Touro Scholar

NYMC Faculty Publications Faculty

5-1-2017

Evidence of Strain Structure in Plasmodium falciparum var Gene Repertoires in Children from Gabon, West Africa

K Day

Y Artzy-Randrup

K Tiedje

V Rougeron

Donald S. Chen
New York Medical College

See next page for additional authors

Follow this and additional works at: https://touroscholar.touro.edu/nymc_fac_pubs

Part of the Genetic Structures Commons, and the Parasitic Diseases Commons

Recommended Citation

Day, K., Artzy-Randrup, Y., Tiedje, K., Rougeron, V., Chen, D. S., Rask, T., Rorick, M., Migot-Nabias, F., Deloron, P., Luty, A., & Pascual, M. (2017). Evidence of Strain Structure in Plasmodium falciparum var Gene Repertoires in Children from Gabon, West Africa. *Proceedings of the National Academy of Sciences of the United States of America, 114* (20), E4103-E4111. https://doi.org/10.1073/pnas.1613018114

This Article is brought to you for free and open access by the Faculty at Touro Scholar. It has been accepted for inclusion in NYMC Faculty Publications by an authorized administrator of Touro Scholar. For more information, please contact touro.scholar@touro.edu.

Authors

K Day, Y Artzy-Randrup, K Tiedje, V Rougeron, Donald S. Chen, T Rask, M Rorick, F Migot-Nabias, P Deloron, A Luty, and M Pascual

This article is available at Touro Scholar: https://touroscholar.touro.edu/nymc_fac_pubs/523

PNAS PLUS

Evidence of strain structure in *Plasmodium falciparum* *var* gene repertoires in children from Gabon, West Africa

Karen P. Day[a,b,1], Yael Artzy-Randrup[c,d], Kathryn E. Tiedje[a,b], Virginie Rougeron[b,e], Donald S. Chen[b,f], Thomas S. Rask[a,b], Mary M. Rorick[d,g], Florence Migot-Nabias[h,i], Philippe Deloron[h,i], Adrian J. F. Luty[h,i], and Mercedes Pascual[g,j]

[a]School of Biosciences, The University of Melbourne, Parkville, VIC 3052, Australia; [b]Department of Microbiology, New York University, New York, NY 10016; [c]Theoretical Ecology Group, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, 1090 GE Amsterdam, The Netherlands; [d]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109; [e]Laboratoire Maladies Infectieuses et Vecteurs: Ecologie, Génétique, Evolution et Contrôle, UMR 224-5290 CNRS, Institut de Recherche pour le Développement–Université de Montpellier, Centre Institut de Recherche pour le Développement de Montpellier, 34394 Montpellier, France; [f]Department of Medicine, New York Medical College, Valhalla, NY 10595; [g]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [h]Institut de Recherche pour le Développement, UMR 216 Mère et Enfant Face aux Infections Tropicales, 75006 Paris, France; [i]Communautés d'Universités et Établissements, Sorbonne Paris Cité, Université Paris Descartes, Faculté des Sciences Pharmaceutiques et Biologiques, 75006 Paris, France; and [j]Santa Fe Institute, Santa Fe, NM 87501

Edited by Burton H. Singer, University of Florida, Gainesville, FL, and approved April 4, 2017 (received for review August 6, 2016)

Existing theory on competition for hosts between pathogen strains has proposed that immune selection can lead to the maintenance of strain structure consisting of discrete, weakly overlapping antigenic repertoires. This prediction of strain theory has conceptual overlap with fundamental ideas in ecology on niche partitioning and limiting similarity between coexisting species in an ecosystem, which oppose the hypothesis of neutral coexistence. For *Plasmodium falciparum*, strain theory has been specifically proposed in relation to the major surface antigen of the blood stage, known as *Pf*EMP1 and encoded by the multicopy multigene family known as the *var* genes. Deep sampling of the DBLα domain of *var* genes in the local population of Bakoumba, West Africa, was completed to define whether patterns of repertoire overlap support a role of immune selection under the opposing force of high outcrossing, a characteristic of areas of intense malaria transmission. Using a 454 high-throughput sequencing protocol, we report extremely high diversity of the DBLα domain and a large parasite population with DBLα repertoires structured into nonrandom patterns of overlap. Such population structure, significant for the high diversity of *var* genes that compose it at a local level, supports the existence of "strains" characterized by distinct *var* gene repertoires. Nonneutral, frequency-dependent competition would be at play and could underlie these patterns. With a computational experiment that simulates an intervention similar to mass drug administration, we argue that the observed repertoire structure matters for the antigenic *var* diversity of the parasite population remaining after intervention.

Plasmodium falciparum | *var* genes | parasite diversity | strain structure | Gabon

Malariologists have understood since Koch's observation of 1905 that individuals living in malaria endemic areas develop a nonsterilizing immunity that protects against clinical disease in children after several years of repeated exposure (1–3). This slow acquisition of immunity has been proposed to be due to the existence of many antigenically diverse parasites or "strains" (2).

Genetic evidence to explore this "diversity hypothesis" has been collected over the past 30 y. Most recently, genome sequencing of *Plasmodium falciparum* has shown that there are many diverse single-copy antigen-encoding genes as well multigene families encoding variant antigens (4). The highest numbers of SNPs occur in antigen loci, particularly the major variant surface antigen-encoding genes called *var* (4). Each parasite genome has up to 60 *var* genes encoding variants of the major blood stage antigen known as *P. falciparum* erythrocyte membrane protein 1 or *Pf*EMP1. Analysis of *var* gene diversity in seven sequenced genomes has shown that different genomes have distinct repertoires of *var* genes (5). From molecular epidemiology studies, we now understand that the transmission system is composed of antigenically distinct parasite genomes defined by diverse repertoires of *var* genes (6) with varying levels of overlap in repertoires seen in endemic areas of Africa, South America, and Papua New Guinea (7, 8).

Given that extensive parasite variation has been described by genome sequencing, the question of whether there is a "strain structure" in an organism like *P. falciparum* remains to be answered. This pathogen undergoes conventional meiosis each time it passes through the mosquito and frequent outcrossing occurs in nature where carriage of diverse genomes is the norm (9, 10). The classic microbiological paradigm for strain structure, even in organisms that recombine such as influenza A, HIV, and *Neisseria meningitides*, is based on the population genetics of the major surface antigens of a pathogen to which the dominant immune response occurs. The obvious strain-structuring, immune-dominant antigen in *P. falciparum* is *Pf*EMP1 encoded by the *var* multigene family. *Pf*EMP1 is also a virulence factor where expression of certain variants of *Pf*EMP1 can lead to a variety of clinical outcomes in susceptible hosts (11–15). Even though *var* genes that encode *Pf*EMP1 are prime candidates for disease surveillance to detect immune selection, they have not traditionally

Significance

This paper aims to discover how diverse malaria parasites are in children from an African village. DNA sequencing shows that they are highly diverse with respect to the genes encoding the surface coat. Indeed, every child has a malaria infection with a different set of these genes. Importantly, this paper shows by computational methods that the pattern of this diversity is not random but structured to enhance the parasites' chance to evade host immunity and has implications for the success of malaria control programs.

Author contributions: K.P.D., Y.A.-R., T.S.R., and M.P. designed research; Y.A.-R., K.E.T., V.R., D.S.C., T.S.R., F.M.-N., P.D., A.J.F.L., and M.P. performed research; K.P.D., Y.A.-R., K.E.T., D.S.C., T.S.R., M.M.R., and M.P. analyzed data; and K.P.D., Y.A.-R., K.E.T., D.S.C., T.S.R., and M.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The nucleotide sequences reported in this paper have been deposited in the GenBank database (accession nos. KY328840–KY341897).

[1]To whom correspondence should be addressed. Email: karen.day@unimelb.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1613018114/-/DCSupplemental.

POPULATION BIOLOGY

been used as genetic markers for malaria surveillance. This is due to the complexity of this multicopy gene family that can diversify by mitotic and meiotic recombination. In contrast, microsatellites (16) and SNPs (17, 18) have been widely used in malaria surveillance as putatively neutral markers to track parasite population structure but obviously cannot specifically track immune selection.

