An antigenic diversification threshold for falciparum malaria and its control at high endemicity

Qixin He[a,] **and Mercedes Pascual**[a,b,1]

[a]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [b]Santa Fe Institute, Santa Fe, NM, 87501

This manuscript was compiled on March 6, 2020

In malaria and several other important infectious diseases, high prevalence occurs concomintantly with incomplete immunity. This apparent paradox poses major challenges to malaria elimination in highly endemic regions, where asymptomatic *Plasmodium falciparum* infections are present across all age classes creating a large reservoir that maintains transmission. This reservoir is in turn enabled by extreme antigenic diversity of the parasite and turnover of new variants. We present here the concept of a threshold in local pathogen diversification that defines a sharp transition in transmission intensity below which new variants generated by either recombination or migration cannot establish. Transmission still occurs below this threshold but parasite diversity can neither accumulate, nor recover, from interventions that further reduce it. An analytical expectation for this threshold is derived and compared to numerical results from a stochastic individual-based model of malaria transmission that incorporates the major antigen-encoding multigene family known as *var*. This threshold we call R_{div}; it is complementary to the one defined by the classic basic reproductive number of infectious diseases, R_0, which does not easily apply under large and dynamic strain diversity. This new threshold concept should be exploited for effective malaria control and applied more broadly to other pathogens with large multilocus antigenic diversity.

antigenic diversification | epidemiological measure | malaria control

The reproductive number of an infectious disease, R_0, quantifies its epidemic growth potential and provides a threshold condition for the spread and control of pathogens (1). This number has been applied extensively to pathogens with either no antigenic variation or a typically low number of genetically defined strains that are relatively stable in space and time. It becomes problematic how to evaluate and even to define it when antigenic variation is large and dynamic, challenging the actual definition of a strain. This is particularly the case for *Plasmodium falciparum* as well as for other pathogens with multigene and multilocus encoding of antigens, combined with extensive recombination (2–5). Furthermore, R_0 does neither predict the turnover rate of new genes encoding antigens, nor explain how transmission characteristics and accumulation of these new genes influence each other and responses to control.

Under high transmission of *P. falciparum*, the major antigen-encoding gene family of the blood stage of malaria infection, known as *var*, acquires vast antigenic diversity via large gene copy numbers as well as ectopic recombination. Laboratory experiments have shown that a naïve infection can generate about sixty new recombinants per year (6, 7) although this has not yet been demonstrated to occur in nature. Within local populations, parasites share only a few common *var* genes between different strains (8–10), and across seasons (11). Spatial diversity in *var* genes has been documented (12–14) indicating that migration from surrounding areas also contributes to new diversity and to the immunological challenge.

The accumulation and turnover of new antigenic variants constitutes a major impediment to control in regions of high endemism, where it underlies the large reservoir of chronic asymptomatic infections that sustains transmission. Today the global burden of *Plasmodium falciparum* is concentrated in these high transmission endemic areas within fifteen countries, mainly in sub-Saharan Africa (WHO 2017). A similar reservoir is found in other vector-borne diseases that exhibit high prevalence of infection with no clinical symptoms in domestic and wildlife hosts (15–17). Nonsterile specific immunity is common to all these pathogens as a result of extreme antigenic variation encoded by multigene families (18, 19).

We present here a reproductive number complementary to R_0 that defines a threshold for parasite antigenic diversification below which the accumulation of new variants no longer occurs even though they are consistently produced. We derive an analytical expectation for the rate of generation of "successful" new genes, and numerically compute the analytical threshold associated with this rate using simulations of a stochastic agent-based model that incorporates *var* genes and the acquisition of immunity by individual hosts. We then investigate the epidemiological and evolutionary factors that

Significance Statement

The vast diversity of falciparum malaria as seen by the immune system underlies only partial protection to re-infection and many infections lacking symptoms in high transmission regions. Large antigenic parasite diversity challenges control efforts and the associated application of the classic reproductive number (known as R0) because strains cannot be clearly defined. We propose a complementary number, Rdiv, measuring the potential of transmission to accumulate new antigenic variation over time. When Rdiv is pushed below one, new genes encoding this variation cannot accumulate resulting in a low number of strains and facilitating further intervention. The reproductive number for antigenic diversification can be applied to other infectious diseases with fast turnover of antigens, where large standing diversity similarly opposes successful control.

Q.H. and M.P. designed the research. Q.H. performed the research, proposed and derived the threshold concept. Q.H. and M.P. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence should be addressed. E-mail: pascualmm@uchicago.edu

DRAFT

influence this rate analytically. We show that this threshold for the accumulation of novelty, we call R_{div}, maps onto one in transmission intensity, separating a regime in which new genes are able to accumulate from one in which they are unable to do so, despite transmission still occurring (R_0 remaining above one). We discuss implications for malaria control and elimination and for the monitoring of the effect of these efforts in high transmission endemic regions.

