THE ROYAL SOCIETY

doi 10.1098/rspb.2001.1759

Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*

Austin L. Hughes[1*] **and Federica Verra**[2]

[1]*Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA*
[2]*Istituto di Parassitologia, Università di Roma 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy*

It has been proposed that the virulent human malaria parasite *Plasmodium falciparum* underwent a recent severe population bottleneck. In order to test this hypothesis, we estimated the effective population size of this species from the patterns of nucleotide substitution at 23 nuclear protein-coding loci, using a variety of methods based on coalescent theory. Both simple methods and phylogenetically based maximum-likelihood methods yielded the conclusion that the effective population size of this species has been of the order of at least 10^5 for the past 300 000–400 000 years.

Keywords: effective population size; genetic polymorphism; malaria; *Plasmodium falciparum*

1. INTRODUCTION

The malaria parasite *Plasmodium falciparum* is a major human disease agent, responsible for approximately two million deaths annually worldwide (World Health Organization 1997). Rich *et al*. (1998) proposed that this species underwent a recent severe population bottleneck and that all living members of the species are descended from a single haploid genotype that lived between 24 500 and 57 500 years ago. The hypothesis that *P. falciparum* underwent a recent severe bottleneck has significant implications for attempts to control this important human pathogen. A species descended from a single ancestor in the past few thousand years is expected to have an extremely low level of genetic polymorphism. If this were true of *P. falciparum*, it would imply that there is negligibly small pre-existing genetic variation in the parasite population with respect to responses to vaccines and other therapeutic agents, considerably facilitating the task of control. On the other hand, strategies for the control of a genetically highly polymorphic pathogen will necessarily be more complex.

Coalescent theory provides methods for estimating effective population size (N_e) on the basis of neutral genetic polymorphism (Watterson 1975; Tajima 1983; Nei 1987; Felsenstein 1992; Kuhner *et al*. 1995, 1998). We applied a number of these methods to DNA-sequence data from alleles at 23 nuclear protein-coding loci of *P. falciparum* in order to test the hypothesis of a recent population bottleneck. We chose only loci for which there was no evidence of positive Darwinian selection. Positive Darwinian selection is expected to influence the extent of nucleotide polymorphism at a locus in two different ways, depending on the nature of the selection. In the case of balancing selection (such as overdominant selection) polymorphism is maintained for much longer than in the case of selective neutrality (Takahata & Nei 1990). As a result, alleles will accumulate a large number of nucleotide differences, as seen in the case of the major-histocompatibility-complex genes of vertebrates (Hughes & Hughes 1995*a*). On the other hand, directional selection, in which a single allele is favoured, will lead to a reduction in polymorphism in the region of a selectively favoured mutation because linked sites will 'hitch-hike' along with the selected site, a phenomenon referred to as a 'selective sweep' (Charlesworth 1992).

Using only loci at which we found no evidence of either type of positive selection, we applied methods for estimating N_e both to all nucleotide sites and to synonymous (silent) sites alone. Because synonymous substitutions are likely to be selectively neutral or nearly so in most organisms (Nei 1987), estimates of N_e on the basis of synonymous sites would seem to be preferable to those based on all sites in coding regions. However, in this case, the results using synonymous sites and those using all nucleotide sites showed only slight differences (see § 3); this suggests that much of the non-synonymous polymorphism observed at the loci we analysed is selectively neutral or nearly so.

2. METHODS

(a) *Sequences analysed*

We analysed published sequence data for partial or complete coding regions from two or more alleles at 23 loci for which we could find no evidence of positive Darwinian selection (Genbank accession numbers in table 1). As a criterion of positive Darwinian selection at a locus, we compared the number of synonymous nucleotide substitutions per synonymous site (d_S) with the number of non-synonymous substitutions per non-synonymous site (d_N) (Nei & Gojobori 1986). When d_N in the whole coding region, or in a particular domain of the gene, significantly exceeded d_S, we took this as evidence of positive selection favouring amino-acid replacements (Hughes & Nei 1988). In this case, since only comparisons within *P. falciparum* were involved, positive selection could take one of two forms: balancing selection favouring polymorphism at the amino-acid level or recent directional selection favouring an allele that is not yet fixed. A pattern of $d_N > d_S$ has previously been reported for a number of loci encoding surface proteins of *P. falciparum* that are immunogenic to the host, suggesting that host immune

*Author for correspondence (austin@biol.sc.edu).