The "diversity hypothesis," although substantiated, has had limited impact in the development of epidemiological theory of malaria. It was formulated for the first time in a population dynamic theoretical framework in the early 1990s by Gupta and Day (19–21). They proposed that immunity to variants of the major antigen of the blood stages of *P. falciparum* (*Pf*EMP1) was a key driver in the transmission dynamics of *P. falciparum* due to the role of these variants in prolonging the duration of infection by the mechanism of clonal antigenic variation. Consequently, they demonstrated that the basic reproduction number (R_0) used to measure the transmission potential of a disease in a naive population, would be considerably lower if anti-*Pf*EMP1 immunity structured the system into independently transmitted strains or antigenic types (19). They showed serologic data demonstrating nonoverlapping specificities to *Pf*EMP1 variants in five randomly chosen isolates of *P. falciparum* to support their theory. This serologic experiment assumed that strains had different sets of variants in the absence of genetic proof, as *var* genes had not been discovered at that time. Buckee et al. (22) used similar serological data to infer population structure of the *P. falciparum* parasite population in Kilifi, Kenya. They proposed a parasite population where conserved variant-encoding genes would be partitioned and rare variant-encoding genes would have less structure.

"Strain theory" proved to be controversial (23). Interestingly, Saul (24) argued that, if this theory were to be true, then a very large parasite population size would be needed in a relatively small human population for malaria to be endemic at the level of village communities. Furthermore, the theory was rejected by some geneticists who argued that distinct antigenic strains could not exist for organisms such as *P. falciparum* that were shown to recombine frequently in nature. However, Gupta et al. (21) showed that this is not necessarily the case. Mathematical models of competition between pathogen strains have demonstrated that immune selection can lead to maintenance of strain structure of discrete, nonoverlapping antigenic repertoires (21, 25), although these studies address a finite number of variants at each locus in a closed system and much less diversity than that of the *var* genes in natural populations of *P. falciparum*. Artzy-Randrup et al. (26) explored the special case of population structuring of the *var* multigene family. They used a computational model that simulates the dynamics of unique combinations of *var* genes in a population of hosts, which shows that, even with high recombination rates, the system can self-organize into a limited number of coexisting strains: the distinct *var* gene repertoires of these strains only weakly overlap, suggesting that the immune response of the host population has been partitioned into distinct niches (26). To date, there has been no deep sampling of the population genomics of *var* genes in local populations with sufficient coverage to define whether patterns of repertoire overlap support the existence of such "strains" under conditions of high outcrossing where meiotic recombination will constantly reassort *var* repertoires.

This study reports deep sampling of *var* genes from the parasite population of the West African village of Bakoumba, Gabon, where high levels of outcrossing can be expected from previously reported high transmission in the wet season, high carriage of multiple infections (27), and linkage equilibrium among microsatellite markers (26). Specifically, we sampled the Duffy binding-like (DBL) α adhesion domain of *var* genes in the reservoir of asymptomatic *P. falciparum* infections in the majority of microscopy slide-positive children from Bakoumba using a 454 high-throughput sequencing protocol. The DBLα domain is present in all *var* genes except the atypical, placental adhesion VAR2CSA. Typical of high transmission areas in West Africa, asymptomatic infections represent the majority of the *P. falciparum* parasite population across all age groups (28–30). While carrying parasites most of the time, children experience only one to five clinical episodes per year with 1–2% of these infections leading to severe clinical disease (31). Furthermore, asymptomatic infections contribute significantly to the *P. falciparum* reservoir that fuels transmission and therefore need to be prioritized for elimination strategies. The size and persistence of this reservoir in Africa present a major drawback for malaria elimination. Hence, the target of our study, with the rationale being to better understand the genetics of the *P. falciparum* reservoir, is to ultimately inform the theory of malaria control. We report a pattern of low overlap in the *var* DBLα repertoires consistent with a population-structuring role of immune selection. Frequency-dependent competition for hosts, mediated by the immune system, could underlie this pattern of limiting similarity between parasites. We discuss potential implications for both the theory of malaria control and molecular surveillance as well as open questions for epidemiology.

Results

Summary of Sequencing Results. Among the four pools for sequencing, a total of 372,000 sequence reads were obtained. A total of 341,891 of the sequence reads was from Bakoumba isolates, and the remainder was from the control laboratory clones. The mean read length was 400 bp. Following application of quality control measures, there were 200 isolates with DBLα sequence reads available for further analyses, with a median of 663 sequence reads per isolate. Results on the sequence reads can be found in Table S1. A comparison of the isolates sequenced in a prior study and then resequenced in this study can be found in *Supporting Information*.

Assembly of Reads into DBLα Sequences. Within each isolate, quality sequence reads were clustered into nonredundant DBLα sequences using flowgram clustering (*Materials and Methods*). This process resulted in a total of 13,058 DBLα sequences among the 200 Bakoumba isolates, representing the dataset on which analyses were performed. Methods and parameters for quality control and clustering were validated on control laboratory reference genomes (Table S2).

Definition of DBLα Types, Frequency Distribution, and Richness Estimates. To subsequently determine DBLα types shared between isolates, we clustered nonredundant DBLα sequences from all samples by average linkage using a sequence identity threshold of 96%. Clustering by average linkage resulted in 6,404 unique DBLα types among the 200 isolates (Table S1). The majority of DBLα types were rare and seen only once among the Bakoumba isolates (3,917, 61.2%), although there were a small number of more common DBLα types found in ≥30 of the 200 isolates (13, 0.2%) (Fig. 1*A*). The minimum and maximum number of times a DBLα type were seen was 1 and 76, respectively (Fig. 1*A*). For 10 most common DBLα types in this population, they could be grouped into those matching GenBank *var* gene sequences from global isolates, and those matching GenBank *var* gene sequences only from Gabon (Table S3). Some of the matches represent DBLα types from identical isolates previously cloned and sequenced by Sanger methodology. Application of the *Chao1* richness estimator to the observed frequency distribution resulted in a prediction of a minimum of 12,536 DBLα types in the population. The *Chao1* richness estimate at this level of sampling displayed a slight sensitivity to sample size (Fig. S1). The type accumulation curve displayed high levels of richness; although there is a bend in the curve, we did not reach saturation in sampling of DBLα types in the population (Fig. 1*B*). In sensitivity analyses, these results remain

