University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Tan, MH;Tiedje, KE;Feng, Q;Zhan, Q;Pascual, M;Shim, H;Chan, YB;Day, KP

Title:
A paradoxical population structure of var DBLα types in Africa

Date:
2025-02-01

Citation:
Tan, M. H., Tiedje, K. E., Feng, Q., Zhan, Q., Pascual, M., Shim, H., Chan, Y. B. & Day, K. P. (2025). A paradoxical population structure of var DBLα types in Africa. Plos Pathogens, 21 (2), pp.e1012813-. https://doi.org/10.1371/journal.ppat.1012813.

Persistent Link:
https://hdl.handle.net/11343/359822

License:
CC BY

PLOS PATHOGENS

RESEARCH ARTICLE

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[4], Heejung Shim[2], Yao-ban Chan[2], Karen P. Day[1*]

1 Department of Microbiology and Immunology, Bio21 Institute and The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, Australia, **2** School of Mathematics and Statistics / Melbourne Integrative Genomics, The University of Melbourne, Melbourne, Australia, **3** Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, Illinois, United States of America, **4** Department of Biology, New York University, New York, New York, United States of America

☯ These authors contributed equally to this work.
* karen.day@unimelb.edu.au (KPD)

Check for updates

OPEN ACCESS

Citation: Tan MH, Tiedje KE, Feng Q, Zhan Q, Pascual M, Shim H, et al. (2025) A paradoxical population structure of *var* DBLα types in Africa. PLoS Pathog 21(2): e1012813. https://doi.org/10.1371/journal.ppat.1012813

Editor: Zbynek Bozdech, Nanyang Technological University, SINGAPORE

Received: February 5, 2024

Accepted: December 6, 2024

Published: February 4, 2025

Copyright: © 2025 Tan et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data tables of calculated frequencies from this study are available online at https://github.com/mh-tan/Paradoxical_DBLa_popstructure. The cUps algorithm is available at https://github.com/qianfeng2/cUps. Publicly available datasets were analysed in this study. Targeted amplicon sequencing datasets from GenBank (Bongo - BioProject accession PRJNA396962, Uganda - BioProject accession PRJNA385208, Gabon - accessions KY328840–KY341897). Assembled var gene data from https://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/PF3K/varDB/.

Abstract

The *var* multigene family encodes *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1), central to host-parasite interactions. Genome structure studies have identified three major groups of *var* genes by specific upstream sequences (upsA, B, or C). *Var* with these ups groups have different chromosomal locations, transcriptional directions, and associations with disease severity. Here we explore temporal and spatial diversity of a region of *var* genes encoding the DBLα domain of PfEMP1 in Africa. By applying a novel ups classification algorithm (*cUps*) to publicly-available DBLα sequence datasets, we categorised DBLα according to association with the three ups groups, thereby avoiding the need to sequence complete genes. Data from deep sequencing of DBLα types in a local population in northern Ghana surveyed seven times from 2012 to 2017 found variants with rare-to-moderate-to-extreme frequencies, and the common variants were temporally stable in this local endemic area. Furthermore, we observed that every isolate repertoire, whether mono- or multiclonal, comprised DBLα types occurring with these frequency ranges implying a common genome structure. When comparing African countries of Ghana, Gabon, Malawi, and Uganda, we report that some DBLα types were consistently found at high frequencies in multiple African countries while others were common only at the country level. The implication of these local and pan-Africa population patterns is discussed in terms of advantage to the parasite with regards to within-host adaptation and resilience to malaria control.

Author summary

The World Health Organisation reported 233 million clinical cases in the African region in 2022, accounting for 94% of global malaria cases. The *var* multigene family encodes an important virulence factor of the dominant malaria parasite, *Plasmodium falciparum*, and is often reported with extreme genetic diversity, particularly in areas with high malaria transmission. Here, we report on the diversity and prevalence of a *var* fragment known as

Funding: This study was funded by Fogarty International Center at the National Institutes of Health through the joint NIH-NSF-NIFA Ecology and Evolution of Infectious Diseases award R01-TW009670 to KPD and MP; and the National Institute of Allergy and Infectious Diseases, National Institutes of Health through the joint NIH-NSF-NIFA Ecology and Evolution of Infectious Diseases award R01-AI149779 to KPD and MP. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Salary support for MT was provided by R01-AI149779 and for KET from R01-TW009670 and R01-AI149779.

Competing interests: The authors have declared that no competing interests exist.

DBLα types in several populations in Africa, using a novel algorithm developed to classify DBLα types into three major ups groups (A, B, C). We found that there were individual DBLα types that persist in a local area through time, occurring at stable high, moderate, or low frequencies in the population. We showed that this frequency structure was possible because every infection also exhibited balanced structures, suggesting that there is pressure for a parasite to maintain common and rare types in its genome. We also identified "local" DBLα types that were present predominantly in a single location but were absent/rare in other locations. By uncovering stable patterns within this complex system, our study provides new insights into the genetics of these important genes to bridge our understanding of interactions of parasites with human hosts.

1. Introduction

Antigenic variation is an immune evasion strategy that has evolved in viral, bacterial, fungal, and protist pathogens. In many taxa, it is mediated by differential expression of hyperdiverse variant antigen genes to rapidly change surface antigens. These antigenic switches facilitate within- and between-host transmission.

Plasmodium parasites of the subgenus Laverania infecting gorillas, chimps, and humans have a multigene family designated *var* [1,2]. In *Plasmodium falciparum*, the most virulent species infecting humans, the *var* multigene family encodes the major variant surface antigen of the blood stages, designated PfEMP1 [3]. This molecule undergoes clonal antigenic variation such that a single infected red blood cell expresses only one of the 40 to 60 *var* genes in a parasite's genome at the trophozoite stage [4,5]. PfEMP1 facilitates sequestration by cytoadhesion to several host receptors to avoid splenic mechanisms of parasite clearance [6–8]. Consequently, certain variants and *var* gene motifs have been identified as virulence factors [9–16].

Var genes exist in multiple architectural types that encode the specific structural arrangements of PfEMP1 domains [17]. These genes evolve by proposed mechanisms involving homologous recombination where sequences encoding specific domains recombine with a homologous domain rather than heterologous [18,19], maintaining a cassette-based structure [20]. Despite the importance of *var* genes to the biology of *P. falciparum*, there have been limited studies of the population genetics of these genes. Currently, they are excluded from genome projects due to challenges in the assembly of this hyperdiverse multigene family. To date there has been only one large-scale *var* gene assembly project publicly available [21] describing aspects of the population genetics of *var* gene domains at the continent level, showing global population structure.

To facilitate data collection on the diversity of these genes in local endemic areas, where tens of thousands of variants have been shown to exist in relatively limited sampling of infected individuals, a strategy has emerged to circumvent the need for complete *var* gene sequences by exploiting the cassette-like structure of these genes [22]. This approach focuses on the Duffy-binding-like alpha (DBLα) domain that is encoded by all *var* genes (Fig 1A), with the exception of one specific *var* gene involved in pregnancy-associated malaria [20]. The DBLα domain is one of the most diverse [19] and has been shown to be immunogenic [14,23]. This domain also potentially plays a role in adhesion, either in itself or in linkage with a proximal domain. Specifically, except for the isolate-transcendent *var* genes (i.e., *var1*, *var2csa*, *var3*), the extracellular N-terminal PfEMP1 head structure consists of a DBLα domain and a cysteine-rich interdomain region (CIDR) (i.e., a DBLα-CIDR tandem) [17,20] and can influence ligand binding and disease pathogenicity [24,25]. Analysis of the relationship between DBLα types

