Title: Networks of genetic similarity reveal non-neutral processes shape strain structure in *Plasmodium falciparum*

Qixin He, Shai Pilosof, Kathryn E. Tiedje, Shazia Ruybal-Pesántez, Yael Artzy-Randrup, Edward B. Baskerville, Karen P. Day & Mercedes Pascual

Author information:

Affiliations
Department of Ecology and Evolution, University of Chicago, Chicago IL, USA 60637
Qixin He, Shai Pilosof, Edward B. Baskerville, Mercedes Pascual

School of BioSciences, Bio21 Institute/University of Melbourne, Melbourne, Australia
Kathryn E. Tiedje, Shazia Ruybal-Pesántez, Karen P. Day

IBED, University of Amsterdam
Yael Artzy-Randrup

Contributions
Q.H., S.P. and M.P. conceived and designed the study. Y.A.R. and Q.H. worked on the model. E.B.B. and Q.H. wrote the simulation code. K.E.T. and K.P.D. designed and led the data collection in Ghana. K.T. and S.R.P. performed the molecular experiments and genetic sequencing. Q.H. and S.P. analyzed the data. Q.H., S.P. and M.P. wrote the paper. All authors contributed to the final version of the manuscript.

Competing financial interests
The authors declare no competing financial interests.

Corresponding author
Correspondence to Mercedes Pascual: pascualmm@uchicago.edu.

Document order:
Main text, figure and figure legends (3), online-only methods, References, Acknowledgements, Extended Data tables (2), Extended Data figure and figure legends (7), and Supplementary Information

Summary Paragraph

Pathogens compete for hosts through patterns of cross-protection conferred by immune responses to antigens[1]. To evade the immune system, several pathogens possess hypervariable multi-copy gene families encoding large pools of antigenic variants[2]. In the malaria parasite *Plasmodium falciparum, var* genes encoding the major blood-stage antigen, *Pf*EMP1, are one such family, with repertoires of 50-60 genes in an individual parasite[3], and tens of thousands of gene variants in local populations of high transmission regions[4] generated through ectopic recombination[5,6] and mutation. Deep sampling of asymptomatic children in a West African population has recently revealed non-random structure in this enormous diversity, with extremely low genetic overlap among *var* gene repertoires even in multi-genome *P. falciparum* isolates[7]. This is consistent with previous strain theory, postulating co-existence of discrete non-overlapping pathogen strains[8–10] as a result of selection against recombinants due to cross-immunity. However, the combinatorial complexity of the *var* system in high transmission regions remains beyond the reach of existing strain theory, and neutral models do not yet exist to differentiate signatures of immune selection from those of pure transmission dynamics. Here, we present theory to identify signatures of immune selection that reveal non-neutral structures both in simulated systems and in an extensively sampled population in Bongo District (BD), Ghana. We develop two neutral models that encompass malaria epidemiology but exclude competitive interactions between parasites. We then present an analytical framework based on genetic similarity networks appropriate for the recombination mechanisms that generate diversity. Network patterns harbor distinctive signatures of selection structuring antigenic diversity in this highly recombinant gene family through frequency-dependent competition for hosts. This unique population structure created by non-neutral forces, likely immune selection, underlies the ability of the parasite to multiply infect individual hosts with long lasting chronic infections[11], constituting a large reservoir of transmission in highly endemic regions of Africa. To be successful, elimination strategies must move beyond prevalence of infection and target this diversity which is at the heart of malaria transmission and pathology.

A central question in ecology and evolution regards the extent to which non-neutral processes structure diversity[12–15]. It remains a challenge to identify signature patterns that reveal an important role of ecological interactions in facilitating and stabilizing species coexistence in ecosystems with high diversity, such as tropical rain forests[16,17]. Here we address whether competitive interactions act as a non-neutral stabilizing force that promotes coexistence in another highly diverse system: *Plasmodium falciparum* populations as an ensemble of diverse strains in regions of high malaria transmission.

Recurrent malaria infections in endemic regions do not generate sterilizing immunity towards subsequent infection[18]; this suggests the existence of a large number of strains of the pathogen. A vast reservoir of *Plasmodium falciparum* exists in local human populations in Africa in the form of asymptomatic infections, hosts that carry the parasite without manifestation of the disease[19]. An understanding of the antigenic diversity of the parasite in such reservoirs, including whether and how this diversity is structured into strains, is fundamental to understanding immunity patterns and developing intervention strategies in the transmission dynamics of *falciparum* malaria.

The high transmission rates of endemic regions suggest frequency-dependent competition among parasites for hosts, through the cross-protection conferred by the adaptive immune

system[20,21]. Since the success of an infection depends on the immunological memory of a host, new and rare antigenic types have a fitness advantage in the transmission system relative to common ones. "Immune selection", a form of balancing selection, is recognized as an important force promoting the diversification and persistence of the *var* gene family, whose ancient origin predates the speciation of *P. falciparum*[22,23]. However, the role of immune selection is much less recognized and understood at a higher organizational level, in shaping both the repertoires of *var* genes that constitute a parasite and the population structure of such coexisting strains[7].

In high transmission regions, the extensive diversity of the *var* gene family[24] exhibits low amino acid similarity encoded by different *var* genes and a very low percentage of genes shared between parasites, both locally and regionally (e.g., less than 0.3% in Africa[4,7]). Previous work, known as strain theory, has posited that the non-random association of gene variants results from selection against recombinants through cross-immunity[9], akin to emergent niches of limiting similarity[25] or selection towards divergent local adaptations[26].

Existing models for strain theory[10,27] incorporate limited *var* gene diversity compared to observed numbers for *P. falciparum*. It is unclear whether structure can emerge at such vast diversity, especially under high recombination rates. We developed an individual-based stochastic model with realistic mutation and recombination processes that generates levels of diversity comparable to those of the *var* gene family (Extended Data Fig. 1; Methods). In the model, mitotic recombination and mutation generate new *var* genes, making the overall pool of alleles effectively open to innovation, whereas meiotic recombination shuffles the composition of *var* genes of two or more repertoires in co-infections during the vector stage of transmission (Extended Data Fig. 1b). The overall system is composed of a pool of gene variants and a local population open to immigration, in which we track transmission (Extended Data Fig. 1c).

Existing strain theory also lacks a neutral counterpart, a null hypothesis to disentangle patterns generated by the acquisition of specific immunity from those resulting simply from the basic demography of the system, specifically transmission dynamics and generalized immunity. We explicitly developed two such null models and compared the repertoire structures they generate with those from the model with specific immunity. The first model assumes generalized immunity in which protection is acquired as a function of the number of previous exposures irrespective of their specific antigenic identity; the second one is a completely neutral model in which infections propagate and recover, but hosts are blind to any history of exposure so that repertoires do not compete for hosts. Thus, the two null models include all the epidemiological processes, except for specific immunity towards the *var* genes that a host has been exposed to, and the resulting cross-protection. The epidemiological phenotype under immune selection is the duration of infection, this directly influences the basic reproduction number (or fitness) of the parasite, R_0, we therefore match the infection periods in the two null models to that of the corresponding immune selection model (Methods).

The reticulate evolutionary pattern of *var* genes, generated by frequent mitotic and meiotic recombination within and between parasite genomes[5,6], respectively, precludes the application of traditional population genetics tests for balancing selection (e.g., Tajima's *D*). Hence, we develop an application of network theory to study the evolution of *var* repertoire structures, and show that the structure of genetic similarity networks contains clear signatures of neutral versus non-neutral processes. We analyzed the genetic structure of the parasite

population using networks in which nodes are *var* repertoires, weighted edges encode the degree of overlap between the alleles of these repertoires, and the direction of an edge indicates whether one node can out compete the other (Methods). Comparisons of structure across the similarity networks generated under the three models reveal distinctive features of immune selection, although the specific features that distinguish immune selection vary under different epidemiological settings (described below).

Because *var* genes exhibit different diversity levels across different endemic regions (Chen et al. 2011), we investigated the influence of *var* gene pool size (i.e., the number of *var* genes in a given population) on the immune selection signatures (Extended Data Fig. 1d). Because the two most relevant epidemiological parameters, transmission intensity and duration of a naïve infection, determine the intensity of competition among *var* repertoires, we vary them systematically to address their influence on signatures of immune selection. Higher transmission and longer duration of a naïve infection intensify competition among repertoires, they also increase the rate of recombination in the mosquito vector between different repertoires (Extended Data Fig. 1d). It follows that signature patterns of immune selection should be most evident with increasing values of duration, diversity and transmission, for conditions representative of high endemicity.

To explore selection signatures in networks generated under regimes of strong competition between strains, we use a suite of network metrics (see complete list in Extended Data Table 1c, and Extended Data Fig. 2 for the low competition scenarios). If a process akin to limiting similarity underlies population structure, networks are expected to be partitioned into disconnected clusters of highly similar repertoires that occupy separate niches in antigenic/genetic space. One way to quantify the partitioning of a network is by calculating maximum modularity (Q)[28]. When the local *var* gene pool is of a medium size characteristic of endemic regions of Asia/Pacific (~1200-2400 different *var* genes)[29], the selection case differs notably from those of the two null models: repertoires are typically grouped into well-defined clusters (exemplified by high Q and module F_{ST} values, Fig. 1a, b), whereas in networks resulting from the two null models, nodes are typically connected to form star-shaped or tree-like structures. This qualitative difference in structure resembles the prediction of non-overlapping strains in classic strain theory, where the disconnected clusters are analogous to niches in immune space consisting of highly similar repertoires.

In addition, because competition at the repertoire level in the selection case promotes equal competitiveness of two given connected repertoires, it results in reciprocally-connected directed edges of similar weights. In contrast, in the two null models, repertoires with lower number of unique genes are not removed by selection, and when one repertoire outcompetes another, there is only one directed edge between the pair. We use 3-node motifs to capture this variation in competitiveness. For example, a binary in-tree motif (A->B<-C) reflects that repertoire B is outcompeted by A and C, whereas a complete graph motif in which three repertoires are all reciprocally connected (A<->B<->C<->A) indicates a balanced, reciprocal competition. We find that networks of the selection model have a high proportion of such reciprocal motifs compared to those of the two null models. Binary in-tree or out-tree motifs are instead the most common in the null models, reflecting (parent-offspring) evolutionary relationship (Fig. 1c, f).