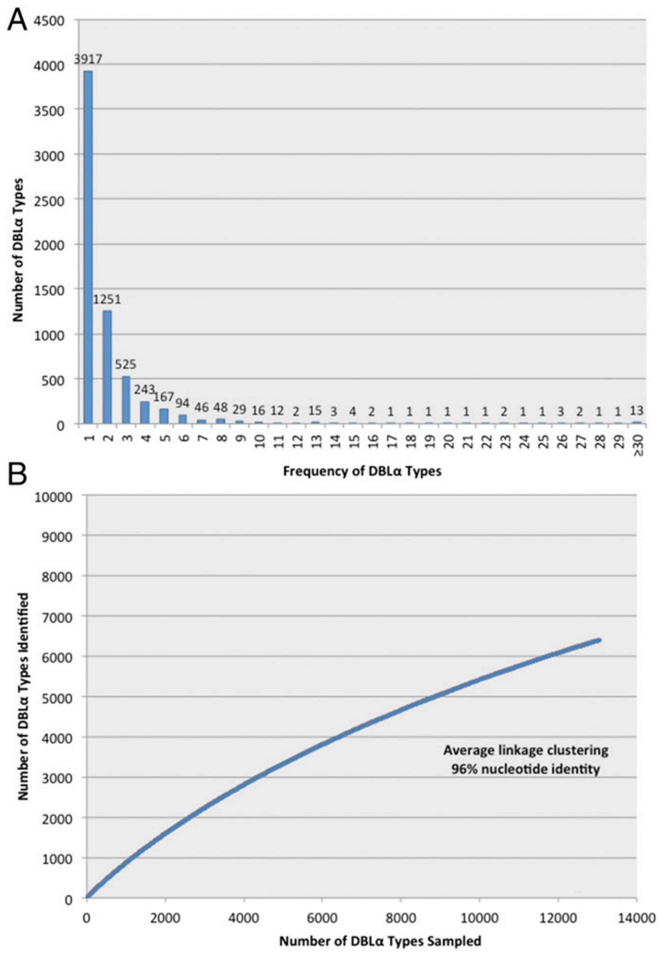


Fig. 1. Frequency distribution and diversity of DBLα types in the population. (*A*) For the Bakoumba isolates, clustering of the 13,058 DBLα sequences in the dataset at 96% nucleotide identity resulted in 6,404 DBLα types. As depicted in the frequency distribution of DBLα types for Bakoumba, 61.2% DBLα types (3,917) were encountered in only 1 of the 200 isolates, whereas 0.2% DBLα types (13) were found in ≥30 of the 200 isolates with the maximum DBLα type frequency being 76. Identical or near-identical matches to the 10 most "common" DBLα types were found in the public sequence databases (Table S3). (*B*) The Bakoumba type accumulation curve for the sampling reveals a large number of DBLα types in the population. Despite in-depth sampling of the local population, the up-sloping curve demonstrates that more DBLα types are to be found with further sampling of the population.

relatively stable regardless of clustering method (single, average, or complete linkage) and throughout a range of percent nucleotide identity clustering thresholds (0.90–0.98) (Fig. S2).

Diversity Measures. Diversity is most commonly measured by "Richness" (32), defined as the number of distinct types (or species in an ecosystem) observed in a location. Intuitively, this measure is relatively simple to grasp as it gives equal weight to all types, regardless of their relative abundance. Qualitatively, however, this approach is not always meaningful. For example, in the case of ecological interactions, an ecosystem with 10 equally common species is likely to be considered much more "diverse" than one that has a single dominant species and nine others that are vagrant. However, Richness assigns both ecosystems identical scores of diversity. Valuable information can be lost when relative frequencies of the different types are excluded from how diversity is defined. Here, we take a broader and more inclusive approach in our definition of diversity by considering a set of different measures. To do this, we make use of the so-called "Hill numbers" (33, 34), an ordered set of diversity indices that give different degrees of importance to the frequent vs. rare DBLα types. By doing so, the DBLα type diversity in this study is better characterized. In particular, one benefit of this approach is that it provides visual means for viewing the relationship between the different indices in a continuous manner, revealing, for example, the degree of unevenness in a location (Fig. 2).

The Hill numbers, H_q, are ordered by index, $q = \{0, 1, \ldots\}$, where for $q = 0$, H_0 is the classic species richness index, as mentioned above; for $q = 1$, H_1 is the exponential of "Shannon's entropy index," where species are weighed in proportion to their frequency, a measure that can roughly be interpreted as the number of "typical species." Then for $q = 2$, H_2 is the inverse of Simpson's concentration index, where the weight toward the most common species in the assemblage increases, providing roughly the number of "very abundant species." As the order of q grows, Hill numbers give higher weight to the more abundant types, which make them less sensitive to sample size (*Materials and Methods*). Calculation of the Hill numbers in our data reveals an extreme unevenness in the frequencies of DBLα types, with only a few very abundant types and a majority of extremely rare ones (Fig. 2).

Analysis of the DBLα Repertoires. The size of DBLα repertoires as sampled by our PCR-based methods ranged from a minimum of 1 DBLα type to a maximum of 258 DBLα types (Fig. S3), with a median size of 56 DBLα types per isolate (Fig. S3). Using the number of DBLα types per isolate, it can be inferred that 87 (43.5%) of the 200 isolates were polygenomic as they had >60 DBLα types identified. Overall, the overlap in DBLα repertoires was minimal, as evidenced by our calculation of pairwise type sharing (PTS), a similarity index (6). PTS comparisons among the 200 isolates resulted in 19,900 comparisons. In 7,495 (37.7%) of the pairwise comparisons, there were no shared DBLα types. The median and mean PTS score were 0.017 and 0.025, respectively, with a maximum PTS score of 0.984 (Fig. 3 *A* and *B*). There were, however, smaller clusters of isolates with higher DBLα repertoire overlap as indicated by darker shading on the PTS heat map (Fig. 3*B*). Because of the high diversity of DBLα types and their skewed frequency distribution, it is not possible,

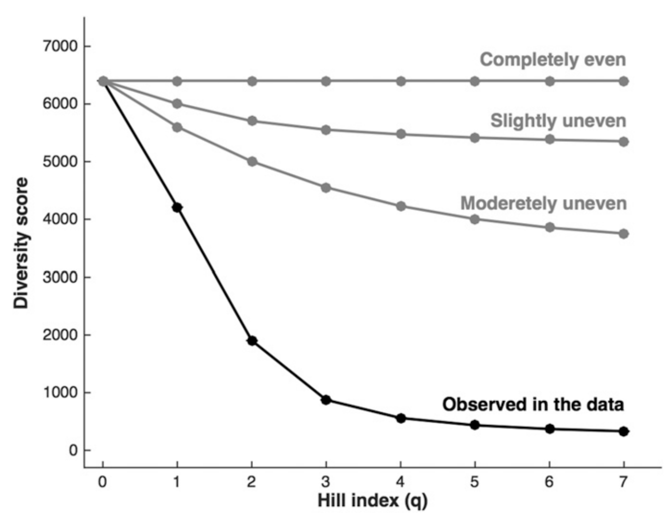


Fig. 2. Diversity profiles. Hill diversity numbers of the observed data show high levels of unevenness in the frequency of DBLα types (black). The *x* axis is the index of the Hill number (see text), and the *y* axis is the associated level of diversity for each of the measures. Examples of more common forms of diversity profiles are illustrated in gray. The calculation of the Hill numbers follows the procedure of Chao et al. (34) (*Materials and Methods*).

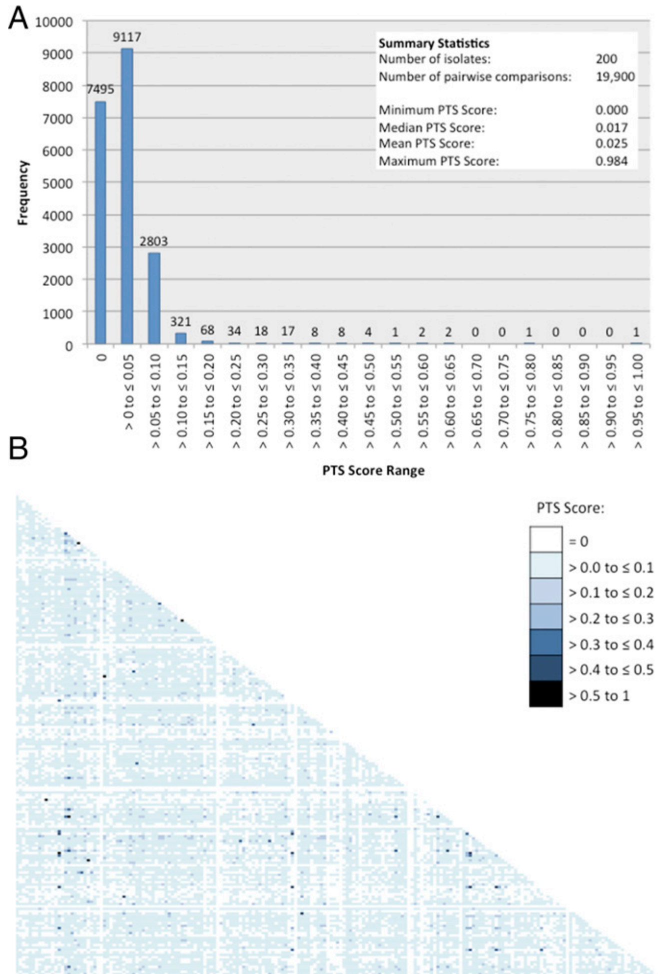


Fig. 3. Pairwise comparisons of repertoire overlap. The pairwise type-sharing (PTS) scores (representing the proportion of DBLα types shared between two isolates) was calculated for all possible pairwise comparisons in the studies conducted in Bakoumba. A score of 0 represents no shared DBLα types between two isolates; a score of 1 represents complete identity of all DBLα types between two isolates. For Bakoumba, there were 19,900 possible pairwise comparisons of the 200 isolates in the study. (*A*) PTS scores for Bakoumba ranged from 0 to 0.984, with median and mean scores of 0.017 and 0.025, respectively. Overall, there was little to no overlap between most pairs of isolates, with 83.5% of scores being less than or equal to 0.05. (*B*) Bakoumba PTS scores depicted on a heat map reveal this low overlap among most pairs of isolates (light shading). A minority of isolates, however, appear more closely related to each other based on relatively higher levels of DBLα repertoire overlap (>0.50 PTS, darker shading). These clusters are apparent in the dendrogram in Fig. S4.

however, to evaluate whether these patterns of overlap and especially their low values, are not simply the result of randomly assembled repertoires. To evaluate the significance of these patterns, we compare them next to those expected in randomizations of the data that conserve both the length of the isolates and the frequency of the DBLα types themselves (*Materials and Methods*).

