol Sci. 1994 Jul 12;346(1317):379–85.
50. Lima-Junior J da C, Pratt-Riccio LR. Major Histocompatibility Complex and Malaria:

Focus on Plasmodium vivax Infection. Vol. 7, Frontiers in Immunology. 2016.
51. Peterson TA, Bielawny T, Kimani M, Ball TB, Plummer FA, Luo M, et al. Diversity and frequencies of HLA class I and class II genes of an East African population. 2014;
52. Hill AVS, Elvin J, Willis AC, Aidoo M, Allsopp CEM, Gotch FM, et al. Molecular analysis of the association of HLA-B53 and resistance to severe malaria. Nature. 1992;360(6403):434–9.
53. Goeury T, Creary LE, Brunet L, Galan M, Pasquier M, Kervaire B, et al. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa. HLA. 2018 Jan 1;91(1):36–51.
54. Sanchez-Mazas A. African diversity from the HLA point of view: influence of genetic drift, geography, linguistics, and natural selection. Hum Immunol. 2001;62(9):937–48.
55. Taylor HM, Kyes SA, Harris D, Kriek N, Newbold CI. A study of var gene transcription in vitro using universal var gene primers. Mol Biochem Parasitol. 2000;105(1):13–23.
56. Bull PC, Berriman M, Kyes S, Quail MA, Hall N, Kortok MM, et al. Plasmodium falciparum Variant Surface Antigen Expression Patterns during Malaria. PLOS Pathog. 2005 Nov 18;1(3):e26.
57. Eddy SR. Accelerated Profile HMM Searches. PLOS Comput Biol. 2011;7(10):e1002195.
58. He Q, Pilosof S, Tiedje KE, Ruybal-Pesántez S, Artzy-Randrup Y, Baskerville EB, et al. Networks of genetic similarity reveal non-neutral processes shape strain structure in Plasmodium falciparum. Nat Commun. 2018;9(1):1817.
59. Otto TD, Böhme U, Sanders MJ, Reid AJ, Bruske EI, Duffy CW, et al. Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres [version 1; peer review: 3 approved]. Wellcome Open Res. 2018;3(52).
60. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Jan 1;7(1):539.
61. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998 Jan 1;14(9):755–63.
62. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.
63. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.

FUNDING

This study was funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health through the joint NIH-NSF-NIFA Ecology and Evolution of Infectious Disease award R01-AI149779 to KPD. Salary support for MT was provided by R01-AI149779.

790 **FIGURES**



791

792 **Figure 1. Classification of DBLα types into ups groups (upsA, upsB, upsC) [Malaria Reservoir**
793 **Study (MRS)].** Figure shows (A) DBLα richness (horizontal dashed lines show mean richness
794 per ups group) and (B) genetic similarity (i.e., overlaps in isolate repertoire by pairwise type
795 sharing (PTS)), in each of the seven MRS surveys in Bongo, Ghana.