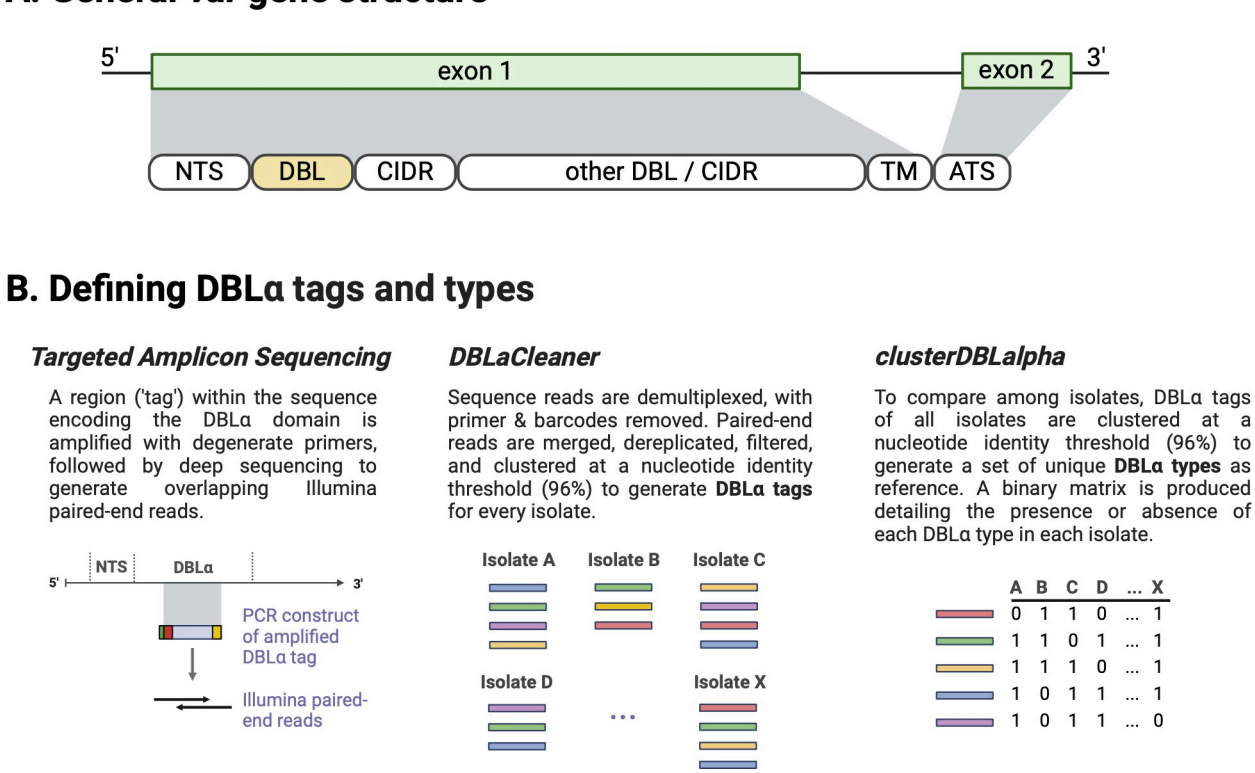

Fig 1. *Var* genes and workflow to generate *var* DBLα sequences. (A) The general *var* gene structure consists of two exons, which encode the extracellular and intracellular portions of the PfEMP1 protein. The first exon typically consists of an N-terminal segment (NTS) on the 5' end, followed by sequences encoding multiple semi-conserved domains such as Duffy binding-like (DBL) and cysteine-rich interdomain region (CIDR) domains that, when combined, make up various *var* domain compositions and structures. Based on the upstream promoter sequence, *var* genes in this multigene family can be further divided into four subgroups of upsA, upsB, upsC, and upsE. **(B)** Defining the DBLα tag region and a description of the analysis workflow to generate DBLα types.

https://doi.org/10.1371/journal.ppat.1012813.g001

and *var* exon 1 sequences has shown that majority of DBLα types (especially in the non-upsA group, >75%) represent a unique *var* gene each in areas with high malaria transmission [26], which serves as the basis for a laboratory protocol for targeted amplicon sequencing of "DBLα tags" (i.e., *var*coding [22]), ultimately generating reference "DBLα types" [22,27] (Fig 1B).

Global studies have shown high diversity and geographic variation in DBLα types [21,28] but with a minority of DBLα types or *var* genes found to be conserved globally at high frequencies (e.g., the top 100 frequent DBLα types in 1,248 isolates across Africa, Asia/Oceania, and South America [28]; conserved *var* gene in 36 of 714 African and Asian parasites [29]). Here, we describe the population frequencies of individual DBLα types in African *P. falciparum* populations in high transmission, where most of the global malaria burden is concentrated. We focused on sampling from the African continent where we have the oldest parasite populations with the greatest diversity of DBLα types [21,28] and extensive geographic variation in the human genome, which may result in selection of local patterns of parasite diversity. Analyses in high-transmission African settings, where we have observed a parasite population structure of largely non-overlapping DBLα repertoires [22,30], also avoids confusing stable frequencies of individual DBLα types that result from clonality or high relatedness as selection.

Genome structure studies of the *var* multigene family have shown that *var* genes can be divided into groups of A, B, C, and E, with a minority of genes grouped into two intermediate groups of B/A or B/C, based on upstream promoter sequences [31]. The three major 'ups'

groups of upsA, upsB, and upsC are associated with different chromosomal locations, transcriptional directions, and sequences [1,4,31,32]. Associations of expression of specific ups groups and disease outcomes have also been found, e.g. upregulated expressions of upsA *var* genes have been shown in cerebral and severe malaria patients, compared to uncomplicated malaria [33–36].

Given these differences, we sought to explore the population genetics of DBLα types by analysis of spatial and temporally collected datasets stratified by ups groups. This was done using a novel ups classification algorithm (*cUps*) introduced in this paper, capable of classifying DBLα types further into upsA, upsB, and upsC groups. Using these short sequences avoids the challenge and expense of long-range sequencing of hyperdiverse *var* genes in multiclonal infections. We analysed the temporal patterns of 62,158 DBLα types and their associated frequencies from seven serial cross-sectional surveys of a local parasite population in Ghana over six years. Within this Ghanaian population, we identified a paradoxical population structure of DBLα types where types in all three major ups groups were maintained at different levels of "common-ness" at low, moderate, high, and extreme frequencies in the population, and this pattern also persisted through time. This observation was possible because every isolate and parasite repertoire also comprised DBLα types occurring at low-to-extreme population frequencies. Based on a further frequency analysis of 79,192 DBLα types from locations in Ghana, Gabon, Malawi, and Uganda, we noted that there were DBLα types that were common in a specific country but not necessarily as common across wider geographical scales, which can suggest local adaptation of the parasite to geographically-diverse receptors and immune response genes of humans 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., referred to as the "Malaria Reservoir Study" (MRS)) (Fig A in S1 Text) [22,37–41]. This MRS dataset consists of sampling at seven time points from 2012 to 2017 at the end of wet (October) or dry (May/June) seasons. Each time point represented an age-stratified cross-sectional survey of approximately 2,000 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, located in northern Ghana. Surveyed participants (i.e., isolates) represented repeat sampling of ~15% of the total population that reside in the two catchment areas in Bongo District at a time (Table A in S1 Text). This area is characterised by high, seasonal malaria transmission and has undergone several types of malaria control interventions, including long-lasting insecticidal nets (LLINs) and indoor residual spraying (IRS) that reduced transmission [22,40], as well as seasonal malaria chemoprevention (SMC) that reduced the burden of infection in children younger than 5 years old [22]. A total of 62,158 representative DBLα types were found in 3,166 asymptomatic isolates from seven surveys (S1 to S7) and served as the dataset for exploring DBLα type frequencies in the parasite population in Bongo (Table A in S1 Text). The number of DBLα types per isolate was not impacted by sequencing depth (Fig B in S1 Text).

To clearly define terminologies, in a high-transmission setting, the asymptomatic "parasite 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. Conversely, at MOI $>$ 1, an isolate's DBLα repertoire would encompass $>$ 1 parasites' DBLα repertoire.

2.2 A novel ups classification algorithm based on DBLα sequences

There are limitations to current approaches to classify DBLα types into their respective ups groups. Ups grouping of *var* genes with phylogeny-based methods typically require 5' UTR sequences that are not generated from targeted amplification [20]. Without access to assembled genomes, as is common in many large-scale targeted amplicon sequencing projects, a current approach exists to classify DBLα types into ups groups, but is limited to only differentiating upsA from non-upsA types by the DBLα domains identified [42]. This study introduces ***cUps***, a novel algorithm for classifying DBLα types further into the different groups of upsA, upsB, and upsC (S2 Text), reporting higher levels of richness of upsB and upsC types, relative to upsA at the population level (Fig 2A). At the isolate level, the average isolate repertoire consisted of 20.9%, 48.6%, and 30.5% of upsA, upsB, and upsC DBLα types, respectively (Fig C in S1 Text). These proportions differ from those reported in [20] that estimated higher proportions of upsB and lower proportions of upsC in isolate repertoires, based on the average of seven genomes. The *cUps* algorithm showed a tendency to classify more upsB types as upsC types, and this is in line with validation results on the algorithm's specificity and sensitivity, elaborated in S2 Text. A reduced analysis that involved thresholding for DBLα types with higher confidence in classification yielded similar patterns of observation we report in this manuscript. Genetic similarity of isolate repertoires by pairwise type sharing (PTS) remained low for all ups groups (median PTS: 0.0455 (upsA), 0.0094 (upsB), and 0.0215 (upsC)) (Fig 2B), where PTS values from 0 to 1 represent the range from unrelated to identical isolate repertoires. The 62,158 representative DBLα types from the seven combined MRS surveys were