Results

The parasite population will accumulate new genetic variants when these are produced at a sufficient rate for their lifespans to overlap with each other (Fig. S1). Novelty per se guarantees neither their establishment nor their persistence. Even under high absolute fitness, new variants need to survive initial drift to establish in the parasite population (20). Their accumulation further requires that the rate at which they are generated, G_{new}, be on average larger than that of their loss, given by the inverse of their lifespan, T_{new}. In other words, at least one beneficial gene needs to be produced and become established in the population during the typical lifespan of a previously generated new gene. We denote by R_{div} the expected number of new genes produced during the average lifespan. This reproductive or innovation number should be greater than 1 for new variants to accumulate, namely

$$R_{div} = G_{new} T_{new} = N \mu p_{inv} T_{new} > 1 \qquad [1]$$

where N denotes the population size of the parasite, μ, the mutation rate of the genes, and p_{inv}, the invasion probability of a low-frequency variant. The average lifespan of a new gene T_{new} is under frequency-dependent selection from acquisition of immunity by the host, with immune selection conferring an advantage to the rare and a disadvantage to the common (10, 21). New variants are generated through either mutation or ectopic recombination (Methods), and therefore μ generically refers to the rate of novelty generation regardless of specific mechanism. Equation (1) establishes an expectation for the existence of a threshold, whose expression we proceed to further develop next, to be able to verify it computationally.

The factors that determine the spread and establishment of a new antigenic variant consist of its selective advantage and the rates of innovation and transmission. We first consider the probability p_{inv} that a new gene survives its initial low frequency and invades. Based on birth-death processes in the Moran model with selection (22), the establishment probability of a low frequency variant is largely determined by its relative fitness advantage over other genes. The fitness of genes in a transmission model is essentially given by their effective reproductive number R_{eff} or number of copies they produce via transmission events during their lifetime. Thus,

$$p_{inv} \approx (R_{new}/\bar{R}_{eff}) - 1 \qquad [2]$$

Here, R_{eff} for a given *var* gene is the product of the epidemiological contact rate of the disease (β) and the typical infection duration (τ) of parasites that carry the given gene (Methods). If we consider that genes are equivalent in transmissibility (i.e., their products are functionally equivalent in their ability to bind host receptors), an assumption we later relax in numerical results, fitness differences between variants are only determined by the duration of infection these genes can typically sustain.

Equation Eq. (2) holds in general for any infectious disease that generates new antigens. We consider here the specifics of *Plasmodium falciparum* and its multicopy *var* genes (typically about 40-60 per genome), whose expression is sequential during the blood stage of infection (23). Because only those genes towards which the host has not yet built immunity are expressed, the average duration of infection will equal the number of genes per genome times the average proportion of susceptible (non-immune) hosts per gene,

$$\bar{\tau} = dg \sum_{i=1}^{k} S_i f_i \qquad [3]$$

where d is the duration of infection for a given gene in a naive host, g is the number of genes per genome, and f_i is the population frequency of a given gene. We can rewrite equation Eq. (2) using Eq. (3) (Methods), to obtain

$$p_{inv} \approx \frac{S_{new} - \bar{S}}{\bar{S} g} \qquad [4]$$

where the mean number of susceptible hosts for a gene is given by $\bar{S} = \sum_{i=1}^{k} S_i f_i$. This expression for the invasion probability shows that a new gene is likely to invade when it affords a wider host niche than that of older genes (by encoding for epitopes that are new given the immunity of the host population). In other words, the available number of hosts for its expression should be higher on average than that for existing genes. In addition, the invasion probability of a single gene decreases with increasing genome size g, as the importance of a single gene also decreases.

We can now evaluate the existence of the threshold behavior indicated by the above analytical argument. To this end, we compute R_{div} from a stochastic agent-based model of malaria transmission (10, 24) that tracks *var* evolution and immunity and is described in detail in (10). A number of extensions are also considered here to address the generality of the argument, including distinct major *var* gene groups with associated differential fitness and constrained recombination (Methods). We specifically examine how the accumulation of new variants over a given period of time, varies as a function of R_{div}. We calculate G_{new} according to Eq. Eq. (2) and Eq. (4) by obtaining $\bar{S}$, N, and μ directly from the simulations, and T_{new} from the average lifespan of all the new variants that are produced during this time period.

Results show that the transmission system naturally falls into two regimes separated by a threshold at $R_{div} = 1$ (Fig. 1A). Below this transition, new antigenic variants are generated but do not accumulate or persist (Fig. 1C), whereas above it, they are able to accumulate and experience a continuous turnover rate (see shifting shades of colors in Fig. 1D). The transition between these regimes occurs around the proposed boundary where the rate of generation of genes surpasses the average lifespan of new antigenic variants ($R_{div} > 1$). This threshold is robust to differences in specific assumptions about the transmission and genetic systems (including processes of within-host dynamics, functional differences between genes, values of the recombination and biting rates), as each point in Fig. 1A represents a simulation with different model assumptions and parameter combinations (Methods; Table S1-2).