Under a regime with a larger initial gene pool that matches the diversity levels of endemic regions in Africa (~12,000-24,000 different *var* genes)[7], repertoires have a lower genetic overlap compared to medium diversity (see Extended Data Fig. 3). This pattern follows

naturally from increased diversity because repertoires can be formed from a larger number of gene combinations. Although such low overlap can indicate a non-random structure[7], it cannot per se distinguish selection from the two null models. Accordingly, module F_{ST} is low in all three cases and is not a good indicator of selection (Fig. 1e), despite the selection model possessing more separate components than null models (Fig. 1d). Nonetheless, selection networks can still be differentiated from those generated under null models in terms of motif composition (Fig. 1f), as well as other features (see Fig. 2 and Extended Data Fig. 3). In particular, the weaker similarity between repertoires means a less clear network partitioning than that of lower diversity systems. We therefore use betweenness centrality as a property reflecting their limiting similarity: this metric measures the importance of a repertoire in a network by calculating the proportion of shortest paths connecting any pair of nodes that go through it[30]. For the networks generated with the null models, betweenness centrality varies little among repertoires, with no highly central ones (Fig. 1d). This is because the persistence of each repertoire is independent of the antigenic composition of other repertoires given the lack of specific competition. By contrast, in the selection case some repertoires are clearly more central than others (Fig. 1d), reflecting the non-random persistence of antigenic niches, connected through these networks' hubs via a series of recombination events.

To apply these findings to empirical data, we first asked whether networks produced by the models can be classified using a set of networks properties into the processes that generate them, immune selection vs. generalized immunity or complete neutrality (Extended Data Table 1). With a medium diversity gene pool, there is a positive correlation between the strength of competition and our ability to classify networks correctly, reflecting an increasing divergence between the networks (Fig. 2a-d). With a high diversity gene pool, the classification always differentiates the selection scenario correctly (Fig. 2e), whereas it often fails to differentiate generalized immunity from complete neutrality (Fig. 2f-h).

In addition to the effects of immune selection at the repertoire level, frequency-dependent competition also works at the level of *var* genes in the population. Specifically, frequency-dependent competition will limit the abundance of genes that are similar to many others — and are thus readily recognized by the immune system — while favoring the abundance of genes with a unique composition of alleles. We can test this prediction using a network in which nodes are genes and edges encode similarity in allelic composition (Extended Data Fig. 4a). In the immune selection case, we find a characteristic negative correlation between node degree (number of genes similar to a focal gene) and the frequency of genes in the host population. This effect is absent for the null models (Extended Data Fig. 4b-d).

Deep genetic sampling of local populations in BD, Ghana allows application of these theoretical findings to examine the role of immune selection in nature. Gene similarity networks were built from *var* DBLα domain tags sequenced from 1,248 *P. falciparum* isolates in Ghana (Methods). We restrict our analyses to the upsB/upsC group of the DBLα domain because this subset is known to exhibit frequent ectopic recombination within itself relative to the more conserved upsA[26]. This group rather than the whole set of genes is therefore less prone to generate the above negative correlation spuriously out of differences in recombination rates, and provides a more appropriate counterpart to our theory, which does not consider functional differences between *var* gene variants. The resulting gene similarity networks exhibit a strong negative correlation between *var* DBLα type frequency and number of similar neighbors, providing evidence for frequency-dependent competition (Fig. 3a). We then examine the strain similarity network by calculating shared DBLα types between different repertoires in the subsample of isolates (=161) whose multiplicity of

infection is equal to one (i.e., whose *var* genes most likely consist of a single infection; Methods; Extended Data Fig. 5). We applied our network classification method to ask whether immune selection played an important role in shaping the empirical population structure. We generated a library of simulated networks under parameter ranges representative of Ghana (Methods), which resulted in corresponding values of the annual entomological inoculation rate (number of infective bites per person per year, EIR) [~25-170] that encompass empirical observations for the region[19]. Classification of the empirical network using discriminant functions indicates its resemblance to networks generated with the immune selection scenario (Fig. 3b).

Interestingly, the network signatures we have identified present conceptual similarity to traditional tests of balancing selection developed in population genetics or community ecology (as summarized in Extended Data Table 2), thus filling the gap of available tests for highly recombinant gene families. It follows that these network properties can be adapted for application to other multicopy gene repertoires for antigenic variation, such as *vsg* genes in *Trypanosoma brucei* or *msg* genes in *Pneumocystis carinii*[1].

These findings provide unequivocal evidence for frequency-dependent competition structuring antigen composition in a natural population of *P. falciparum*: these patterns differ from those expected from the simple demography of transmission. We show that under extreme diversity and intense recombination, immune selection profoundly shapes repertoire diversity into a distinct population structure that can be detected using network metrics. Open areas include consideration of functional differences among *var* genes, phenotypic mapping of sequence diversity to immunity, how parasite population structure changes in time and how it influences responses to interventions.

Early motivation for strain theory was the recognition that organization of *Pf*EMP1 variants (and their underlying genes) into persistent largely non-overlapping sets can deeply alter our understanding of epidemiology and control, for example by viewing *P. falciparum*'s apparent large reproductive number (R_0) as resulting from a large ensemble of strains with much lower reproductive numbers[8]. With the sheer number of existing and ever-changing variants, previous definitions of strains as long-lived entities do not apply at high endemicity. The resulting population structure nevertheless exhibits limited similarity, in the form of sparse small clusters and/or isolated individual repertoires interspersed into voids in antigenic/genetic space, instead of well-defined niches. This emergent structure provides an image of competition at the 'limit' of limiting similarity because of immense diversity. The resulting coexistence and diversity at the different hierarchical levels of genes and repertoires enables the large reservoir of asymptomatic infections that makes malaria so resilient to elimination in high transmission regions. Control strategies that target this diversity are needed, and may have positive feedback mechanisms that enhance intervention efforts by facilitating recognition by the immune system of strains and antigenic variants that would otherwise escape detection.

Figure and Figure Legends

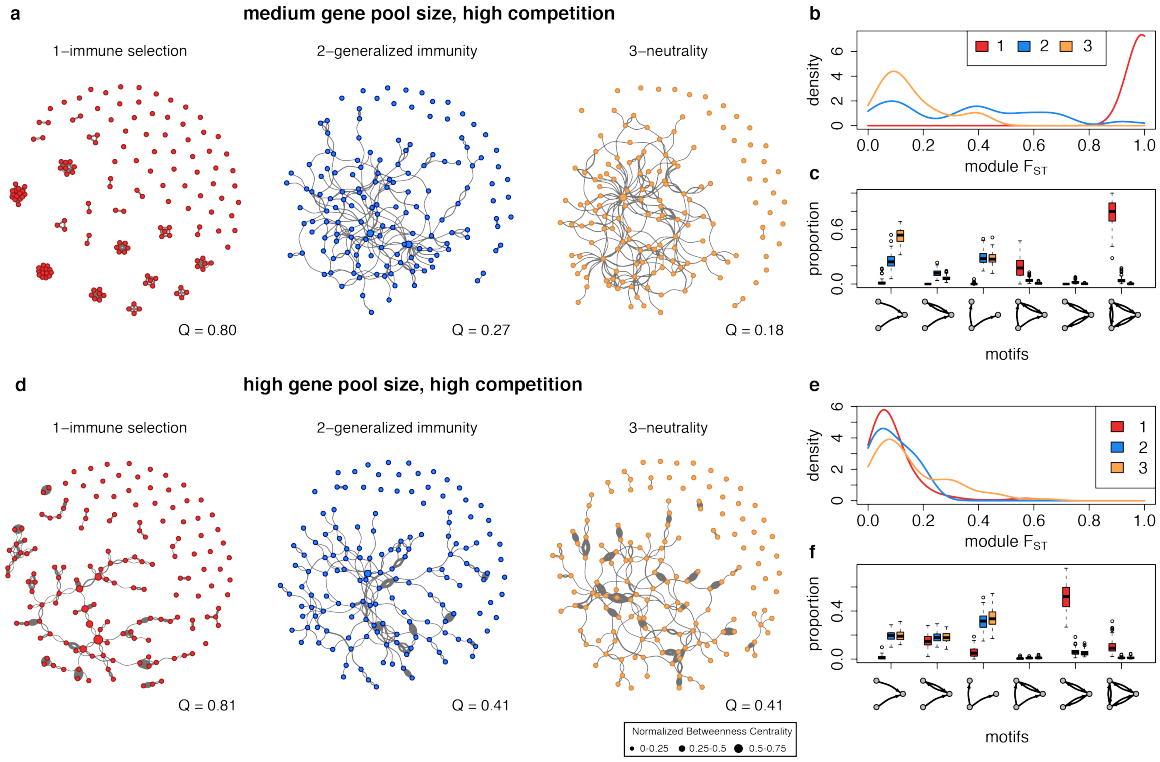


Figure 1. **Repertoire similarity networks and representative network metrics across scenarios under different diversity regimes generated from model simulations with high competition (high duration of naive infection and high transmission rates).** Upper panel, medium diversity (*gene pool size of* 1,200) and lower panel, high diversity (*gene pool size of* 24,000). **a, d**, comparisons of strain similarity networks of 150 randomly sampled parasite *var* repertoires from one time point under three scenarios. Edge width is relative to the strength of genetic similarity between pairs of repertoires. Only the top 1% of edges are drawn and used in the analysis (see Extended Data Fig. 3 for distribution of edge weights). Within the largest component of each network, the size of each repertoire indicates its normalized betweenness centrality. The value of maximum modularity Q is calculated using edge-betweenness[28] and shown at the lower right corner of each network. The modules obtained in these networks represent groups of highly similar repertoires (strain modules), which are conceptually similar to geographically isolated populations with limited gene flow. We therefore calculate the pairwise F_{ST} of strain modules identified by the Girvan–Newman algorithm[28], to quantify how different strain modules are from each other, providing a measure of limiting similarity that compares within-module and between-module diversity[31]. **b, e**, pairwise module F_{ST} distributions. **c, f**, proportion of occurrence of 3-node graph motifs for the three models.