Received 12 April 2001 *Accepted* 1 June 2001

Table 1. Summary of nucleotide polymorphisms at *Plasmodium falciparum* loci.

locus	number of alleles	number of codons compared	segregating sites	segregating sites (four-fold degenerate)	π_S	π_N	d_{Smax}	θ (all sites)	θ (four-fold degenerate sites)
aldolase	3	326	14	2	0.0106	0.0081	0.0159	0.0086	0.0086
calmodulin	2	146	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
cyclophilin 1	2	210	1	0	0.0000	0.0020	0.0000	0.0016	0.0000
DHFR-TS[a]	29	220	3	0	0.0000	0.0027	0.0000	0.0012	0.0000
EBA-175	2	1435	18	0	0.0037	0.0043	0.0037	0.0042	0.0000
EBL-1[a]	2	699	2	0	0.0024	0.0024	0.0024	0.0024	0.0000
enolase	2	446	4	0	0.0068	0.0019	0.0068	0.0030	0.0000
falcipain 2	7	484	8	4	0.0060	0.0014	0.0141	0.0022	0.0013
falcipain 3	2	488	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
GBP[a]	2	199	4	1	0.0084	0.0000	0.0084	0.0067	0.0189
GLURP[a]	43	416	24	0	0.0019	0.0045	0.0088	0.0044	0.0000
GRP78	2	652	13	2	0.0095	0.0059	0.0095	0.0066	0.0086
G6PD	2	734	10	0	0.0057	0.0043	0.0057	0.0045	0.0000
HSP60	2	577	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
HSP90	3	745	2	0	0.0000	0.0007	0.0000	0.0006	0.0000
LSA-1	20	280	18	0	0.0047	0.0012	0.0063	0.0060	0.0000
ookinete antigen	2	217	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
Pf27/25	2	217	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
RAP-1	7	782	7	1	0.0013	0.0027	0.0041	0.0026	0.0017
SOD[a]	27	198	14	1	0.0025	0.0021	0.0343	0.0061	0.0046
SSA[a]	6	191	6	0	0.0028	0.0051	0.0085	0.0046	0.0000
STARP	2	559	11	0	0.0029	0.0076	0.0029	0.0066	0.0000
TPI	2	222	0	0	0.0000	0.0000	0.0000	0.0000	0.0000
mean ± s.e.m.					0.0030 ± 0.0007	0.0026 ± 0.0005	0.0057 ± 0.0016	0.0031 ± 0.0006	0.0019 ± 0.0009

[a] Indicates partial coding region.

Genbank accession numbers: aldolase (AF179421, J03084, M28881); calmodulin (M59349, M99442); cyclophilin 1 (AF177281, U10322); DHFR-TS (AF248503–18, AF248527–29, AF248531–5, AF248537, J03028, J03772, J04043, M22159); EBA-175 (AF258781, X52524); EBL-1 (AF131999, L38450); enolase (AB026051, U00152); falcipain 2 (AF239801, AF251193, AF282975–9); falcipain 3 (AF258791, AF282974); GBP (M12897, M27438); GLURP (AF247634, AJ269898–901, AJ269903–9, AJ269911–40, M59706); GRP78 (L02822, X69121); G6PD (M80655, X74988); HSP60 (U38963, U94594); HSP90 (L34027–8, Z29667); LSA-1 (L40834–7, L40884–93, L40908–10, L40947, X56203, Z30320); ookinete antigen (AF154117, X07802, *Plasmodium reichenowi* M36915); Pf27/25 (AF179422, X84904); RAP-1 (AF205282–4, AF206631, M32853, M80807, U20985); SOD (AF113142–67, Z49819); SSA (AF177634–6, AF206630, X81648, Z22145, *P. reichenowi* L33882); STARP (AF209925, Z26314, *P. reichenowi* Z30339); TPI (L01654–5).