DRAFT

Importantly, the quantity R_{div} scales monotonically with the intensity of transmission measured here as the entomological inoculation rate (or EIR, the number of infectious bites per person per year) (Fig S2, 1B), a practical empirical measure from field epidemiology. The association with R_{div} should hold more generally with any other measure of transmission intensity. This association implies that the transition between regimes also occurs as a function of transmission intensity (Fig 1A, C, D), and therefore, that the malaria system can be pushed below threshold by changing this control variable.

The transition examined so far represents the behavior of the system for different values of R_{div} or transmission intensity. Its existence should influence the temporal response of the malaria system to intervention events that reduce transmission at a given point in time. In particular, interventions that take the transmission system above threshold should lead to distinct responses than those that fail to do so. This is illustrated in Fig. 1 (G and H) where we numerically introduce a transient reduction of the biting rate to lower levels (respectively 30 and 50% of its original value). New genes cease to accumulate only when R_{div} goes below threshold (Fig. 1G), whereas they continue to invade and accumulate following a temporary decrease otherwise (Fig. 1H).

In order to further understand how epidemiological and genetic factors influence R_{div}, we examine with a simple theoretical model the equilibrium values of $\bar{S}$ and N, which enter prominently in the expression for G_{new} (Methods). At equilibrium, increasing contact rates (β) via mosquito bites, result in higher parasite population sizes (N) and a lower average number of susceptible hosts ($\bar{S}$, Fig. S2). More specifically, $\bar{S}$ is mostly determined by β and g, whereas parasite population size strongly scales with the diversity ratio (i.e., gene pool size G over genome size, G/g). A lower $\bar{S}$ favors invasion of a new variant (i.e., increases p_{inv}), and a higher parasite population size (N) and a genome with a higher number of unique genes (g) generate new variants faster. Theoretical predictions underestimate $\bar{S}$ because they neglect the higher-level of organization of the genes into different genomes, also under immune selection (10). This indicates that such strain structure can significantly reduce the percentage of genes that a host is immune to, especially under high competition and high diversity (see $g = 60$, and $G/g = 100$ in Fig. S3B).

An explicit expression for T_{new} cannot be obtained analytically because this quantity continuously changes as new genes enter the system and influence the nonlinear dynamics of N, G and $\bar{S}$. To gain nevertheless an understanding for how these variables affect T_{new}, we approximate the average lifespan of a new gene on the basis of an adapted diffusion equation (22) under the assumption that the system remains constant and only this lifespan, we call $\bar{t}$, varies (SI). The diffusion equation for $\bar{t}$ requires consideration of how the frequency of a new gene $x(t)$ varies in time (SI). When applied to our model, the resulting analytical approximation for $\bar{t}$ (Eq. S8-11) shows that the expected lifespan of a new gene grows faster than exponential with decreasing $\bar{S}$, and surpasses the average time to fixation of a neutral gene ($2N$) when $\bar{S}$ is below a given value (40%) (Fig. 2). The mean T_{new} evaluated numerically will always be shorter than that predicted from the diffusion approximation $\bar{t}$, especially as $\bar{S}$ becomes smaller and persistence times rapidly increase. This is because the stochastic simulations can only track lifespan within a finite time period (which places an upper bound on its value), and because the assumption of constant $\bar{S}$ does not apply. Nevertheless, the theoretical trend of increasing T_{new} with decreasing $\bar{S}$ and therefore, higher transmission intensity, does apply to the numerical system.

In summary, by evaluating whether $G_{new} < 1/T_{new}$ (or equivalently, $R_{div} < 1$), one can predict whether the system has a relatively stable antigen-encoding gene pool, or whether alternatively, new variants continuously enter into it. We have shown that whether new genes can successfully establish in the population is most tightly linked with the average proportion of susceptible hosts (or the niche) available for existing genes ($\bar{S}$). When transmission rate is low, $\bar{S}$ is large and new genes do not have a significant advantage over older ones. New genes experience a small invasion probability, and even when they invade, they experience strong drift, functioning as effectively neutral. As transmission intensity increases, the selective advantage of new genes also increases as $\bar{S}$ decreases. Once $\bar{S}$ is below a given value (0.4 in our simulations), new genes are most likely to be maintained in the population indefinitely. Concomitantly, the increase in gene diversity results in higher parasite population sizes N (Fig. S3A). The system thus enters a regime of positive feedback for new variants, as elevated diversity boosts N and therefore also, G_{new}, before reaching equilibrium.

Discussion. The concept of a threshold behavior in the accumulation of antigenic variants has implications for overcoming the resilience of highly endemic falciparum malaria to intervention efforts. Although a decreasing trend in antigenic diversity with transmission intensity is well known and expected from both biogeography and epidemiology, the actual form of this reduction is much less clear. Our results indicate the existence of a sharp transition below which the disease system should effectively respond as a typical low transmission region, not just because of reduced transmission intensity but also because of much lower antigenic diversity no longer able to rebuild. Failure to push transmission intensity below this threshold would lead to a fast rebound in new variation, despite an overall diversity reduction. Crossing of the threshold would instead provide an indication that the system is now poised for further intervention with enhanced results.