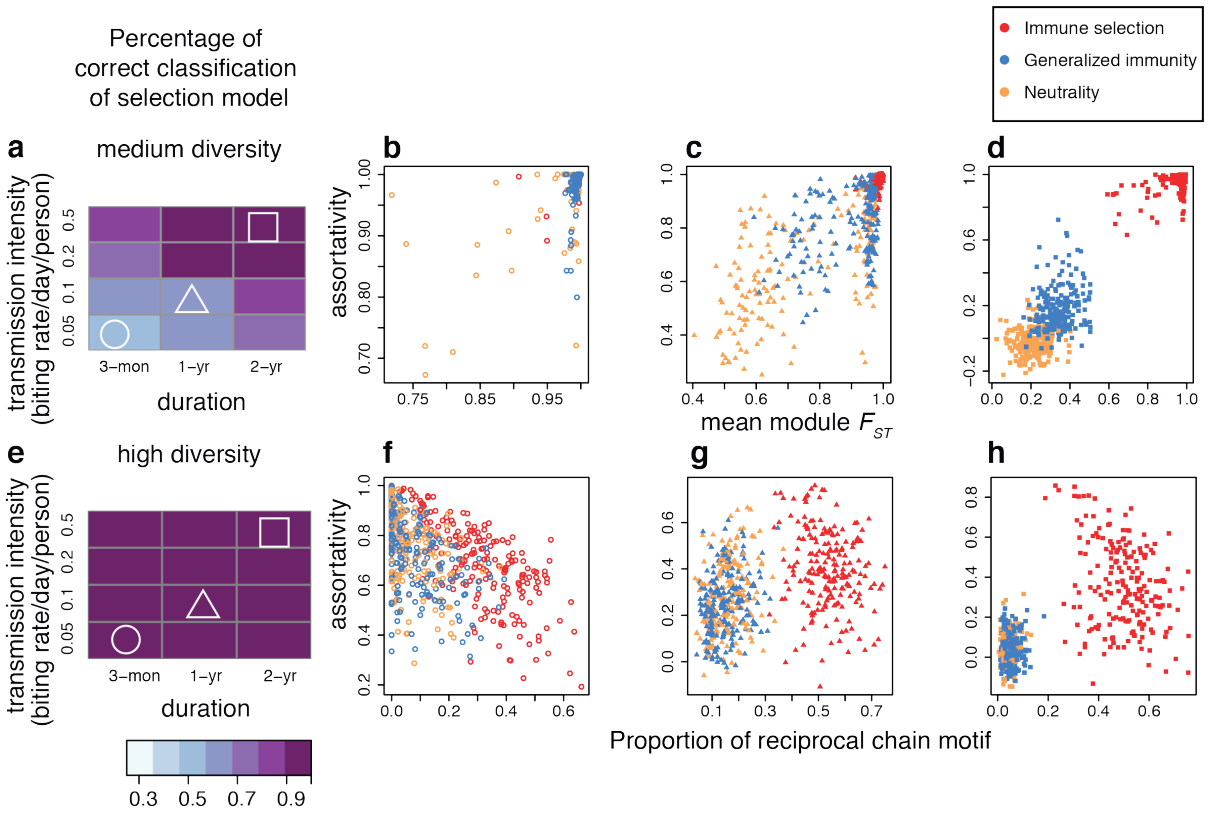


Figure 2. Divergence in network metrics between the three scenarios and the power of correct classification of the selection model for different levels of competition (as a function of infection duration and biting rate) and *var* gene pool size (medium diversity [1,200-2,400], **a-d**, and high diversity [12,000-24,000], **e-h**). **a**, **e**, The shade of colored squares indicates the proportion of correct assignments for the immune selection scenario. The relationship between selected network properties are shown for low (**b**, **f**), medium (**c**, **g**) and high competition (**d**, **h**) (with the corresponding point shapes indicated in **a**, **e**). In simulations with a gene pool of medium diversity (**a**), the proportion of correct assignments of the selection model increases with infection duration and biting rate. The divergence between the three models increases with increasing level of competition (**b-d**). When the genetic pool is of high diversity, the selection model is almost always perfectly assigned (**e**), while neutral and generalized immunity models are harder to differentiate, even under high competition levels, as shown by the relationship between assortativity and the proportion of reciprocal chain motifs (A<->B<->C) (**f-h**) (see Extended Data Table for definitions of these network metrics). Therefore, high diversity per se provides enough variation for selection to operate and leave a signature, even when transmission and duration of infection are low.

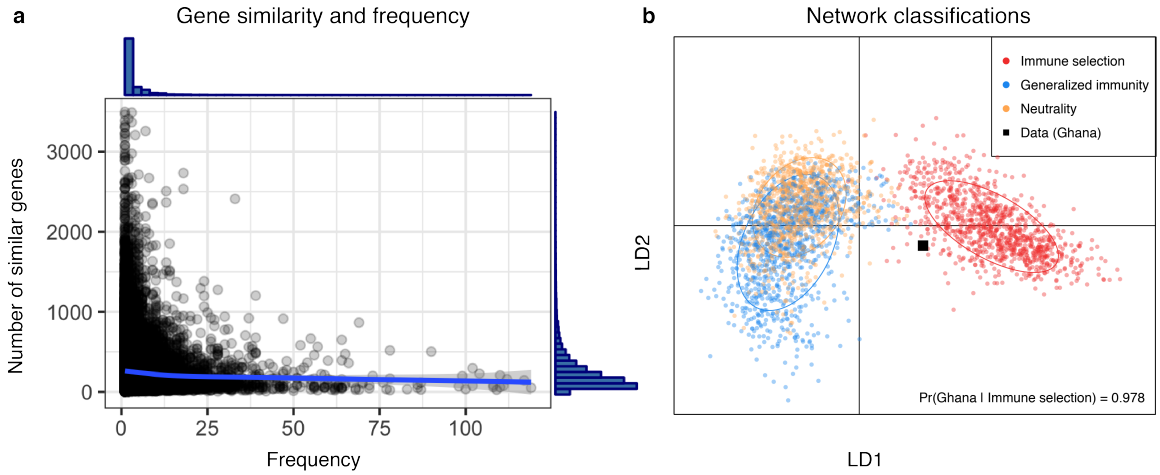


Figure 3. Empirical investigations of Ghana data. (**a**) Negative correlation between DBLα type frequencies and their number of similar genes for the upsB/upsC *var* genes in the parasite population (r = -0.040, p<2.2e-16) (This number is calculated as the degree [k] of the focal gene in the gene similarity network, for amino acid similarities above 0.6). Histograms on the top and right of the plot show the distributions of k and DBLa-type frequencies. (**b**) Discriminant Analyses of Principle Components[32] show the classification of the training networks onto the 2-D space formed by two linear discriminant (LD) functions, and classify the empirical similarity network as more likely to be generated under an immune selection regime (posterior probability [PP] = 0.978), as opposed to neutrality (PP = 0.008) or generalized immunity (PP = 0.014). The classification relies on comparisons of 39 network properties (see Extended Data Table 1) calculated for the 1000 simulated networks (from 100 combinations of different parameter settings and 10 random networks sampled at different times per simulation run).

Methods

The extended *var* evolution model

In the model, each parasite genome is a combination of 60 copies of the *var* genes, and each gene is in turn a linear combination of loci encoding epitopes (we used two in the current implementation). Immune selection in the model derives from specific immunity to epitope variants (alleles), which represent components of the *Pf*EMP1 molecule that are recognized and remembered by the immune system of the host[33]; In effect, these epitopes serve as traits that mediate competition for hosts at the population level, since individuals gain protection against specific alleles expressed by the parasite during an infection. (Extended Data Fig. 1a).

Model parameters and symbols are summarized in Extended Data Fig. 6a. Diversity of *var* genes is represented at three levels: alleles (epitopes), genes, and strains. Specifically, each parasite genome consists of *g var* genes. The specific combination of the *var* genes is referred to as a strain or *var* repertoire throughout the paper. Each *var* gene is composed of *l* epitopes that are connected linearly and each epitope can be viewed as a multi-allele locus with *n* possible alleles.

Initiation of the simulation

To initiate the *var* gene pool *G*, a random allele for each epitope is chosen from the allele pool to form each gene. In the simplest case, if there are two epitopes in a *var*, then a particular *var* $g_i = \{L_{i1}, L_{i2}\}$, where L_{i1}, L_{i2} are random numbers from $U(1,n)$. With n_i possible alleles at each epitope, the total number of possible genes is $\prod n_i$. However, we choose *G* at least five times less than $\prod n_i$ so that not all combinations of alleles as a gene are available. This relates to the fact that not all combinations of alleles form viable proteins. In the beginning of the simulation run, twenty hosts are selected and infected with a distinctive parasite genome that consists of a set of *g var* genes randomly selected from the pool *G*. The size of the host population, *H,* is kept at a constant size (i.e., when a host dies, a new host is born). Each host has a death rate of $d = 1/30$ per year.

Repertoire transmission

Vectors (mosquitoes) are not explicitly modeled. Instead, we set a biting rate *b* so that the average waiting time to the next biting event is equal to $1/(b*H)$. When a biting event occurs, two hosts are randomly selected, one donor and one recipient. If the donor is infected with malaria strains that have passed the liver stage, then the receiver will be infected with a probability of *p* (i.e., transmission probability). If the donor is infected with multiple strains in blood stage, then the transmission probability of each strain is *p* divided by the number of active strains.

Meiotic recombination

Meiotic recombination occurs between strains in the sexual stage of the parasite's life cycle. When multiple strains are transmitted to the donor, these strains have a $(1- P_r)$ probability to remain as the original strain, and a P_r probability to become a recombinant strain, with P_r calculated as follows,

$$P_r = 1 - 1/(N_s) \qquad (1)$$

, where Ns is the number of strains transmitted to the donor. Although the association of physical locations and major groups of *var* genes is established, orthologous gene pairs between two strains are often unknown. Therefore, we implement recombination between

strains as a process in which *g* genes are randomly selected out of all the original genes from the two strains pooled together. This is an oversimplification of the real process. However, because physical locations of *var* genes can be mobile, this assumption is a reasonable simplification of the meiotic recombination processes.

Ectopic recombination within the strain in the asexual blood stage

Var genes often change their physical locations through ectopic recombination and gene conversions. These processes occur at both sexual and asexual stages. However, ectopic recombination is observed more often in the asexual stage where the parasites spend most of their life cycle[6]. Therefore, we only model ectopic recombination among genes within the same genome during asexual stage. Two genes are selected from the repertoire. When the location of the breakpoint (i.e., between which loci recombination occurs) is decided, loci to the right of the breakpoint between the genes are either swapped (recombination) or copied (gene conversion). The ratio of these two outcomes is controlled by the parameter P_c. Newly recombined genes will have a probability P_f to be functional (i.e., viable) defined by the similarity of the parental genes.

$$P_f(x) = \tau^{\frac{x(\delta-x)}{\delta-1}} \quad (2)$$

(Eq.3 in Drummond et al.[34]), where x is the number of mutations between the recombined gene and one of the parental genes, δ is the difference between the two parental genes and τ is the recombinational tolerance. If the recombined gene is selected to be non-functional, then the parental gene will be kept. Otherwise, the recombined gene will substitute the parental gene so that a new strain is formed.