796

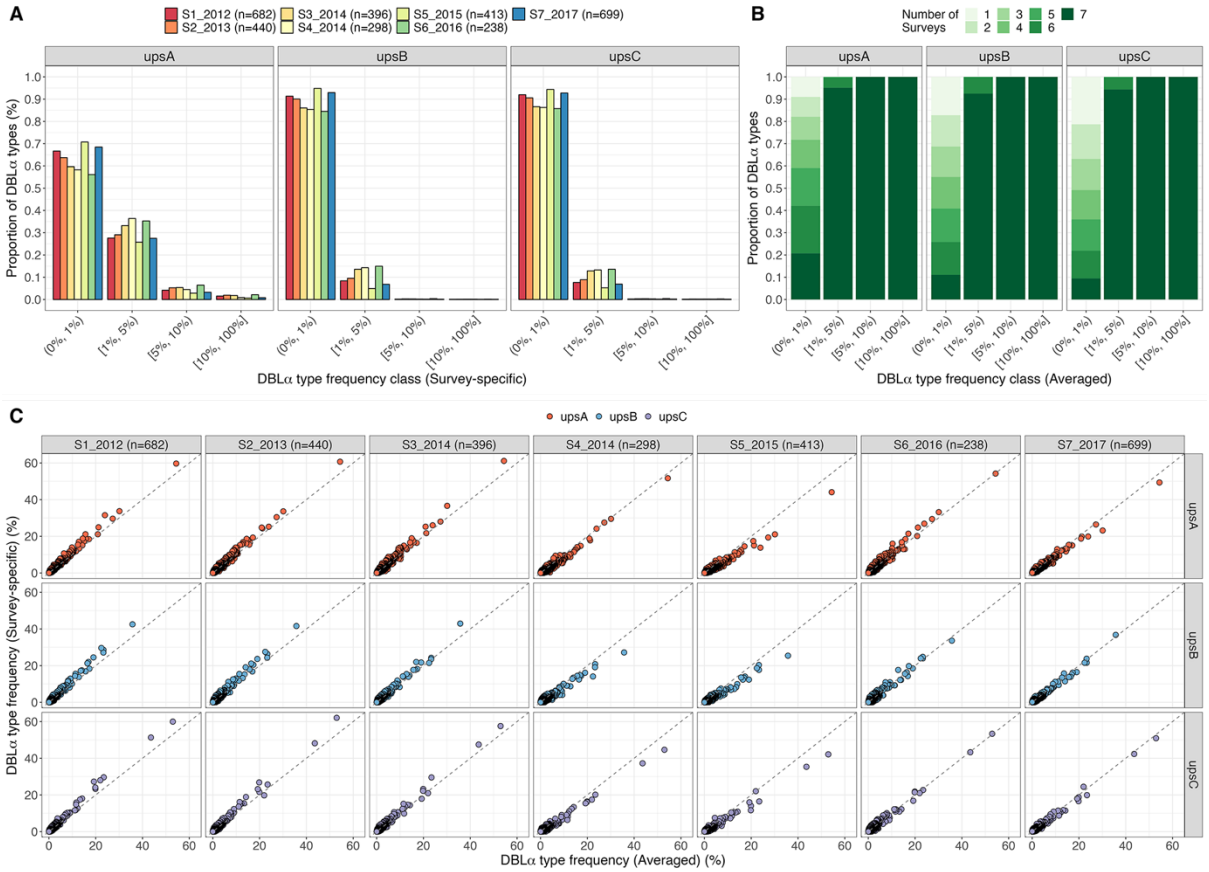


Figure 2. Conservation of DBLα types and frequencies in a local population and through time [Malaria Reservoir Study (MRS)]. (A) Distribution of survey-specific frequencies of DBLα types, binned into categorical frequency classes. (B) Number of surveys DBLα types were observed in, showing that DBLα types found at ≥1% survey-averaged frequencies were seen to also persist through most surveys. This is also true for DBLα types found at ≥ 1% survey-specific frequencies (Figure IV in Data S1). (C) Survey-specific frequencies (y-axis) of DBLα types are plot against survey-averaged frequencies (x-axis), showing positive correlation between both frequencies. Points represent individual DBLα types, coloured by ups groups.