Control and even elimination efforts are indeed known to be most successful in biogeographical regions of low transmission, such as those at the edge of the distribution of the disease in Africa and in other continents (25). Arresting the fast turnover of the local antigenic pool typical of high endemism would significantly repress disease burden and facilitate its further reduction. Concomitant control efforts at a regional level are critical to stem immigration, as migrant genomes would exhibit higher invasion probabilities than local ones given their higher likelihood of harboring new antigens. Monitoring the turnover of *var* gene diversity through molecular epidemiology in response to control efforts should inform intervention evaluation in high transmission regions.

Although the importance of host immune selection in shaping the antigenic variation of *P. falciparum* and other pathogens is recognized (21, 26–28), mathematical and computational models typically evaluate intervention efficacy without explicit consideration of antigenic diversity (e.g. (29, 30)) and openness of the system to innovation (31). Our results underscore the importance of these aspects.

Estimation of the diversification number, R_{div}, would pro-

DRAFT

vide general guidelines for intervention evaluation where traditional application of R_0 is unable to do so for highly diverse pathogens (2, 21). Future work should consider how to obtain this number from estimation of key parameters, such as parasite population size, transmission rates and gene pool size, based on combined data from molecular and field epidemiology. Parameterization of an agent-based stochastic transmission model that implements immune selection and recombination explicitly (e.g., (10, 32)) could be used, which represents a computational challenge (but see (33)). Estimating the viability of new recombinants will require bioinformatic analyses of population-level *var* sequence data.

For simplicity, our analytical derivations treated each gene independently, even though *var* gene composition in parasite genomes of local populations in regions of high transmission has been shown to be non-randomly and non-neutrally structured, exhibiting low overlap as the result of immune selection (10, 12, 24, 34). Hence, the fate of a viable new antigen-encoding gene depends on its genomic background, which ultimately determines the strength of competition among parasites. Comparisons of analytical expectations with numerical simulations revealed an influence of such population structure on the fate of new genes, and therefore, on components of R_{div}. Future work should examine extensions of this work that account for this further complexity of immune selection operating at different levels of organization.

The concept of R_{div} arises from the interplay of immune memory and antigenic variation at the population level, as a result of frequency-dependent selection. As such, it differs from the antigenic diversity threshold previously proposed for the HIV virus and its transition to AIDS, arising from the race between viral replication and immune responses at the within-host level (Nowak and May, 1991). The concept itself and the associated transition regime described here should apply more generally to other infectious diseases with antigen-encoding multigene families, such as *vsg* genes in *Trypanosoma brucei* and *msg* genes in *Pneumocystis carinii* (18). Because the basic concept is independent from specific consideration of multigene families and their properties, it should be adaptable to other pathogens in which large standing antigenic diversity at the population level results from multilocus genetic variation (5). Explicit consideration of a reproductive number for antigenic diversification will enhance our understanding of transmission dynamics where large standing pathogen diversity represents a main challenge to control efforts.

Materials and Methods

Analytical derivation of R_{div}. We consider a population of hosts whose number is denoted by N_{host}, receiving malaria infections from a diverse set of parasites, each composed of g genes from a constant gene pool of size G. One of the two main components of R_{div} is the rate at which new genes are generated, G_{new}. Besides the mutation rate μ and the equilibrium parasite population size N, its expression requires the invasion probability p_{inv}, we derive below.

Invasion probability of a new variant, p_{inv}. From birth-death processes according to the Moran model with selection (22), the probability of establishment of a low frequency variant is determined by its fitness advantage relative to that of other genes, and by the parasite population size. That is,

$$p_{inv} = \frac{1-(W/W_{new})^n}{1-(W/W_{new})^N} \qquad [5]$$

where n denotes the number of copies of new genes, and W, the fitness of a gene. In our case, $n = 1$ as new genes originate from a unique mutation or an ectopic recombination event. When $N >> 1$, the invasion probability p_{inv} is approximately its initial selective advantage relative to established gene variants, provided the selection coefficient remains the same. Since the fitness of each individual gene in a transmission model is essentially given by their effective reproductive number R_{eff}, we have

$$p_{inv} \approx (R_{new}/\bar{R}_{eff}) - 1. \qquad [6]$$

R_{eff} for a given *var* gene is in turn the product of the epidemiological contact rate of the disease (β) and the typical infection duration (τ) of parasites that carry the given gene,

$$R_{eff} = \bar{\beta}\bar{\tau} \qquad [7]$$

The contact rate β is equal to the product of the transmission rate (b) and the 'transmissibility' or infectivity of the given *var* gene (i.e., the functionality of the gene, f). Because we do not model vectors explicitly in the numerical stochastic model, the contact rate (β) refers to the rate at which a transmission event occurs, with a 'donor' host transmitting infection to a 'recipient' one (detailed description in section on "the modified *var* evolution model").

Different groups of *var* genes may vary in their binding affinities to host receptors and therefore in their transmissibility. For simplicity, we consider that all genes exhibit the same transmissibility and therefore, the same absolute fitness, as we are most interested in estimating the fate of a new variant as a result of immune (frequency-dependent) selection. (We do explore later the effect of fitness differences numerically with the agent-based stochastic model, described in the section on "the modified *var* evolution model").