Mutation

Mutations occur at the level of epitopes. While infecting a host, each epitope has a rate of mutation, μ, to mutate to a new allele so that n increases by one. New mutations can die from lack of new transmissions, proliferate through new transmissions of the same strain, incorporate into other genes through ectopic recombination, or recombine into a different repertoire.

Within-host dynamics

Each strain is individually tracked through its entire life cycle, encompassing the liver stage, asexual blood stage, and the transmission and sexual stages. Because we do not explicitly model mosquitoes, we delay the expression of each strain in the receiver host by 14 days to account for the time required for the sexual and liver stage. Specifically, the infection of the host is delayed 7 days to account for the time required for gametocytes to develop into sporozoites in mosquitoes. When a host is infected, the parasite remains in the liver stage for additional 7 days[35] before being released as merozoites into the bloodstream, invading red blood cells and starting expressing the *var* repertoire. The expression of genes in the repertoire is sequential and the infection ends when the whole repertoire is depleted. During the expression of the repertoire, the host is considered infectious with the active strain. The expression order of the repertoire is randomized for each infection, while the deactivation rates of the genes are controlled by the host immunity. When one gene is actively expressed, host immunity 'checks' whether it has seen any epitopes in the infection history. The deactivation rate changes so that the duration of active period of a gene is proportional to the number of unseen epitopes. After the gene is deactivated, the host gains immunity to all the new epitopes. A new gene from the repertoire is immediately active, and the strain is cleared

from the host when the whole repertoire of *var* genes is depleted. The immunity towards a certain epitope wanes at a rate $w = 1/100$ per day[36].

Implementation of the simulation

The simulation is an individual-based, discrete-event, continuous-time stochastic model in which all known possible future events are stored on a single event queue along with their putative times, which may be at fixed times or drawn from a probability distribution. When an event is triggered, it may trigger the addition or removal of future events on the queue, or the modification of their rates, thus causing a recalculation of their putative time. This approach is adapted from the next-reaction method[37], which optimizes the Gillespie first-reaction method[38] by storing all events on an indexed binary heap. This data structure is simple to implement and sufficiently fast and compact to store all events in the system, down to individual state transitions for each infection course within each host. Specifically, modifying the putative time for an event on the queue is O(log N), and heap storage is O(N), where N is the number of events.

Statistical analyses

Selection versus neutral models

In order to disentangle signatures of immune selection from those of transmission per se in parasite population structures, we ran null models in which hosts do not build specific immunity towards alleles or genes, in addition to the selection model described above. In the complete neutrality model (Extended Data Fig. 6b), when hosts are infected, the duration of infection is determined by the deactivation rate of each gene, which is kept constant; thus, hosts do not build immunity after an infection. The rate of deactivation is calculated to match the average duration of infection of the corresponding selection model, while maintaining the rest of the parameters (e.g., G, b). In the generalized immunity model, the duration of infection decreases as the number of infections increases, similarly to the selection model. However, the identity of the alleles does not play a role and we therefore matched the average curve of duration of infection vs. number of infections to that of the corresponding selection scenario.

Diversity metrics, as well as epidemiological parameters, are calculated after each run to compare between scenarios. Diversity is quantified using common measures from ecology, including Shannon diversity[39], Simpson's diversity and evenness[40], beta diversity (i.e., turnover in composition of *var* genes or repertoires among parasite samples in time), as well as within-repertoire diversity at the allelic and genetic levels. Within-repertoire diversity is calculated by the number of unique alleles/genes divided by the potential maximum number of unique alleles/genes (e.g., 60 if the genome size is 60). Entomological inoculation rate (EIR), prevalence, and multiplicity of infection (MOI) are also compared among model runs under different parametric settings and scenarios.

Building of similarity networks

In addition to diversity, similarity networks based on allelic composition at the gene or repertoire levels are built to investigate parasite population structure. For this purpose, 150 parasites are sampled at 120-day intervals in the hosts, to subsample the simulations in a way that is meaningful for later empirical application of network analyses. Directional similarity networks for *var* genes or parasite genomes (i.e., *var* repertoires) are built with edges represented by the proportion of shared unique alleles. Specifically,

$$S_{ij} = \frac{a}{N_i}\,;\, S_{ji} = \frac{a}{N_j}\,,$$

where a is the shared number of unique alleles between i and j repertoires, and N_i and N_j are total number of unique alleles of repertoires i and j respectively. This directional index of genetic similarity is designed in the system to represent the relative competition between two repertoires, as explained in Extended Data Fig. 6c.

Calculation of Network properties

For repertoire similarity networks, 39 network properties are calculated to detect selection signatures, as well as to distinguish these from patterns generated by pure transmission dynamics or generalized immunity. These properties include diagnostics of transitivity, degree distributions, component sizes, diameters, reciprocity, and proportion of 3-node graph motifs (see Extended Data Table 1 for a complete list of parameters and definitions). One additional metric is introduced and named "module F_{ST}". This metric quantifies to what extent the strain modules inferred from repertoire similarity networks are genetically different from each other, by comparing the genetic diversity within and between communities[31,41].

Simulations and machine learning algorithms for classification

For each combination of parameters (i.e., initial gene pool size G, biting rate b, and duration of infection D), 100 simulations were run to calculate the distribution of the network properties under immune selection, neutral and generalized immunity scenarios. The properties are then transformed into non-correlated principle components. Discriminant analyses[32] were performed on the principle components, to design functions that maximize the differences among networks generated under different scenarios while minimizing the within-scenario variance. The accuracy of the discriminant functions is assessed by the proportion of correct classifications (i.e., power) as well as false positive rates. A similar approach is conducted for building a classifier for empirical networks. Details are documented below.

Comparisons with empirical data

Data sampling

The empirical data used was collected from a study performed across two catchment areas in Bongo District (BD), Ghana located in the Upper East Region near the Burkina Faso border. Malaria in BD is hyperendemic and is characterized by marked seasonal transmission of *P. falciparum* during the wet season between June and October. This age-stratified serial cross-sectional study was conducted over two sequential seasons. The first survey was completed at the end of the wet season in October 2012, followed by a second survey at the end of the dry season between mid-May and June 2013. Details on the study population, data collection procedures and epidemiology have been published elsewhere[19]. Briefly, after obtaining informed consent, finger prick blood samples were collected for parasitological assessment for *P. falciparum* by blood smears, and dried blood spots for molecular genotyping[19]. The study was reviewed and approved by the ethics committees at the Navrongo Health Research Center, Ghana; Noguchi Memorial Institute for Medical Research, Ghana; New York University, United States; University of Melbourne, Australia; University of Michigan, United States; and the University of Chicago, United States.

PCR amplification and var DBLα sequence analysis

The DBLα domain of *P. falciparum var* genes were amplified from genomic DNA using universal degenerate primers, as previously described[42]. Amplicons were pooled and barcoded libraries were sequenced on an Illumina MiSeq sequencer using the 2x300 paired-end cycle protocol, MiSeq Reagent kit v3 chemistry (NYUGTC, New York USA; AGRF, Melbourne Australia). A custom pipeline was developed to de-multiplex and remove PCR and sequencing artefacts from the DBLα sequence tags. Reads were demultiplexed into individual fastq files for each isolate using flexbar v2.5[43] and paired based on valid combinations of MID tags in the forward and reverse reads. A minimum read length of 100nt and a maximum uncalled bases threshold of 15 were used. The resulting paired fastq files were then merged using PEAR v.0.9.10[44] to ensure the resulting merged fastq files had appropriate base quality scores allowing for filtering of low quality reads. The minimum assembly length was set to 100nt and the minimum overlap required between a read pair was set to 20nt. Low quality reads were filtered if they had more than one expected error using the fastq_filter option of Usearch v8.1.1832[45,46]. Next, chimeras were filtered using Uchime denovo[47] and then the filtered reads were clustered using the cluster_fast function of Usearch[45] after the removal of singletons to reduce the impact of errors. A threshold of 96% identity[7] was used to cluster the reads. To increase the overall quality of the sequences, the resulting clusters were removed if they contained less than 15 reads to remove low support reads. The representative read from each cluster was kept for the remaining stages of the pipeline. Next, any non-DBLα sequences were filtered out using Hmmer[48] with a domain score threshold of 80. Finally, as a quality check the remaining reads were aligned to the reference *var* DBLα sequences of the 3D7, Dd2 and HB3 laboratory clones from experimental sequence data. To subsequently determine DBLα types shared between isolates, the cleaned DBLa reads were clustered using a pipeline based on the USEARCH software suite version 8.1.1831[45]. Initially duplicate reads were removed and the remaining reads were sorted by how many duplicates were present using the derep_prefix command. The remaining reads were then clustered at 96% pairwise identity using the usearch cluster_fast command. Finally, the original unfiltered reads were aligned back to the centroids of the clusters and an OTU table was generated using the usearch_global command before a binary version of the table was generated.

Building of empirical networks and model prediction

Empirical networks were built from *var* DBLα types sequenced and processed from 1284 *P. falciparum* isolates from individuals residing in BD, Ghana. Following a previously published analysis framework, the DBLα types were translated into all six reading frames and classified into either upsA or upsB/upsC (i.e., non-upsA) groups[42]. Gene networks are built based on pairwise similarities of unique upsB/upsC DBLα types that are above 0.6. The choice of the threshold is based on the average within-class sequence similarity of the 24 DBLα subclasses (see %ID in Fig. 2 of Rask et al.[16]). Since infections by multiple parasite genomes (multiplicity of infection [MOI] larger than 1) are very common in malaria endemic regions, we selected isolates with a total number of upsB/upsC DBLα types ranging from 40-55 copies to maximize the probability of selecting hosts with single-clone infections, which reduced the number of isolates to 161 (see main text for rationale of focusing on upsB/upsC DBLα types). The repertoire similarity network is built among these isolates (Extended Data Fig. 6). In order to build a classifier for the empirical network, a library of simulated networks was generated for parameter ranges representative of Ghana: global *var* gene pools from 10,000 to 20,000, duration of naïve infection equal to 1 year, and biting rates ranging from 0.1 to 0.5 person per day. The simulated networks were then constructed by sampling

161 random isolates from two periods seven months apart from each other, resembling the sampling regime of the empirical data. The trained discriminant functions from simulated networks[32] are then applied on empirical networks to predict whether the network is generated from immune selection dominant forces or pure neutral forces. The Bayesian posterior probability of classification is calculated by assuming Gaussian densities of prior distributions of each class.

