



Assessing the effect of natural selection in malaria parasites

Ananias A. Escalante^{1,2}, Omar E. Cornejo¹, Ascanio Rojas¹, Venkatachalam Udhayakumar² and Altaf A. Lal²

¹Instituto Venezolano de Investigaciones Científicas, Apartado 21827, Caracas 1020-A, Venezuela

²Division of Parasitic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Public Health Service, U.S. Department of Health and Human Services, Mail Stop F-12, 4770 Buford Highway, Chamblee, GA 30341, USA

There are few concepts that have been used across disciplines; one of them is natural selection. The impact that this process has on parasite genetic diversity is reviewed here by discussing examples on drug resistance and vaccine antigens. Emphasis is made on how mechanisms need to be addressed rather than associations, and how such investigations were out of reach of biomedical researchers only a decade ago.

Malaria is endemic in most of the tropical and subtropical ecosystems worldwide, and exhibits great geographical diversity. This diversity includes not only ecological and epidemiological characteristics, but also extensive polymorphisms in the genes encoding antigenic proteins. This expressed phenotypic variation is the raw material on which natural selection operates.

Natural selection is brought up in the malaria literature whenever an elicited immune response is detected, or linkage is found between alleles and a phenotype, such as drug resistance. However, evidence to support such statements might not be easy to obtain.

Evolution by natural selection is the outcome of differences in reproduction of phenotypic variants in a given environment; this differential reproduction affects the frequencies of the associated genotypes in the next generation. Thus, there are three necessary conditions for evolution by natural selection to occur: (i) phenotypic variation; (ii) differences in the reproductive capacity of those phenotypic variants, given the environment; and (iii) genetic variation associated with the phenotypic variation.

The processes driving natural selection (also called selective forces or selective pressure) are those capable of affecting the reproduction of the exposed phenotypes. In the context of malaria parasites, studies of natural selection have focused on antimalarial drugs and immune responses as selective pressures; however, natural selection will occur as the result of any process that differentially affects the reproduction of the parasite population, based on phenotypes linked to genetic variation. As a result, genetic variants favored by a selective pressure will increase in frequency or be maintained in the

population (positive selection), whereas those negatively selected will decrease or have been eliminated (negative selection). These apparently simple processes represent what we try to capture wherever we suspect that natural selection is important in explaining the observed genetic diversity.

Current investigations point out that phenotypes frequently associated solely to parasite genetic variants are a result of interactions among vertebrate, mosquito and parasite populations [1,2]. Virulence, for example, is associated with transmission dynamics that could select variants locally as a result of competition among lineages and/or by selecting those with a high replication rate [2]. Point mutations known to be associated with drug resistance *in vitro*, however, do not necessarily predict clinical outcome [3,4] because we try to infer the effect of the drug on a parasite subpopulation (the infection) by looking in the host (the clinical outcome). Thus, the parasite intrahost dynamics are essential in understanding the origin of phenotypes of public health relevance. Natural selection is part of this dynamic.

It is not debatable, for example, that part of the phenotypic diversity exhibited by malaria worldwide is associated with the host genetic make up and immunological conditions; however, commonly, the genetic basis of complex phenotypes are investigated solely in the parasite. This 'parasite determinism' is usually a consequence of the logistic complications of studies that include human genetics rather than a conceptual issue. Historically, the importance of the interaction between host genetics and the parasite has received attention [5,6], and evidence supporting a role for host genetics is continuously gathered in issues such as disease severity [7–9] or drug resistance [10].

This review covers only part of the story – the role played by natural selection in shaping the genetic diversity of parasite population and its implications in biomedical research. Nevertheless, the role of the host and transmission dynamics will be addressed in those cases where its inclusion is necessary.

Formal analyses for detecting evidence of natural selection acting on the parasite population are relatively new. As an example, the diversity observed in genes encoding antigens, especially those in the merozoite and

sporozoite, were attributed to the action of natural selection imposed by the host immune system [11,12]. The assumption is that genes encoding antigenic proteins accumulate mutations that allow the parasite to hamper the host capacity of building up efficient immune responses; thus, accumulation of mutations will be favored by natural selection. This idea was originally tested in only a few studies [13–17]. Nevertheless, there is an increasing interest in studying the polymorphism and molecular evolution of genes encoding malaria antigens [18–23]. Nowadays, studies on malaria parasite populations include a broad range of issues from loci to genome-based analyses. Indeed, current studies on malaria drug resistance [24,25] could become textbook examples of selective sweeps.

Regardless of the increasing amount of information and sophistication of the current population studies, formal evolutionary analyses are still rare in malaria literature. Investigating the effect of natural selection on parasite genetic diversity goes beyond a statistical technicality; it is testing biological models that allow us to make inferences on the mechanisms behind the genetic diversity rather than simply talking about correlations. The practical side of the problem is that ignoring mechanisms could lead to spurious associations of genotypes with phenotypes of public health relevance.

In general, questions lead us to collect samples (the opposite order is usually not recommended). How we perceive the organization of the parasite genetic diversity in the population affects how we collect the samples and generate genetic data. In this context, the question, is it a strain or not a strain, is not a rhetorical one.