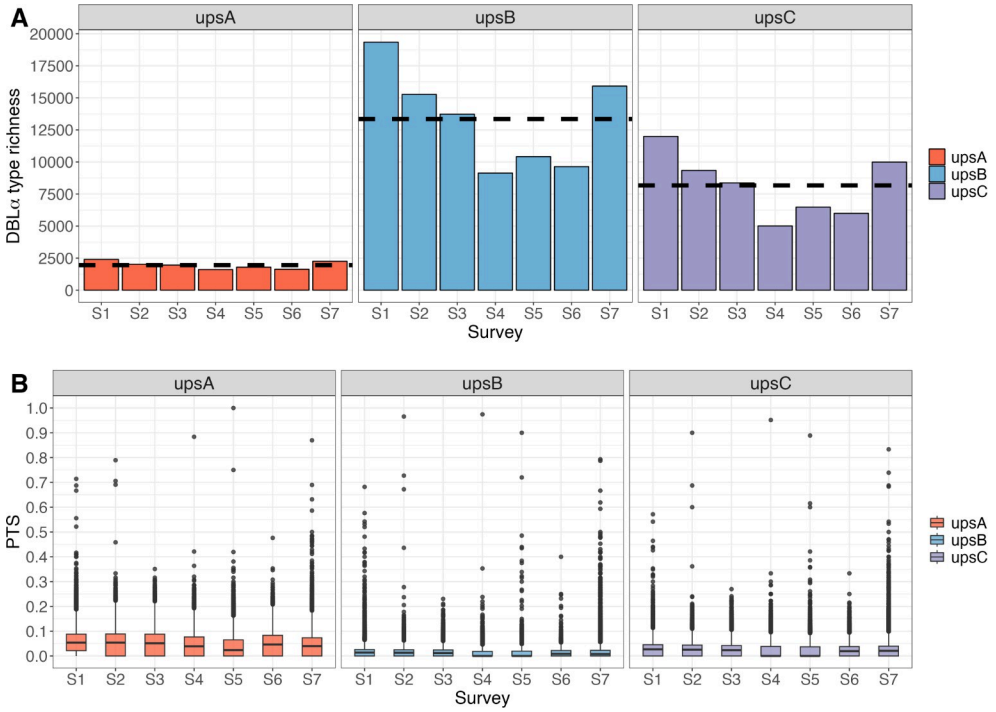

Fig 2. Classification of DBLα types into ups groups (upsA, upsB, upsC) for each of the seven MRS surveys in Bongo, Ghana [Malaria Reservoir Study (MRS)]. Higher DBLα type richness (i.e., number of unique DBLα types) and lower repertoire overlap is observed in upsB and upsC groups, relative to upsA. **(A)** DBLα type richness, where the horizontal dashed lines show mean richness per ups group. **(B)** Genetic similarity / overlap of isolate repertoires by pairwise type sharing (PTS). Box plots indicate the median value (centre line), interquartile ranges (IQR, upper and lower quartiles), 1.5× IQR (whiskers), and outliers (points).

https://doi.org/10.1371/journal.ppat.1012813.g002

classified into upsA, upsB, and upsC groups (Table A in S1 Text). The differences in proportions of ups groups at the isolate *vs* population levels are attributed to the negative relationship between PTS and richness, as a higher level of upsA DBLα type sharing among isolates will result in a lower proportion of unique representative upsA DBLα types in the population.

2.3 Stability of DBLα types and frequencies in the local Bongo population

The frequency of a DBLα type is defined as the proportion of isolates with this DBLα type and was calculated in the context of individual surveys (i.e., "survey-specific frequency") or of the average of seven surveys (i.e., "survey-averaged frequency"). These frequencies were further categorised into four classes to indicate different levels of "common-ness" with ranges given in interval notations: ***low*** (0%, 1%), ***moderate*** [1%, 5%), ***high*** [5%, 10%), and ***extreme*** [10%, 100%] frequencies. While it is well reported that DBLα types in the upsA group are generally more commonly shared relative to DBLα types in the upsB and upsC groups, this study identified individual DBLα types that were stable in the context of sequences and frequencies in all three ups groups at the population level, as described in the following subsections.

2.3.1. A subset of DBLα types was present in many isolates at each surveyed time point. Most DBLα types occurred at low frequencies, found in <1% of isolates. For all three ups groups, the second largest subsets of DBLα types were found to be moderately common, present at 1% to 5% frequencies in each survey, followed by smaller subsets of highly or extremely common DBLα types, exceeding 5% or 10% frequencies, respectively (Fig 3A). The most common 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 DBLα types with moderate-to-extreme frequencies in all three ups groups (upsA: 526 to 801 types per survey, upsB: 540 to 1,918 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: 29.3% to 43.7% per survey, upsB: 5.2% to 15.5% per survey, upsC: 5.6% to 14.3% per survey) (Table B in S1 Text).

2.3.2. Moderate-to-extremely common DBLα type sequences persisted in the population through time. *Var* genes were thought to be inherently unstable, based on *in vitro* evolution experiments, with the DBLα domain reported with the highest rate of recombination [19]. Contrary to this report, this study observed that specific DBLα types with moderate-to-extreme frequencies in each survey were seen persisting through time for all ups groups (Fig 3B, Fig D in S1 Text). All highly or extremely common DBLα types in a survey were also present across all seven surveys. The majority of those moderately common DBLα types were found in all seven surveys, with a smaller subset found in four to six surveys. Notably, DBLα types found in six surveys were missing mostly in the post-IRS intervention survey (S5). On the other hand, the remainder of DBLα types present at low frequencies were found in the range of one to seven surveys.

2.3.3. DBLα type frequencies were stable through time. The population frequencies of DBLα types were also maintained across 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 present at high frequencies in individual surveys were also present at high frequencies consistently through time (Fig 3C, Fig E in S1 Text). Some statistically significant variation across surveys of the frequencies of the 500 most common types was observed; however, this variation appears to come exclusively from surveys S4 and S5, which were affected by the IRS intervention (Fig F in S1 Text). Likewise, most moderately common DBLα types maintained stable frequencies across multiple surveys; no

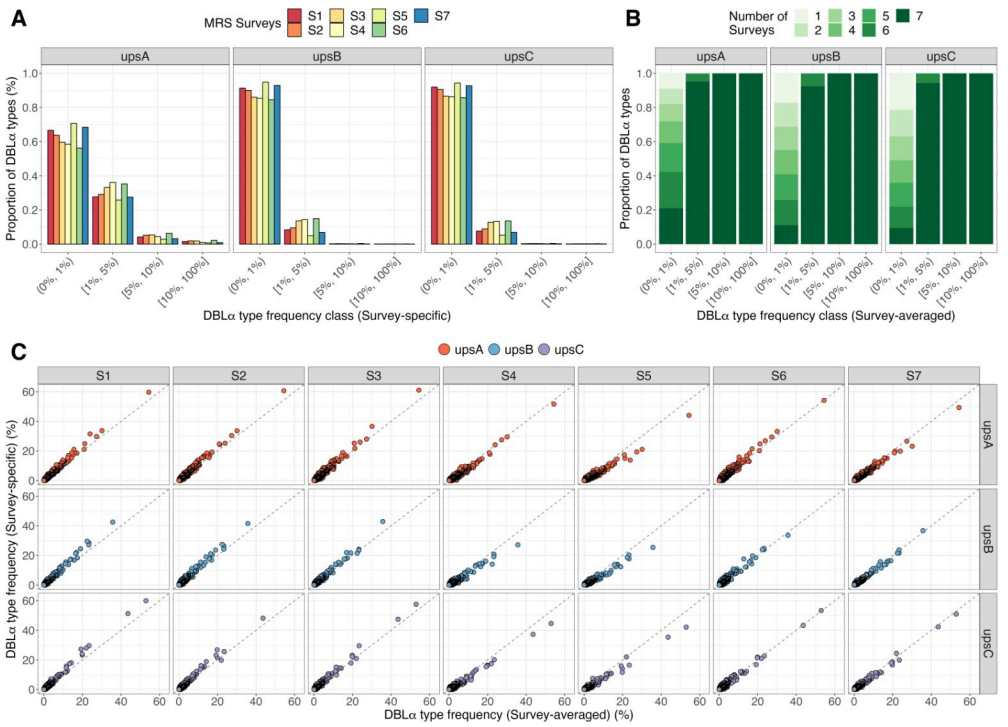

Fig 3. Stable DBLα type sequences and frequencies are observed in a local population and through time [Malaria Reservoir Study (MRS)]. (A) Binned into categorical frequency classes with ranges given in interval notations of low (0%, 1%), moderate [1%, 5%), high [5%, 10%), and extreme [10%, 100%], distributions of survey-specific DBLα type frequencies show that most DBLα types are present at low frequencies, followed by those at moderate frequencies, and a minority of types present at high or extreme frequencies. These proportions are listed in Table B in S1 Text. 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. **(B)** Based on the number of surveys DBLα types are observed in, DBLα types that are present at moderate-to-extreme survey-averaged frequencies (≥1%) are shown to also persist through time. This was observed for all three ups groups. This is also true for DBLα types present at moderate-to-extreme survey-specific frequencies (Fig C in S1 Text). **(C)** Strong, positive correlation is shown between survey-specific frequencies and survey-averaged frequencies of individual DBLα types (points), coloured by ups groups, indicative of stable DBLα type frequencies through time.

https://doi.org/10.1371/journal.ppat.1012813.g003

significant variation could be detected even in surveys S4 and S5. For these surveys affected by the IRS intervention, while rare types were lost and other types shifted to lower frequency classes, the rank of DBLα types by frequency did not change.