References

1. Deitsch, K. W., Lukehart, S. A. & Stringer, J. R. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat. Rev. Microbiol.* **7,** 493–503 (2009).
2. Deitsch, K. W., Moxon, E. R. & Wellems, T. E. Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol. Mol. Biol. Rev.* **61,** 281–293 (1997).
3. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419,** 498–511 (2002).
4. Chen, D. S. *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).
5. Freitas-Junior, L. H. *et al.* Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature* **407,** 1018–1022 (2000).
6. Claessens, A. *et al.* Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLoS Genet.* **10,** e1004812 (2014).
7. Day, K. P. *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).
8. Gupta, S. *et al.* The maintenance of strain structure in populations of recombining infectious agents. *Nat. Med.* **2,** 437–442 (1996).
9. Gupta, S., Ferguson, N. & Anderson, R. Chaos, Persistence, and Evolution of Strain Structure in Antigenically Diverse Infectious Agents. *Science* **280,** 912–915 (1998).
10. Buckee, C. O., Recker, M., Watkins, E. R. & Gupta, S. Role of stochastic processes in maintaining discrete strain structure in antigenically diverse pathogen populations. *Proc. Natl. Acad. Sci.* **108,** 15504–15509 (2011).
11. Chen, Q. *et al.* Developmental selection of var gene expression in Plasmodium falciparum. *Nature* **394,** 392–395 (1998).
12. Rosindell, J., Hubbell, S. P. & Etienne, R. S. The Unified Neutral Theory of Biodiversity and Biogeography at Age Ten. *Trends Ecol. Evol.* **26,** 340–348 (2011).
13. Levine, J. M. & HilleRisLambers, J. The importance of niches for the maintenance of species diversity. *Nature* **461,** 254–257 (2009).
14. Fijarczyk, A. & Babik, W. Detecting balancing selection in genomes: limits and prospects. *Mol. Ecol.* **24,** 3529–3545 (2015).
15. Chisholm, R. A., Fung, T., Chimalakonda, D. & O'Dwyer, J. P. Maintenance of biodiversity on islands. *Proc. R. Soc. B Biol. Sci.* **283,** 20160102 (2016).
16. Cavender-Bares, J., Kozak, K. H., Fine, P. V. A. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecol. Lett.* **12,** 693–715 (2009).
17. Levine, J. M., Bascompte, J., Adler, P. B. & Allesina, S. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature* **546,** 56–64 (2017).

18. Day, K. P. & Marsh, K. Naturally acquired immunity to Plasmodium faldparum. *Immunol. Today* **12,** A68–A71 (1991).
19. Tiedje, K. E. *et al.* Seasonal Variation in the Epidemiology of Asymptomatic Plasmodium falciparum Infections across Two Catchment Areas in Bongo District, Ghana. *Am. J. Trop. Med. Hyg.* **97,** 199–212 (2017).
20. Barry, A. E. *et al.* The Stability and Complexity of Antibody Responses to the Major Surface Antigen of Plasmodium falciparum Are Associated with Age in a Malaria Endemic Area. *Mol. Cell. Proteomics MCP* **10,** (2011).
21. Bull, P. C. *et al.* Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat. Med.* **4,** 358–360 (1998).
22. Zilversmit, M. M. *et al.* Hypervariable antigen genes in malaria have ancient roots. *BMC Evol. Biol.* **13,** 110 (2013).
23. Larremore, D. B. *et al.* Ape parasite origins of human malaria virulence genes. *Nat. Commun.* **6,** 8368 (2015).
24. Rask, T. S., Hansen, D. A., Theander, T. G., Pedersen, A. G. & Lavstsen, T. Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer. *PLOS Comput. Biol.* **6,** e1000933 (2010).
25. Macarthur, R. & Levins, R. The Limiting Similarity, Convergence, and Divergence of Coexisting Species. *Am. Nat.* **101,** 377–385 (1967).
26. Nosil, P., Funk, D. J. & Ortiz-Barrientos, D. Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* **18,** 375–402 (2009).
27. Artzy-Randrup, Y. *et al.* Population structuring of multi-copy, antigen-encoding genes in Plasmodium falciparum. *eLife* **1,** e00093 (2012).
28. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69,** 26113 (2004).
29. Barry, A. E. *et al.* Population Genomics of the Immune Evasion (var) Genes of Plasmodium falciparum. *PLOS Pathog.* **3,** e34 (2007).
30. Freeman, L. C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* **40,** 35–41 (1977).
31. Takahata, N. & Satta, Y. Footprints of intragenic recombination at HLA loci. *Immunogenetics* **47,** 430–441 (1998).
32. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11,** 94 (2010).
33. Blomqvist, K. *et al.* A Sequence in Subdomain 2 of DBL1α of Plasmodium falciparum Erythrocyte Membrane Protein 1 Induces Strain Transcending Antibodies. *PLoS ONE* **8,** e52679 (2013).
34. Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O. & Arnold, F. H. On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 5380–5385 (2005).
35. Hermsen, C. C. *et al.* Detection of Plasmodium falciparum malaria parasites in vivo by real-time quantitative PCR. *Mol. Biochem. Parasitol.* **118,** 247–251 (2001).
36. Collins, W. E., Warren, M., Skinner, J. C. & Fredericks, H. J. Studies on the relationship between fluorescent antibody response and the ecology of malaria in Malaysia. *Bull. World Health Organ.* **39,** 451–463 (1968).
37. Gibson, M. A. & Bruck, J. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A* **104,** 1876–1889 (2000).
38. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22,** 403–434 (1976).

39. Shannon, C. E. A Mathematical Theory of Communication. *SIGMOBILE Mob Comput Commun Rev* **5,** 3–55 (2001).
40. Pielou, E. C. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13,** 131–144 (1966).
41. Charlesworth, D. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLOS Genet.* **2,** e64 (2006).
42. Ruybal-Pesántez, S. *et al.* Population genomics of virulence genes of Plasmodium falciparum in clinical isolates from Uganda. *Nat. Sci. Rep.* (under review).
43. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology* **1,** 895–905 (2012).
44. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (2014).
45. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
46. Edgar, R. C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31,** 3476–3482 (2015).
47. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
48. Rask, T. S., Petersen, B., Chen, D. S., Day, K. P. & Pedersen, A. G. Using expected sequence features to improve basecalling accuracy of amplicon pyrosequencing data. *BMC Bioinformatics* **17,** 176 (2016).
49. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442 (1998).
50. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 3747–3752 (2004).
51. Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications*. (Cambridge University Press, 1994).
52. Newman, M. E. J. Assortative Mixing in Networks. *Phys. Rev. Lett.* **89,** 208701 (2002).
53. Costa, L. da F., Rodrigues, F. A., Travieso, G. & Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Adv. Phys.* **56,** 167–242 (2007).
54. Harary, F. *Graph theory*. (Addison-Wesley, 1994).
55. Latora, V. & Marchiori, M. Efficient Behavior of Small-World Networks. *Phys. Rev. Lett.* **87,** 198701 (2001).
56. Cordella, L. P., Foggia, P., Sansone, C. & Vento, M. An improved algorithm for matching large graphs. in *In: 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen* 149–159 (2001).
57. Hudson, R. R., Kreitman, M. & Aguadé, M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* **116,** 153–159 (1987).
58. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709 (1993).
59. Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123,** 585–595 (1989).
60. Alonso, D. & McKane, A. J. Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.* **7,** 901–910 (2004).
61. DeGiorgio, M., Lohmueller, K. E. & Nielsen, R. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. *PLOS Genet.* **10,** e1004561 (2014).
62. Koleff, P., Gaston, K. J. & Lennon, J. J. Measuring beta diversity for presence–absence data. *J. Anim. Ecol.* **72,** 367–382 (2003).

63. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
64. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLOS Biol.* **4,** e72 (2006).
65. Buckee, C. O. & Recker, M. Evolution of the Multi-Domain Structures of Virulence Genes in the Human Malaria Parasite, Plasmodium falciparum. *PLoS Comput. Biol.* **8,** e1002451 (2012).

Acknowledgements

This research was supported by the Fogarty International Center at the National Institutes of Health [Program on the Ecology and Evolution of Infectious Diseases (EEID), Grant number: R01-TW009670]. S.P. was supported by a James S. McDonnell Foundation 21st Century Science Initiative -- Postdoctoral Program in Complexity Science-Complex Systems Fellowship Award and by a Fulbright Fellowship from the U.S. Department of State. We wish to thank the participants, communities and the Ghana Health Service in Bongo District Ghana for their willingness to participate in this study. We would also like to thank the personnel at the Navrongo Health Research Centre for sample collection and parasitological assessment/expertise. We are grateful to Abraham R. Oduro, Anita Ghansah, and Kwadwo Koram for their helpful input related to the field study, to Gerry Tonkin-Hill for the development of the Illumina sequencing cleaning and clustering pipelines, and to Michael F. Duffy and Andrew P. Dobson for their 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.

