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

Qixin He[1], Mercedes Pascual[1,2*]

[1]Department of Ecology and Evolution, University of Chicago, USA
[2]Santa Fe Institute, USA

*To whom correspondence should be addressed; E-mail: pascualmm@uchicago.edu

In malaria and several other important infectious diseases, high prevalence occurs under widespread but incomplete immunity. The large reservoir of transmission constituted by asymptomatic *Plasmodium falciparum* infection of all ages challenges malaria elimination in highly endemic regions, and is enabled by extreme antigenic diversity of the parasite and turnover of new variants. We present 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. An analytical expectation for this threshold is derived and compared to numerical results from a stochastic agent-based model of malaria transmission that incorporates the major antigen-encoding multigene family known as *var*. The threshold can be exploited for malaria control and should apply to other pathogens with large multilocus antigenic diversity.

High transmission endemic areas present a significant challenge to the control and elimination of falciparum malaria due to the large reservoir of chronic asymptomatic infections sustaining transmission. The global burden of *Plasmodium falciparum* is concentrated today in these 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 (*1–3*). Nonsterile specific immunity is common to all these pathogens as a result of extreme antigenic variation encoded by multigene families (*4, 5*).

Under high transmission of *P. falciparum*, the major antigen-encoding gene family of the blood stage of infection, known as *var*, acquires vast antigenic diversity via large gene copy numbers as well as ectopic recombination, with laboratory experiments showing that a naïve infection can generate about sixty new recombinants per year (*6, 7*). Local parasite populations share only a few common *var* genes between different strains (*8, 9*), and across seasons (*10*). Spatial diversity in *var* genes has been shown in (*11, 12*), indicating that migration from surrounding areas will contribute to new diversity and to the immunological challenge. A major impediment to control is therefore the accumulation and turnover of new antigenic variants.

The well-known reproductive number of an infectious disease known as R_0 provides a threshold condition for the spread and control of pathogens with either no antigenic variation or a typically low number of defined strains. It remains unclear however how to apply and even define R_0 when antigenic variation is large and dynamic, challenging the definition of strain itself, such as in *P. falciparum* and other pathogens with multigene and multilocus encoding in addition to extensive recombination (*13–16*). Moreover, R_0 does neither predict the turnover rate of new genes nor explain how transmission characteristics and accumulation of new genes influence each other.

We investigate 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 influence this rate analytically. We show that this threshold for the accumulation of novelty, we call R_{div}, maps to one in transmission intensity, separating a regime in which new genes cannot accumulate from one in which they do. We discuss implications for malaria control and elimination, and for the monitoring of the effect of these efforts, in high transmission endemic regions.

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 (*17*). 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

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

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 (*9, 18*). 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 (*19*), 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,

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

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 (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 (*20*). 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,

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

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 (2) using (3) (Methods), to obtain

$$p_{inv} \approx \frac{S_{new} - \bar{S}}{\bar{S}g} \tag{4}$$

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 (*9, 21*) that tracks *var* evolution and immunity and is described in detail in (*9*). 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. (2) and (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).

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 (*9*). 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 (*19*) under the assumption that the system remains constant and only this lifespan, we call $\bar{t}$, varies (Methods). The diffusion equation for $\bar{t}$ requires consideration of how the frequency of a new gene $x(t)$ varies in time (Methods). 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.

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 re-

build. 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 (*22*). 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 (*18, 23–25*), mathematical and computational models typically evaluate intervention efficacy without explicit consideration of antigenic diversity (e.g. (*26, 27*)) and openness of the system to innovation (*28*). Our results underscore the importance of these aspects.

Estimation of the diversification number, R_{div}, would provide general guidelines for intervention evaluation where traditional application of R_0 is unable to do so for highly diverse pathogens (*13, 18*). 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., (*9, 29*)) could be used, which represents a computational challenge (but see (*30*)). Esti-

mating 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 (*9, 11, 21, 31*). 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* (*4*). 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 (*16*). 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.

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.

