ELSEVIER

Gene 261 (2000) 161–170

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Genetic variation and the recent worldwide expansion of *Plasmodium falciparum*

Francisco J. Ayala[a,*], Stephen M. Rich[b]

[a]*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525, USA*
[b]*Division of Infectious Disease, Tufts University School of Veterinary Medicine, North Grafton, MA 01536, USA*

Received 22 June 2000; accepted 25 September 2000
Received by G. Bernardi

Abstract

Plasmodium falciparum, the agent of human malignant malaria, diverged from *Plasmodium reichenowi*, the chimpanzee parasite, about the time the human and chimpanzee lineages diverged from each other. The absence of synonymous nucleotide variation at ten loci indicates that the world populations of *P. falciparum* derive most recently from one single strain, or 'cenancestor,' which lived a few thousand years ago. Antigenic genes of *P. falciparum* (such as *Csp, Msp*-1, and *Msp*-2) exhibit numerous polymorphisms that have been estimated to be millions of years old. We have discovered in these antigenic genes short repetitive sequences that distort the alignment of alleles and account for the apparent old age of the polymorphisms. The processes of intragenic recombination that generate the repeats occur at rates about 10^{-3} to 10^{-2}, several orders of magnitude greater than the typical mutational process of nucleotide substitutions. We conclude that the antigenic polymorphisms of *P. falciparum* are consistent with a recent expansion of the world populations of the parasite from a cenancestor that lived in tropical Africa a few thousand years ago.

Keywords: Coalescent theory; Malaria; Nucleotide polymorphism; Neutral variation; Circumsporozoite protein; MSP-1; MSP-2

1. Evolutionary history of *Plasmodium* species

There are 300–500 million clinical cases of malaria per year, more than 1 million children die in sub-Saharan Africa, and more than 2 billion people are at risk throughout the world (World Health Organization, 1995). Four species of *Plasmodium* are parasitic to humans: *P. falciparum, P. malariae, P. ovale,* and *P. vivax*; *P. falciparum* is the most malignant. At least 80% of the mortality and most of the malignant cases occur in Africa.

Fig. 1 is a phylogenetic tree of *Plasmodium* species derived from *Csp* (circumsporozoite protein) gene sequences (Escalante et al., 1995; for very similar trees based on other genes see Escalante and Ayala, 1994; Ayala et al., 1998; 1999). Estimates of divergence times are shown in Table 1.

It is apparent that the three human parasites, *P. falciparum, P. malariae,* and *P. vivax* are very remotely related to each other, so that the evolutionary divergence of these three human parasites greatly predates the origin of the hominids. *P. ovale*, a fourth human parasite, is also remotely related to the other three, although considerably closer to *vivax* than to *malariae* or *falciparum* (Qari et al., 1996). These remote relationships are consistent with the diversity of physiological and epidemiological characteristics of these four *Plasmodium* species.

Plasmodium falciparum is more closely related to *P. reichenowi*, the chimpanzee parasite, than to any other *Plasmodium* species. The time of divergence between these two *Plasmodium* species is estimated at 8–12 million years ago (Table 1), which is roughly consistent with the time of divergence between the two host species, human and chimpanzee. A parsimonious interpretation of this state of affairs is that *P. falciparum* has been associated with our ancestors since the divergence of the hominids from the great apes. Fig. 1 shows that *P. malariae*, a human parasite, is genetically indistinguishable from *P. brasilianum*, a parasite of New World monkeys; similarly, human *P. vivax* is genetically indistinguishable of *P. simium*, also a parasite of New World monkeys. It follows that lateral transfer between hosts has occurred in recent times. Whether in these two cases the transfer has been from humans to monkeys or vice

Abbreviations: GPDH, glycerol-3-phosphate dehydrogenase; SOD, Cu–Zn superoxide dismutase

* Corresponding author. Tel.: +1-949-824-8293; fax: +1-949-824-2474.
E-mail address: fjayala@uci.edu (F.J. Ayala).