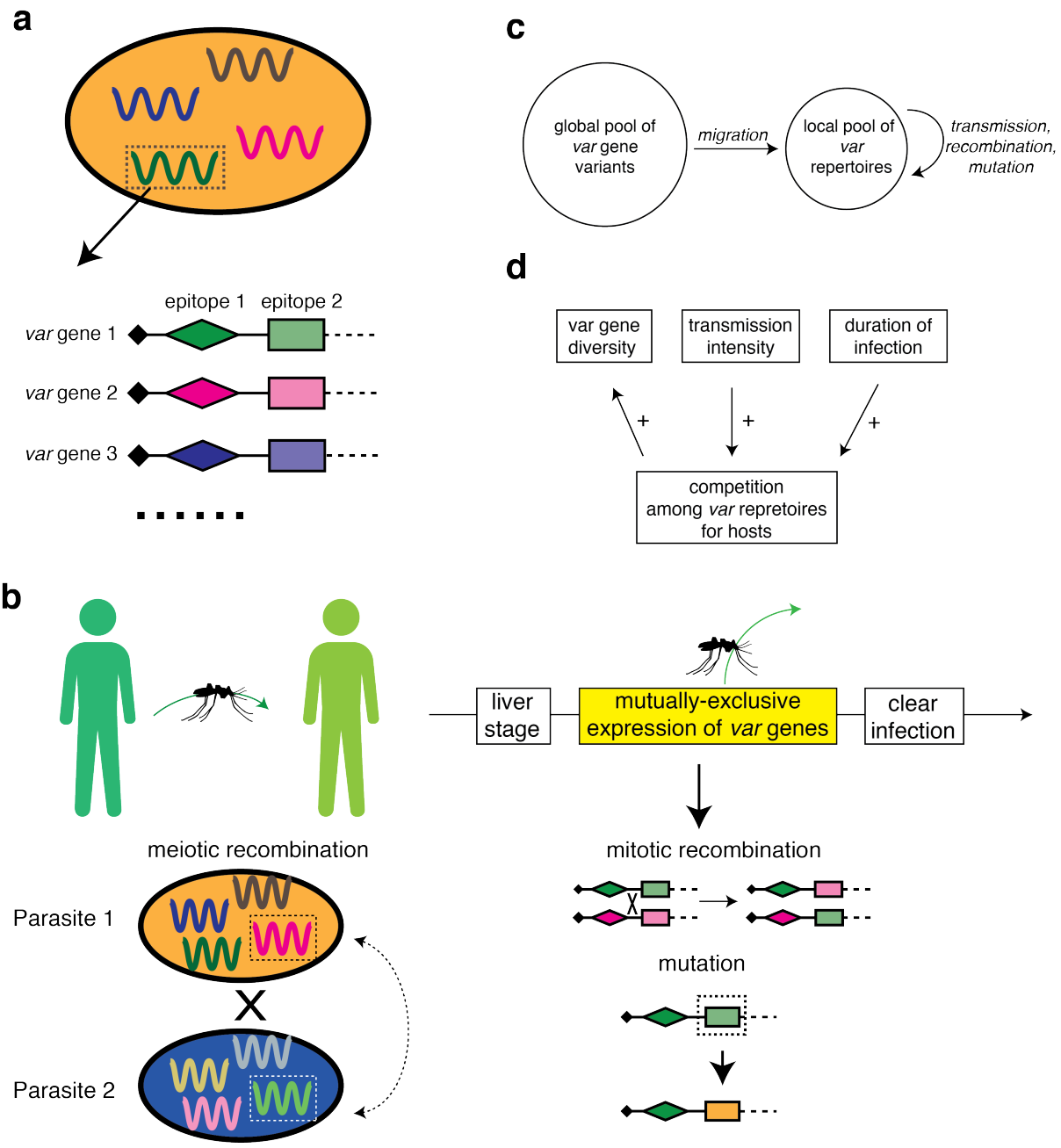
Extended Data Table 1 | List of network properties used in network classification.

No	Network Properties	Description
	Transitivity/clustering coefficient	The degree to which nodes in a graph tend to cluster together
1	Average local clustering coefficient	Ratio of the triangles connected to a node and the triples centered on the node (undirected unweighted networks)[49]
2	Average weighted local clustering coefficient	Local clustering coefficient in weighted network[50]
3	Global clustering coefficient	The ratio of the triangles and the connected triples in the network[51]
	Degree	Number of edges per node
4	Graph density	The ratio of the number of edges and the number of possible edges[51]
5	Proportion of nodes with degree 0	
6	Proportion of nodes with degree 1	
7	Average degree	
8	Assortativity	Pearson correlation coefficient of the degrees at either ends of an edge[52]
9	Average strength	Sum of edge weights of the adjacent edges for each node
10	Straightness (Power law test)	Pearson coefficient of a power-law degree distribution[53]
11	Average measurement of the heterogeneity of the network	Entropy of the degree distribution[53]
12	Number of components relative to network size	
13	Average component size	
14	Entropy of component size distribution	
	Geodesic distance	Shortest path between two vertices
15	Central point dominance	Average difference in centrality between the most central node and all others[30]
16	Average eccentricity	Average maximum shortest paths from a node to all other nodes[54]
17	Diameter	Length of the longest distance between any pairs of nodes
18	Mean diameter of all components	
19	Reciprocal of global efficiency	Harmonic mean of the geodesic distances[53,55]
20	Average closeness centrality	Average steps to access every other node from a given node[30]
	Graph motifs proportions	Percentage of each motif among all 3-node motifs within the network[56]
21	A->B<-C, the binary in-tree.	
22	A->B->C, the directed line.	
23	A<->B<-C.	
24	A<-B->C, the binary out-tree.	
25	A->B<-C, A->C.	
26	A<-B->C, A<->C.	
27	A<->B->C.	
28	A<->B<->C.	
29	A->B->C, A<->C.	
30	A->B<-C, A<->C.	
31	A->B<->C, A<->C.	
32	A<->B<->C, A<->C, the complete graph.	
	In and out edges	
33	Reciprocity	Proportion of mutual connections
34	In-out correlation	Correlation between numbers of in and out edges for each node
	Modules and F_{ST}	Community detected with Newman-Girvan algorithm[28]
35	Community Size Evenness	Gini index of community sizes
36	Number of common communities	
37	Ratio of biggest community	
38	Q	Maximum modularity[28]
39	Maximum F_{ST}	Maximum diversity within community compared with between communitites[31,41]

Extended Table 2 | Characteristics of immune selection (balancing selection) captured by different network metrics compared with traditional approaches from community ecology and population genetics. Network properties provide more information on population structure compared to the diversity measures commonly used in ecology (see Supplementary results and discussion; Extended Data Fig. 7). Moreover, these properties are more appropriate for describing similarity patterns in the presence of frequent recombination than population genetics measures that assume the existence of a 'tree'-like phylogeny. The table summarizes general patterns that arise from balancing selection, and includes a specific pattern relevant for gene families. The correspondence between common tests or indices in ecology and population genetics is shown, together with their predictions relative to those of neutral scenarios (in parentheses). Network metrics are for the most part explained in Methods and in the Extended Data table 1. Remaining ones are referred to specific figures or references in this table.

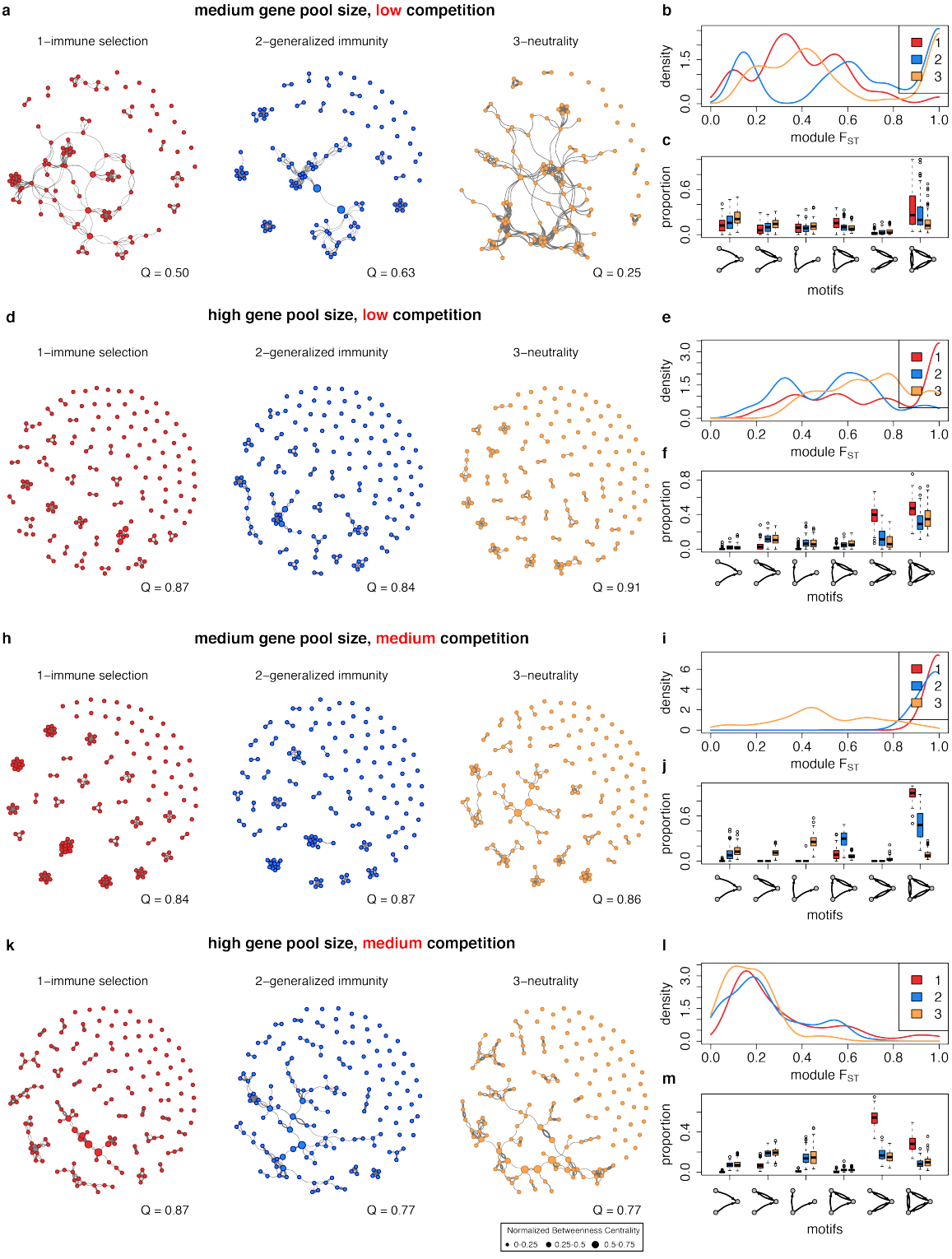
Patterns	Community ecology	Population genetics*	Network metrics
Increased diversity around targets of selection	Shannon diversity *H* (higher)[39]	Hudson–Kreitman–Agaudé test (much higher polymorphism than divergence)[57]	Greater number of components (see Fig. 1 and Extended Data Fig. 2)
Evenly sized niches (ecology); Excess of common polymorphisms (evolution)	Evenness (higher)[40]	Fu and Li's F[58], Tajima's D[59] (positive)	Negative relationships between frequency and relatedness (see text and Extended Data Fig. 4)
	Abundance distribution (skewed toward intermediate-abundance)[60]	Allele frequency spectrum (skewed toward intermediate-frequency alleles)[61]	Even component/community sizes (see Fig. 1 and Extended Data Fig. 2)
Persistence of niches (ecology); Shared variants across species (evolution)		F_{ST} among species compared to neutral sites (lower)	Persistence of strain modules (not tested in this paper)
		Persistence of similar gene variants across species	
Shared diversity across location	Beta diversity across locations (lower)[62]	F_{ST} among locations compared to neutral genes (lower)	Module identities have less correlation with locations than neutral genes (not tested in this paper)
Limiting similarity (ecology); Linkage disequilibrium among genes (evolution)	Pairwise type sharing (low)[29]	Long Range Haplotype test (longer haplotype than expected)[63]	Modularity (see Fig. 1 and Extended Data Fig. 2)[28]
		Integrated Extended Haplotype Homozygosity (negative, long haplotypes associated with derived alleles)[64]	Module F_{ST} (see Methods for definition and references, Fig. 1 and Extended Data Fig. 2)
			Transitivity
			Higher proportion of complete graph in motif compositions (see Fig. 1 and Extended Data Fig. 2)
High within-genome diversity of gene families			High reciprocity (see Fig. 1 and Extended Data Fig. 2)