Overlap Indices. Studying the frequency distribution of PTS values between all isolate couples, we found that the median PTS is significantly lower than expected at random ($p_v = 0.002$), whereas the maximum PTS is significantly larger than expected at random ($p_v = 0.017$). Fig. 4*A* shows the distribution of the median scores of a set of 1,000 random samples, and Fig. 4*B* shows the distribution of the maximum scores of this set (*Materials and Methods*). This is consistent with the patterns in the PTS heat map (Fig. 3*B*), where a small number of higher overlap PTS values coexist against a background of mainly weaker ones. The results were found to be robust also when considering alternative pairwise indices, namely the well-known Pianka index (35) commonly used in ecology to test for nonrandom structure in communities of species (Fig. S5), as well as the Czechanowski and Jaccard indices (36, 37) (Table S4). We also tested whether the results would remain after excluding all of the relatively small isolates (i.e., those smaller than 20) and/or all DBLα types that were identified only once. These subsets of the original data also generated similar findings for all four-overlap indices (Table S4).

Associations Between DBLα Types. The level of co-occurrence of the different DBLα types in the observed data reveals that there is significantly higher co-occurrence among low-frequency DBLα types (Fig. 5), as well as significantly higher co-occurrence among the higher-frequency DBLα types than expected at random. Namely, the very abundant types are more likely to co-occur in an isolate than expected by chance, and similarly the very rare types are more likely to co-occur in an isolate than expected by chance. In contrast, co-occurrence of rare types with very highly abundant types in an isolate is less likely than expected by chance (*Materials and Methods*).

Computational Removal Experiment. The existence of a nonrandom structure of the DBLα repertoires would imply significant differences in the *var* diversity remaining after a simulated "intervention" that removes a given percentage of the parasite population. We implemented such a computational experiment and compared the diversity profiles measured by the Hill numbers of the data to those in a set of randomizations (*Materials and*

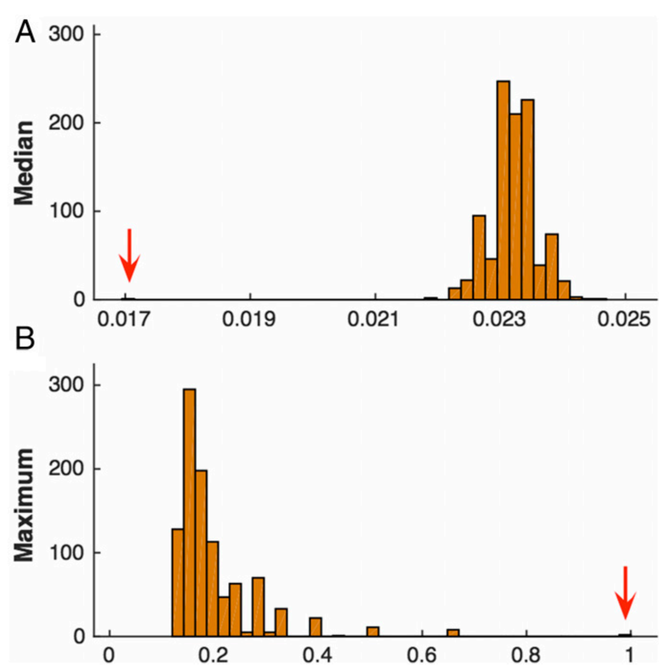
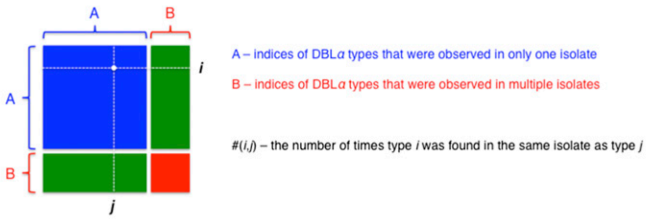


Fig. 4. Distribution of median and maximum PTS isolate similarity scores. Histograms showing the distribution of the median and maximum scores are found in a random ensemble of 1,000 samples. (*A*) The mean of the median scores in the sample is 0.023, with SD (std) = 0.0004, whereas the observed median in the data is 0.0169 (indicated by the red arrow). This is significantly lower than expected by chance ($p_v = 0.002$, and *z* score = −16.70). Similarly, (*B*) the mean of the maximum scores found in the sample is 0.199, with std = 0.088, whereas the observed maximum in the data is 0.983 (indicated by the red arrow). This is significantly higher than expected by chance ($p_v = 0.017$ and *z* score = 9.286). See *Materials and Methods* for information on the randomization process.



	Observed	Mean	Std	Z-score
The number of times two types that were observed only once were found together in the same isolate	67284	55396	914.5	12.99
The number of times a type observed only once was found together with a type that was observed multiple times	237701	254048.5	1115.7	-14.65
The number of times two types that were observed multiple times were found together in the same isolate	295813	291353.6	1763.7	2.53

Fig. 5. An illustration of the co-occurrence matrix of DBLα types in the data. The DBLα types in the matrix are sorted according to their observed frequency (i.e., the number of isolates they were observed in). Sum "A" represents the total number of DBLα couples, each of which was observed exactly once, but both in the same isolate. These couples were observed a significantly higher number of times than expected by chance using our random ensemble of 1,000 matrices (*Materials and Methods*). The *z* score equals 12.99 [the number of SDs the observed data departs from the mean; *z* score = (observed − mean)/std].

Methods). We found a significant difference between the datasets for all removal percentiles (based on four different statistics). Moreover, a multicomparison Tukey's post hoc test between datasets confirms a significant difference of marginal means between our observed data and the reshuffled datasets. Further implications of these experiments for changes in parasite antigenic diversity with intervention are discussed below.

Discussion

By deep sampling of the DBLα region of *var* genes in children in a local population from Gabon, extensive diversity was observed, and yet a clear pattern of *var* population structure has emerged with limited overlap in *var* DBLα repertoires. Given a parasite population with extensive diversity in DBLα types where most types appear once, we had to prove that this pattern did not occur as a result of the frequency distribution dominated by rare DBLα types. Several statistical methods were used to prove that this structure of limited overlap in DBLα repertoires was nonrandom. These results are surprising given the high rates of meiotic recombination observed in areas where individuals carry multiple genomes and transmission is high in the wet season (9). Mating patterns would be expected to be characterized by high levels of cross-mating (9). We found an absence of highly related (PTS > 0.20) parasites in terms of DBLα overlap in this transmission system. This suggests selection against recombinant repertoires as would be expected by immune selection of related strains (26).

A 96% cutoff was used to define unique DBLα types. The question may arise that, if we used a less stringent cutoff for a DBLα type, we might see more overlap, as the number of types would be reduced. This cutoff has proven to be robust in that we can define DBLα types with identical sequences (excluding minor sequence errors) globally and within sites. Another question to be asked of our data is whether the overlap is missing because of the failure to distinguish DBLα repertoires with significant overlap in multiple infections where we have counted DBLα types once when they may occur more than once. Our overlap analyses reach similar conclusions even when isolates with only single infections are considered: the typical DBLα type overlap is less than that expected at random.