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00478-9

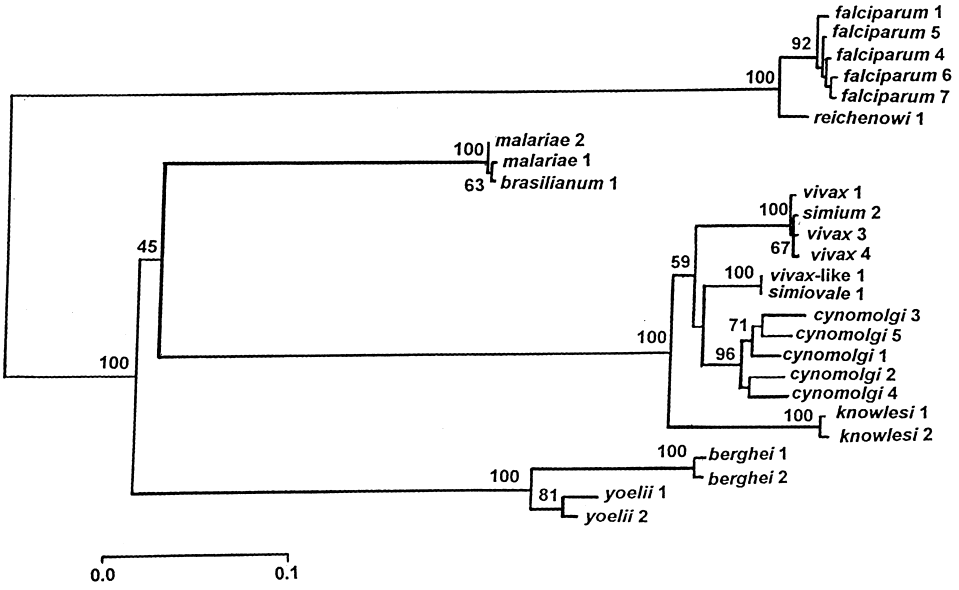

Fig. 1. Phylogeny of 12 *Plasmodium* species inferred from *Csp* gene sequences (from Ayala et al., 1998). *P. falciparum, malariae,* and *vivax* are human parasites; *berghei* and *yoelii* are rodent; all others are primate parasites. The numbers after the species names refer to different strains. Bootstrap values above branches assess the reliability of the branch cluster to the right; values above 70 are considered to indicate statistically reliable clusters.

versa is a moot question (for discussion, see Ayala et al., 1999).

2. Dearth of neutral polymorphism

Table 2 summarizes the nucleotide variation observed in ten genes of *P. falciparum*. The gene sequences analyzed are geographically representative of the global malaria endemic regions (Rich et al., 1997; 1998). Notable is the total absence of silent (synonymous) polymorphisms. Silent polymorphisms may be assumed to be adaptively neutral (or very nearly so) and thus reflect the mutation rate and the time elapsed since divergence from a common ancestral nucleotide sequence. On the contrary, non-synonymous (amino acid replacement) nucleotide polymorphisms may respond to natural selection, so that with mass selection a new, highly adaptive mutant may rise to high frequency in one of very few generations. Some of the genes in Table 2 are known to be antigenic or involved in drug resistance (Rich et al., 1998) and, not unexpectedly, exhibit amino acid polymorphisms. In a separate study of ten *P. falciparum* genes, most of them antigenic, Escalante et al. (1998) also observed a scarcity of synonymous polymorphisms.

Table 1
Time (million years) of divergence between *Plasmodium* species, based on genetic distances at two gene loci, the small subunit ribosomal RNA (*rRNA*) and the circumsporozoite protein (*Csp*) genes (see Fig. 1; adapted from Escalante and Ayala, 1994; Escalante et al., 1995)

Plasmodium	rRNA	Csp
Falciparum vs. reichenowi	11.2 ± 2.5	8.9 ± 0.4
vivax vs. monkey[a]	20.9 ± 3.8	25.2 ± 2.1
vivax vs. malariae	75.7 ± 8.8	103.5 ± 0.6
Falciparum vs. vivax/malariae	75.7 ± 8.8	165.4 ± 1.6

[a] *Brasilianum* not included.