Strains as units of selection

The term strain has been used when parasite genetic diversity has been implicated in a clinical end point. Strains have become a unit on which selection might operate. The term usually refers to a group of parasites with a particular phenotypic trait. The implicit assumption is that these groups of parasites share the genetic make up behind the observed phenotype and that they are stable entities, although their stability in time or space is seldom addressed.

The term strain is properly used with two major formal connotations: one based on population genetic structure, where the parasite population is subdivided into linked multi-locus genotypes that have some temporal and spatial stability [26]; and the other based on phenotypes, in which the phenotypes are transmitted independently [27]. The phenotypic definition was originally used to describe groups of parasites based on the specificity of their immune responses (without cross-immunity). These two concepts are related if linkage among loci is generated by natural selection [28]. Despite these two formal strain definitions, the term is still used in a casual way.

The application of a formal definition for the term strain is not easy. If, for example, we applied the population structure concept of strain when we refer to strain-specific immunity, we are implying that linkage among loci encoding antigens was observed and that it is maintained by the host immune response [28]. Following this

argument, because linkage groups can originate by inbreeding if transmission is low [29], evidence is needed to confirm that the observed linkage cannot be explained by the population demographics alone. This can be tested by studying linkage among neutral genetic markers that are not linked themselves with the antigens under consideration (loci that can freely segregate during meiosis and that are not themselves under a selective pressure). If linkage is not found in neutral markers, then the linkage among the antigens can be used as evidence of selection by the host immune response; such evidence is yet to be seen in studies on malaria parasites. Given this context, the term strain is usually not an observation or a hypothesis – it only expresses confusion.

At this point, it is noteworthy that detectable linkage among genetic markers could increase with control measures that reduce transmission (and the parasite population size), for example, the use of bednets [30]. Careful epidemiological and population genetic designs are needed to rule out spurious associations of genetic variants with phenotypes of biomedical interest.

The term strain could also refer to phenotypes that are independently transmitted [27]. This is a robust use of terms such as strain-specific immunity. However, the term applied to any single antigen generates unnecessary confusion because information on cross-reactivity among alleles is usually limited or absent. A term such as allele-elicited immunity, specific or not, might better describe the immunological implications of the genetic polymorphism in an antigen. Indeed, more often, we only have limited data that just begins to unveil the interaction of the observed polymorphism with the host immune response [12,31–33].

The term strain is also widely used in the case of drug resistance. As we look for specific mutations [3,4], it seems inappropriate to talk about strains when we are talking about alleles.

After all these arguments, probably the most conservative approach is to eradicate the term strain if we do not mean, or provide data supporting, any of the two formal definitions described above, the one based on population structure and the other based on independently transmitted phenotypes.

Why should we worry about using the term strain as similes of multi-loci genotype, phenotype, isolate and group of parasites? These casual uses sometimes make researchers miss the point that, in most cases, we detect transient associations of genetic variants of phenotypes, given a process rather than entities with their own existence. This misunderstanding distracts our attention from the processes, which is what matters, thus reducing the problem to correlate typologically defined strains with a phenotype of interest. This approach tells us little about how phenotypes originated and dispersed.

The strain-free space: documenting selective sweeps

Biomedical scientists working on drug resistance often look for evidence of selective sweeps [24,25], that is, the process of a favorable mutation increasing in frequency (positive selection) carrying with it the neutral mutations linked by proximity. The expected pattern after a selective

sweep is a reduction in the overall genetic diversity around the mutations targeted by the selective force [34,35] because linked neutral mutations will also increase in frequency (see Figure 1). Finding evidence of a selective sweep is a demonstration that selection is acting by favoring the fixation of an allele; thus, it is also a method to find the genetic basis of a phenotype [24].

Linking genotypes with phenotypes is seldom a short and clean story. Conflicting or non-conclusive results are found to some extent in the literature. Examples can be found when looking for the role of specific point mutations associated with malaria protection [36]. Generally, the discrepancies are explained by differences in experimental design, case definition (how the phenotype is defined), reagents or sample sizes. These are all valid concerns. However, beyond the methodological differences, a phenotype could represent the convergence of genotypes from different genetic background precisely because natural selection is operating. An example of convergence on a specific phenotype, chloroquine (CQ) resistance, has been documented.

The gene encoding the digestive vacuole transmembrane protein, *pfcr*, was suspected to be linked with CQ resistance in *Plasmodium falciparum* [37]. Molecular studies of population genetics have revealed that CQ resistance originated independently three times, by showing three selective sweeps in the *pfcr* region [24]. This convergent phenotype, from different genetic backgrounds, provides exceptional evidence incriminating this gene with CQ resistance, but also indicate that the same phenotype could emerge independently because it is under natural selection. In this specific case, the same locus was involved, however, this might not be the case in more-complex phenotypic traits.

Sometimes evidence of selection is not observed because the experimental data are collected without consideration of the underlying mechanism. In the case of mutations in the dihydrofolate reductase thymidylate synthase (DHFR) gene of *P. falciparum*, associated with resistance to sulfadoxine–pyrimethamine (SP), the parasite population

dynamics hamper our capacity to find associations in some epidemiological settings [3,4]. The DHFR mutations confer resistance to SP *in vitro* [38]; however, their association with a clinical drug failure *in vivo* might depend on their frequency in the host. If the surviving parasites carrying the DHFR mutants replace the parasite subpopulation participating in the infection after drug application, we could see a clinical failure, but this scenario depends on the capacity of the host immune response to control the remaining parasitemia and other aspects of the intrahost dynamics [39]. Thus, to understand the data, we need to look at the process rather than relying solely on correlations. If we consider the process, then it becomes clear why the association of point mutations with drug resistance varies in a case-by-case basis [3,4]. However, the overall effect on the parasite population can be observed in the form of selective sweeps [25]. The challenge for public health researchers is to integrate the dynamics of the process [39], which affect short-term clinical correlations, with the fact that natural selection is favoring the fixation of mutations [25] into a meaningful drug resistance surveillance system that supports public health policy.