With Eq. (7) for R_{eff}, we can write

$$\begin{aligned} p_{inv} &\approx \frac{\tau_{new}}{\bar{\tau}} - 1 \\ &= \frac{\bar{S}\times(g-1)+S_{new}}{\bar{S}\times g} - 1 \end{aligned} \qquad [8]$$

Numerical evaluation of R_{div}. With equation (8) (or its equivalent (4)), we can now compute $G_{new} = N\mu p_{inv}$ in equation (1) from the output of our numerical simulations (described below under the *The modified var evolution model*). We also obtained from the simulations the other major component of R_{div}, the mean lifetime of the genes in the system, T_{new}, by directly tracking their fate individually. Although an analytical expression for T_{new} was not achievable for this nonlinear and stochastic transmission system, we considered gene lifetime under simplifying assumptions, as explained next.

Analytical derivation of $\bar{t}$. The simplifying assumptions are that the system has reached an equilibrium ((i.e., parasites get transmitted and die at the same rate), and that only the average lifetime of a gene varies, with all other variables remaining unchanged, including N and the average proportion of hosts $\bar{S}$ susceptible to an average gene. To differentiate gene lifetime under these conditions from T_{new} itself, we call it $\bar{t}$. An expression for $\bar{t}$ is derived by considering the frequency-dependent selection experienced by a new gene variant entering the system at equilibrium. We specifically approximate the dynamics of $\bar{t}$ on the basis of an adapted diffusion equation (22) (Supplementary Text).

The modified *var* evolution model. We used an extended implementation of the agent-based model developed in (10), where a complete description can be found. Here, we first briefly summarize the main features of the computational model, and then document the specific changes implemented in this study, including different transmission scenarios and rules of within-host dynamics. (Parameter combinations and specific rules are listed in Table S1-2).

The computational model is an individual-based, discrete-event, continuous-time stochastic system in which the infection and immune history of each host are tracked individually. In the numerical