recognition maintains a balanced polymorphism at these loci (Hughes 1991, 1992; Hughes & Hughes 1995*b*; Verra & Hughes 2000). In examining loci for possible inclusion in the present analyses, we found mean d_N to be significantly greater than mean d_S at three additional loci: SPAM (seven alleles; mean ± s.e.m. $d_S = 0.0000 \pm 0.0000$, mean ± s.e.m. $d_N = 0.0059 \pm 0.0019$, $p < 0.01$), chloroquine resistance transporter (five alleles; mean ± s.e.m. $d_S = 0.0000 \pm 0.0000$, mean ± s.e.m. $d_N = 0.0056 \pm 0.0017$, $p < 0.001$) and dihydropteroate synthetase (34 alleles; mean ± s.e.m. $d_S = 0.0000 \pm 0.0000$, mean ± s.e.m. $d_N = 0.0028 \pm 0.0013$, $p < 0.05$). At the latter two loci there is presumably ongoing directional selection of recent origin in response to human chemotherapeutic agents (Basco *et al*. 2000).

(b) *Divergence time and mutation rate*

We use the notations π_S and π_N to denote, respectively, the mean pairwise numbers of synonymous and non-synonymous substitutions per site among alleles at a locus. The mean time to coalescence (last common ancestor) of pairs of alleles at a locus can be estimated as $\pi_S/2\lambda_S$, where λ_S is the rate of synonymous substitution (per site per year). We estimated λ_S and μ (the neutral-mutation rate per site per generation) by comparison of synonymous sites between genes in our sample (ookinete antigen, Pf27/25, RAP-1 and SSA) for which sequences were available for the chimpanzee malaria parasite *Plasmodium reichenowi* (Genbank accession numbers provided in table 1). Phylogenetic analyses have shown that *P. falciparum* and *P. reichenowi* are sister taxa (Escalante & Ayala 1994; Hughes & Hughes 1995*b*; Escalante *et al*. 1998), and are consistent with the hypothesis that these two species diverged when their host species diverged (Escalante *et al*. 1998). Mean ± s.e.m. d_S between *P. falciparum* and *P. reichenowi* for the four loci was 0.0511 ± 0.0083. On the basis of life-cycle and epidemiological information (Molineaux 1988), we used three generations per year as a conservative estimate of the average generation time in *P. falciparum*. Using five million years ago as the divergence of humans and chimpanzees (and thus of *P. falciparum* and *P. reichenowi*), λ_S was estimated at 5.1×10^{-9} substitutions per site per year, and μ was estimated at 1.7×10^{-9} substitutions per site per generation. If chimpanzees and humans diverged seven million years ago, then λ_S was estimated at 3.9×10^{-9} substitutions per site per year, and μ was estimated at 1.2×10^{-9}

Table 2. Estimates of effective population size (N_e) in *Plasmodium falciparum* for constant population-size models.
(Estimates of θ and N_e are given as mean ± s.e.m.)

		N_e	
method	θ	$\mu = 1.2 \times 10^{-9}$	$\mu = 1.7 \times 10^{-9}$
equation (2.1)	—	$8.39 \pm 2.04 \times 10^5$	$5.93 \pm 1.44 \times 10^5$
equation (2.2), all sites	0.0031 ± 0.0006	$6.46 \pm 1.25 \times 10^5$	$4.56 \pm 0.88 \times 10^5$
equation (2.2), four-fold degenerate sites	0.0019 ± 0.0009	$3.96 \pm 1.88 \times 10^5$	$2.79 \pm 1.32 \times 10^5$
maximum likelihood, all sites	0.0056 ± 0.0012	$1.17 \pm 0.25 \times 10^6$	$8.24 \pm 1.76 \times 10^5$
maximum likelihood, four-fold degenerate sites	0.0092 ± 0.0040	$1.91 \pm 0.83 \times 10^6$	$1.35 \pm 0.59 \times 10^6$

substitutions per site per generation. To provide upper and lower bounds for λ_S and μ, both values were used in our analyses.