*** Tests listed under population genetics are adapted from Fijarczyk and Babik[14], where more tests and references are reviewed and discussed.**



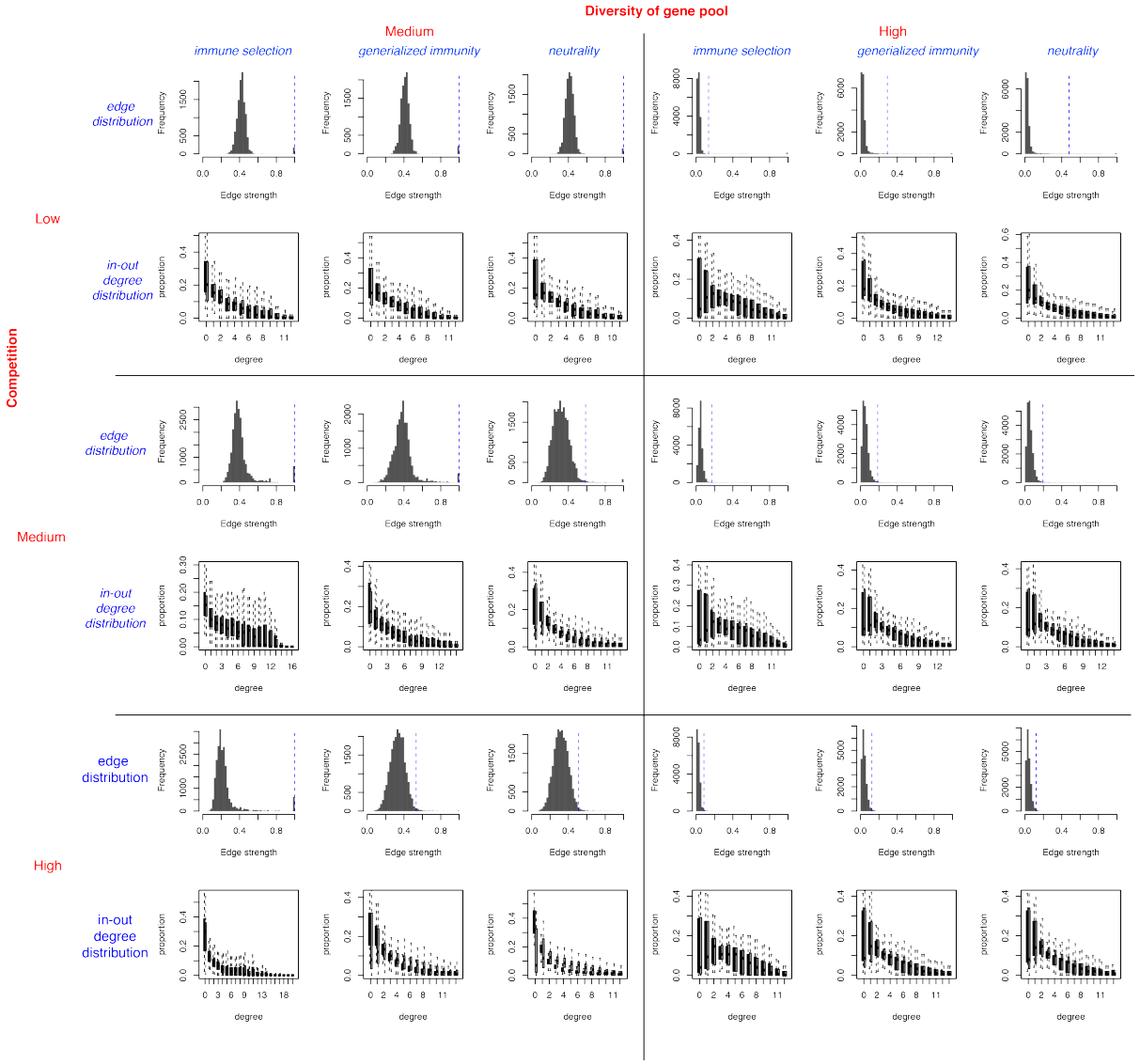
Extended Data Figure 1. Schematic illustration of the *var* gene model. **a**, Each parasite genome (ovals) consists of a repertoire of *g* copies of *var* genes. Each *var* gene (depicted by different colors within each parasite) is in turn represented as a linear combination of epitopes (depicted by different shapes), with each epitope having many possible variants (alleles, shown in different colors). **b**, At each transmission event, one donor and one receiver host are selected at random from the host pool. Each parasite genome in the donor host is transmitted to the mosquito with probability of 1/(number of repertoires). During the sexual stage of the parasite (within mosquitoes), different parasite genomes can exchange *var* repertoires through meiotic recombination to generate novel recombinant repertoires. The receiver host can receive either recombinant genomes or original genomes. During the asexual reproduction stage of the parasite (within the blood stage of infection), *var* genes within the same genome exchange epitope alleles through mitotic (ectopic) recombination. Also, epitopes can mutate. These two processes generate new *var* genes. Each *var* gene is expressed sequentially and the infection ends when all the *var* genes in the repertoires have

been expressed. A new transmission event may occur throughout the period of expression of *var* genes as the result of biting events. **c**, The local population receives *var* repertoires from a fixed global *var* gene pool through migration. **d**, Gene pool size (regional *var* gene diversity), biting rate (transmission intensity), and duration of naïve infections all interact to influence the level of competition among *var* repertoires for human hosts.

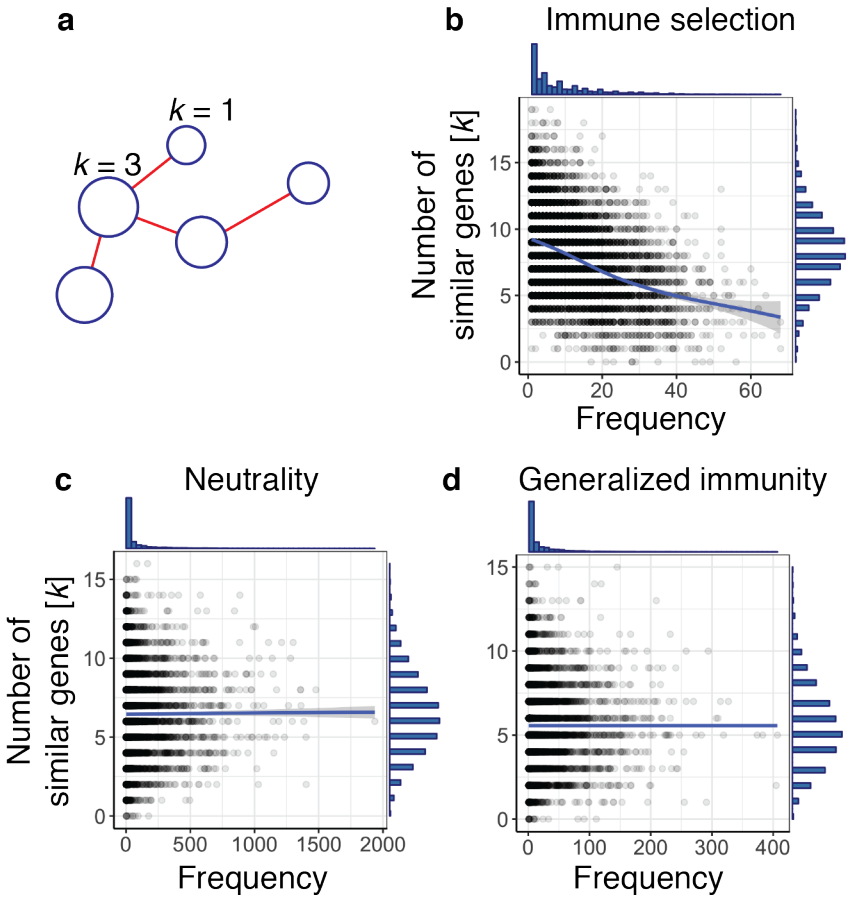


Extended Data Figure 2 | Repertoire similarity networks and representative network metrics across scenarios for different diversity regimes generated with low and medium competition (shorter duration of naive infection and lower transmission rates). First and third panels, medium gene pool size (G = 1,200) and second and fourth panels, high gene pool size (G = 24,000). **a, d**, **h**, **k**, comparisons of strain similarity networks of 150 randomly sampled parasite *var* repertoires from one time point under the three scenarios. Only the top 1% of the strongly connected links are drawn and used in the analyses, with the thickness of the edges representing the relative strength of connections within the network (see Extended

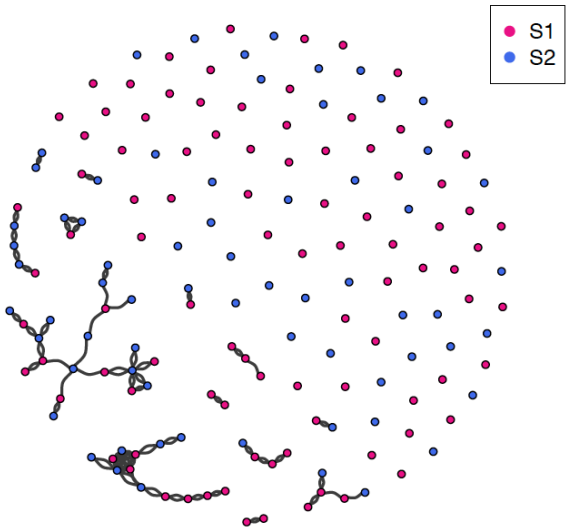
Data Fig. 3 for distribution of edge weights and degrees). Within the largest component of each network, the size of each node indicates its normalized betweenness centrality. The value of maximum modularity Q is calculated using edge-betweenness and shown at the lower right corner of each network. **b, e**, **i**, **l**, pairwise module F_{ST} distributions. **c, f**, **j**, **m**, proportion of occurrence of 3-node graph motifs for the three models.