DRAFT

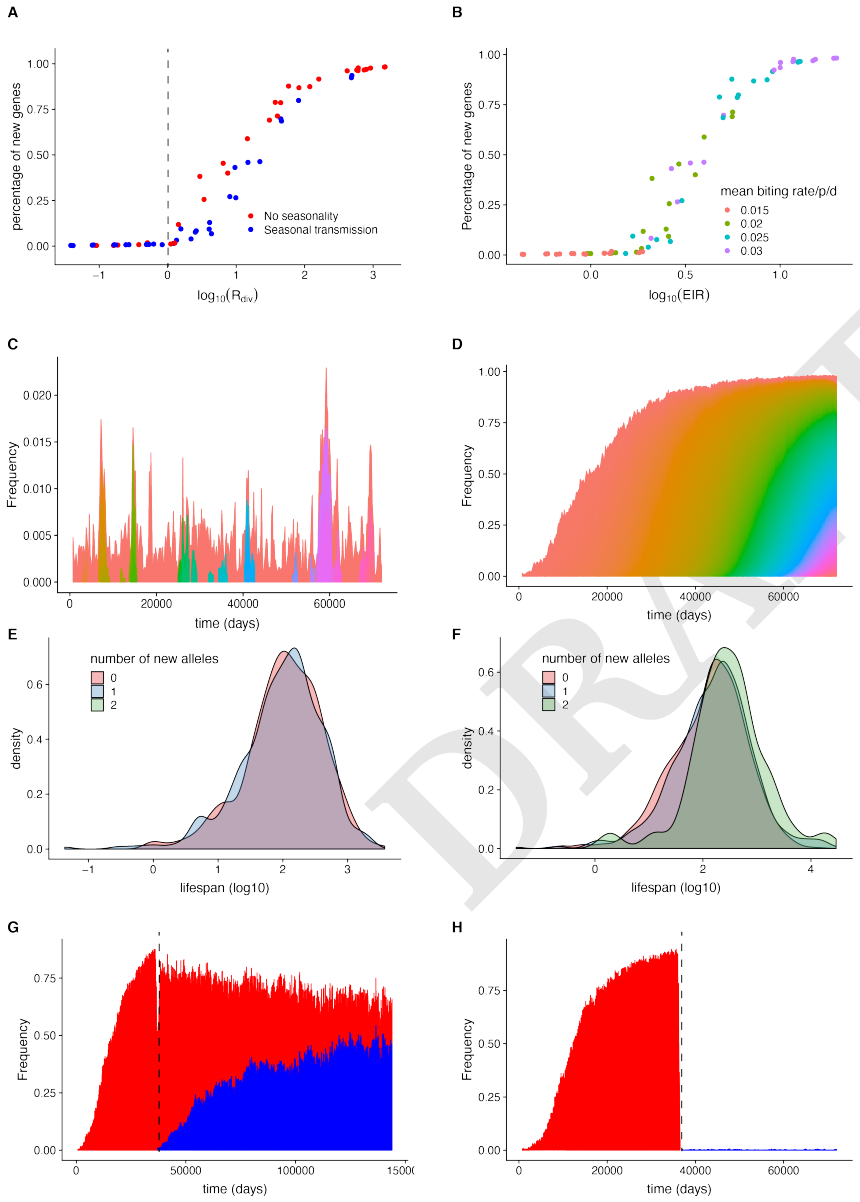

Fig. 1. Numerical simulations reveal a transition between two regimes of antigenic diversity accumulation. (**A**) The percentage of new genes in the local parasite population at the end of a given simulation period (200 years) remains negligible when the reproductive number R_{div} for antigen-encoding new genes is lower than one. By contrast, this percentage increases rapidly above this threshold. Because the time interval over which we computed $R_{div} = G_{new}T_{new}$ concerns long transients, we evaluated the rate of generation of new genes G_{new} as a mean over this interval (by averaging the values of N and $\bar{S}$ every 180-day interval), and the lifespan T_{new}, as an average for all the new genes that invaded during this time (with this interval placing an upper bound on individual lifespans). Each point represents a simulation with different combinations of parameters and assumptions (including variation in rules of within-host dynamics, in strength of the trade-off between transmissibility and duration of infection, and in values and seasonality of the transmission rates, Table S1-2). (**B**) Because R_{div} increases monotonically with transmission intensity (Fig. S2), the percentage of new genes also exhibits the threshold behavior with this variable, measured here by the entomological inoculation rate (EIR, the number of infectious bites per person per year). For simplicity, when a transmission event occurs, our model considers that all bites of an infected 'donor' host generate an infectious bite, and that all infectious bites of the 'recipient' host result in infection. This implies that the EIR values in the graph should be adjusted for comparison to actual field values (by dividing by the product of the competence/transmissibility probabilities, which will raise EIR). (**C**) New genes do not accumulate below the transmission threshold where they essentially follow neutral dynamics. In contrast, they do accumulate and turn over at a constant rate under frequency-dependent selection above this threshold (**D**). Each color in these two panels refers to a new gene in the population. New genes that account for less than five infections over the entire period are combined and represented in red. (**E**) The average lifespan of new genes is shorter below the threshold than above it (**F**). (**F**) In addition, new genes with a greater number of new alleles (epitopes) live longer above the threshold, whereas below the threshold, they experience similar lifespans. Interventions that push the system below the threshold are effective at stopping the accumulation of new genes (**H**), whereas those that do not, result in the rebound and rebuilding of diversity (**G**). Red and blue colors indicate genes that originate respectively before and after the intervention.

DRAFT

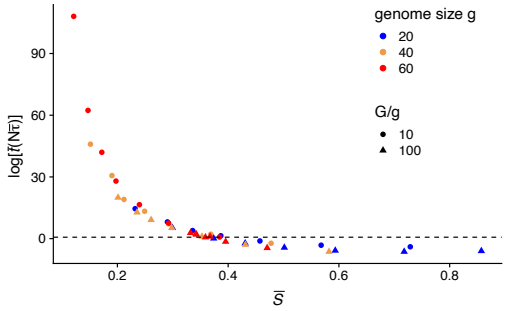

Fig. 2. Theoretical expectation of the average lifespan of a new gene $\bar{t}$. The analytical expression shows that $\bar{t}$, measured in units of $N\bar{\tau}$, increases faster than exponential as the average number of available hosts $\bar{S}$ decreases. The dashed line represents the time to fixation of a neutral gene, which means that under small $\bar{S}$, once established, the gene can be maintained in the population for much longer than the typical epidemiological timescale (or for much longer than the simulation period of 200 years in our model). The average lifespan T_{new} obtained from the computational model will always be considerably smaller than the theoretical expectation $\bar{t}$ derived under the assumption that other factors remain constant, in particular the average number of hosts $\bar{S}$ that are susceptible to the invading gene. The general trend of increased persistence with lower $\bar{S}$ will hold however for the full numerical system and for finite time windows.

implementation of the simulation, all possible future events are stored in a single event queue along with their putative times, which may be fixed or drawn from a probability distribution. When an event occurs, it may trigger the addition or removal of future events on the queue, or changes of their rates, leading to a recalculation of their putative time. The implementation is adapted from the next-reaction method (35), which optimizes the Gillespie first-reaction method (36) and allows for faster simulation times.

Transmission events are sampled at the rate, $N_{host}\beta$, in which a donor and a recipient host are sampled randomly from the host population. If the donor harbors parasites, then each parasite has a probability of being transmitted to the mosquito that is proportional to the functionality of the *var* gene that is currently under expression. *var* genomes picked up by the mosquito will recombine with another genome to produce sporozoites. Specifically, if there are n parasite genomes, each genome has probability $1/n$ of recombining with itself, producing the same offspring genome, and probability $1-1/n$ of recombining with a different genome, producing recombinants. The total number of *var* genomes passed onto the receiver is kept the same than that received from biting the infectious host. Parasites in a human host are not infectious until the completion of a delay, representing altogether oocysts development in the mosquito during the sexual stage and the initial liver stage in the receiver host. Since we do not model mosquitoes explicitly, we implement this delay as a 14-day gap between a transmission event and the genome becoming infectious. *var* genes within a genome express sequentially in a random order, or according to their functional level from high to low, depending on the specific rule (Table S1), with a switching rate of $1/d$ (with d denoting the mean duration of expression). When the expression switches to a new gene, immunity to the gene previously expressed is added to the host's immune memory. The infection ends when the expression of the whole *var* genome is completed. Mutation μ and ectopic recombination r occur randomly during the infectious stage (see below).