Inferring positive natural selection on vaccine candidates

Another way to look for the effect of natural selection in the parasite population is the one used by researchers working on antigens. In the case of drug resistance, we look for positive selection that reduces genetic diversity around the mutation targeted by the selective pressure, whereas for antigens we usually expect positive natural selection to maintain genetic diversity. The term balancing selection is used when selection is maintaining genetic polymorphism for a longer time than the one expected under a pure random genetic drift model [40], that is, alleles are maintained in the population for a longer period of time than expected by the random loss or fixation of genetic variants as a result of the limited population size. The term is also used as a synonym of heterozygote advantage, a type of balancing selection that does not apply to malaria antigens expressed in the vertebrate host because the host immune system is acting on a haploid stage.

As stated earlier, high levels of genetic diversity observed in antigens have been attributed to the action of natural selection imposed by the host immune system. However, an alternative scenario is that such variation could be neutral or irrelevant in terms of the host immune response.

The neutral theory is currently the null hypothesis against which patterns of genetic variation are contrasted. Under the neutral theory, mutations are seen as deleterious or selectively neutral [40,41]. Those belonging to the first category are expected to be eliminated by negative natural selection; thus, they contribute little to the genetic polymorphism in a population. Under the neutral theory, the vast majority of the observed polymorphism in populations is considered a transient distribution of neutral alleles toward getting lost or fixed by genetic drift. The theory has been enriched with new models that include a third category of mutations considered semi-deleterious,

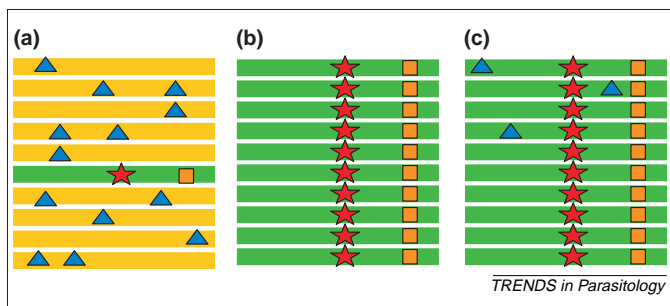


Figure 1. A selective sweep. Colored bars represent chromosomes; orange bars are chromosomes that only have neutral mutations and green bars represent chromosomes that have a selected allele at a specific locus (positions in the chromosome). (a) Before an event of selective sweep, the genetic diversity is as high as would be expected under neutrality (neutral mutations are represented by blue triangles). (b) After an event of selective sweep, an advantageous mutation (red star) rises to fixation by positive natural selection; thus, the genetic diversity around the locus targeted by positive selection is reduced considerably and the neutral linked mutations (orange square) are carried out together with the selected variant. (c) After a period of recovery, the genetic diversity in the flanking regions increase as result of recombination or new mutations, but the region around the locus under selection remains with low diversity. The different parasite populations are indicated (P1, P2 and P3).

which behave as neutral or deleterious depending on the effective population size [42].

Under neutrality, it is expected that genes on which negative selection operates differently will differ in their levels of genetic diversity, so diversity alone, in the absence of any other criteria, does not imply that a gene is more or less neutral. Thus, silent substitutions in a coding protein could be taken as neutral regardless of whether the protein itself is under constraint by negative selection. The implications of the theory are discussed elsewhere (see Refs [40–42]). The neutral theory does not imply that natural selection is not taking place; indeed, negative natural selection is an important part of the model. The model implicates that positive selection maintaining the genetic polymorphism will be a rejection of the neutral model in a given dataset.

There are a variety of tests for detecting departures from neutrality. It is noteworthy that each test has its own assumptions and power limitations. The fact that a polymorphism is not explained by neutrality (reject the null hypothesis) does not imply that the only alternative is positive natural selection.

To discuss this issue, we separate the tests that have been used in the malaria literature into two groups: (i) those based on the distribution of allele frequencies and/or segregating sites; and (ii) those that explicitly try to associate the pattern of polymorphisms with a phenotypic change. In the first group, we include the D test developed by Tajima [43], and the tests D^* and F^* , developed by Fu and Li [44]. In the second group, we include the F_{ST} , used as a measure of genetic differentiation around a gene under selection [32,45–47], the McDonald–Kreitman test (MK) [48] and the rate of synonymous versus nonsynonymous substitutions [49]. It is important to note that these are not the only tests available for detecting departure from neutrality.

As an example of tests on the first group, we will discuss Tajima's D test [43]. This test is based on the neutral model prediction that there are two different ways of estimating the expected heterozygosity under the neutral model from a random sample of alleles obtained from a population (summarized in the parameter θ). One of the estimates is based on the number of segregating sites and the other on pairwise differences among alleles. If these two estimates of θ show a statistically significant discrepancy, then the null hypothesis, the neutral model, is rejected.