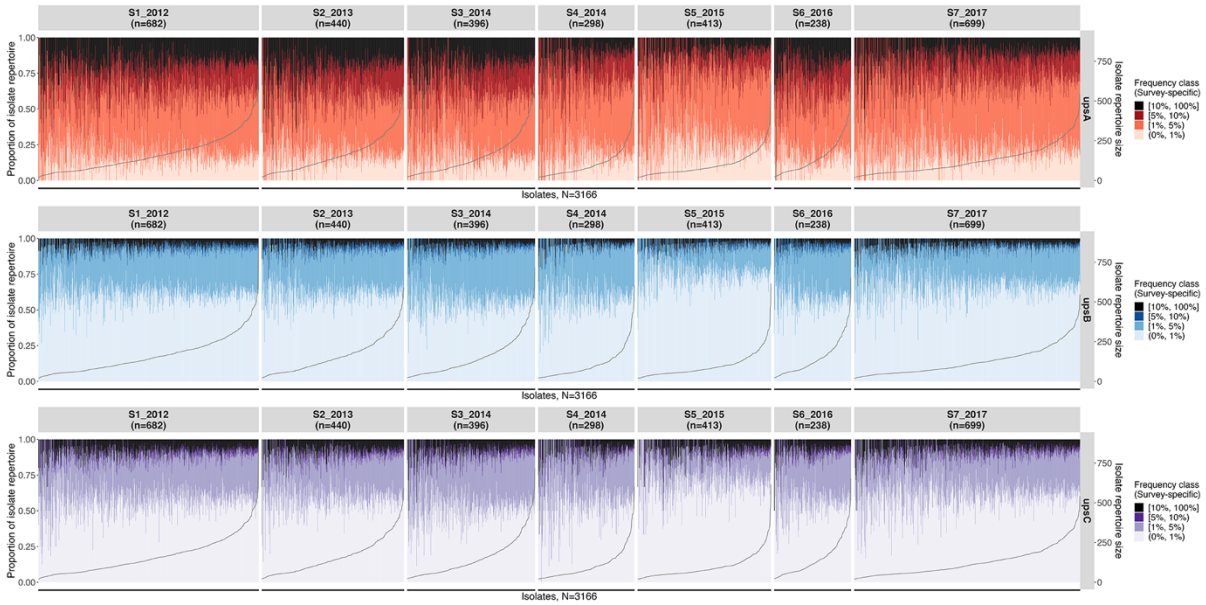


Figure 3. Per-isolate frequency profiles show the composition of survey-specific frequency classes in every isolate repertoire [Malaria Reservoir Study (MRS)]. Vertical bars represent individual isolates, showing the composition of frequency classes in each isolate (left y-axis) by ups group (horizontal panels). Isolates are sorted by isolate repertoire size in increasing order, with isolate repertoire sizes indicated by the grey line (right y-axis).