Genome structure. The genome of an individual parasite is a combination of g var genes. Each *var* gene, in turn, is a linear combination of two loci encoding epitopes that are connected linearly, and each epitope can be viewed as a multi-allele locus with n possible alleles. The initial conditions for the simulation include g*2 alleles per epitope and g*20 combinations of these genes in the gene pool. A typical simulation starts by initializing the local parasite population via a given number of transmission events with migrant genomes whose composition is sampled randomly from a regional pool of genes, G. Specifically, 20 hosts are infected with randomly assembled genomes, and one migrant genome is introduced every day into the population to simulate exposure to all the variants in the genome pool quickly.

Ectopic recombination during the asexual blood stage of infection. *var* genes often change their physical location and form new variants through ectopic recombination and gene conversions. These processes occur during both sexual and asexual stages. As ectopic recombination is observed more often during the asexual stage where parasites spend most of their life cycle, and our model does not represent the mosquito stage explicitly, we consider ectopic recombination among genes within the same genome during the asexual stage. Two genes are selected from the genome repertoire, with a breakpoint located along the gene randomly. Newly recombined genes have a probability P_f to be functional (i.e., viable), defined by the similarity of the new variant with its parental genes as in He et al. (10). If the recombined gene is selected to be non-functional, then the parental gene is kept. Otherwise, the recombined gene substitutes the parental one and a new strain is formed. In the current implementation, each recombination has a 50% chance to generate a new allele.

var gene groups and trade-offs. In the previous model, genes differed antigenically but not functionally. For increased realism, each gene is assigned to either *var* upsGroup A or *var* upsGroup B/C to represent existing differences in recombination rates and functionality of *var* gene groups (37). Ectopic recombination is only allowed to occur within each group, and genes in upsGroup B/C have higher recombination rates than those in upsGroup A (6). Each gene is also assigned an intrinsic growth rate of the parasites f that express it (Table S2), because antigens that better bind to host receptors result in a higher parasite growth rate (38, 39). In an additional implementation of the model, genes with higher growth rates are expressed first, followed in decreasing order by genes of lower growth rates. Also, genes with higher growth rates are expected to be cleared faster by the immune system, translating into a higher switch rate to the next gene, which is controlled by the trade-off parameter *tff*.

Data Availability. The original C++ code for the var evolution model is available on Github (https://github.com/pascualgroup/VarModel2).

ACKNOWLEDGMENTS. We are grateful to the funding provided by the joint NIH-NSF-NIFA Ecology and Evolution of Infectious Disease (award R01 AI149779). We acknowledge valuable discussions at the Santa Fe Institute, as part of a working group supported by the James S. McDonnell Foundation (JSMF). We thank Karen P. Day for her insightful comments on an earlier version of the manuscript. We appreciate the support of the University of Chicago through computational resources at the Midway cluster.

1. MJ Keeling, P Rohani, *Modeling Infectious Diseases in Humans and Animals.* (Princeton University Press), (2008).
2. S Gupta, K P. Day, A theoretical framework for the immunoepidemiology of Plasmodium falciparum malaria. *Parasite Immunol.* **16**, 361–370 (1994).
3. CO Buckee, M Recker, Evolution of the Multi-Domain Structures of Virulence Genes in the Human Malaria Parasite, Plasmodium falciparum. *PLOS Comput. Biol.* **8**, e1002451 (2012).
4. KW Deitsch, ER Moxon, TE Wellems, Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol. Mol. Biol. Rev.* **61**, 281–293 (1997).
5. M Georgieva, CO Buckee, M Lipsitch, Models of immune selection for multi-locus antigenic diversity of pathogens. *Nat. Rev. Immunol.* **19**, 55 (2019).
6. A Claessens, et al., Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLOS Genet.* **10**, e1004812 (2014).
7. M Frank, et al., Frequent recombination events generate diversity within the multi-copy variant antigen gene families of Plasmodium falciparum. *Int. J. for Parasitol.* **38**, 1099–1109 (2008).
8. KP Day, et al., Evidence of strain structure in plasmodium falciparum var gene repertoires in children from gabon, west africa. *Proc. Natl. Acad. Sci.* **114**, E4103–E4111 (2017).
9. S Ruybal-Pesántez, et al., Population genomics of virulence genes of Plasmodium falciparum in clinical isolates from Uganda. *Sci. Reports* **7**, 11810 (2017).
10. Q He, et al., Networks of genetic similarity reveal non-neutral processes shape strain structure in Plasmodium falciparum. *Nat. Commun.* **9**, 1817 (2018).
11. S Ruybal-Pesántez, et al., Why do adults harbor Plasmodium falciparum infections in high transmission African settings? *eLife* (2020).
12. KP Day, et al., Evidence of strain structure in Plasmodium falciparum var gene repertoires in children from Gabon, West Africa. *PNAS* **114**, E4103–E4111 (2017).
13. S Ruybal-Pesántez, et al., Population genomics of virulence genes of Plasmodium falciparum in clinical isolates from Uganda. *Sci. Rep.* **7**, 11810 (2017).