Under which scenarios could we expect rejection of the null hypothesis? The number of alleles in low frequency affects the number of segregating sites in the sample, whereas the number of alleles in intermediate frequency affects the pairwise differences. A negative value of D will imply an excess of alleles in low frequency, whereas a positive value will imply an excess of alleles in intermediate frequencies. Balancing selection (heterozygote advantage or frequency-dependent selection) will give an excess of alleles in intermediate frequency (positive D), whereas a selective sweep will give an excess of alleles in low frequency (negative D). However, an excess of alleles in intermediate frequencies is expected in a structured population, whereas an excess of alleles in low frequency is expected if the population is growing.

Nevertheless, this test has been successfully applied in studies of the *P. falciparum* apical membrane antigen 1 (AMA-1) [18,19,23].

How can one separate the effect of history or demography from natural selection? As in the case of linkage, historical and demographic processes affect complete genomes, whereas selection is usually assessed at a specific locus in organisms that are not clonal. Thus, we can investigate neutral markers to differentiate historical–demographic processes from selection at a specific locus. Neutral and selected markers are usually included to document selective sweeps [24,25]. However, neutral markers are seldom included in the same study in cases where balancing selection is suspected. This brings us to the discussion of the second group of tests, where a link with a phenotype is made, rather than simply investigating the pattern of heterozygosity.

Conway *et al.* [32] used F_{ST} to point out that the polymorphism in block 2 of the merozoite surface protein 1 (MSP-1) is maintained by balancing selection. The F_{ST} is not by itself a test for natural selection; however, it can be used if there is a clear notion of the expectations of our model. The patterns expected under balancing selection and population structure in Conway's study are depicted in Figure 2. In this case, under balancing selection, allele frequencies are expected to be maintained at a given locus because genetic diversity will be advantageous (Figure 2), whereas the population will differentiate at neutral loci as a result of genetic drift (Figure 2). Then, a measure of genetic differentiation at the population level, such as F_{ST} , can be used because genetic differentiation among populations at the loci under balancing selection is expected to be less (Figure 2) than the population differentiation by genetic drift (population structure, Figure 2) estimated on other independent genetic markers. This is exactly the observation made on the MSP-1 block-2; there was not a meaningful differentiation on the allele frequencies of the MSP-1 block 2 among populations, whereas there was evidence of population structure using other genetic markers. Conway's approach worked perfectly because: (i) there was enough gene flow among populations, so they shared the same alleles; and (ii) balancing selection on the block 2 was strong enough to stabilize the allele frequencies reducing the genetic differentiation.

Keeping in mind the expectations under balancing selection, another way to look at these results is to acknowledge that there is no evidence of local adaptations in the case of MSP-1 block 2 (Figure 2). Local adaptations are genetic variants that have been selected in a given environment by local processes. A local adaptation could be, for example, an association between alleles encoding a cytotoxic T lymphocyte (CTL) epitope in the parasite population and the local frequency of human leukocyte antigen (HLA) alleles in the human population. As a result, the selected alleles of the CTL epitopes will change, depending on the frequencies of the HLA alleles in the human populations. This process will favor genetic differentiation among the parasite populations at the loci under selection (Figure 2). Thus, F_{ST} estimates on the gene under selection are expected to be higher than those

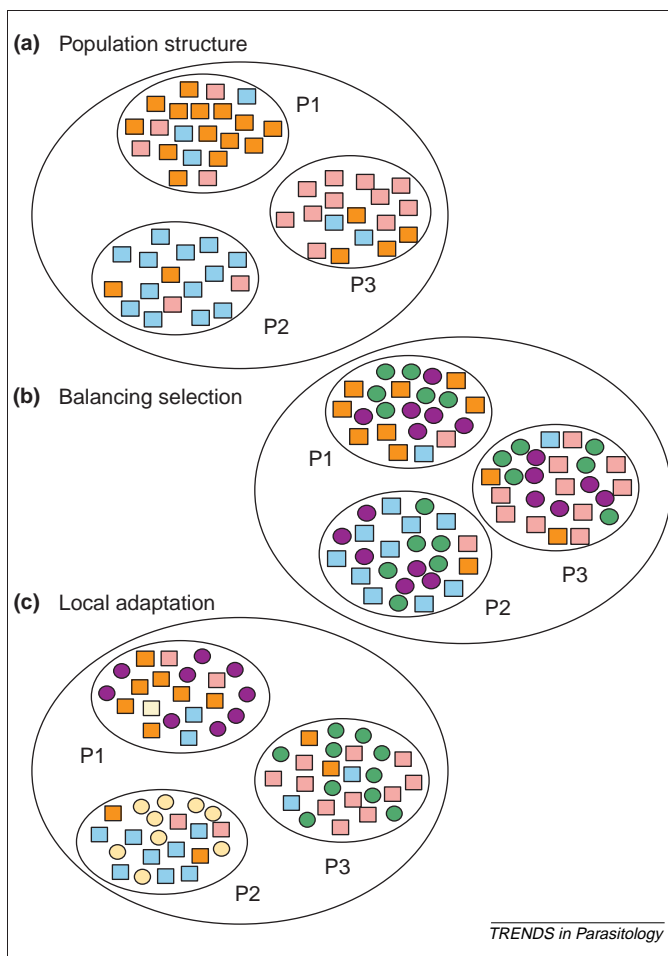


Figure 2. Expected genetic differentiation among parasite populations. The different colored squares represent alleles at a neutral locus and the colored circles are alleles of a selected locus. (a) Under a population genetic structure scenario without selection, populations differentiate as a result of changes in allele frequencies caused by genetic drift. (b) Under a scenario of population structure with a locus under balancing selection, the locus under selection (circles) shows a lower degree of differentiation in its allele frequencies among populations than expected by genetic drift in a neutral locus (squares). (c) By contrast, under the local adaptation scenario, the locus under selection (circles) has a higher degree of differentiation in its allele frequencies among populations than would be expected for a neutral locus (squares) as a result of differentiation by genetic drift.