The theory of coalescence assumes that the allelic sequences of a particular gene that are found at a given time in a population all derive from a single ancestral sequence that existed in the past (this last common ancestor to all existing sequences is called the 'cenancestor'). Mutations appear in the allelic lineages and may increase in frequency as a consequence of drift or natural selection. If we focus on synonymous substitutions and assume these to be neutral, the number of polymorphisms in a sample of allele sequences will be a function of the rate of neutral mutation (equivalently, neutral substitution), the time elapsed since the cenancestor, and the size of the sample. If we ignore the possibility of multiple hits, the observed number of neutral polymorphisms in the sample will have a Poisson distribution, with the expected mean

$$\lambda = \mu_a t \sum n_i l_i + \mu_b t \sum n_i m_i$$

where μ_a and μ_b are the neutral mutation rates at the third position of 4-fold and 2-fold degenerate codons, respectively; t is the time since the bottleneck; n_i is the number of lineages sampled at the i^{th} locus; and l_i and m_i are, respectively, the number of 4-fold and 2-fold synonymous sites examined at the i^{th} locus. The time of the cenancestor, obtained by solving for t and replacing λ by S, the observed number of polymorphisms, is given by

$$\hat{t} = \frac{S}{\mu_a \sum n_i l_i + \mu_b \sum n_i m_i} \quad (1)$$

Table 2
Polymorphisms in ten loci of *P. falciparum*. (Modified from Ayala et al., 1999)[a]

Gene	Chromosome location	Length (bp)	Number of Sequences (n_i)	Polymorphisms		Synonymous sites analyzed	
				Non-synonymous (D_n)	Synonymous (D_s)	4-fold ($n_i l_i$)	2-fold ($n_i m_i$)
Dhfr	4	609	32	4	0	2144	4128
Ts	4	1215	10	0	0	1250	2640
Dhps	8	1269	12	5	0	1536	2724
*Mdr*1	5	4758	3	1	0	1350	2088
*Rap*1	–	2349	9	8	0	1092	1668
Calm	14	441	7	0	0	364	602
G6pd	14	2205	3	9	0	726	1404
*Hsp*86	7	2241	2	0	0	532	910
Tpi	–	597	2	0	0	180	262
*Csp*5′end	3	387	25	7	0	688	2010
*Csp*3′end	3	378	25	17	0	1050	1625
Total	–	–	–	51	0	10912	20061

[a] *Dhfr* and *Ts* encode the bifunctional dihydrofolate reductase-thymidylate synthetase (DHFR-TS) domain; *Dhps* codes for dihydropteroate synthetase; *Mdr*1 for a gene conferring multidrug resistance; *Rap*1 encodes a rhoptry-associated protein; *Calm* codes for calmodulin; *G6pd* for glucose-6-phosphate dehydrogenase; *Hsp*86 for heat-shock protein 86; *Tpi* for triose phosphate isomerase; only the two terminal (5′ and 3′) non-repetitive sequences are included for the *Csp* gene, which encodes a circumsporozoite surface protein.

In our sample $S = 0$, so $\hat{t} = 0$. Because S is assumed to be Poisson-distributed, we can estimate an upper 95% confidence limit, t_{95}, for the time of the bottleneck by finding the value of t such that the probability of no polymorphism ($e^{-\lambda}$) equals 0.05. Since $e^{-2.996} = 0.05$, we calculate the t_{95} by writing 2.996 in the numerator of Eq. (1).

Estimates of the neutral mutation rates, μ_a and μ_b, may be obtained by comparing species for which the time of divergence is known: the number of neutral substitutions between species divided by the time elapsed is an estimator of the neutral mutation rate. We have obtained four estimates of neutral mutation rates, based on two comparisons: *P. falciparum* with *P. berghei* and *P. falciparum* with *P. reichenowi* (Rich et al., 1998). A summary of the results is shown in Table 3. We estimate the 95% confidence interval for the *falciparum* cenancestor as 0–24,511 or 0–57,481 years. We give in Table 3 also the t_{50} values, which represent the time such that, if the cenancestor would have been older, there is a probability greater than 50% that we had observed greater neutral variation than actually observed (which is zero).

How can we account for the recent origin (thousands of years) of the world populations of *P. falciparum*? We have concluded earlier that *P. falciparum* has been a human (hominid) parasite for millions of years. One possible hypothesis to account for the recent cenancestor is that such a conclusion is erroneous; rather *P. falciparum* has become a human parasite in recent times, by lateral transfer from some other host species (Waters et al., 1991). This hypothesis is contrary to available evidence, because it demands that the *Plasmodium* parasites in the ancestral host species would be extremely similar to *P. falciparum* (i.e. that there be no synonymous or non-synonymous substitutions between *P. falciparum* and the unknown parasite, except for amino acid replacements recently arisen in response to drugs or the host's immune system). The closest known relative of *P. falciparum* is the chimpanzee parasite, *P. reichenowi*, which on the basis of nucleotide sequence differences is estimated to have diverged several million years ago from *P. falciparum*. There are no grounds to believe that a parasite exists of some non-human organism (perhaps a primate), which is more closely related to *P. reichenowi* than to any other known *Plasmodium* species, as well as identical to *P. falciparum*. Much more parsimonious is our interpretation of a long-term association of *P. falciparum* with human ancestral populations, since their divergence from the ape lineages (Escalante and Ayala, 1994; Escalante et al., 1995).