(c) *Estimation of N_e*

The coalescent theory predicts the following relationship:

$$\bar{t} = 4N_e(1 - 1/n), \tag{2.1}$$

where $\bar{t}$ is the mean number of generations to the coalescence time of n alleles at a locus sampled at random from the population (Tajima 1983; Nei 1987, p. 395). For each locus, we estimated the coalescence time using d_{Smax} (the d_S value between the most distant pair of alleles) and μ, and then computed N_e from equation (2.1). For an overall estimate of N_e, the mean of the 23 individual N_e estimates was used. The parameter $\theta = 4N_e\mu$ can be estimated as follows:

$$\hat{\theta} = s^*/(a_1 - c_2 s^*), \tag{2.2}$$

where s^* is the minimum number of mutations per nucleotide site, n is the number of sequences examined, $a_1 = \sum_{i=1}^{n-1}(1/i)$ and $a_2 = \sum_{i=1}^{n-1}(1/i^2)$, $a_3 = \frac{1}{2}(a_1^2 - a_2)$ and $c_2 = (4a_1/3) - (7a_3/3a_1)$ (Tajima 1996). Tajima (1996) and Misawa & Tajima (1997) give modified versions of equation (2.2) taking into account rate variation between sites, assuming that rates vary according to a gamma distribution. In preliminary analyses, we applied these methods after estimating the gamma parameter following Tamura & Nei (1993); however, the results were virtually identical to those obtained using equation (2.2). This was expected since the correction for rate variation between sites has negligible effect if θ is low (Misawa & Tajima 1997), as was true in this case. Therefore, we present results using equation (2.2). This method does not assume the infinite-sites model, which is not strictly applicable to DNA-sequence data (Tajima 1996).

In addition to equations (2.1) and (2.2), we used the maximum-likelihood methods incorporated in the FLUCTUATE program of Kuhner *et al.* (1998) to estimate θ. This program provides both a method assuming constant N_e, the results of which are directly comparable to the results of equations (2.1) and (2.2), and a method in which exponential population growth or decline is assumed. The latter simultaneously estimates both θ and g, where g is the rate of population growth (or decline) per μ per generation. Because these methods are based on a phylogeny, a minimum of three alleles at a locus is required.

3. RESULTS

At the 23 loci for which the hypothesis of neutral evolution could not be rejected, 10 479 codons were examined (table 1). Mean ± s.e.m. π_S for the 23 loci was 0.0030 ± 0.0007 (table 1), a value significantly different from zero (*t*-test, $p < 0.001$). Mean ± s.e.m. π_N (0.0026 ± 0.0005) was slightly lower than mean π_S, although the difference was not statistically significant (paired-sample *t*-test). Assuming five million years for the *P. falciparum–P. reichenowi* divergence, the mean ± s.e.m. pairwise divergence time of alleles at these loci was estimated at $2.94 \pm 0.69 \times 10^5$ years. Assuming seven million years for the *P. falciparum–P. reichenowi* divergence, the mean ± s.e.m. pairwise divergence time was estimated at $3.85 \pm 0.90 \times 10^5$ years. These times are an order of magnitude or more lower than the coalescence times that have been estimated for loci in *P. falciparum* that are subject to balancing selection arising from host immune recognition (Hughes 1992; Hughes & Verra 1998; Verra & Hughes 2000). Thus, these estimates of mean coalescence time are consistent with the hypothesis that polymorphisms at the 23 loci analysed are selectively neutral or nearly so.

Applying equations (2.1) and (2.2) to the 23 loci produced estimates of N_e between about 300 000 and 800 000 (table 2). Equation (2.1) provided slightly higher estimates of N_e than equation (2.2), while applying equation (2.2) to four-fold degenerate sites provided slightly lower estimates than applying equation (2.2) to all sites (table 2).