The nonrandom pattern of DBLα repertoire overlap provides evidence of strain structure where individuals have isolates of *P. falciparum* composed of one or more genomes with largely nonoverlapping sets of *var* genes. The reduced overlap documented here from deep population sampling and sequencing of DBLα, complements previous studies addressing a role of immune selection based on serology and cross-reactivity assays (19, 22). In particular, Buckee et al. (22) construct serological networks of parasite recognition and compare their properties to those generated in theoretical models encapsulating different hypotheses on random vs. nonrandom structure of *var* repertoires. Interestingly, they find evidence for a role of immune selection in structuring repertoires but also propose that different levels of immune selection occur within different groups of the *var* multigene family, leading to mixed population structures in which only the most conserved types would exhibit a nonrandom organization into discrete strains. No such distinction was found in the patterns described here.

These considerations bring us to limitations of current theory and future directions. Documented recombination hierarchies (38) should be included in the form of *var* gene groups that differ in functional properties and span a gradient from highly conserved to much more variable (39). The immense diversity of DBLα types described here goes well beyond that so far considered by theoretical computational models (26, 39–41). Characterization of emergent population structure at such high levels of complexity should be addressed, together with changes under decreasing endemicity. Moreover, theory has so far considered largely "closed" systems in the sense that either the host population has been exposed to most genes in the pool (26) or that mutation acts within a limited set of possible variants (40). Consideration of more open systems would be of interest, in which immigration of new repertoires and new genes, or mutation and mitotic recombination create new types in the overall pool. Importantly, theory should also help develop better statistical approaches to interrogate population genetic data on the processes underlying the nonrandom structure of strains. Statistical null models such as the randomizations applied here provide a starting point as they control for given quantities (e.g., frequency of DBLα types). To specifically address the role of immune selection, process-based null models should be developed that dynamically take into account epidemiology and neutral evolution of the transmission system. Moreover, the implications of parasite population structure for estimating epidemiological parameters and for the consequences of intervention remain an important open area (42). Inference of key evolutionary and epidemiological parameters from *var* gene data that can inform immunity considerations in stochastic neutral and nonneutral models is of particular significance for high transmission areas.

Finally, we have considered here asymptomatic infections. Future work should develop comparisons to infections that cause uncomplicated and/or severe clinical episodes even though they represent a small fraction of the parasite population at any point in time. Clearly understanding the genetic basis of these infections is also vital to save lives and to address aspects of parasite diversity that are critical to treatment and prevention. Empirical findings show that the parasites that cause severe disease express subsets of group A and B *var* genes; thus, it is not the *var* haplotype of the genome but the pattern of expression of *var* genes in a host that leads to clinical disease (12, 15, 43–53). This pattern of expression is itself a function of the immunological history of the host and hence age of exposure (22). It would follow that parasites responsible for symptomatic infections are likely to be the same parasites that cause asymptomatic malaria with diverse expression patterns influenced by host immunity.

In relation to epidemiology, strain theory originally postulated that the apparent high R_0 of *P. falciparum* is the result of a high number of strains with low R_0 (19, 20), a hypothesis with far-reaching consequences for the impact of interventions. We understand overall R_0 for pathogens with well-defined persistent strains; we understand even better R_0 for systems without population structure in which one can effectively consider the transmission dynamics of a single parasite. It is less clear how

particular patterns of overlap emerging from immune selection translate into R_0 and resilience to intervention, for example, in relation to changes in antigenic diversity. Limited overlap could facilitate persistence of the parasite in small host populations. It should also facilitate the faster acquisition of immunity than in the absence of parasite population structure, with no strains.

Thus, an important motivation for systematically mapping the diversity of DBLα types in these local settings is to address whether underlying genetic composition of *P. falciparum* influences the outcome of disease control. On regional scales, for example, DBLα diversity loss is expected following major interventions such as mass drug administration (MDA) (54, 55). It is unknown, however, whether the intraspecific repertoire structuring of *P. falciparum* has a role in how this diversity is lost. A simple computational experiment on our data makes it possible to gain some insights into this question. Taking the inclusive Hill diversity approach that was used to characterize diversity in our data, we can track the patterns of diversity loss following the random removal of isolates (e.g., clearance of infection). We compared these observations to what would be expected in a null scenario, where patterns of isolate overlap are random but where DBLα type frequency distribution remains identical (*Materials and Methods*, *Removal Experiment*). This ensures that significant differences identified between the data and the null scenario can be attributed to the underlying repertoire structure, irrespective of the observed DBLα type frequency distribution. Hence, before the removal of isolates, we can expect that the Hill diversity profiles across all our null realizations are identical to the observed profile in our data (Fig. 2). Our results reveal two interesting trends following the removal of isolates at different levels. With moderate to intermediate removal (±40–60%), a more rapid loss of DBLα richness is observed in our data than would be expected at random (Fig. 6*A*). This observation is likely to be associated with the discordant nature of the DBLα repertoires in our data. With higher removal of isolates (±60–80%), we observe a significantly slower decrease in the higher Hill indices, possibly implying that the remaining DBLα types were more evenly distributed than expected at random (Fig. 6*B*). This outcome would result from the significantly high co-occurrence of the abundant DBLα types in the same isolates. Thus, as we remove more and more isolates, the outcome of the removal experiment shifts from a lower to a higher diversity than expected at random, as the influence of the *var* repertoire structure transitions from emphasizing the many rare DBLα types that do not overlap to the fewer common ones that do overlap.

These computational experiments simulating MDA show that *var* DBLα structure and diversity could alter in response to malaria control programs in complex ways. Results provide additional evidence on the existence of higher-order repertoire structuring of *P. falciparum*, which cannot be simply explained by the highly skewed frequency distribution of DBLα types. They also underscore that the natural repertoire structuring of *P. falciparum* can have direct effects on outcomes of intervention.

Despite the obvious importance of the *var* genes in driving population dynamics, there is so much complexity seen in even the limited genome and field studies that the *var* system was previously thought to be intractable for disease surveillance. However, deep sampling of these genes is beginning to reveal structure that is very different to that seen by microsatellite markers (26). Importantly, this structure of limited overlap was shown here to have potential epidemiologic significance in terms of malaria control such as MDA. For the purpose of disease surveillance, loss of *var* diversity as well as changes in type sharing could be measured as indicators of changes in parasite fitness in response to control, where reduced diversity and higher type sharing would show reduced fitness to evade host immunity and may impact duration of infection. Further research on the effect of interventions on the *var* system is required to establish these genes as surveillance tools.

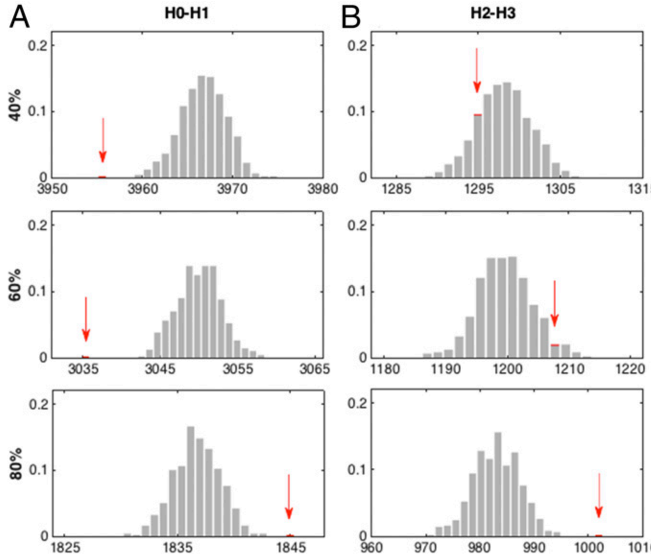


Fig. 6. Distribution of Hill measures for different removal percentiles. Histogram distributions of marginal means of the random ensemble in gray for Hill measures: (*A*) H_0 and H_1, and (*B*) H_2 and H_3 (*Materials and Methods*), for removal of 40%, 60%, and 80% of the isolates. The corresponding values observed in the data are marked in red (indicated by the red arrows).