predicted by population structure [45–47]; this is the opposite pattern to that expected under balancing selection.

Testing for local adaptations is intrinsically difficult in that we need to separate differentiation by selection from the expected differentiation because of the population structure, which is generated by genetic drift. Separating the sources of variation demands very careful sampling designs [45–47]. As an example, a phenotype found worldwide, such as CQ resistance, will be almost impossible to study by using an approach based on F_{ST} . These are the kind of problems faced when explaining whether the genetic differentiation observed in antigens worldwide are local adaptations or an outcome of population structure [18,23].

Other tests that include an explicit use of the phenotype are those based in the comparison of the synonymous and nonsynonymous substitution rates. A synonymous substitution is a nucleotide replacement that does not change the amino acid and a nonsynonymous substitution is a

nucleotide replacement that changes the amino acid. Thus, nonsynonymous substitutions are more often subject to negative natural selection (they change the protein) and are eliminated from the population. Therefore, under neutrality, a faster accumulation of synonymous substitution is expected [49]. An excess of nonsynonymous substitutions can be taken as evidence that these mutations might confer an adaptive advantage because they are maintained by natural selection in the population.

The MK test [48] compares patterns of polymorphism (diversity within species) and divergence (substitutions between species) by subdividing them into synonymous and nonsynonymous substitutions. The substitutions are counted, and the synonymous and nonsynonymous of the divergence is compared with the number of synonymous and nonsynonymous substitutions in the polymorphism by using a 2×2 contingency table. When we compare the synonymous and nonsynonymous polymorphism (within a species), we are comparing the synonymous and nonsynonymous mutation rates, as in the case of divergence (between species). The MK test is very powerful in detecting selection because it is less sensitive to departures of neutrality such as population structure and population growth. The species used to estimate the divergence need to be outside the observed polymorphism; that is, the time to the most recent common ancestor (MRCA) of the alleles within the species should be less than the time to the MRCA in alleles between species; hence, fixed changes between species (divergence) can be used for comparisons. However, the time to the MRCA should not allow the neutral accumulation of nonsynonymous mutations [50] (i.e. very distant species are not useful for comparison). In the case of *P. falciparum*, *Plasmodium reichenowi* can be used to estimate the divergence [17,51], whereas for *Plasmodium vivax* there are several closely related species, such as *Plasmodium cynomolgi* [52]. In addition to choosing an appropriate species for comparison, the test could be affected by changes in the effective population size [53] and does not consider mutations with small effects on fitness, or slightly deleterious mutations that are selected against in large populations, but behave as neutral in small populations [54]. Despite these limitations, we believe that the MK test is still a valuable tool for studying genetic polymorphism. This test has been used for several genes encoding antigens in *P. falciparum*, including AMA-1 [17–19] and erythrocyte-binding antigen (EBA) [21].

The most widely used test for detecting positive natural selection in malaria parasites is the synonymous (Ds) to nonsynonymous (Dn) substitution rates ratio [15,17]. The Dn:Ds ratio has been widely used in other organisms for testing neutrality (for review, see Ref. [55]). It is a simple test, which is little affected by the population demographics and other departures from the neutral model that do not relate with positive selection. Synonymous substitutions are assumed to be neutral, and the null hypothesis of neutrality is rejected if Dn:Ds is >1 . It should be noted that in a gene under negative selection, Dn:Ds is expected to be <1 ; however, this is accounted by the neutral theory. This is a conservative test in that selection should be strong enough to allow for the

accumulation of nonsynonymous substitutions at a higher rate than synonymous substitutions. D_s and D_n are calculated by pairwise comparison in a set of aligned sequences, and $D_n:D_s$ is estimated per pairwise comparison. Significance of the ratio can be tested by bootstrap methods; however, a more adequate way is to compute $D = D_n - D_s$, and its variance which can be approximated to a Z test based on the normal distribution [56]. Significantly, explicit hypotheses are usually tested whenever the $D_n:D_s$ test is used. For example, comparisons between the 5' and 3' portions of the circumsporozoite protein (CSP) gene and studies on the regions encoding its T-cell epitopes in *P. falciparum* [13,20], the domains of *P. falciparum* AMA-1 and MSP-1 [14,18], or specific regions in the thrombospondin-related adhesive protein (TRAP) antigen of *P. vivax* [57]. Thus, using the $D_n:D_s$ test goes beyond a simple ratio calculation on a set of aligned sequences.