2.4 Isolate and parasite genome repertoires consisted of a mix of common and rare DBLα types

While isolate DBLα repertoires in high-transmission populations have been reported to be unrelated and largely non-overlapping [22,30,41–43], there has not been a detailed exploration of the composition of DBLα types within an isolate and their associated frequencies in the population (i.e., 'per-isolate frequency profiles'). These per-isolate frequency profiles comprised of a mix of all frequency classes, based on survey-specific frequencies, and were consistent across isolates within the same survey, regardless of isolates' infection complexities (Fig 4, Fig G in S1 Text). Fitting a binomial regression to per-isolate proportions of DBLα types in each frequency class did not detect any underdispersion, which would arise from a force of balancing selection to maintain these proportions at a fixed level in each isolate. Importantly, observing these per-isolate frequency profiles for monoclonal isolates (i.e., MOI = 1) indicate that these per-isolate

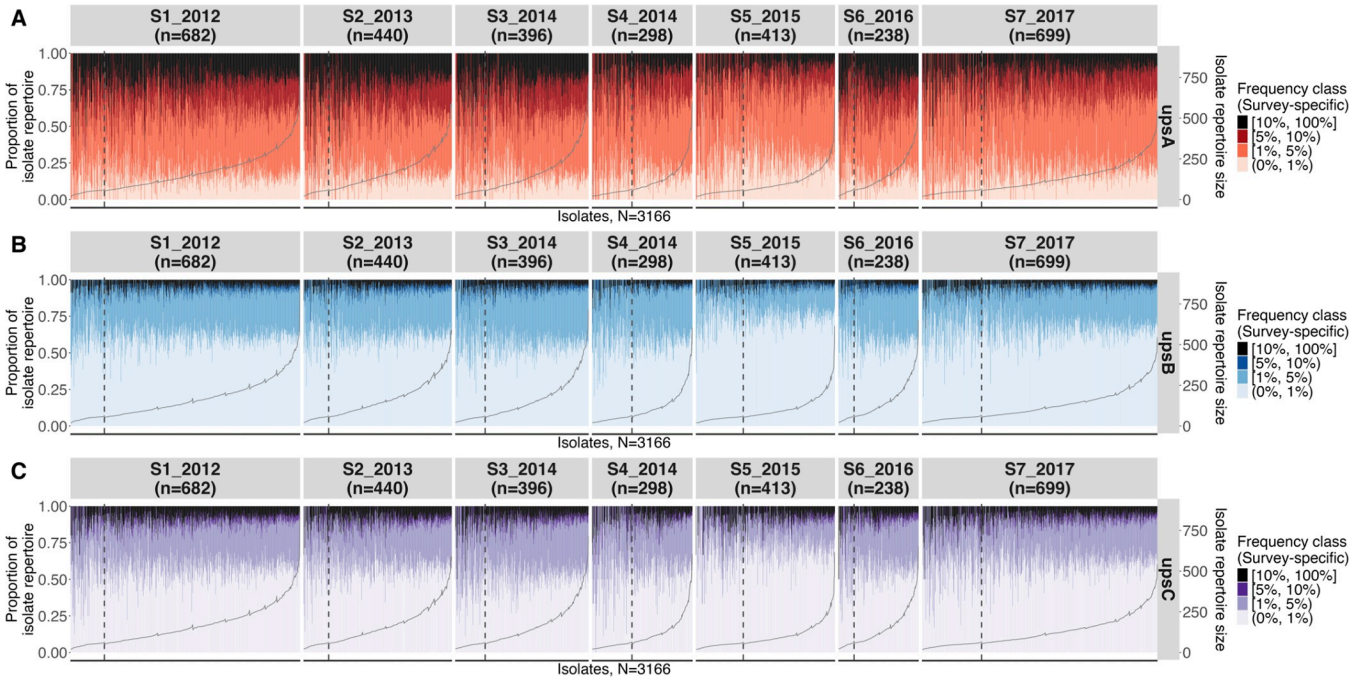

Fig 4. A consistent pattern in per-isolate frequency profiles for upsA, upsB, and upsC DBLα types reveals a new aspect of genome structure [Malaria Reservoir Study (MRS)]. For all ups groups, the per-isolate frequency profiles comprised of a mix of low-to-extreme survey-specific frequency classes and are consistent across isolates within the same survey, regardless of isolates' infection complexities. Per-isolate frequency profiles are shown here by ups group **(A)** upsA, **(B)** upsB, and **(C)** upsC and by survey (vertical panels), where the 'n' value in the label represents the number of isolates per survey. Vertical bars represent individual isolates, ordered in increasing MOI and isolate repertoire size (grey line, secondary y-axis). Colours indicate survey-specific frequency classes with ranges given in interval notations of low (0%, 1%), moderate [1%, 5%), high [5%, 10%), and extreme [10%, 100%], and the proportions of these frequency classes within each isolate is shown on the primary y-axis. Within each panel, dashed vertical lines separate monoclonal (MOI = 1, left of line) and multiclonal (MOI > 1, right of line) isolates, whereby the monoclonal infections reflect the composition within actual parasite repertoires.

https://doi.org/10.1371/journal.ppat.1012813.g004

frequency profiles reflect the repertoire composition within actual parasite genomes (Fig 4). We also confirmed this observation of per-isolate frequency profiles using an independent DBLα sequence dataset [21,26,44], extracted from *var* genes of isolates sampled from Navrongo in Ghana, situated ~30 km adjacent to Bongo District (Fig H in S1 Text). The use of data obtained from two different methods (i.e., targeted amplicon sequencing and whole genome sequencing) yielded similar observations, providing confidence that these patterns were not introduced from the PCR protocol used in *var*coding.

Isolates' upsA frequency profiles consisted of mostly moderate-to-extremely common DBLα types and relatively small proportions of rare or unique DBLα types. In contrast, isolates' upsB and upsC frequency profiles consisted of large proportions of DBLα types in the low frequency class, followed by those in the moderate frequency class. The introduction of malaria control interventions did not perturb these frequency profiles, which maintained the composition of different frequency classes but in different proportions. In surveys of the population affected by interventions (e.g., S4 and S5, during and post-IRS), per-isolate frequency profiles generally trended toward a larger proportion of relatively rare DBLα types and smaller proportions of relatively common types within each isolate (Fig 4). However, as noted above, the rank of DBLα types by population frequencies were maintained.

To investigate the extent of sharing of DBLα types within each of the four frequency classes, genetic similarity among pairwise isolate repertoires was calculated (i.e., pairwise type sharing (PTS) values). As expected, PTS values increased as rare DBLα types were excluded (Fig I in S1 Text). Interestingly, even when considering only extremely common DBLα types, median PTS

values remained generally low (median PTS of 0.02, 0.05, 0.13, and 0.20 when considering DBLα types at >0%, ≥1%, ≥5%, ≥10% survey-averaged frequencies, respectively, across all surveys). This indicates that, even though every isolate repertoire contained a proportion of types that were present in many isolates in the population, identical sets of common DBLα types were rarely observed. When evaluating DBLα types exclusively in the different ups groups, shifts in PTS distributions were more substantial for the upsA or upsC groups relative to the upsB group, consistent with the lower DBLα richness in the two former groups.

2.5 Local DBLα types were detected in individual African countries

A separate geographical study of DBLα types and frequencies in multiple African countries (i.e., "locations") representing West Africa (Ghana, Gabon), Central Africa (Malawi) and East Africa (Uganda) was conducted based on 79,192 DBLα types found in 4,561 isolates (Table A and Fig A in S3 Text) [21,26,44]. Similarly, this geographical study showed that the majority of DBLα types in each location were present at low frequencies with smaller proportions seen at relatively higher frequencies. While most of the more common upsA DBLα types were present at relatively stable frequencies in most of the analysed countries, DBLα types in the upsB and upsC groups were common at either the continent or local levels (Fig 5). This was evident from finding common upsB and upsC DBLα types that were present predominantly in a single location but were absent or uncommon in other locations, suggesting local selection (Fig 5, Fig B in S3 Text). As observed for isolates in the Ghana MRS study, per-isolate frequency profiles in these different locations also consisted of a mix of common and rare DBLα types (Fig C in S3 Text).

It is worth remembering that these common DBLα types made up the minority of all DBLα types in every ups group, especially in the upsB and upsC groups that formed the majority of DBLα types in a population as well as in an isolate's repertoire. An exploration of the relationship between DBLα types and *var* exon 1 sequences revealed that DBLα types at high or extreme frequencies tended to be associated with multiple different *var* exon 1 sequences (Fig D in S3 Text), 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, analysis of sequence similarity between *var* exon 1 sequences that shared the same DBLα type suggested that the majority of these *var* exon 1 sequences exhibited low shared identity and therefore appeared to represent actual different genes (Fig D in S3 Text), with a smaller subset potentially being alleles of a same gene [26]. In a highly-dynamic system where sequences encoding the DBLα domain have been shown *in vitro* to exhibit the highest recombination rate [19], the maintenance of specific DBLα types at high frequencies and through extensive durations is characteristic of balancing selection. We make clear that this study describes the conservation of DBLα types, where we envisage evolution of domains or cassettes, but not necessarily the conservation of *var* genes.