If "strain theory" based on *var* gene diversity is correct, it would necessitate significant revision of the theory of malaria control as current theory is largely based on mathematical models of transmission dynamics that consider a genetically homogeneous parasite population and do not explicitly incorporate its population structure and diversity as related to major variant antigens. More generally, parasite populations can provide powerful systems to test fundamental ideas at the interface of ecology and evolution pertaining to the structure of diversity, the forces that shape it, and its implications for persistence. There are unexploited but promising connections between questions in strain theory, and those in community ecology and macroevolution.

Materials and Methods

Study Design and Data Collection. The study was performed in Bakoumba in southeast Gabon near the border with the Republic of the Congo. In this region, malaria is highly endemic with peaks in transmission at the end of the rainy seasons (September to December and March to June) (56). A cross-sectional survey was conducted in May to June 2000 in a cohort of 641 asymptomatic children between the ages of 1 and 12 y. Further details on the study population and data collection procedures have been published elsewhere (57). After obtaining informed consent from all parents, venous blood samples were collected for parasitological assessment for *Plasmodium* spp. by blood smears and dried blood spots for genotyping (58). For the present study, 264 children from the Bakoumba cohort were found to be smear positive for *P. falciparum*. We successfully sampled DBLα types from 211 (79.9%) of these isolates, and following the application of quality control measures the DBLα types from 200 (94.8%) isolates were analyzed. The study was reviewed and approved by the ethics committees at the International Center for Medical Research of Franceville, Gabon; New York University School of Medicine, United States; and the University of Melbourne, Australia.

DNA Extraction and Genotyping. Genomic DNA for each isolate was extracted from the dried blood spots on filter paper using the QIAamp DNA Mini Kit (Qiagen) according to the procedure as described by the manufacturer.

PCR Amplification for *var* DBLα Typing. The *P. falciparum var* DBLα domain from genomic DNA was amplified using fusion primers for multiplexed 454 amplicon sequencing to the DBLα domain as previously described (6, 7). The DBLα domain has been used previously as a marker of *var* gene diversity

in other investigations (6–8). From each isolate of genomic DNA, a ~550- to 700-bp region of the DBLα domain was amplified using a degenerate primer set (forward, 5′-CMTGYGCDCCRTWYMGAMG; reverse, 5′-TCKGCCCATTCYTCRAACCA) designed against the semiconserved blocks B and H of DBLα (14, 59). Each of the DBLα primers were barcoded with a 10-nt sequence Multiplex Identifier (MID) tag published by Roche (Roche 454 Sequencing Technical Bulletin No. 013-2009; 454 Sequencing Technical Bulletin No. 005-2009), which was used to code and distinguish the *var* genes amplified from each unique isolate once all isolates were pooled and sequenced (60). Intervening primer sequence necessary for the 454 titanium platform was included in these fusion primers. These primers were validated for amplification of sequences of the appropriate length using *P. falciparum* 3D7 genomic DNA.

The PCR conditions for the DBLα primers were as follows: 2 μL of isolate genomic DNA, 0.2 mM dNTPs, 1 μM of each primer, 1× reaction buffer, and 1.25 units of HotMaster Taq polymerase in 50 μL of total reaction volume. PCR cycling was carried out on an Eppendorf EP Gradient Mastercycler and involved an initial denaturing step of 94 °C for 2 min, 35 cycles of 94 °C × 5 s, 50 °C × 20 s, and 60 °C × 45 s, followed by a final extension step of 60 °C for 2 min. PCR amplification was confirmed visually by gel electrophoresis (1.5% agarose in 0.5× TE buffer) with nucleic acid staining demonstrating a band of the appropriate size (~550–700 bp). Positive controls (laboratory genomic *P. falciparum* DNA) and negative controls (no template) were performed for quality assurance.

Finally, the isolate amplicons were pooled and sequenced at SeqWright DNA Technology Services through next-generation 454 sequencing (Roche) using titanium chemistry. The 454 sequencing provides average read lengths of 400 bp, therefore lending itself to the assembly of the individual *var* DBLα amplicons of 550- to 700-bp lengths using the forward and reverse sequence reads from each direction. Individual *P. falciparum* field isolates were distinguished by the unique MID barcodes added to the fusion primer.

***var* DBLα Sequence Analysis.** A custom pipeline was developed to demultiplex, de-noise, and remove PCR and sequencing artifacts from the DBLα domain reads. The first part of the pipeline is available as the Multipass web server: www.cbs.dtu.dk/services/MultiPass-1.0, and the following cleaning steps described below are implemented in a Python script available here: https://github.com/454data/postprocess. The sff files obtained from each region on the 454 plate were divided into smaller isolate-specific sff files by identification of reads with exact matching MID sequences in both ends using BioPython, version 1.57. Ambiguous primer sites were then identified (exact match) and trimmed off the flowgrams, reverse reads were reverse complemented, and a dat file (AmpliconNoise format) with the resulting flowgrams was created for each isolate, using BioPython, version 1.57. By combining the forward and reverse reads, this method takes advantage of bidirectional amplicon sequencing, because the forward reads will have highest quality in the 5′-end of the target sequence, and the reverse reads will improve the 3′-end quality. Flowgram clustering was performed using PyroDist, FCluster, and PyroNoiseM from the AmpliconNoise package, version 1.25 (61). The flowgram clusters produced by AmpliconNoise were base called using Multipass to obtain the most likely *var* DBLα sequences given the flowgrams and a high ORF likelihood, as described in Rask et al. (62). The nucleotide sequences generated by Multipass were clustered by 96% identity using Usearch, version 5.2.32 (63) with seeds (cluster member with support from highest number of reads after dereplication) as output. Chimeras were removed using Uchime implemented in Usearch, version 5.2.32 (63, 64), first in de novo mode where chimera detection is based on read abundance, all parents are expected to be present in the sequence set, and candidate parents must be at least 2× more abundant than the chimera candidate sequence; and subsequently in database mode, where sequences are searched against self and chimeras are found irrespective of the abundance of the parents. To increase overall quality of the sequences remaining at this point, a minimal coverage threshold of three reads per sequence type was applied to remove the least supported sequences. Next, we screened for and removed nontarget-amplified human sequences by local alignment search against the BLAST human genomic databases (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using the blastn feature of BLAST+ 2.2.25 [National Center for Biotechnology Information (NCBI)], with expectation value criteria of 1e-50. Sequences were also BLASTed against the remaining 3D7 genome, to remove any potential contamination, and searched using a DBLβ HMM with HMMer, version 3.1 (hmmer.org). After the human and nontarget *P. falciparum* check, a small number of sequences remained (0.089%) that had no similarity to a DBLα-tag HMM. When these sequences were BLASTed against NCBI databases, homology was found to among other several bacteria and so these sequences were removed. The pipeline was validated and optimized on experimental sequence data generated on the laboratory clones (3D7, Dd2, and HB3) for which published genome sequence is available. More than 90% of the sequences obtained from the control samples had no errors compared with the known reference, and the deviating sequences had maximally five errors. To subsequently determine *var* DBLα types shared between isolates, we clustered nonredundant sequences from all samples within each sentinel site by average linkage using a sequence identity threshold of 96%.