Table 3
Estimated years to the cenancestor of the world populations of *P. falciparum*. (Adapted from Ayala et al., 1998; Rich et al., 1998) t_{95} and t_{50} are the upper boundaries of the confidence intervals. μ_a and μ_b are two estimates of the neutral mutation rate of four-fold and two-fold degenerate codons, respectively

Estimated mutation rate $\times 10^{-9}$		t_{95}	t_{50}
(μ_a)	(μ_b)		
7.12	2.22	24511	5670
3.03	0.95	57481	13296

We propose the hypothesis that human parasitism by *P. falciparum* has long been highly restricted geographically (in the African tropics), but has dispersed throughout the Old World continents only recently, probably within the last 10,000 years, after the Neolithic revolution (Coluzzi, 1994; 1997; 1999). Three possible factors may have led to this recent rapid dispersion.

One factor that may have impacted the widespread distribution of *P. falciparum* in human populations from its original focus in tropical Africa, may have been changes in human living patterns, particularly the development of agri-

cultural societies and urban centers that increased human population density (Livingstone, 1958; Wiesenfeld, 1967; de Zulueta, 1973, 1994; Coluzzi, 1997, 1999; Sherman, 1998). The multiple independent origins of the sickle-cell trait have been cited in support of this hypothesis (Pagnier et al., 1984; Stine et al., 1992).

A second explanation for the recent expansion considers possible genetic changes that have increased the affinity within the parasite-vector-host system. The high rate of phenotypic change in the parasite in recent times (Garnham, 1996; Waters et al., 1991) is consistent with this suggestion. Coluzzi (1997, 1999) has cogently argued that the worldwide distribution of *P. falciparum* is recent and has come about, in part, as a consequence of a recent dramatic rise in vectorial capacity due to repeated speciation events in Africa of the most anthropophilic members of the species complexes of the *Anopheles gambiae* and *A. funestus* mosquito vectors. The biological processes implied by this account may have, in turn, been associated with, and even dependent on the onset of agricultural societies in Africa and climatic changes.

A third factor that could account for the recent world expansion of *P. falciparum* is the gradual increase in ambient temperatures that followed the Würm glaciation, so that by about 6,000 years ago climatic conditions in the Mediterranean region and the Middle East made possible the spread of *P. falciparum* and its vectors beyond tropical Africa (de Zulueta, 1973, 1994; Coluzzi, 1997, 1999).

The three factors suggested may have been causally interdependent. Once the demographic and climate conditions became suitable for the propagation of *P. falciparum*, natural selection would have facilitated the evolution of *Anopheles* species that as highly anthropophilic and effective *falciparum* vectors (de Zulueta, 1973; Bruce-Chwatt and de Zulueta, 1980; Coluzzi, 1997, 1999).

One historical observation that supports a recent dispersal of *P. falciparum* from Africa has been made by Sherman (1998). Hypocrates (460–370 B.C.) describes quartan and tertian fevers, but he does not mention severe malignant tertian fevers, which suggests that *P. falciparum* infections did not yet occur in classical Greece, as recently as 2,400 years ago. (In recent centuries, *P. falciparum* was present in the northern Mediterranean region and persisted in spots in countries such as Spain, Greece, and Italy until it was finally eradicated around 1950, see Sherman, 1998.)

3. Alternatives to the recent expansion hypothesis

We have concluded that the world populations of *P. falciparum* derive from a small population and, ultimately, from a single strain that lived, probably in equatorial Africa, a few thousand years ago. Are there alternative hypotheses that could account for the dearth of synonymous polymorphisms in the current populations of *P. falciparum*? Four possible hypotheses are the following: (i) persistent low effective population size, (ii) low rates of spontaneous mutation, (iii) strong selective constraints on silent variation, and (iv) one or more recent selective sweeps affecting the genome as a whole, or most of it (Rich et al., 1998; Ayala et al., 1999).