Extended Data Figure 3 | Edge weight distributions and in-out degree distributions under different gene pool diversity and competition regimes. Distributions are ordered from medium diversity (left) to high diversity (right), and low competition (upper) to high competition (lower). Under each regime, scenarios are shown for immune selection (left), generalized immunity (middle) and neutrality (right). The blue dotted lines in the edge distribution plots shows the repertoire similarity cutoff (for the top 1% of edges used in building the networks). Bar plots of in and out degree are shown across 100 networks that are generated under these same regimes (respectively in black and grey).



Extended Data Figure 4. The relationship between gene frequencies and number of similar genes. **a**, In a gene similarity network, each node represents a unique gene transmitted in the population and the edges encode the sharing of at least one allele between genes. The size of the node is proportional to its frequency in the population and the node degree k depicts the number of genes that share at least one allele with the focal gene. There is a negative correlation in the immune selection scenario (**b**, r = -0.412, p-value < 2.2e-16), and no statistically significant relationship for complete neutrality (**c**, r = 0.016, p-value = 0.17) and for generalized immunity (**d**, r = -0.002, p-value = 0.89). G = 24,000, b = 0.5, D = 1 year.



Extended Data Figure 5 | Strain similarity network of var upsB/C DBLα types in the Ghana samples. The color or each node represents the season in which the isolate was sampled. The top 1% of edges (i.e., $S_{ij} > 0.0755$) is shown in the graph and used in the analyses.

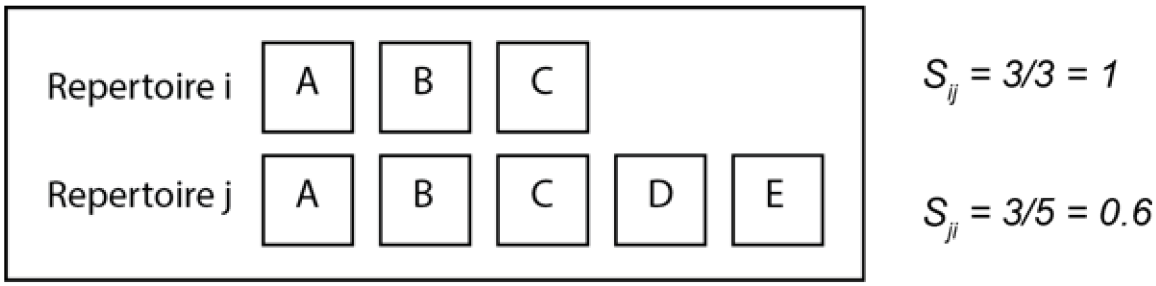
a

Symbol	Type	Description	Values
H	Integer	Host population size	10,000
G	Integer	Pool of *var* genes	1,200 – 24,000
g	Integer	Size of each repertoire (i.e., number of *var* genes per strain)	10 - 60
l	Integer	Number of epitopes per *var* gene	2, 5
n	Integer	Number of allele variants per epitope	120 – 2,400
b	Rate	Biting event rate per host	0.05-0.5
p	Probability	Transmission probability	0.5
ρ	Rate	Mitotic recombination rate per gene	1.8e-07/day
μ	Rate	Mutation rate per epitope per gene	1.42e-08/day
w	Rate	Rate of specific epitope immunity wane per day	0.001
m	Rate	Migration rate of new strains	1 genome per day
D	Rate	Duration of infections	3 months to 2 years

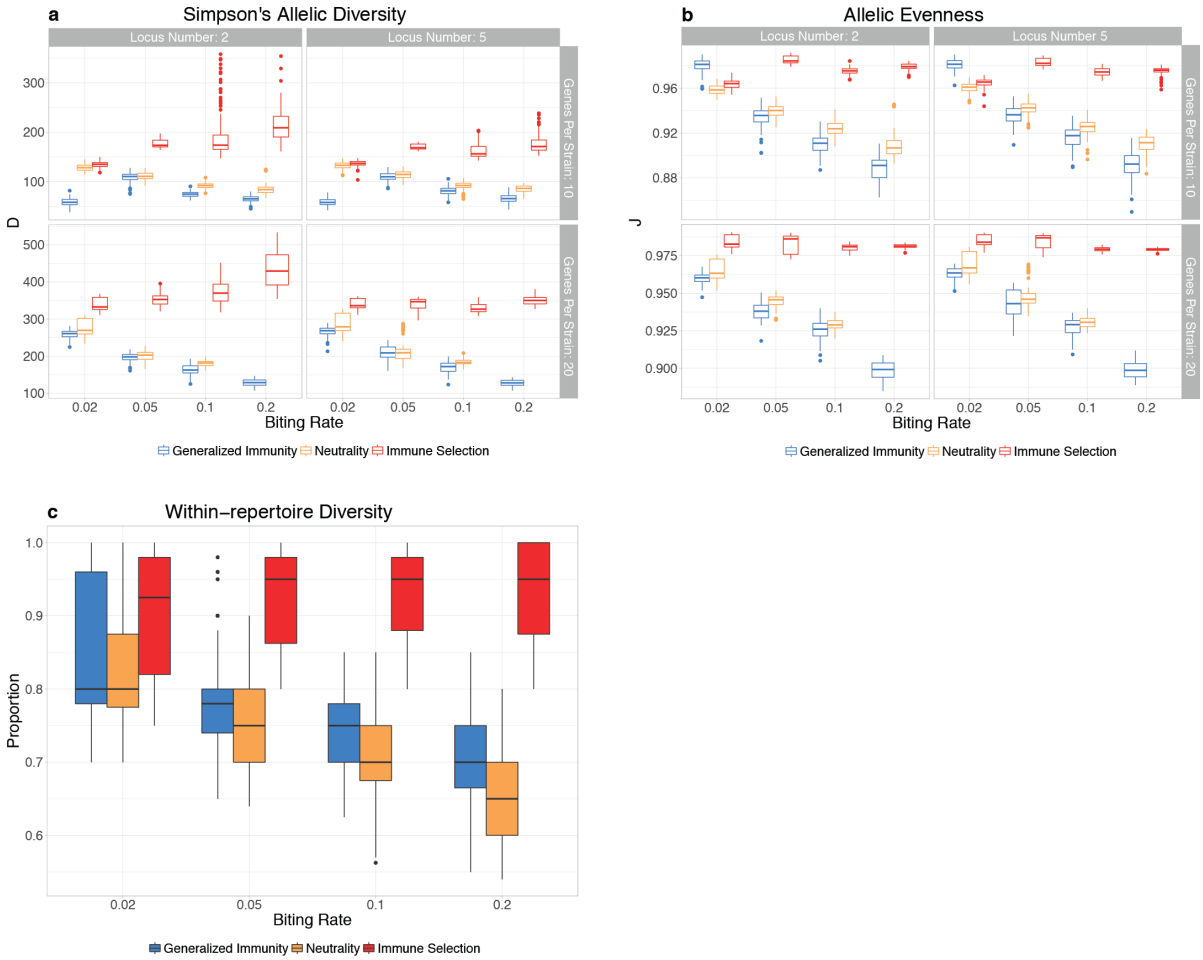
b

Scenarios	Specific immunity	Infection duration reduced after reinfection	Average duration of infection
Immune selection	Yes	Yes	Number of epitopes seen
General immunity	No	Yes	Number of infections
Complete neutrality	No	No	Constant

c



Extended Data Figure 6 | Model parameters, differences between model scenarios, and computation of edge strength in network construction. a, Description of parameters and selected ranges. **b**, Model comparisons for immune selection, generalized immunity and complete neutrality. **c**, Illustration of the similarity index used in the directional similarity networks. We use a directional network because of the asymmetric competition resulting from different numbers of unique variants in a repertoire. In the example, strain *i* has 3 unique alleles, while strain *j* has 5. Together they share 3 alleles. Therefore, strain *i* can be substituted by strain *j* completely, while strain *j* can only be substituted by strain *i* partially. Therefore, strain *j* will have a prolonged expression in a host which is immune to strain *i*, while strain *i* will not be able to cause infection in a host which is immune to strain *j*.



Extended Data Figure 7 | Comparisons of allelic diversity across immune selection (red), generalized immunity (blue) and neutral (yellow) scenarios. Simpson's allelic diversity *D* (**a**) and allelic evenness *J* (**b**) are compared across biting rates. The number of epitopes per *var* gene (columns) vs. the number of *var* genes per genome (rows). Within repertoire diversity (**c**) is presented as the proportion of unique alleles divided by the length of the genome. Values are combined for all parameter combinations for a given biting rate.

Supplementary information:
Supplementary results and discussion

Selection signature based on standard ecological diversity measures

We investigated whether standard ecological diversity measures can be used to differentiate selection signatures. As expected from standard genetic signatures of frequency-dependent selection, the parasite *var* gene population has higher and more even epitope (allelic) diversities in the selection than in the neutral models for the same parameter ranges (Extended Data Fig. 7a, b). Diversity patterns under generalized immunity, although different from those under complete neutrality, nevertheless resemble those of complete neutrality more than those of immune selection. What is more unique to the system as a result of within-strain competition is a higher within-genome diversity than that of the null models (also see Buckee and Recker[65] on the evolution of multi-domains in gene structures) (Extended Data Fig. 7c). Except for the allelic diversity indices, most of the other diversity indices (such as beta diversity, genetic or repertoire diversity) do not show clear trends differentiating the underlying processes. Because these differences are relative, a given value of these indices would not provide information about underlying processes. In this sense, they are un-informative and would require comparisons across endemicity gradients to provide evidence for non-neutrality in empirical systems.

Supplementary file:

1. Gene compositions of Ghana isolates. (To be provided as a separate file, as table is too large to be included here).