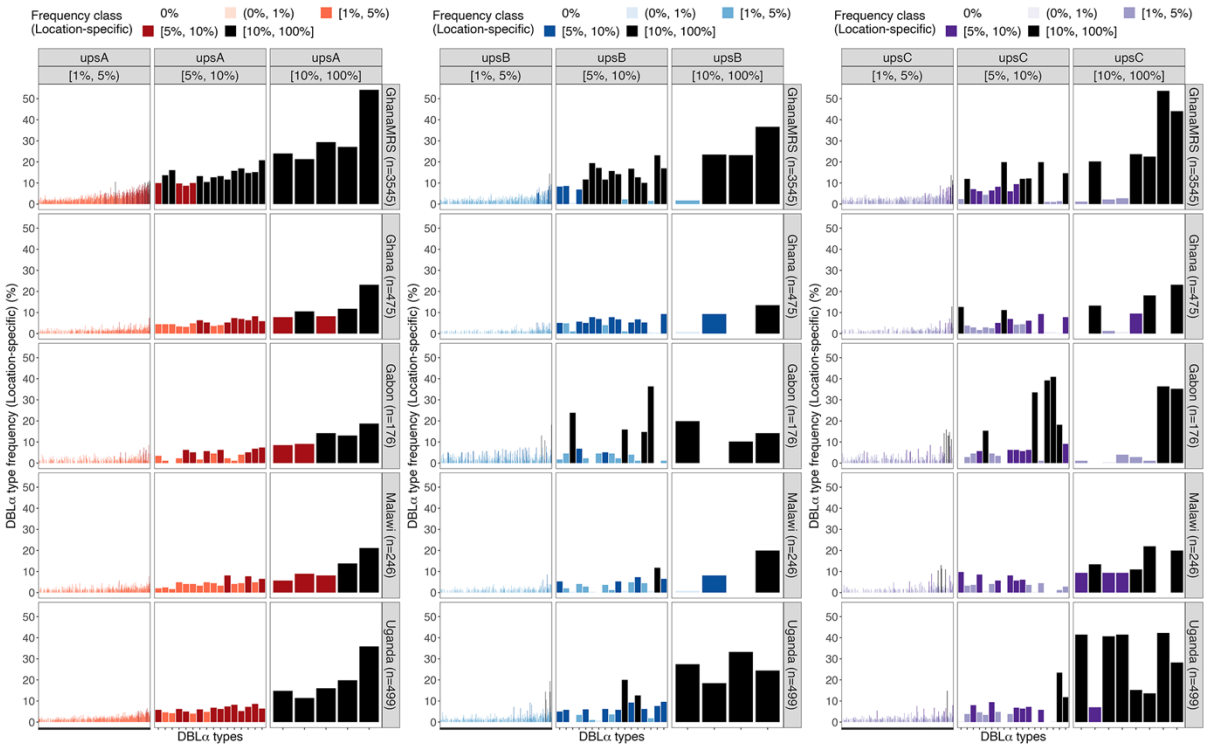


Figure 4. Conservation of DBLα types and frequencies at local and continent levels [spatial analysis]. Location-specific frequencies of individual DBLα types by ups groups (left to right: upsA, upsB, upsC). Shown here are DBLα types with ≥1% location-averaged frequencies, ordered in increasing frequencies.

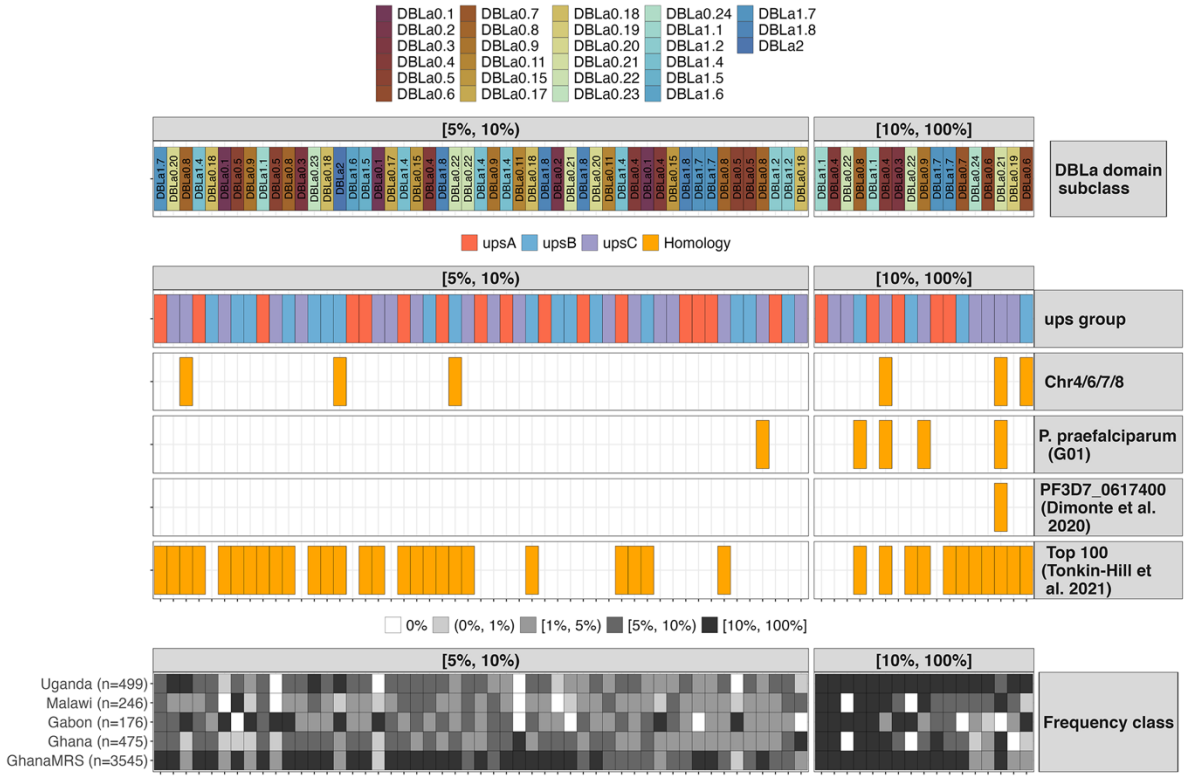


Figure 5. Annotation of possible factors maintaining conserved DBLα types [spatial analysis]. Vertical panels indicate DBLα types with high or very high frequencies (≥5% location-averaged frequencies) and sequence homology (presence/absence) to published DBLα tag and *var* sequence data.