DRAFT

14. DS Chen, et al., A Molecular Epidemiological Study of var Gene Diversity to Characterize the Reservoir of Plasmodium falciparum in Humans in Africa. *PLOS ONE* **6**, e16629 (2011).
15. JW Barnwell, RJ Howard, HG Coon, LH Miller, Splenic requirement for antigenic variation and expression of the variant antigen on the erythrocyte membrane in cloned Plasmodium knowlesi malaria. *Infect. Immun.* **40**, 985–994 (1983).
16. SM Handunnetti, KN Mendis, PH David, Antigenic variation of cloned Plasmodium fragile in its natural host Macaca sinica. Sequential appearance of successive variant antigenic types. *The J. Exp. Medicine* **165**, 1269–1283 (1987).
17. A Stern, M Brown, P Nickel, TF Meyer, Opacity genes in Neisseria gonorrhoeae: Control of phase and antigenic variation. *Cell* **47**, 61–71 (1986).
18. KW Deitsch, SA Lukehart, JR Stringer, Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Micro* **7**, 493–503 (2009).
19. JM Levine, J Bascompte, PB Adler, S Allesina, Beyond pairwise mechanisms of species coexistence in complex communities. *Nature* **546**, 56–64 (2017).
20. J Hermisson, PS Pennings, Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics* **169**, 2335–2352 (2005).
21. S Gupta, et al., The maintenance of strain structure in populations of recombining infectious agents. *Nat Med* **2**, 437–442 (1996).
22. WJ Ewens, Mathematical population genetics, i. theoretical introduction. 2 (2004).
23. M Recker, et al., Antigenic Variation in Plasmodium falciparum Malaria Involves a Highly Structured Switching Pattern. *PLOS Pathog.* **7**, e1001306 (2011).
24. Y Artzy-Randrup, et al., Population structuring of multi-copy, antigen-encoding genes in Plasmodium falciparum. *Elife* **1**, e00093 (2012).
25. J Hemingway, et al., Tools and Strategies for Malaria Control and Elimination: What Do We Need to Achieve a Grand Convergence in Malaria? *PLOS Biol.* **14**, e1002380 (2016).
26. CO Buckee, et al., Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *PNAS* **105**, 15082–15087 (2008).
27. S Gupta, N Ferguson, R Anderson, Chaos, Persistence, and Evolution of Strain Structure in Antigenically Diverse Infectious Agents. *Science* **280**, 912–915 (1998).
28. M Recker, et al., Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature* **429**, 555 (2004).
29. JT Griffin, et al., Reducing Plasmodium falciparum Malaria Transmission in Africa: A Model-Based Evaluation of Intervention Strategies. *PLOS Medicine* **7**, e1000324 (2010).
30. TK Nkuo Akenji, et al., Evaluating a malaria intervention strategy using knowledge, practices and coverage surveys in rural Bolifamba, southwest Cameroon. *Transactions The Royal Soc. Trop. Medicine Hyg.* **99**, 325–332 (2005).
31. T Holding, JJ Valletta, M Recker, Multiscale Immune Selection and the Transmission-Diversity Feedback in Antigenically Diverse Pathogen Systems. *The Am. Nat.* **192**, E189–E201 (2018).
32. CO Buckee, M Recker, ER Watkins, S Gupta, Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. *PNAS* **108**, 15504–15509 (2011).
33. J Ozik, NT Collier, JM Wozniak, C Spagnuolo, From desktop to Large-Scale Model Exploration with Swift/T in *2016 Winter Simulation Conference (WSC)*. pp. 206–220 (2016) ISSN: 1558-4305.
34. S Pilosof, et al., Competition for hosts modulates vast antigenic diversity to generate persistent strain structure in Plasmodium falciparum. *PLOS Biol.* **17**, e3000336 (2019).
35. MA Gibson, J Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A* (year?).
36. DT Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
37. JA Rowe, A Claessens, RA Corrigan, M Arman, Adhesion of Plasmodium falciparum-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. *Expert. Rev. Mol. Medicine* **11** (2009).
38. A Bachmann, et al., Mosquito Passage Dramatically Changes var Gene Expression in Controlled Human Plasmodium falciparum Infections. *PLOS Pathog.* **12**, e1005538 (2016).
39. A Bachmann, et al., Controlled human malaria infection with Plasmodium falciparum demonstrates impact of naturally acquired immunity on virulence gene expression. *PLOS Pathog.* **15**, e1007906 (2019).

DRAFT