Because the maximum-likelihood methods require at least three polymorphic sequences, they could be applied only to a smaller set of loci than the other methods. There were nine loci for which the maximum-likelihood method could be applied to all sites, and only four loci for which maximum-likelihood methods could be applied to four-fold degenerate sites (table 3). Table 2 shows estimates of N_e on the basis mean θ values for the maximum-likelihood method assuming a constant population size. Using either all sites or only four-fold degenerate sites, this method estimated N_e at around one million (table 2). Thus, these estimates of N_e were somewhat higher than those based on equations (2.1) and (2.2) (table 2). Because the loci to which maximum-likelihood methods could be applied were among the most polymorphic loci (tables 1 and 3), this method may have somewhat overestimated N_e.

For the maximum-likelihood method allowing for population growth or decline, estimates of g were greater than zero for all nine loci used in analyses based on all sites and for three out of four loci used in analyses based on four-fold degenerate sites (table 3). Using *z*-tests based on the approximate estimates of the standard error of g for each locus, g was significantly different from zero in six out of nine loci used in analyses based on all nucleotide sites

Table 3. Estimates of θ using maximum-likelihood models.

(θ_C: maximum-likelihood estimate of θ assuming constant population size; θ_V: maximum-likelihood estimate of θ in model allowing for exponential population growth or decline.)

	all sites			four-fold degenerate sites		
locus	θ_C	θ_V	g	θ_C	θ_V	g
aldolase	0.0076	0.0123	151	0.0075	0.0167	197
DHFR-TS	0.0015	0.0021	1365*	—	—	—
falcipain 2	0.0021	0.0024	352	0.0203	0.0118	−1
GLURP	0.0108	0.0248	1031**	—	—	—
HSP90	0.0005	0.0013	6089	—	—	—
LSA-1	0.0101	0.1768	3518**	—	—	—
RAP-1	0.0055	0.1282	4216**	0.0014	0.0455	7810
SSA	0.0054	0.0296	1271**	—	—	—
SOD	0.0073	0.0770	1692**	0.0075	0.0211	669
Mean ± s.e.m.	$0.0057 \pm 0.0012^{\dagger\dagger}$	$0.0505 \pm 0.0212^{\dagger}$	$2187 \pm 664^{\dagger}$	0.0092 ± 0.0040	0.0238 ± 0.0075	2169 ± 1886

z-tests of the hypothesis that $g = 0$: $^{*}p < 0.05$, $^{**}p < 0.001$; t-tests of the hypothesis that mean $= 0$: $^{\dagger}p < 0.05$, $^{\dagger\dagger}p < 0.005$.

and in two out of four loci used in analyses based on four-fold degenerate sites (table 3). In addition, the mean estimate of g for the nine loci for which all sites were used was significantly different from zero (table 3). Thus, overall, the results supported the hypothesis that the population of *P. falciparum* has been increasing.

The mean estimate of θ assuming population growth was slightly higher when all sites were used, but the mean estimate of g was very similar (table 3). Given our estimates of μ, the mean estimates of g were expressed as estimates of the instantaneous growth rate per generation (r) (table 4). Using these data, the current N_e for *P. falciparum* was estimated to be of the order of 10^6–10^7. Extrapolating backwards in time and assuming exponential population growth, estimates for N_e were of the order of 10^6 200 000 years ago and of the order of 10^5 300 000–400 000 years ago (table 4). As in the case of equation (2.2) (table 2), slightly lower estimates of N_e were obtained when only four-fold degenerate sites were used than when all sites were used (table 4). Note that estimates for 300 000 years ago showed good agreement with the estimates based on equation (2.1). Since the mean coalescence time for alleles at the 23 loci was around 300 000 years ago, this agreement revealed remarkable consistency between these different methods of estimating N_e.

Table 4. Maximum-likelihood estimates of effective population size (N_e) assuming exponential growth.

(Estimates of r and N_e are given in the form estimate ± standard error.)