In the case of *P. falciparum*, there is a complication in applying the $D_n:D_s$ ratio in that synonymous substitutions could be under selection by the strong codon bias and A + T richness of its genome [17]. Nevertheless, it has been shown that the observed differences cannot be accounted by codon bias under the models used to estimate D_s and D_n to date [13,17].

In the malaria literature, there have been discrepancies on how to estimate the rate of synonymous substitutions per synonymous sites [58–60]. Although these discrepancies have been outside the discussion of testing the null hypothesis of $D_n:D_s \leq 1$, the incongruities highlighted the fact that estimating substitution rates goes beyond simply counting them. Small changes in D_s , for example, could affect $D_n:D_s$. There are several methods that can be used to calculate D_n and D_s , and we studied their performance in four antigens in both *P. falciparum* and *P. vivax*. Extensive reviews of the available methods for calculating D_s and D_n have been published elsewhere [40,56,61]. The methods can be separated into two major categories. First, methods based on simple models that count the number of synonymous and nonsynonymous sites, and then estimate

the rate of nucleotide substitutions per each class of sites correcting with a model, usually Jukes and Cantor or Kimura two parameters [56]. These methods have the advantage of being relatively easy to calculate and demand little computing time. This group will be referred to here as simple models.

The second group of methods uses maximum likelihood (ML) models; they estimate D_n and D_s by considering the codon as unity of observation [61]. The advantage of the ML models is that they account for the differences in nucleotide composition and unequal codon usage, extreme characteristics found in the *P. falciparum* genome. The goal of this comparison is to detect if violations in the assumptions of the simple models have an effect on rejecting the null hypothesis of $D_n:D_s \leq 1$.

Using partial sequences of the malaria antigens AMA-1, MSP-1 (42 kDa), and complete non-repetitive sequences of CSP and TRAP genes of *P. falciparum* and *P. vivax* (all the available sequences in the GenBank by March 2003), we estimated D_n and D_s using simple methods, calculating the mean over all pairwise comparisons and their bootstrap standard errors. In the case of ML methods, we first estimated D_n and D_s for all pairwise comparisons, and then estimated the average. Second, we chose the best model by using likelihood ratio tests. We found that for all four genes and both parasites, the models that fit the data better were always complex such as the general time reverse models [62,63].

How does this finding affect the available evidence that $D_n:D_s$ is > 1 in several malaria antigens? As an example, Table 1 shows the results obtained using three of the methods tested for the *P. falciparum* genes. Overall, all methods (simple models or ML) reached comparable estimates of the rate of nonsynonymous substitutions when used to estimate D_n and D_s on the average of pairwise comparisons (Table 1). However, differences were observed in the rate of synonymous substitutions, especially in *P. falciparum* MSP-1, where D_s was systematically underestimated. Nevertheless, the observed deviations between simple and ML models did not change the

Table 1. Synonymous and nonsynonymous substitutions in *Plasmodium falciparum* genes^a

	$-lnL$	D_s (SE)	D_n (SE)
Nei Gojobori model			
AMA-1	NA	0.001 (0.001)	0.032 (0.007)
42-kDa	NA	0.102 (0.015)	0.071 (0.007)
CSP	NA	0.095 (0.012)	0.011 (0.0016)
TRAP	NA	0.001 (0.001)	0.01 (0.002)
Pamilo-Bianchi-Li model			
AMA-1	NA	0.008 (0.006)	0.033 (0.007)
42-kDa	NA	0.116 (0.020)	0.072 (0.007)
CSP	NA	0.105 (0.015)	0.01205 (0.0019)
TRAP	NA	0.001 (0.001)	0.011 (0.002)
Yang model			
AMA-1	– 575.536 ^b	0.000765 (0.00355)	0.02955 (0.01088)
42-kDa	– 1560.04 ^c	0.299114 (0.35101)	0.06501 (0.071923)
CSP	– 1504.98 ^c	0.081972 (0.06074)	0.01222 (0.01158)
TRAP	– 2043.50 ^c	0.001578 (0.00327)	0.008901 (0.00418)

^aThe rates of D_s and D_n substitution are estimated according to the Nei Gojobori model [64], Pamilo-Bianchi-Li model [65] and Yang model with 60 free-parameters and Ts/Tv fixed (F61, [66]). D_s , D_n and the log likelihood are averaged over the range of pairwise comparisons. Abbreviations: AMA-1, apical membrane antigen 1; CSP, circumsporozoite protein; df, degrees of freedom; D_n , nonsynonymous; D_s , synonymous; $-lnL$, negative natural logarithm of the likelihood or likelihood score; NA, not applicable; SE, standard error; TRAP, thrombospondin-related adhesive protein.

^bLikelihood ratio test is significant ($p < 0.01$, $df = 51$).

^cLikelihood ratio test is significant ($p < 0.001$, $df = 51$).

overall result of the Dn:Ds ratio. In summary, simple models appear to be suitable as a first approximation for testing departures from neutrality on *Plasmodium* genes regardless of the fact that their assumptions are violated.

Perspectives

The analysis of malaria genome sequences will allow the discovery of thousands of new genes and proteins. This abundance of molecular data and the diversity of techniques to detect DNA polymorphisms sometimes make us forget that biomedical researchers, especially epidemiologists, usually investigate phenotypes in time and space. Characterizing the factors leading to their maintenance or emergence unveils complex processes where the parasite genetic diversity interacts with the environment. In such cases, the balance of natural selection and demographic processes plays an important role in defining when, how and where a given phenotypic trait will be maintained, dispersed or disappear.