References

1. J. W. Barnwell, R. J. Howard, H. G. Coon, L. H. Miller, *Infection and Immunity* **40**, 985 (1983).

2. S. M. Handunnetti, K. N. Mendis, P. H. David, *The Journal of Experimental Medicine* **165**, 1269 (1987).

3. A. Stern, M. Brown, P. Nickel, T. F. Meyer, *Cell* **47**, 61 (1986).

4. K. W. Deitsch, S. A. Lukehart, J. R. Stringer, *Nat Rev Micro* **7**, 493 (2009).

5. J. M. Levine, J. Bascompte, P. B. Adler, S. Allesina, *Nature* **546**, 56 (2017).

6. A. Claessens, *et al.*, *PLOS Genetics* **10**, e1004812 (2014).

7. M. Frank, *et al.*, *International Journal for Parasitology* **38**, 1099 (2008).

8. S. Ruybal-Pesántez, *et al.*, *Scientific Reports* **7**, 11810 (2017).

9. Q. He, *et al.*, *Nature Communications* **9**, 1817 (2018).

10. S. Ruybal-Pesántez, *et al.*, *eLife* (2020).

11. K. P. Day, *et al.*, *PNAS* **114**, E4103 (2017).

12. D. S. Chen, *et al.*, *PLOS ONE* **6**, e16629 (2011).

13. S. Gupta, K. P. Day, *Parasite Immunology* **16**, 361 (1994).

14. C. O. Buckee, M. Recker, *PLOS Computational Biology* **8**, e1002451 (2012).

15. K. W. Deitsch, E. R. Moxon, T. E. Wellems, *Microbiol. Mol. Biol. Rev.* **61**, 281 (1997).

16. M. Georgieva, C. O. Buckee, M. Lipsitch, *Nature Reviews Immunology* **19**, 55 (2019).

17. J. Hermisson, P. S. Pennings, *Genetics* **169**, 2335 (2005).

18. S. Gupta, *et al.*, *Nat Med* **2**, 437 (1996).

19. W. J. Ewens, Mathematical population genetics, i. theoretical introduction. 2 (2004).

20. M. Recker, *et al.*, *PLOS Pathogens* **7**, e1001306 (2011).

21. Y. Artzy-Randrup, *et al.*, *Elife* **1**, e00093 (2012).

22. J. Hemingway, *et al.*, *PLOS Biology* **14**, e1002380 (2016).

23. C. O. Buckee, *et al.*, *PNAS* **105**, 15082 (2008).

24. S. Gupta, N. Ferguson, R. Anderson, *Science* **280**, 912 (1998).

25. M. Recker, *et al.*, *Nature* **429**, 555 (2004).

26. J. T. Griffin, *et al.*, *PLOS Medicine* **7**, e1000324 (2010).

27. T. K. Nkuo Akenji, *et al.*, *Transactions of The Royal Society of Tropical Medicine and Hygiene* **99**, 325 (2005).

28. T. Holding, J. J. Valletta, M. Recker, *The American Naturalist* **192**, E189 (2018).

29. C. O. Buckee, M. Recker, E. R. Watkins, S. Gupta, *PNAS* **108**, 15504 (2011).

30. J. Ozik, N. T. Collier, J. M. Wozniak, C. Spagnuolo, *2016 Winter Simulation Conference (WSC)* (2016), pp. 206–220. ISSN: 1558-4305.

31. S. Pilosof, *et al.*, *PLOS Biology* **17**, e3000336 (2019).

Figures and Figure Legends

Figs. 1-2

Supplementary materials

Materials and Methods

Figs. S1 to S6

Tables S1 to S2

References *(32-38)*

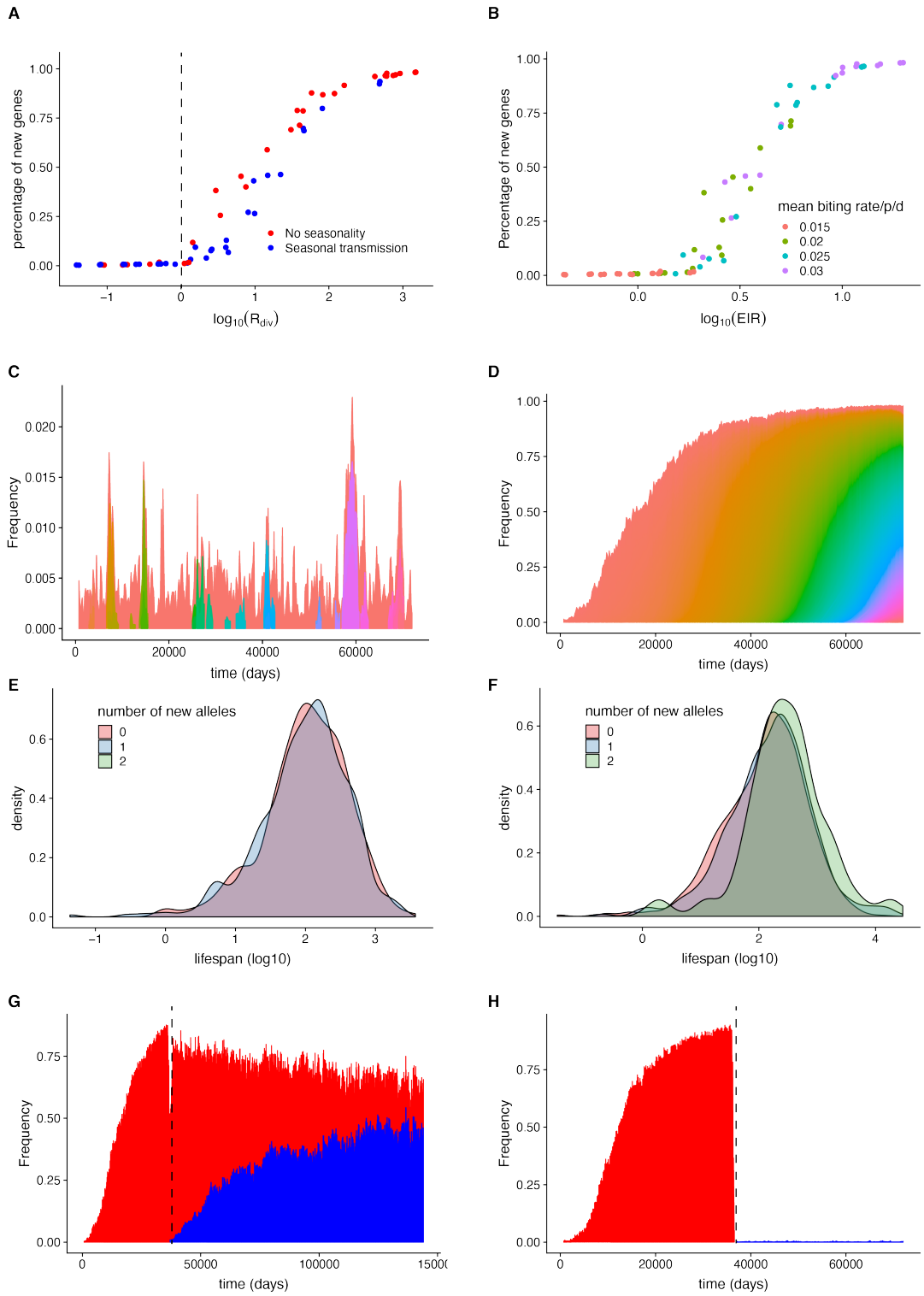
A
percentage of new genes
log10(Rdiv)
No seasonality
Seasonal transmission
B
Percentage of new genes
log10(EIR)
mean biting rate/p/d
0.015
0.02
0.025
0.03
C
Frequency
time (days)
D
Frequency
time (days)
E
density
lifespan (log10)
number of new alleles
0
1
2
F
density
lifespan (log10)
number of new alleles
0
1
2
G
Frequency
time (days)
H
Frequency
time (days)