	$\mu = 1.2 \times 10^{-9}$	$\mu = 1.7 \times 10^{-9}$
all sites		
r	$2.62 \pm 0.80 \times 10^{-6}$	$3.72 \pm 1.13 \times 10^{-6}$
N_e current	$1.05 \pm 0.44 \times 10^{7}$	$7.46 \pm 3.14 \times 10^{6}$
N_e − 200 000 years	$2.18 \pm 0.92 \times 10^{6}$	$2.44 \pm 1.03 \times 10^{6}$
N_e − 300 000 years	$9.93 \pm 4.17 \times 10^{5}$	$8.01 \pm 3.37 \times 10^{5}$
N_e − 400 000 years	$4.53 \pm 1.90 \times 10^{5}$	$2.81 \pm 1.18 \times 10^{4}$
four-fold degenerate sites		
r	$2.60 \pm 2.26 \times 10^{-6}$	$3.69 \pm 3.21 \times 10^{-6}$
N_e current	$4.96 \pm 1.56 \times 10^{6}$	$3.50 \pm 1.10 \times 10^{6}$
N_e − 200 000 years	$1.04 \pm 0.33 \times 10^{6}$	$3.82 \pm 1.20 \times 10^{5}$
N_e − 300 000 years	$4.78 \pm 1.51 \times 10^{5}$	$1.26 \pm 0.40 \times 10^{5}$
N_e − 400 000 years	$2.19 \pm 0.69 \times 10^{5}$	$4.18 \pm 1.32 \times 10^{4}$

4. DISCUSSION

Our results revealed substantial genetic polymorphism in *P. falciparum* at both synonymous and non-synonymous nucleotide sites (table 1). Rich *et al*. (1998) based their conclusion of a recent bottleneck in *P. falciparum* on their observation of no synonymous differences in a small sample of loci. By contrast, in our sample, non-zero values of d_S were obtained for 14 out of 23 loci, and some polymorphism was observed at all but six loci (table 1). A similar study of human polymorphism reported synonymous differences at only eight out of 49 loci examined (Li & Sadler 1991). Values of nucleotide diversity reported for humans on the basis of allelic sequence comparisons at various loci are in the range 0.0005–0.002 (Li & Sadler 1991; Clark *et al*. 1998; Fullerton *et al*. 2000; Przeworski *et al*. 2000). A recent estimate based on single nucleotide polymorphisms across the genome showed a mean of *ca*. 0.0006 (International SNP Map Working Group 2001). Thus, our results for synonymous sites show about a five-fold greater level of nucleotide diversity in *P. falciparum* than in humans.

All methods used indicated that the effective population size of *P. falciparum* is quite high and has been of the order of 10^5 for at least the past 300 000 years. Thus, no support was obtained for the hypothesis of a recent worldwide severe population bottleneck (Rich *et al*. 1998). Our estimates of N_e depend on our estimates of λ_S and μ, which in turn depend on the assumption that *P. falciparum* and *P. reichenowi* diverged around the time that humans and chimpanzees diverged. Escalante *et al*.'s (1998) analysis of *Plasmodium* mitochondrial cytochrome *b* sequences provides support for this hypothesis. The extent of nucleotide divergence between *P. falciparum* and *P. reichenowi* in the cytochrome *b* gene is consistent with the species having diverged at around the time of the human–chimpanzee divergence, given a mutation rate similar to that of other

Plasmodium species parasitic on Asian primates whose dates of radiation have been estimated from the fossil record (Escalante *et al*. 1998). Escalante & Ayala (1994) reached a similar conclusion based on comparison of rRNA sequences. Note also that our estimates of λ_S are similar to those for other eukaryotic nuclear genes (Li 1997).

The assumption of exponential population growth or decline made by Kuhner *et al*.'s (1998) method is no doubt unrealistic in many cases, quite possibly including *P. falciparum*. None the less, there was good agreement between the results produced using this method and methods assuming a constant population size (tables 2 and 4). If there has been population fluctuation, the latter methods provide an estimate of the long-term N_e, taking into account the changes in population size. A long-term N_e of *ca*. 10^5 over the past 300 000–400 000 years provides a good fit with the results of both methods (tables 2 and 4). By the maximum-likelihood method assuming population growth, we estimated that N_e has increased from *ca*. 10^5 to *ca*. 10^7 over the past 300 000–400 000 years (table 4), but in that case the long-term effective population size over the whole time interval is closer to the lower number (Nei 1987, p. 362).