A mechanistic model, regardless of its simplicity, can always offer the opportunity to explore the consequences and limitations of our observations in a general context. The studies included in this review all address the apparently simple phenomena of genetic variants favored by a selective pressure increasing in frequency, whereas those negatively selected decrease. Indeed, we have shown examples of how natural selection is playing a role in shaping the pattern of polymorphism in parasite populations.

Testing for processes, rather than only looking for associations, opens new possibilities for those interested in public health issues. An integration of laboratory and field base studies, and close coordination between evolutionary biologists and biomedical researchers will provide excellent opportunities to understand the phenotypic variation of malaria.

Acknowledgements

A.A.E. is supported by a grant from The National Institutes of Health (R01 GM60740). O.E.C. and A.R. are supported by fellowships of FONACIT-Venezuela. We thank R. Steketee, A. McCollum and anonymous reviewers for helpful comments.

References

- Gupta, S. and Hill, A.V.S. (1995) Dynamic interactions in malaria: host heterogeneity meets parasite polymorphism. *Proc. R. Soc. Lond. B. Biol. Sci.* 261, 271–277
- Read, A.F. and Taylor, L.H. (2001) The ecology of genetically diverse infections. *Science* 292, 1099–1102
- Kublin, J.G. *et al.* (1998) Molecular assays for surveillance of antifolate-resistant malaria. *Lancet* 351, 1629–1630
- Kublin, J.G. *et al.* (2002) Molecular markers for failure of sulfadoxine-pyrimethamine and chlorproguanil-dapsone treatment of *Plasmodium falciparum* malaria. *J. Infect. Dis.* 185, 380–388
- Haldane, J.B.S. (1949) Disease and evolution. *La Ricerca Scientifica. Supplemento A* 19, 2–11
- Livingstone, F.B. (1984) The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. *Hum. Biol.* 56, 413–425
- Aidoo, M. *et al.* (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* 359, 1311–1312
- Hobbs, M.R. *et al.* (2002) A new NOS2 promoter polymorphism associated with increased nitric oxide production and protection from severe malaria in Tanzanian and Kenyan children. *Lancet* 360, 1468–1475
- Zimmerman, P.A. *et al.* (1999) Emergence of FY*A(null) in a *Plasmodium vivax*-endemic region of Papua New Guinea. *Proc. Natl. Acad. Sci. U. S. A.* 96, 13973–13977
- Terlouw, D.J. *et al.* (2002) Increased efficacy of sulfadoxine-pyrimethamine in the treatment of uncomplicated falciparum malaria among children with sickle cell trait in Western Kenya. *J. Infect. Dis.* 186, 1661–1668
- McCutchan, T.F. *et al.* (1988) Antigenic diversity in *Plasmodium falciparum*. *Prog. Allergy* 41, 173–192
- Udhayakumar, V. *et al.* (1995) Identification of T and B cell epitopes recognized by humans in the C-terminal 42-kDa domain of the *Plasmodium falciparum* merozoite surface protein (MSP)-1. *J. Immunol.* 154, 6022–6030
- Hughes, A.L. (1991) Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. *Genetics* 127, 345–353
- Hughes, A.L. (1992) Positive selection and intrallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. *Mol. Biol. Evol.* 9, 381–393
- Hughes, M.K. and Hughes, A.L. (1995) Natural selection on *Plasmodium* surface proteins. *Mol. Biochem. Parasitol.* 71, 99–113
- Conway, D.J. (1997) Natural selection on polymorphic malaria antigens and the search for a vaccine. *Parasitol. Today* 13, 26–29
- Escalante, A.A. *et al.* (1998) Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. *Genetics* 149, 189–202
- Escalante, A.A. *et al.* (2001) Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of *Plasmodium falciparum* VI Asembo bay cohort project. *Mol. Biochem. Parasitol.* 113, 279–287
- Polley, S.D. and Conway, D.J. (2001) Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics* 158, 1505–1512
- Escalante, A.A. *et al.* (2002) A study of genetic diversity in the gene encoding the circumsporozoite protein (CSP) of *Plasmodium falciparum* from different transmission areas. *Mol. Biochem. Parasitol.* 125, 83–90
- Baum, J. *et al.* (2003) Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. *Genetics* 163, 1327–1336
- Cole-Tobian, J. and King, C.L. (2003) Diversity and natural selection in *Plasmodium vivax* Duffy binding protein gene. *Mol. Biochem. Parasitol.* 127, 121–132
- Cortes, A. *et al.* (2003) Geographical structure of diversity and differences between symptomatic and asymptomatic infections for *Plasmodium falciparum* vaccine candidate AMA1. *Infect. Immunol.* 71, 1416–1426
- Wootton, J.C. *et al.* (2002) Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* 418, 320–323
- Nair, S. *et al.* (2003) A selective sweep driven by pyrimethamine treatment in Southeast Asian malaria parasites. *Mol. Biol. Evol.* 20, 1526–1536
- Tibayrenc, M. and Ayala, F.J. (2002) The clonal theory of parasitic protozoa: 12 years on. *Trends Parasitol.* 18, 405–410
- Gupta, S. *et al.* (1994) Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. *Science* 263, 961–963
- Gupta, S. *et al.* (1996) The maintenance of strain structure in populations of recombining infectious agents. *Nat. Med.* 2, 437–442
- Anderson, T.J. *et al.* (2000) Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol.* 17, 1467–1482
- Hawley, W.A. *et al.* (2003) Implications of the western Kenya permethrin-treated bed net study for policy, program implementation, and future research. *Am. J. Trop. Med. Hyg.* 68, 168–173
- Branch, O.H. *et al.* (1998) A longitudinal investigation of IgG and IgM antibody responses to the merozoite surface protein-1 19-kiloDalton domain of *Plasmodium falciparum* in pregnant women and infants: associations with febrile illness, parasitemia, and anemia. *Am. J. Trop. Med. Hyg.* 58, 211–219
- Conway, D.J. *et al.* (2000) A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. *Nat. Med.* 6, 689–692
- Aidoo, M. and Udhayakumar, V. (2000) Field studies of cytotoxic T lymphocytes in malaria infections: implications for malaria vaccine development. *Parasitol. Today* 16, 50–56