Hypothesis (i) can be excluded to the extent that we know that *P. falciparum* occurs in many millions of infected humans. If the effective worldwide population of *P. falciparum* would have been very small (tens or at most hundreds of individuals) for very many generations until not long ago, this would effectively amount to a population bottleneck, which would be consistent with a single, relatively recent, cenancestor of the current world populations of *P. falciparum*.

Hypothesis (ii) proposes that spontaneous mutation rates are exceptional in *P. falciparum*. There are two arguments against it. One is the high incidence of polymorphisms at antigenic and drug-sensitivity sites, both in worldwide samples (Kemp et al., 1987; Bickle et al., 1993; Qari et al., 1994; Escalante et al., 1998) and in laboratory selection experiments with mice (Cowman and Lew, 1989). The other argument is that there is divergence, in synonymous as well as non-synonymous sites, between *P. falciparum* and other *Plasmodium* species (Hughes, 1993; Escalante and Ayala, 1994; Escalante et al., 1995).

Hypothesis (iii) proposes the existence of selective constraints against silent variation due to codon bias and high AT content. This hypothesis, however, cannot account for the total absence of silent variation in *P. falciparum* (Escalante et al., 1998b; Rich and Ayala, 1998; 1999). Two lines of evidence suggest that the paucity of synonymous polymorphism cannot be attributed entirely to codon bias. Firstly, in the case of 4-fold redundant codons, the bias would be for codons with either A or T in the third position (at the cost of G and C). This bias is consistent with the AT richness of the *falciparum* genome (61.1, 70.1, and 83.5 for 1st, 2nd, and 3rd positions, respectively; 71.6% on average). However, the fact that the mean ratio of A/T in 3rd position of 4-fold codons (Leu, Ile, Val, Ser, Pro, Thr, Ala, Arg, and Gly) is 1.1, suggests that, while A/T ↔ G/C changes may be restricted, A ↔ T changes seem not to be, since there is no evidence of one base being favored over the other (Ayala et al., 1999). Secondly, as noted by Ayala et al. (1999; see Table 6) levels of codon bias found in several species of *Plasmodium* are similar to those in *falciparum*. Presumably, these species would, therefore, have the same constraint on synonymous substitutions in 4-fold sites as *P. falciparum*. Yet synonymous polymorphisms occur in these other species, as well as among them and between them and *falciparum*, suggesting that high codon bias does not preclude synonymous substitutions. Comparisons between *P. falciparum* and *P. reichenowi*, at each of five genes for which data are available in both species, indicate high numbers of synonymous substitutions (average $K_s = 0.072$ and $K_n = 0.046$, for synonymous and non-synonymous substitutions, respectively, calculated from Escalante et al., 1998,

Table 4
Composition of the central region of the *Csp* gene in *P. falciparum* (25 alleles) and *P. reichenowi* (bottom sequence). The repeated amino acid motifs NANP, NVDP, and NVNP are represented by 1, 2 and 3, respectively. (Adapted from Ayala et al., 1999)

Sequence	Repeat motifs	Number of repeats		
		1	2	3
M15505	121211111111111111121111111111111111111111111	43	3	0
M83173	121211111111111111121111111111111111111111111	43	3	0
M83149	121212111111111111111111111111111111111111111	41	3	0
M83150	121211111111111111121111111111111111111111111	44	3	0
M83156	1212111111111111111111111111111111111111111111111	49	2	0
M83158	121212121111111111111111111111111111111111111	42	4	0
M83161	121212111111111111111111121111111111111111	39	4	0
M83163	121211111111111111121111111111111111111111111	43	3	0
M83164	121211111111111111121111111111111111111111111	46	3	0
M83165	121212111111111111111111111111111111111111111	43	3	0
M83166	121212121111111111111111111111111111111111111	42	4	0
M83167	1212121111111111111111111111111111111111111111	46	3	0
M83168	121212121111111111111111111111111111111111111	42	4	0
M83169	121212111111111111111111111111111111111111	41	3	0
M83170	1212121211111111111111111111111111111111111111	42	4	0
M83174	121212111111111121111111111111111111111111	39	4	0
M19752	1212121111111111111111111111111111111111111	41	3	0
M83172	121212111111111111111121111111111111111111	38	4	0
K02194	12121211111111111111121111111111111111111	37	4	0
M57499	1212121211111111111111111111111111111111111	40	4	0
U20969	12121211111111111111121111111111111111111	36	4	0
M83886	121212111111111111111112111111111111111111	38	4	0
M22982	1212121111111111111111112111111111111111111	40	4	0
X15363	1212121111111111111111112111111111111111111	40	4	0
M57498	121212111111111111111121111111111111111111	37	4	0
P. reichenowi	12121212131213131311111111111111111	26	5	4