There is some evidence that recent population growth in *P. falciparum* has been episodic rather than exponential. It has long been speculated that the population size of *P. falciparum* may have increased sharply at the time of the introduction of agriculture in West Africa, about 6000 years ago (Livingstone 1958). Presumably this expansion occurred both because of the expansion of the human population resulting from agriculture and because environmental changes caused by agriculture increased the habitat for mosquito vectors. Moreover, the development of agriculture may have been a key event in the emergence of anthropophilic taxa within the vector species complexes *Anopheles funestus* and *Anopheles gambiae*, which in turn may have led to an increase in the transmission rate (Coluzzi 1999). More recently, *P. falciparum* is believed to have spread from its place of origin to other parts of the world, a process in which the African slave trade evidently played a role. Analysis of microsatellite data from *P. falciparum* populations in different geographic regions has provided evidence of this most recent expansion (Anderson *et al*. 2000). The data presented here are consistent with a recent sharp population expansion in a species that had a substantial long-term effective population size prior to the expansion.

There is evidence at a number of polymorphic loci encoding immunogenic surface proteins of *P. falciparum* that polymorphisms have been maintained by balancing selection over millions of years (Hughes & Verra 1998; Verra & Hughes 2000). Given a sufficiently large N_e, long-term maintenance of polymorphism is expected under balancing selection (Takahata & Nei 1990). However, the evidence of ancient selectively maintained polymorphisms in *P. falciparum* is not consistent with the hypothesis that worldwide populations of this species have a recent common ancestor (Hughes & Verra 1998). The present estimates of N_e in *P. falciparum*, on the other hand, are easily compatible with long-term maintenance of balanced polymorphisms.

In summary, we found that a variety of methods for estimating effective population size from DNA-sequence data indicated a substantial long-term effective population size in *P. falciparum*. Remarkable agreement between different methods provided robust support for this conclusion. As a consequence, we predict the overall extent of genetic polymorphism in this species to be substantial. Of course, due to genetic drift or a recent selective sweep, a given genomic region may show little or no polymorphism. Likewise, if a recent population bottleneck occurred during the colonization by *P. falciparum* of a particular geographic region, then polymorphism within that region may be low (Maitland *et al*. 2000). Similarly, spread of a genotype in an isolated outbreak may lead to a substantial local reduction in genetic diversity (Arez *et al*. 1999). None the less, our results suggest that, worldwide, *P. falciparum* will reveal a high degree of genetic diversity, particularly in Africa, where the species originated. This polymorphism is likely to include currently neutral but potentially selectable polymorphisms conferring an ability to respond to selection, including selection imposed by human therapeutic agents.

This research was supported by grant GM34940 from the National Institutes of Health to A.L.H.

REFERENCES

Anderson, T. J. (and 15 others) 2000 Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol*. **17**, 1467–1482.

Arez, A. P., Snounou, G., Pinto, J., Sousa, C. A., Modiano, D., Ribeiro, H., Franco, A. S., Alves, J. & Do Rosario, V. E. 1999 A clonal *Plasmodium falciparum* in an isolated outbreak of malaria in the Republic of Cabo Verde. *Parasitology* **120**, 335–343.

Basco, L. K., Tahar, R., Keundjian, A. & Ringwald, P. 2000 Sequence variations in the genes encoding dihydropteroate synthase and dihydrofolate reductase and clinical response to sulfadoxine–pyrimethamine in patients with acute uncomplicated falciparum malaria. *J. Infect. Dis*. **182**, 624–628.

Charlesworth, B. 1992 New genes sweep clean. *Nature* **356**, 475–476.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengård, J., Salomaa, E., Perola, M., Boerwinkle, E. & Sing, C. F. 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet*. **63**, 595–612.