Hill Numbers. Based on Chao et al. (34), we use a formulation of the Hill numbers that is appropriate for sample-based incidences. Specifically, the sampling unit here is not an individual DBLα type but an isolate, and the presence of a given type is obtained for a given isolate. (This is similar to the case of a sampling unit consisting of a patch or area rather than an individual of a given species when ecologists consider patterns of species diversity in multiple locations.) For each of the S observed DBLα types, M_i is the number of isolates in which type i was positively identified. For each type i, the probability of being identified in an isolate when it is there is α_i (i.e., hence the probability of a false negative is $1 - \alpha_i$). Hence, π_i is the general probability of observing a DBLα type i in an isolate, and we can use the data to calculate that $\pi_i = M_i/S$. The general equation for the Hill numbers is given by the following:

$$\begin{aligned} q \neq 1: \quad {}^qH &= \left(\sum_{i=1}^{s} \left(\pi_i \Big/ \sum_{j=1}^{s} \pi_j \right)^q \right)^{1/(1-q)}, \\ q = 1: \quad {}^qH &= \exp\left(-\sum_{i=1}^{s} \left(\pi_i \Big/ \sum_{j=i}^{s} \pi_j \right) \log\left(\pi_i \Big/ \sum_{j=i}^{s} \pi_j \right) \right). \end{aligned} \quad [1]$$

Null Model Approach. We make use of the "all possible worlds" approach to identify signatures of nonrandom structure in our data. The arena of null hypothesis testing makes it possible to study whether observed data are different from what might be expected had DBLα types been distributed randomly between isolates. Using different statistical measures for characterizing our data, we then compare their values to what would be expected in a random ensemble that was generated subject to given constraints; this makes it possible to check whether particular features are significantly overrepresented or underrepresented compared with the reference null model. Carrying out this procedure requires defining appropriate preimposed constraints, as well as a technique for generating the ensemble of random samples with equal likeliness for each. In the absence of knowledge about any other lower-level structure affecting the distribution of DBLα types, for our analysis in this study we require that the random ensemble conserve the sizes of isolates, ensuring that all samples of the ensemble are neither larger nor smaller than what had been observed. We further require that the frequency distribution of DBLα types is conserved as well.

Random Ensemble. To generate a set of random samples, we present the data as a binary matrix, A, where each row represents a different DBLα type, and each column, an isolate. Here, for example, if DBLα type i was identified in isolate j, matrix cell $A(i,j) = 1$, and otherwise $A(i,j) = 0$. Fulfilling our constraints is equivalent to conserving row and column sums for all of the matrices in our random sample. Beginning with the observed matrix A, the DBLα types are shuffled by using "checkerboard" switches (65). This process is implemented by identifying a random "checkerboard" quadrat that fulfills the following: $A(x,y) = 0$, $A(x,z) = 1$, $A(w,y) = 1$, $A(w,z) = 0$, which is then switched to: $A(x,y) = 1$, $A(x,z) = 0$, $A(w,y) = 0$, $A(w,z) = 1$. The first sample is generated by implementing 100,000 switches, after which the next samples are separated by 50,000 switches. This process continues until the total number of samples has been obtained. Some of the binary matrices being generated have a higher probability of being sampled than others (66), potentially leading to biases when assessing the expected values of statistical measures. This is fixed by using weighted scores that are reversely proportional to the number of "checkerboard" switches of each of the random matrices (for further details, see ref. 66).

Associations Between DBLα Types. We constructed a co-occurrence matrix C for the DBLα types where the value inserted into each cell $C(i,j)$ equals the number of isolates in which both DBLα types i and j were identified together. The number of rows and columns in this matrix equals the total number of unique DBLα types identified. Similarly, co-occurrence matrices were generated for each of the matrices in the random ensemble.

PTS and Other Similarity Indices. The PTS statistics were calculated to quantify the relatedness between the *var* gene repertoires identified from two distinct isolates using the DBLα domain. This methodology has been published

elsewhere (6, 7), and it provides a useful statistic to analyze diversity and determine the number of unique DBLα types shared between isolates. PTS is specifically calculated as a ratio of the number of shared unique DBLα types between two isolates and the sum of the number of unique DBLα types of the two isolates. The ratio ranges between 0 and 1, where a PTS score of 0 signifies no similarity and 1 signifies an identical DBLα repertoire. If isolate A has a repertoire of n_A unique DBLα types, isolate B has a repertoire of n_B unique DBLα types, and a total n_{AB} DBLα types are shared by the isolates A and B, we define PTS as follows:

$$PTS_{AB} = 2n_{AB}/n_A + n_B. \quad [2]$$

For the definitions of other indices borrowed from community ecology and used here to quantify the overlap between isolates, see the legend of Table S4.

Removal Experiment. Our computational experiment consists of comparing the Hill diversity profile of our observed data to those of a random ensemble of datasets following the removal of isolates. A baseline ensemble of 500 randomized datasets was constructed as described above in Random Ensemble. Because the frequency distribution of DBLα types is conserved in this ensemble, the diversity profiles of all datasets in this ensemble are identical to the diversity profile of our observed data. For each of these sets, we removed the same given percentage of isolates and calculated their new diversity profiles. This process was repeated 100 times, with a different combination of isolates being removed each time. For each removal percentile X, 100*(500 + 1) Hill profiles were obtained, where X = [40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%].

The dependency between the Hill numbers obtained for each sample implies that these can be viewed as multivariate responses. To account for the intercorrelation of these dependent variables, we used multivariate analysis of variance (MANOVA) for testing whether there is a significant difference between the diversity of the sets following removal of isolates. To evaluate the difference between the datasets for all removal percentiles, we used four different statistics: Pillai's trace, Wilks' Lambda, Hotelling–Lawley trace, and Roy's maximum root statistic. Our analysis was carried out with MATLAB_R2015a software using the "manova" function with two between-subjects predictor variables: Data Set and Removal Combination. A multi-comparison Tukey's post hoc test between datasets was used to compare marginal means between our observed data and the reshuffled datasets (the marginal mean being the mean of the multivariate response, i.e., the mean of the measured Hill numbers).

ACKNOWLEDGMENTS. We are grateful to the children and their families of Bakoumba for their willingness to participate in this study. We thank Justice Mayombo, Faustin Lekoulou, and Herbert Moukana for their technical expertise and assistance. We thank Jean Bourgeais (Société d'Exploitation des Produits Alimentaires) for logistical support in Bakoumba. Finally, we thank Michael Duffy for helpful input related to this work and to everyone involved for their continued patience as this research was disrupted due to Hurricane Sandy (New York, NY; October 29, 2012). M.P. is an Investigator at the Howard Hughes Medical Institute. This research was supported by Fogarty International Center at National Institutes of Health, Program on the Ecology and Evolution of Infectious Diseases Grant R01-TW009670 and National Institute of Allergy and Infectious Disease, National Institutes of Health Grant R01-AI084156.