2.6 Factors maintaining DBLα type sequences and frequencies remain unknown

The geographical study considered a few possible factors to explain these common DBLα types, focusing specifically on 50 and 17 DBLα types in the high and extreme location-averaged frequency classes, respectively (Fig E in S3 Text). While positional information is unavailable for the DBLα types analysed in this study, sequence alignments showed that only six of the 67 DBLα types are homologous to DBLα tags of *var* genes on specific *P. falciparum* chromosomes 4, 6, 7, and 8 that were previously detected at high frequencies in various populations, potentially attributed to selective sweep events associated with antimalarial drug

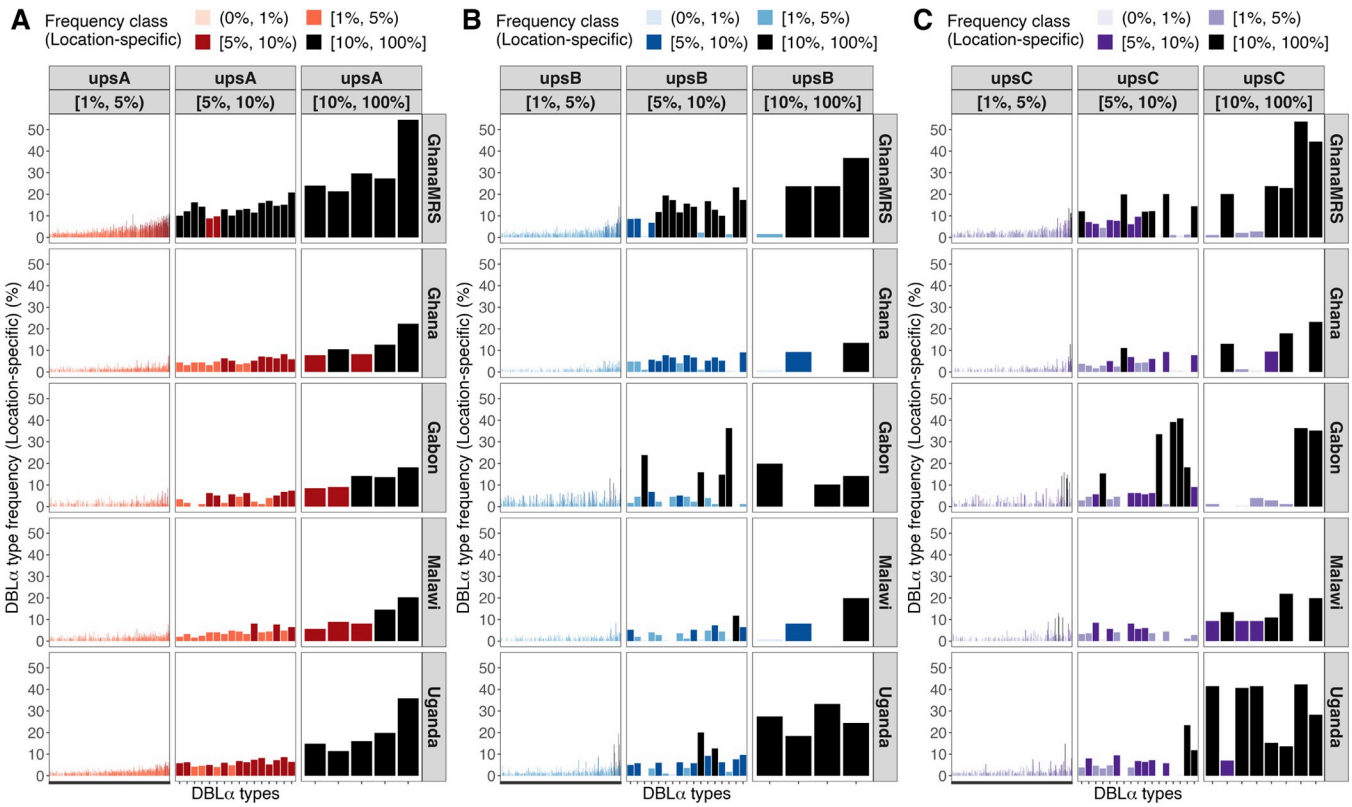

Fig 5. DBLα type frequencies within Ghana, Gabon, Malawi, and Uganda reveal location-specific signatures [geographical analysis]. These location-specific signatures are more apparent in the upsB and upsC groups, with DBLα types that are present predominantly in a single location but are absent or at low frequencies in other locations. Shown here for **(A)** upsA, **(B)** upsB, and **(C)** upsC groups are DBLα types with moderate-to-extreme location-averaged frequencies (≥1%), ordered in increasing location-averaged frequencies. Vertical panels represent location-averaged frequency classes and horizontal panels represent locations in Africa. Bars indicate the location-specific frequencies of individual DBLα types, further coloured according to location-specific frequency classes with ranges given in interval notations of low (0%, 1%), moderate [1%, 5%), high [5%, 10%), and extreme [10%, 100%].

https://doi.org/10.1371/journal.ppat.1012813.g005

resistance [21]. Furthermore, some of these DBLα types with high or extreme frequencies were identified as homologs to the DBLα tags of five *var* genes in *P. praefalciparum*, the closest living sister species of *P. falciparum* that naturally infects gorillas [2, 45]. Homologs to the DBLα tags of two *var* genes of another ancestral Laverania species, *P. reichenowi*, were identified but were present at low-to-moderate frequencies in this geographical dataset (frequencies range from 0.57% to 2.40%). No homologs to the DBLα tags of *P. gaboni var* genes were identified. While some of these factors can explain the reason a few of these sequences were conserved, the majority of these DBLα types with relatively high frequencies were still unaccounted for.

With respect to published studies on globally-conserved DBLα types or *var* genes, homologs to 84 of the globally-conserved DBLα types [28] were identified in this study, with 37 of these DBLα types found in the two highest frequency classes. Furthermore, in the context of general prevalence in the analysed African locations, 30 of these 37 DBLα types were found in all four locations, six in three locations, and only one was found in a single location. The homolog to the DBLα tag of a conserved *var* gene that was reported in a Gabonese parasite isolate [29] was present at high-to-extreme frequencies (ranging from 7.0% to 20.1% in different locations) and present in all locations except, strangely, in Gabon itself. Additionally, the homolog to the DBLα tag of a *var* gene that was reported to be expressed in sporozoites and potentially play a role in hepatocyte infection [46] was present at only <1% frequency in three locations. These annotations are available in the data tables online (see Data Availability section).

3. Discussion

Extensive DBLα type diversity is reported in areas with high malaria transmission, generated by meiotic and mitotic recombination [18,19,47–49] with parasite repertoire diversity driven by frequent outcrossing in the mosquito vector [50,51], such that we would not expect conservation of types. Here we report a paradoxical population structure of DBLα types, where types are seen stable through time and can be found in a population at various frequencies, be it low, moderate, or high, in all three major ups groups. Underlying and maintaining this observed population structure of DBLα types in a local endemic area is the frequency profile of DBLα types within every isolate repertoire, whether mono- or multiclonal. This implies a common genome structure. Whilst multiclonality could potentially lead to the underestimation of sharedness and conversely the overestimation of uniqueness, we perceive this to be of minimal effect due to the reported lack of DBLα repertoire overlap amongst likely monoclonal isolates [30] and the large effective population size in high transmission [22,41]. The consistency of these per-isolate frequency profiles, seen within our Ghana MRS population and the independent African DBLα datasets [21], therefore suggests that each isolate repertoire must have a combination of common and rare types for the three major ups groups while still maintaining limited overlaps with other isolates in the population overall. It is obvious from prior work that rare types would be non-overlapping [22,30], but what is striking in this analysis is that the common types of all three major ups groups were not co-occurring in the same isolates; low PTS was still observed indicating the lack of sharing of common types (Fig I in S1 Text). We hypothesise that temporal stability of DBLα types and frequency structures in the population may be maintained by human host factors involved in host-parasite interactions.

We propose that DBLα types that exist in low-to-moderate frequencies in the parasite population serve to provide the parasite with options to benefit within-host survival by antibody-mediated immune evasion while the minority of DBLα types occurring at higher frequencies in the parasite population may reflect adaptations to maintaining infection in a local human host population. We would interpret this observed structure to be a consequence of different rates of recombination of DBLα types resulting in these common *vs* rare types. Alternatively, there are multiple selective forces to maintain DBLα types at such varying frequencies in a parasite population. These would include a variety of host adhesion receptors and immune response genes. Our finding of local signatures of conservation in multiple African countries supports the possibility of local adaptation of individual DBLα types to the human host population, given that there are reports of geographic variation in host adhesion and immune receptors (e.g., [52–54]) and examples of co-evolution [55,56].