- 34 Galtier, N. *et al.* (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* 155, 981–987
- 35 Smith, J.M. and Haigh, J. (1974) The hitch-hiking effect of a favorable gene. *Genet. Res.* 23, 23–35
- 36 Mazier, D. *et al.* (2000) Cerebral malaria and immunogenetics. *Parasite Immunol.* 22, 613–623
- 37 Fidock, D.A. *et al.* (2000) Mutations in the *Plasmodium falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* 6, 861–871
- 38 Sirawaraporn, W. *et al.* (1997) Antifolate-resistant mutants of *Plasmodium falciparum* dihydrofolate reductase. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1124–1129
- 39 Hastings, I.M. and Mackinnon, M.J. (1998) The emergence of drug-resistant malaria. *Parasitology* 117, 411–417
- 40 Graur, D. and Li, W.H. (2000) *Fundamentals of Molecular Evolution*, 2nd edn, Sinauer
- 41 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
- 42 Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286
- 43 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595
- 44 Fu, Y.X. and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* 133, 693–709
- 45 Lewontin, R.C. and Krakauer, J. (1975) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74, 175–195
- 46 Taylor, M.F. *et al.* (1995) A population genetic test of selection at the molecular level. *Science* 270, 1497–1499
- 47 Kohn, M.H. *et al.* (2003) Locus-specific genetic differentiation at Rw among warfarin-resistant rat (*Rattus norvegicus*) populations. *Genetics* 164, 1055–1070
- 48 MacDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654
- 49 Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267, 275–276
- 50 Smith, J.M. (1994) Estimating selection by comparing synonymous and substitutional changes. *J. Mol. Evol.* 39, 123–128
- 51 Escalante, A. and Ayala, F.J. (1994) Phylogeny of the malarial genus *Plasmodium*, derived from ribosomal gene sequences. *Proc. Natl. Acad. Sci. U. S. A.* 91, 11373–11377
- 52 Escalante, A.A. *et al.* (1998) The evolution of primate malaria parasites based on the gene encoding Cytochrome *b* from the linear mitochondrial genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8124–8129
- 53 Eyre-Walker, A. (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics* 162, 2017–2024
- 54 Ohta, T. (2002) Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16134–16137
- 55 Hughes, A.L. (1999) *Adaptive Evolution of Genes and Genomes*, Oxford University Press
- 56 Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*, Oxford University Press
- 57 Putaporntip, C. *et al.* (2001) Diversity in the thrombospondin-related adhesive protein gene (TRAP) of *Plasmodium vivax*. *Gene* 268, 97–104
- 58 Hughes, A.L. and Verra, F. (1998) Ancient polymorphism and the hypothesis of a recent bottleneck in the malaria parasite *Plasmodium falciparum*. *Genetics* 150, 511–513
- 59 Rich, S.M. and Ayala, F.J. (1998) The recent origin of allelic variation in antigenic determinants of *Plasmodium falciparum*. *Genetics* 150, 515–517
- 60 Saul, A. (1999) Circumsporozoite polymorphisms, silent mutations and the evolution of *Plasmodium falciparum*. *Parasitol. Today* 15, 38–40
- 61 Yang, Z. and Bielawski, J.P. (2000) Statistical Methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503
- 62 Lanave, C.G. *et al.* (1984) A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20, 86–93
- 63 Rodriguez, F. *et al.* (1990) The general stochastic model of nucleotide substitutions. *J. Theor. Biol.* 142, 485–501
- 64 Nei, M. and Gojobori, T. (1986) Simple methods for estimating the number of synonymous and non synonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426
- 65 Li, W.H. (1993) Unbiased estimation of the rates of synonymous and non synonymous substitutions. *J. Mol. Evol.* 36, 96–99
- 66 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556

Reproduction of material from Elsevier articles

Interested in reproducing part or all of an article published by Elsevier, or one of our article figures? If so, please contact our *Global Rights Department* with details of how and where the requested material will be used. To submit a permission request on-line, please visit:

http://www.elsevier.com/wps/find/obtainpermissionform.cws_home/obtainpermissionform

Alternatively, please contact:

Elsevier
Global Rights Department
PO Box 800,
Oxford OX5 1DX, UK.
Phone: (+44) 1865-843830
Fax: (+44) 1865-853333
permissions@elsevier.com