1. Ewers W (1972) Robert Koch, his work in New Guinea and his contributions to malariology. *P N G Med J* 15:117–124.
2. Day KP, Marsh K (1991) Naturally acquired immunity to *Plasmodium falciparum*. *Immunol Today* 12:A68–A71.
3. Koch R (1900) Professor Koch's investigations on malaria. *BMJ* 1:325–327.
4. Gardner MJ, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498–511.
5. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T (2010) *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes—divide and conquer. *PLoS Comput Biol* 6:e1000933.
6. Barry AE, et al. (2007) Population genomics of the immune evasion (*var*) genes of *Plasmodium falciparum*. *PLoS Pathog* 3:e34.
7. Chen DS, et al. (2011) A molecular epidemiological study of *var* gene diversity to characterize the reservoir of *Plasmodium falciparum* in humans in Africa. *PLoS One* 6:e16629.
8. Tessema SK, et al. (2015) Phylogeography of *var* gene repertoires reveals fine-scale geospatial clustering of *Plasmodium falciparum* populations in a highly endemic area. *Mol Ecol* 24:484–497.
9. Paul RE, et al. (1995) Mating patterns in malaria parasite populations of Papua New Guinea. *Science* 269:1709–11.
10. Babiker HA, et al. (1994) Random mating in a natural population of the malaria parasite *Plasmodium falciparum*. *Parasitology* 109:413–421.
11. Kraemer SM, Smith JD (2006) A family affair: *var* genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol* 9:374–380.
12. Warimwe GM, et al. (2012) Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles. *Sci Transl Med* 4:129ra45.
13. Normark J, et al. (2007) PfEMP1-DBL1alpha amino acid motifs in severe disease states of *Plasmodium falciparum* malaria. *Proc Natl Acad Sci USA* 104:15835–15840.
14. Bull PC, et al. (2005) *Plasmodium falciparum* variant surface antigen expression patterns during malaria. *PLoS Pathog* 1:0202–0213.
15. Avril M, et al. (2012) A restricted subset of *var* genes mediates adherence of *Plasmodium falciparum*-infected erythrocytes to brain endothelial cells. *Proc Natl Acad Sci USA* 109:E1782–E1790.
16. Anderson TJ, et al. (2000) Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol Biol Evol* 17:1467–1482.
17. Daniels R, et al. (2008) A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. *Malar J* 7:223.
18. Daniels R, et al. (2013) Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* 8:e60780.
19. Gupta S, Trenholme K, Anderson RM, Day KP (1994) Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. *Science* 263:961–963.
20. Gupta S, Day K (1994) A strain theory of malaria transmission. *Parasitol Today* 446: 3737–3742.
21. Gupta S, et al. (1996) The maintenance of strain structure in populations of recombining infectious agents. *Nat Med* 2:437–442.
22. Buckee CO, Bull PC, Gupta S (2009) Inferring malaria parasite population structure from serological networks. *Proc Biol Sci* 276:477–485.
23. McKenzie FE, Smith DL, O'Meara WP, Riley EM (2008) Strain theory of malaria: The first 50 years. *Adv Parasitol* 66:1–46.
24. Saul A (1996) Transmission dynamics of *Plasmodium falciparum*. *Parasitol Today* 12: 74–79, discussion 82–83.
25. Gupta S, Anderson RM (1999) Population structure of pathogens: The role of immune selection. *Parasitol Today* 15:497–501.
26. Artzy-Randrup Y, et al. (2012) Population structuring of multi-copy, antigen-encoding genes in *Plasmodium falciparum*. *eLife* 1:e00093.
27. Rougeron V, et al. (2013) Epistatic Interactions between apolipoprotein E and hemoglobin S genes in regulation of malaria parasitemia. *PLoS One* 8:e76924.
28. Lindblade KA, Steinhardt L, Samuels A, Kachur SP, Slutsker L (2013) The silent threat: Asymptomatic parasitemia and malaria transmission. *Expert Rev Anti Infect Ther* 11: 623–639.
29. Rek J, et al. (2016) Characterizing microscopic and submicroscopic malaria parasitaemia at three sites with varied transmission intensity in Uganda. *Malar J* 15:470.
30. Stresman GH, et al. (2014) High levels of asymptomatic and subpatent *Plasmodium falciparum* parasite carriage at health facilities in an area of heterogeneous malaria transmission intensity in the Kenyan highlands. *Am J Trop Med Hyg* 91:1101–1108.
31. Greenwood B, Marsh K, Snow R (1991) Why do some African children develop severe malaria? *Parasitol Today* 7:277–281.
32. Colwell RK (2009) Biodiversity: Concepts, patterns, and measurement. *The Princeton Guide to Ecology* (Princeton Univ Press, Princeton), pp 257–264.
33. Hill M (1973) Diversity and evenness: A unifying notation and its consequences. *Ecology* 54:427–432.
34. Chao A, et al. (2014) Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol Monogr* 84:45–67.
35. Pianka ER (1973) The structure of lizard communities. *Annu Rev Ecol Syst* 4:53–74.
36. Feinsinger P, Spears EE, Poole RW (1981) A simple measure of niche breadth. *Ecology* 62:27–32.
37. Jaccard P (1912) The distribution of the flora in the Alpine zone. *New Phytol* 11:37–50.
38. Bull PC, et al. (2008) *Plasmodium falciparum* antigenic variation. Mapping mosaic *var* gene sequences onto a network of shared, highly polymorphic sequence blocks. *Mol Microbiol* 68:1519–1534.
39. van Noort SP, Nunes MC, Weedall GD, Hviid L, Gomes MGM (2010) Immune selection and within-host competition can structure the repertoire of variant surface antigens in *Plasmodium falciparum*—a mathematical model. *PLoS One* 5:e9778.
40. Buckee CO, Recker M, Watkins ER, Gupta S (2011) Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. *Proc Natl Acad Sci USA* 108:15504–15509.
41. Gupta S, Ferguson NM, Anderson R (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* 280:912–915.
42. Daniels RF, et al. (2015) Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc Natl Acad Sci USA* 112:7067–7072.
43. Kaestli M, et al. (2006) Virulence of malaria is associated with differential expression of *Plasmodium falciparum var* gene subgroups in a case-control study. *J Infect Dis* 193: 1567–1574.
44. Falk N, et al. (2009) Analysis of *Plasmodium falciparum var* genes expressed in children from Papua New Guinea. *J Infect Dis* 200:347–356.
45. Rottmann M, et al. (2006) Differential expression of *var* gene groups is associated with morbidity caused by *Plasmodium falciparum* infection in Tanzanian children. *Infect Immun* 74:3904–3911.
46. Kyriacou HM, et al. (2006) Differential *var* gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia. *Mol Biochem Parasitol* 150:211–218.

PNAS PLUS

47. Kalmbach Y, et al. (2010) Differential *var* gene expression in children with malaria and antidromic effects on host gene expression. *J Infect Dis* 202:313–317.
48. Jespersen JS, et al. (2016) *Plasmodium falciparum var* genes expressed in children with severe malaria encode CIDRα1 domains. *EMBO Mol Med* 8:839–850.
49. Bernabeu M, et al. (2016) Severe adult malaria is associated with specific PfEMP1 adhesion types and high parasite biomass. *Proc Natl Acad Sci USA* 113:E3270–E3279.
50. Lau CKY, et al. (2015) Structural conservation despite huge sequence diversity allows EPCR binding by the PfEMP1 family implicated in severe childhood malaria. *Cell Host Microbe* 17:118–129.
51. Claessens A, et al. (2012) A subset of group A-like *var* genes encodes the malaria parasite ligands for binding to human brain endothelial cells. *Proc Natl Acad Sci USA* 109:E1772–E1781.
52. Lavstsen T, et al. (2012) *Plasmodium falciparum* erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children. *Proc Natl Acad Sci USA* 109:E1791–E1800.
53. Avril M, Brazier AJ, Melcher M, Sampath S, Smith JD (2013) DC8 and DC13 *var* genes associated with severe malaria bind avidly to diverse endothelial cells. *PLoS Pathog* 9:e1003430.
54. Poirot E, et al. (2013) Mass drug administration for malaria. *Cochrane Database Syst Rev* 12:CD008846.
55. Newby G, et al. (2015) Review of mass drug administration for malaria and its operational challenges. *Am J Trop Med Hyg* 93:125–134.
56. Elissa N, et al. (2003) Relationship between entomological inoculation rate, *Plasmodium falciparum* prevalence rate, and incidence of malaria attack in rural Gabon. *Acta Trop* 85:355–361.
57. Ntoumi F, et al. (2002) Sickle cell trait carriage: Imbalanced distribution of IgG subclass antibodies reactive to *Plasmodium falciparum* family-specific MSP2 peptides in serum samples from Gabonese children. *Immunol Lett* 84:9–16.
58. Fowkes FJI, et al. (2006) Association of haptoglobin levels with age, parasite density, and haptoglobin genotype in a malaria-endemic area of Gabon. *Am J Trop Med Hyg* 74:26–30.
59. Smith JD, Subramanian G, Gamain B, Baruch DI, Miller LH (2000) Classification of adhesive domains in the *Plasmodium falciparum* erythrocyte membrane protein 1 family. *Mol Biochem Parasitol* 110:293–310.
60. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5:235–237.
61. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12:38.
62. Rask TS, Petersen B, Chen DS, Day KP, Pedersen AG (2016) Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data. *BMC Bioinformatics* 17:176.
63. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
64. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200.
65. Stone L, Roberts A (1990) The checkerboard score and species distributions. *Oecologia* 85:74–79.
66. Artzy-Randrup Y, Stone L (2005) Generating uniformly distributed random networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 72:056708.
67. Colwell RK, et al. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* 5:3–21.
68. Chao A, Chazdon RL, Shen TJ (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* 8:148–159.