Stochastic simulations and network analyses have provided clear evidence for a role of immune selection or negative frequency-dependent selection resulting from specific immune memory, which is a form of balancing selection, in shaping antigenic diversity within natural populations [37]. As antibody-mediated immunity plays a significant role in recognition of PfEMP1 variants, we hypothesise that another possible driver of balancing selection is the arm's race between the parasite PfEMP1 variants and host HLA class II haplotypes [55,57–59]. Similar to our finding of local signatures of DBLα type conservation against a highly-diverse background, there are also geographic differences in HLA class II alleles across the African continent, with allele frequencies also ranging from low to high [60–62]. Immune evasion related to local HLA class II alleles would select for varying frequencies of DBLα types. The paradoxical population structure of DBLα may also be shaped by underlying differences in host receptors of varying spatial niches and if not, this domain could be in linkage disequilibrium with other proximal domains (e.g., CIDR) or genes vital to these roles. Future studies of co-evolution must take into consideration that domains or domain cassettes appear to evolve independently by recombination [20].

The observed parasite genome structure composed of DBLα types with varying frequency classes has significant implications for malaria surveillance and control. The removal of parasite genomes from a population through intervention does not lead to the loss of the same proportion of DBLα types, as the initial removal of parasite genomes involves the loss of rare DBLα types mostly whereas common DBLα types persist until a high proportion of genomes are lost (Fig E in S1 Text). Thus, with interventions, diversity or richness of DBLα types would be seen to decrease in a non-linear manner relative to the removal of parasite genomes from a population (Fig J in S1 Text). Breaking this pattern towards high relatedness or clonality is indicative of a system transitioning into a low-transmission setting and thus could be diagnostic for elimination efforts [63].

In conclusion, the paradoxical population structure of DBLα types created by these consistent per-isolate frequency profile patterns is striking and suggests that maintaining such frequency profiles within a parasite repertoire is advantageous to the parasite. Having a range of rare to common types within each major ups group may allow malaria parasites to adapt to host factors in order to persist through the dynamics and competition within and between hosts.

These observations encourage us to identify the role of host genetic factors in selecting these stable frequencies. Of further interest for investigation is the significance of DBLα frequency profiles within all ups groups in a parasite genome in understanding the hierarchy of *var* gene expression in the human host [64], as well as future studies on gene expression levels and immunity structured according to these specific common *vs* rarer DBLα types within each ups group.

4. Materials and methods

4.1 Data sources and types

Frequency analyses were performed based on a small ~450bp sequence region of a *var* gene that encodes a portion of the DBLα domain of PfEMP1 (i.e., DBLα tags) [65,66]. DBLα tag sequences included in this study were either generated from targeted amplicon sequencing or extracted from assembled *var* gene sequences. This made available DBLα tag datasets of varying sizes from Africa and Asia, which were clustered to generate representative DBLα types. 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, The Gambia, Guinea, Mali, Ghana, Gabon), Central Africa (DR Congo, Malawi) and East Africa (Uganda, Kenya) (Table A in S1 Text and Table A in S3 Text). 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 A in S1 Text and Table A in S3 Text). Amplicon sequencing of DBLα tag sequences typically involves PCR amplification of a small sequence region encoding the DBLα domain of PfEMP1 (Fig 1) using degenerate primers [65,66], followed by high-throughput sequencing on either the Illumina MiSeq platform (GhanaMRS) or on the 454 sequencing platform (Gabon, Uganda). These include sequences from:

i. One area (Bongo) in Ghana: dataset spans seven time points (surveys) from 2012 to 2017 involving sampling of asymptomatic individuals at the end of multiple wet (October) and

dry (May/June) seasons (GenBank BioProject accession number: PRJNA396962) [22,37–41].

ii. One area (Bakoumba) in Gabon: dataset included sampling of asymptomatic children in one year (GenBank accession numbers: KY328840–KY341897) [30].

iii. Six areas (Apac, Arua, Jinja, Kanungu, Kyenjojo, Tororo) in Uganda: dataset included sampling of clinical isolates over two years (GenBank BioProject accession number: PRJNA385208) [42].

4.1.2 DBLα tags from assembled *var* gene sequences. Published *var* gene sequences (from isolates in Africa and Asia) were downloaded from the 'Full Dataset' published by [21]. DBLα tag sequences were identified and extracted from *var* gene sequences (regardless of *var* gene completeness) as described in [26]. Briefly, domain annotations provided by [21] were used to extract nucleotide sequences encoding the DBLα domain. These extracted sequences were further translated into the best reading frames and, using *hmmsearch* [67], the resulting amino acid sequences were further searched against positions 189 to 430 of the PFAM profile alignment (PF05424_seed.txt) to identify the 'tag' region (domain score cut-off of 60 and $\geq$100 aligned positions) and to ultimately extract the DBLα tag sequence that would have been amplified with degenerate primers [65,66]. Isolates were excluded if suspected as laboratory isolate ("Lab" or "Suspected_lab_strain") or incorrectly-designated continent ("Continent_mismatch") based on their metadata.

4.2 Clustering of DBLα tags into DBLα types

DBLα tags (Africa and Asia) were translated into amino acid sequences and any untranslatable sequences (i.e., stop codons in reading frame) were excluded. The remaining DBLα tags were combined and clustered with *clusterDBLa* v1.0 [37] using a 96% nucleotide identity threshold [68] to produce representative DBLα types. This also generated a binary matrix detailing the presence/absence matrix of every DBLα type in every isolate. Initially described in [68], the use of this threshold is further supported by the results shown in Fig 5 of Feng et al. [69] that identified the most frequent recombination breakpoint positions at approximately 0.25, 0.50, and 0.80 relative positions of DBLα types, suggesting that the risk of over-clustering of DBLα tags is higher only when we approach the ~80% threshold point. Previous work shown in Figure S3 (Data Sheet 1) of Tan et al. [26] further reported that varying this similarity threshold from 90% to 100% did not substantially affect the number of total DBLα types generated from isolates in Cambodia, Thailand, Ghana, and Malawi from the dataset of assembled *var* genes [21].

A separate analysis to test the impact of sequencing depth on the number of DBLα types was conducted. For isolates in the GhanaMRS dataset, median read support of DBLα tags was calculated for each isolate, revealing that higher sequencing depth did not lead to a greater number of DBLα types (Fig B in S1 Text).

4.3 Classification of DBLα types into domain classes and ups groups

The *classifyDBLα* v1.0 pipeline [42] was used to classify DBLα types into DBLα domain classes of DBLα0, DBLα1, or DBLα2, in order to confirm that sequences were indeed those encoding the DBLα domain of PfEMP1. In addition, a novel algorithm (***cUps***) described in this study was used to classify DBLα types into the most probable ups group (i.e., upsA, upsB, or upsC), accompanied by assignment probability values. For each DBLα type, ups grouping was assigned according to the prediction with the highest assignment probability. We describe this

novel classification algorithm below and in more detail in S2 Text. An implementation of the algorithm is available at https://github.com/qianfeng2/cUps.

Through the alignment and clustering of 2kb sequences upstream of *var* genes, followed by the classification *var* genes into ups groups by Neighbour-joining (NJ) and Markov clustering (MCL) methods (trees available in S2 Text), a reference dataset of DBLα tag sequences was generated from 846 *var* genes from 16 *P. falciparum* genomes (see S2 Text) [5,20]. We begin with this reference database of DBLα tag sequences with ups groups and DBLα domain subclasses known. For each category (ups group/DBLα subclass combination), we align the reference sequences in the category using Clustal Omega v1.2.4 [70], then fit a profile hidden Markov model [71] using HMMER v3.2.1 [67] 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 domain 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. A threshold may optionally be applied, so that sequences with highest assignment probability below the threshold 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 [72].

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 A in S1 Text and Table A in S3 Text). Specifically for the time-series dataset from the GhanaMRS study in Bongo District, Ghana, submicroscopic or symptomatic isolates were additionally excluded from the dataset. Further, using *blastn* ($\geq$96% nucleotide identity, $\geq$95% query coverage) [73], DBLα types with homology to isolate-transcendent *var1*, *var2csa*, and *var3* sequences (sequences from [20,21]) were excluded to remove putative DBLα types previously reported as isolate-transcendent [1,32]. Finally, given that frequency classes and profiles were calculated based on proportional frequencies, only locations with datasets of $\geq$ 100 isolates were retained. This resulted in the exclusion of six African countries from this study ("*" in Table A in S3 Text).

4.5 Estimation of multiplicity of infection (MOI)

For the GhanaMRS dataset, the number of sequenced non-upsA DBLα types per isolate (i.e., count of upsB and upsC) was converted to its estimated MOI using a published Bayesian approach (prior = "uniform", aggregate = "pool") [22]. This approach relies on the hyperdiversity of DBLα types, particularly those in the non-upsA groups [26], and the limited repertoire similarity [22].

4.6 Genetic similarity between pairwise isolate repertoires

The pairwise type sharing metric (PTS) [68] was used to estimate the overlap between pairwise isolate repertoires (e.g., isolates *i* and *j*). Specifically:

$$PTS = \frac{2*shared_{ij}}{Size_i + Size_j}$$

where $shared_{ij}$ is the number of shared DBLα types between repertoires of isolates *i* and *j*, and $Size_i$ and $Size_j$ are the total number of DBLα types (i.e., repertoire sizes) of isolates *i* and *j*, respectively. A PTS value of 0 indicates the absence of sharing between two isolates whereas a PTS value of 1 indicates completely identical isolate repertoires.