Figure 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**) The percentage of new genes also varies with transmission intensity as measured here by the entomological inoculation rate (EIR, the number of infectious bites per man per year), exhibiting a threshold behavior. 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 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 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. (**G**) Interventions that push the system below the threshold are effective at stopping the accumulation of new genes, whereas those that do not, result in the rebound and rebuilding of diversity (**H**). Red and blue colors indicate genes that originate respectively before and after the intervention.

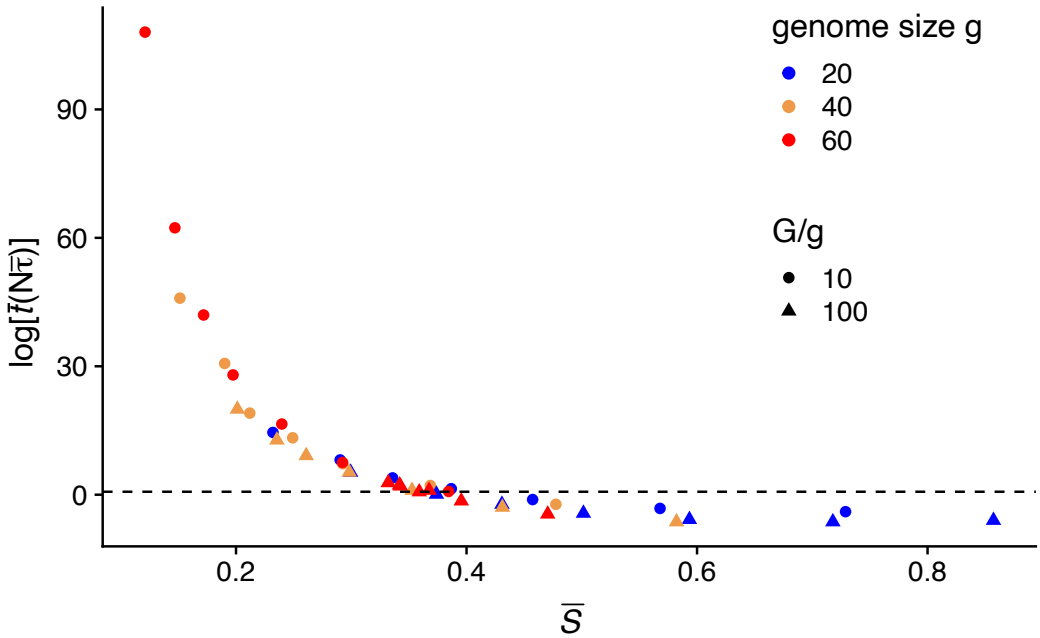

Figure 2: Theoretical expectation of the average lifespan of a new gene $\bar{t}$. The analytical expression shows that $\bar{t}$, measured in units of $N\bar{\tau}$, increases faster than exponential as the average number of available hosts $\bar{S}$ decreases. The dashed line represents the time to 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.