Coluzzi, M. 1999 The clay feet of the malaria giant and its African roots: hypotheses and inferences about origin, spread and control of *Plasmodium falciparum*. *Parassitologia* **41**, 277–283.

Escalante, A. A. & Ayala, F. J. 1994 Phylogeny of the malarial genus *Plasmodium* derived from rRNA gene sequences. *Proc. Natl Acad. Sci. USA* **91**, 11 372–11 377.

Escalante, A. A., Freeland, D. E., Collins, W. E. & Lal, A. A. 1998 The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. *Proc. Natl Acad. Sci. USA* **95**, 8124–8129.

Felsenstein, J. 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res*. **59**, 139–147.

Fullerton, S. M., Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Stengård, J., Salomaa, V., Perola, M., Boerwinkle, E. & Sing, C. F. 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of human polymorphism. *Am. J. Hum. Genet*. **67**, 881–900.

Hughes, A. L. 1991 Circumsporozoite proteins of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. *Genetics* **137**, 345–353.

Hughes, A. L. 1992 Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. *Mol. Biol. Evol.* **9**, 381–393.

Hughes, A. L. & Hughes, M. K. 1995*a* Natural selection on the peptide-binding regions of major histocompatibility complex molecules. *Immunogenetics* **42**, 233–243.

Hughes, M. K. & Hughes, A. L. 1995*b* Natural selection on *Plasmodium* surface proteins. *Mol. Biochem. Parasitol.* **71**, 99–113.

Hughes, A. L. & Nei, M. 1988 Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature* **335**, 167–170.

Hughes, A. L. & Verra, F. 1998 Ancient polymorphism and the hypothesis of a recent bottleneck in the malaria parasite *Plasmodium falciparum*. *Genetics* **150**, 511–513.

International SNP Map Working Group 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

Kuhner, M. K., Yamato, J. & Felsenstein, J. 1995 Estimating effective population size and mutation rate from sequence data using Metropolis–Hasting sampling. *Genetics* **140**, 1421–1430.

Kuhner, M. K., Yamato, J. & Felsenstein, J. 1998 Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429–434.

Li, W. K. 1997 *Molecular evolution*. Sunderland MA: Sinauer.

Li, W.-H. & Sadler, L. A. 1991 Low nucleotide diversity in man. *Genetics* **129**, 513–523.

Livingstone, F. B. 1958 Anthropological implications of sickle cell gene distribution in West Africa. *Am. Anthropol.* **60**, 531–561.

Maitland, K., Kyes, S., Williams, T. N. & Newbold, C. I. 2000 Genetic restriction of *Plasmodium falciparum* in an area of stable transmission: an example of island evolution? *Parasitology* **120**, 335–343.

Misawa, K. & Tajima, F. 1997 Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites. *Genetics* **147**, 1959–1964.

Molineaux, L. 1988 The epidemiology of human malaria as an explanation of its distribution, including some implications for its control. In *Malaria: principles and practice of malariology*, vol. 2 (ed. W. H. Wernsdorfer & I. McGregor), pp. 913–998. Edinburgh, UK: Churchill Livingstone.

Nei, M. 1987 *Molecular evolutionary genetics*. New York: Columbia University Press.

Nei, M. & Gojobori, T. 1986 Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426.

Przeworski, M., Hudson, R. R. & Ayala, F. J. 2000 Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302.

Rich, S. M., Licht, M. C., Hudson, R. R. & Ayala, F. J. 1998 Malaria's Eve: evidence of a recent population bottleneck throughout the world population of *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **95**, 4425–4430.

Tajima, F. 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tajima, F. 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* **143**, 1457–1465.

Takahata, N. & Nei, M. 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967–978.

Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Verra, F. & Hughes, A. L. 2000 Evidence for ancient balanced polymorphism at the apical membrane antigen-1 (AMA-1) locus of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **105**, 149–153.

Watterson, G. A. 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**, 256–276.

World Health Organization 1997 World malaria situation in 1994. *Weekly Epidemiol. Rec.* **72**, 269–276.