4.7 Calculation of DBLα type frequencies and assignments into frequency classes

Depending on the analysis, a population can be the collection of isolates sampled at a specific survey or time point in the time-series analyses (i.e., by year or survey in the GhanaMRS dataset) or the collection of isolates sampled from a specific region or location/country in the geographical analyses. Raw frequencies of DBLα types were defined at the survey or location level in counts (i.e., number of isolates with a particular DBLα type in each survey or location). Raw frequencies were converted into proportional frequencies through division of count frequencies by the total number of isolates at a corresponding time point or location, leading to "survey-specific frequencies" or "location-specific frequencies". Subsequently, these frequencies were further categorised into frequency classes with ranges given in interval notations: ***low*** (0%, 1%), ***moderate*** [1%, 5%), ***high*** [5%, 10%), and ***extreme*** [10%, 100%] frequencies.

Given the substantial differences in dataset sizes across surveys or locations (e.g., 499 isolates for Uganda *versus* 176 isolates for Gabon), simply summing isolates across datasets of multiple surveys or locations would bias total frequencies to reflect those of larger datasets. Hence, averaged frequencies were used instead as a means to normalise total frequencies by isolate counts in each survey or location ("survey-averaged frequencies" or "location-averaged frequencies"). For example, a DBLα type found in 10 out of 100 isolates for location A and 10 out of 500 isolates for location B would be reported to have 10% and 2% frequencies for locations A and B, respectively. A crude total frequency of 3.33% (20 of 600 isolates) would be more reflective of the frequency observed in location B even though the DBLα type was found at relatively high frequency at location A. In this instance, with normalisation, an averaged frequency of 6% would be estimated (12 of 200 isolates), reducing the bias towards larger dataset sizes. This normalisation method provides a less biased approach in identifying DBLα types that are found at high frequencies in one or more datasets but not necessarily uniformly across all datasets.

4.8 Statistical analysis of frequency maintenance

To determine if DBLα type frequency varied across GhanaMRS surveys, we performed (for each DBLα type) a chi-squared test of association between the presence/absence of that type and the survey. We used a Bonferroni correction to control the family-wise error rate at 0.05; DBLα types that had a p-value below this threshold exhibited significant variation between surveys.

To determine if there was any evidence for balancing selection in the frequency profiles of a single isolate, we fit a binomial regression to the counts of each rarity (low, medium, high, extreme) against the total number of DBLα types in each isolate, separated by ups group and surveys. Significant underdispersion would then indicate the presence of a balancing selection force maintaining the proportions of each frequency at fixed levels.

4.9 Determination of DBLα-*var* relationships

For two locations (Ghana and Malawi), *var* gene sequences were available from assemblies generated by [21]. DBLα-*var* relationships were determined using complete *var* exon 1 sequences that are bounded by an N-terminal segment (NTS) and a transmembrane region (TM) on the 5' and 3' ends of exon 1, respectively [26]. Briefly, using *vsearch* [74], DBLα types

were globally aligned to *var* exon 1 sequences from the same location (e.g., Malawi DBLα types to Malawi *var* exon 1). Given that DBLα types were generated from clustering at a 96% nucleotide identity threshold, these DBLα types were aligned to *var* exon 1 sequences, retaining alignments that meet the same threshold of 96% identity, calculated over the alignment length and excluding terminal gaps *(--iddef 2*). The relationship between a DBLα type and distinct *var* exon 1 was determined based on the number of unique *var* exon 1 sequences sharing a same DBLα type (e.g., a 1-to-*n* DBLα-*var* relationship is defined as a DBLα type found in *n* unique *var* exon 1).

For each group of *var* exon 1 that share a same DBLα type, an all *vs* all sequence alignment of *var* exon 1 sequences in the group was performed using the *allpairs_global* option within *vsearch* [74] and set to include all pairwise alignments *(--acceptall*). Pairwise nucleotide identities were estimated based on calculations over whole alignment lengths, including terminal gaps (*--iddef 1*), to account for differences in pairs of *var* exon 1 of variable lengths.

4.10 Search for sequence homology to other DBLα types or *var* genes

4.10.1 *Var* genes in association with selective sweeps on specific chromosomes. Published work reported conserved *var* genes on chromosomes 4, 6, 7, and 8 associated with selective sweep events, potentially due to antimalarial drug resistance or other factors. Accession numbers of these genes were obtained from the author [5,21] and used as reference. Using *blastn* [73], DBLα types were searched against these reference sequences and hits from alignments were reported (≥96% nucleotide identity, ≥95% query coverage).

4.10.2 *Var* genes in primate *Plasmodium* species. *Var* genes from three *Plasmodium* species, *P. praefalciparum*, *P. reichenowi*, and *P. gaboni*, were downloaded from PlasmoDB [2] and used as reference. Using *blastn* [73], DBLα types were searched against these reference sequences and hits from alignments were reported (≥96% nucleotide identity, ≥95% query coverage).

4.10.3 Globally-conserved DBLα types or *var* genes. The 100 most frequent DBLα sequences reported in the global analysis by Tonkin-Hill et al. [28] was used as reference. Using *blastn* [73], 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 sequences of PFGA01_060022400 and PF3D7_0617400, representing conserved *var* genes that are almost identical in sequence as reported by Dimonte et al. [29], as well as to *var* gene sequence PF3D7_0809100, shown by Zanghì et al. [46] to be expressed at the sporozoite stage.

Supporting information

S1 Text. Study of DBLα types and frequencies in Bongo, Ghana.
(PDF)

S2 Text. Description of the *cUps* algorithm for classifying malaria *var* genes into ups groups.
(PDF)

S3 Text. Study of DBLα types and frequencies in Africa.
(PDF)

Acknowledgments

This publication used *var* gene sequences 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.

Author Contributions

Conceptualization: Mun Hua Tan, Karen P. Day.

Data curation: Mun Hua Tan, Kathryn E. Tiedje.

Formal analysis: Mun Hua Tan, Kathryn E. Tiedje, Qian Feng, Qi Zhan.

Funding acquisition: Mercedes Pascual, Karen P. Day.

Investigation: Mun Hua Tan.

Methodology: Qian Feng, Heejung Shim, Yao-ban Chan.

Software: Qian Feng, Heejung Shim, Yao-ban Chan.

Supervision: Heejung Shim, Yao-ban Chan, Karen P. Day.

Visualization: Mun Hua Tan.

Writing – original draft: Mun Hua Tan, Qian Feng, Yao-ban Chan, Karen P. Day.

Writing – review & editing: Kathryn E. Tiedje, Qi Zhan, Mercedes Pascual, Heejung Shim.

References

1. 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.
2. 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. https://doi.org/10.1038/s41564-018-0162-2 PMID: 29784978
3. Bull PC, Lowe BS, Kortok M, Molyneux CS, Newbold CI, Marsh K. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. Nat Med. 1998; 4(3):358–60. https://doi.org/10.1038/nm0398-358 PMID: 9500614
4. 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.
5. 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).
6. Chen Q, Schlichtherle M, Wahlgren M. Molecular Aspects of Severe Malaria. Clin Microbiol Rev. 2000; 13(3):439–50.
7. Baruch DI. Adhesive receptors on malaria-parasitized red cells. Best Pract Res Clin Haematol. 1999; 12(4):747–61. https://doi.org/10.1053/beha.1999.0051 PMID: 10895262
8. 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.
9. Tonkin-Hill GQ, Trianty L, Noviyanti R, Nguyen HHT, Sebayang BF, Lampah DA, et al. The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen–encoding var genes. PLoS Biol. 2018; 16(3):e2004328.
10. Bernabeu M, Danziger SA, Avril M, Vaz M, Babar PH, Brazier AJ, et al. Severe adult malaria is associated with specific PfEMP1 adhesion types and high parasite biomass. Proceedings of the National Academy of Sciences. 2016; 113(23):E3270. https://doi.org/10.1073/pnas.1524294113 PMID: 27185931
11. Lennartz F, Adams Y, Bengtsson A, Olsen RW, Turner L, Ndam NT, et al. Structure-Guided Identification of a Family of Dual Receptor-Binding PfEMP1 that Is Associated with Cerebral Malaria. Cell Host Microbe. 2017; 21(3):403–14. https://doi.org/10.1016/j.chom.2017.02.009 PMID: 28279348
12. Bengtsson A, Joergensen L, Rask TS, Olsen RW, Andersen MA, Turner L, et al. A Novel Domain Cassette Identifies *Plasmodium falciparum* PfEMP1 Proteins Binding ICAM-1 and Is a Target of Cross-Reactive, Adhesion-Inhibitory Antibodies. The Journal of Immunology. 2013; 190(1):240.
13. Magallón-Tejada A, Machevo S, Cisteró P, Lavstsen T, Aide P, Rubio M, et al. Cytoadhesion to gC1qR through *Plasmodium falciparum* Erythrocyte Membrane Protein 1 in Severe Malaria. PLoS Pathog. 2016; 12(11):e1006011. https://doi.org/10.1371/journal.ppat.1006011 PMID: 27835682

14. Tessema SK, Nakajima R, Jasinskas A, Monk SL, Lekieffre L, Lin E, et al. Protective Immunity against Severe Malaria in Children Is Associated with a Limited Repertoire of Antibodies to Conserved PfEMP1 Variants. Cell Host Microbe. 2019; 26(5):579–590.e5. https://doi.org/10.1016/j.chom.2019.10.012 PMID: 31726028

15. 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. https://doi.org/10.1038/nature12216 PMID: 23739325

16. Lavstsen T, Turner L, Saguti F, Magistrado P, Rask TS, Jespersen JS, et al. *Plasmodium falciparum* erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children. Proceedings of the National Academy of Sciences. 2012 Jun 26; 109(26):E1791–800.

17. Kraemer SM, Smith JD. A family affair: *var* genes, PfEMP1 binding, and malaria disease. Curr Opin Microbiol. 2006; 9(4):374–80.

18. 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.

19. 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.

20. 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.

21. 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).

22. Tiedje KE, Zhan Q, Ruybal-Pésantez 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. Elife [Internet]. 2023; Available from: http://dx.doi.org/10.1101/2023.05.18.23290210.

23. 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. Molecular & Cellular Proteomics. 2011; 10(11).

24. Walker IS, Rogerson SJ. Pathogenicity and virulence of malaria: Sticky problems and tricky solutions. Virulence. 2023 Dec 31; 14(1):2150456. https://doi.org/10.1080/21505594.2022.2150456 PMID: 36419237

25. 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.

26. Tan MH, Shim H, Chan Y ban, Day KP. Unravelling var complexity: Relationship between DBLα types and var genes in Plasmodium falciparum. Frontiers in parasitology. 2023; 1:1006341.

27. Rask TS, Petersen B, Chen DS, Day KP, Pedersen AG. Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data. BMC Bioinformatics. 2016; 17(1):176. https://doi.org/10.1186/s12859-016-1032-7 PMID: 27102804

28. 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 [Internet]. 2021; 17(2):e1009269–e1009269. Available from: https://doi.org/10.1371/journal.pgen.1009269.

29. 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.

30. 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. Proceedings of the National Academy of Sciences. 2017; 114(20):E4103–11.

31. 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.

32. 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.

33. Jensen ATR, Magistrado P, Sharp S, Joergensen L, Lavstsen T, Chiucchiuini A, et al. Plasmodium falciparum Associated with Severe Childhood Malaria Preferentially Expresses PfEMP1 Encoded by Group A var Genes. Journal of Experimental Medicine [Internet]. 2004 May 3; 199(9):1179–90. Available from: https://doi.org/10.1084/jem.20040274 PMID: 15123742

34. Rottmann M, Lavstsen T, Mugasa JP, Kaestli M, Jensen ATR, Müller D, et al. Differential expression of var gene groups is associated with morbidity caused by *Plasmodium falciparum* infection in Tanzanian children. Infect Immun. 2006; 74(7):3904–11.

35. Bertin GI, Lavstsen T, Guillonneau F, Doritchamou J, Wang CW, Jespersen JS, et al. Expression of the Domain Cassette 8 Plasmodium falciparum Erythrocyte Membrane Protein 1 Is Associated with Cerebral Malaria in Benin. PLoS One [Internet]. 2013 Jul 29; 8(7):e68368–. Available from: https://doi.org/10.1371/journal.pone.0068368 PMID: 23922654

36. Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, Thera MA, et al. Differential *var* gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia. Mol Biochem Parasitol. 2006; 150(2):211–8.

37. 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.

38. Pilosof S, He Q, Tiedje KE, Ruybal-Pesántez S, Day KP, Pascual M. Competition for hosts modulates vast antigenic diversity to generate persistent strain structure in *Plasmodium falciparum*. PLoS Biol. 2019 Jun 24; 17(6):e3000336.

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. The American Society of Tropical Medicine and Hygiene. 2017; 97(1):199–212.

40. 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 Global Public Health. 2022 May 18; 2(5):e0000285.

41. Ruybal-Pesántez S, Tiedje KE, Pilosof S, Tonkin-Hill G, He Q, Rask TS, et al. Age-specific patterns of DBLα *var* diversity can explain why residents of high malaria transmission areas remain susceptible to *Plasmodium falciparum* blood stage infection throughout life. Int J Parasitol [Internet]. 2022; 52(11):721–31. Available from: https://www.sciencedirect.com/science/article/pii/S0020751922000030.

42. 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 [Internet]. 2017; 7(1):11810. Available from: https://doi.org/10.1038/s41598-017-11814-9.

43. Chen DS, Barry AE, Leliwa-Sytek A, Smith TA, Peterson I, Brown SM, et al. A Molecular Epidemiological Study of var Gene Diversity to Characterize the Reservoir of *Plasmodium falciparum* in Humans in Africa. PLoS One. 2011 Feb 9; 6(2):e16629.

44. MalariaGen. The Pf3K Project (2015): pilot data release 2 [Internet]. 2015. Available from: http://www.malariagen.net/data/pf3k-5.

45. Sharp PM, Plenderleith LJ, Hahn BH. Ape Origins of Human Malaria. Annu Rev Microbiol. 2020 Sep 8; 74(1):39–63. https://doi.org/10.1146/annurev-micro-020518-115628 PMID: 32905751

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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 (1979). 1995; 269(5231):1709–11. https://doi.org/10.1126/science.7569897 PMID: 7569897

52. Cockburn IA, Mackinnon MJ, O'Donnell A, Allen SJ, Moulds JM, Baisor M, et al. A human complement receptor 1 polymorphism that reduces *Plasmodium falciparum* rosetting confers protection against severe malaria. Proceedings of the National Academy of Sciences. 2004 Jan 6; 101(1):272–7.

53. Aitman TJ, Cooper LD, Norsworthy PJ, Wahid FN, Gray JK, Curtis BR, et al. Malaria susceptibility and CD36 mutation. Nature. 2000; 405(6790):1015–6. https://doi.org/10.1038/35016636 PMID: 10890433

54. Casals-Pascual C, Allen S, Allen A, Kai O, Lowe B, Pain A, et al. Short report: codon 125 polymorphism of CD31 and susceptibility to malaria. The American journal of tropical medicine and hygiene Am J Trop Med Hyg Am J Trop Med Hyg. 2001; 65(6):736–7.

55. Hill AVS, Yates SNR, Allsopp CEM, Gupta S, Gilbert SC, Lalvani A, et al. Human Leukocyte Antigens and Natural Selection by Malaria. Philosophical Transactions: Biological Sciences. 1994 Jul 12; 346 (1317):379–85. https://doi.org/10.1098/rstb.1994.0155 PMID: 7708832

56. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. The American Journal of Human Genetics. 2005; 77(2):171–92. https://doi.org/10.1086/432519 PMID: 16001361

57. 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. https://doi.org/10.1038/352595a0 PMID: 1865923

58. Lima-Junior J da C, Pratt-Riccio LR. Major Histocompatibility Complex and Malaria: Focus on Plasmodium vivax Infection. Vol. 7, Frontiers in Immunology. 2016. https://doi.org/10.3389/fimmu.2016.00013 PMID: 26858717

59. 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;

60. 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. https://doi.org/10.1038/360434a0 PMID: 1280333

61. 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. https://doi.org/10.1111/tan.13180 PMID: 29160618

62. 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. https://doi.org/10.1016/s0198-8859(01)00293-2 PMID: 11543896

63. Zhan Q, He Q, Tiedje KE, Day KP, Pascual M. Hyper-diverse antigenic variation and resilience to transmission-reducing intervention in falciparum malaria. Nat Commun [Internet]. 2024; 15(1):7343. Available from: https://doi.org/10.1038/s41467-024-51468-6 PMID: 39187488

64. Bachmann A, Bruske E, Krumkamp R, Turner L, Wichers JS, Petter M, et al. Controlled human malaria infection with *Plasmodium falciparum* demonstrates impact of naturally acquired immunity on virulence gene expression. PLoS Pathog. 2019 Jul 11; 15(7):e1007906.

65. 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.

66. 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.

67. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

68. 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.

69. Feng Q, Tiedje KE, Ruybal-Pesántez S, Tonkin-Hill G, Duffy MF, Day KP, et al. An accurate method for identifying recent recombinants from unaligned sequences. Bioinformatics [Internet]. 2022 Apr 1; 38 (7):1823–9. Available from: https://doi.org/10.1093/bioinformatics/btac012 PMID: 35025988

70. 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.

71. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998 Jan 1; 14(9):755–63. https://doi.org/10.1093/bioinformatics/14.9.755 PMID: 9918945

72. Feng Q. Analysing malaria DBLα sequences with Hidden Markov models. University of Melbourne; 2024.

73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10(1):421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

74. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016; 4:e2584. https://doi.org/10.7717/peerj.2584 PMID: 27781170