Table 4). The high AT content of these two lineages has not impeded the accumulation of synonymous substitutions between them over the 8 million years since their divergence.

The last hypothesis (iv) proposes that natural selection may account for the rapid spread of one favored genotype throughout populations, particularly when the population is large and/or the selection is strong. The repeated appearance throughout global malaria endemic regions of drug-resistant phenotypes, determined by non-synonymous substitutions at the *Dhfr, Dhps,* and other loci, is most likely due to natural selection. Selection sweeps are known in other organisms, such as *Drosophila melanogaster* (Hudson et al., 1997). Natural selection at any one locus can account for the absence of synonymous variation at that locus and in closely-linked loci if the selective sweep is recent; that is, if not sufficient time has elapsed for the accumulation of new synonymous mutations. The ten genes shown in Table 2 are located on, at least, six different chromosomes, and thus this hypothesis would seem to claim that six or more independent selective sweeps have occurred recently and more or less concurrently, which seems unlikely. A selective sweep simultaneously affecting all chromosomes could happen if the population structure of *P. falciparum* were predominantly clonal, rather than sexual. This hypothesis is, in a way, a special case of the 'demographic sweep' hypothesis we have proposed, because it would equally claim that all extant populations of *P. falciparum* have recently derived from a single strain. The one important difference is that, in the case of mass selection, the global spread of the selected strain could happen very quickly, in a span of years rather than thousands of years. A prevailingly clonal population structure for *P. falciparum* would seem, at first, unlikely given that the life cycle of *P. falciparum* includes an obligatory sexual stage that occurs in the gut of the mosquito vector. We have, nevertheless, uncovered evidence consistent with a predominantly clonal mode of reproduction in *P. falciparum* (Rich et al., 1997), although the matter is far from settled.

4. The *Csp* repeat region

We have shown that there is no silent variation in the $5'$ and $3'$ terminal regions of the *Csp* gene (Table 2). The central region (CR) of this gene is variable in length, although nearly 600 nt-long. It consists of a number of tandem repeats of two different amino acid motifs, each made up of four residues. There is silent variation among the repeats of either one of the two motifs. Hughes (1991); Hughes and Verra (1998) have estimated the age of the cenancestor of the *Csp* alleles as 2.1 ± 1.5 million years.

We'll show that this is not so. Rather, the CR of any one of 25 *Csp* alleles sequenced includes all the amino acid variation and most of the silent variation present in all 25 alleles combined (Rich et al., 1997; Ayala et al., 1999). We proceed in two steps, examining in turn the amino acid and the silent polymorphisms. The two amino acid motifs in the *Csp* CR of *P. falciparum* are NANP and NVDP, represented by 1 and 2, respectively, in Table 4, which shows the organization of the region in 25 *Csp* sequences. A parsimonious interpretation of these arrangements is that length variation originates by duplication of the motif doublet 1–2 or simply of motif 1.

We have introduced the concept of repeat allotype (RAT) to refer to particular 12-nucleotide-long sequences coding for a given amino acid motif. In *P. falciparum* there are ten different RATs (A–J) coding for motif 1 and 4 RATs (M–D) coding for motif 2 (Table 5). The number of RATs present in the sample is very uneven. One particular RAT coding for NANP has an incidence greater than 50% over the whole set, whereas a few RATs are present each only 1–3 times. Among the 25 gene sequences of *P. falciparum*, there is an average of about ten different RATs per gene sequence (range 9–11). The only known sequence of *Csp* in *P. reichenowi* is somewhat shorter than those of *falciparum* (35 rather than about 45 repeats per sequence, on average), but has a similar number of distinct RATs (ten, the same as the *falciparum* average) and three rather than only two amino acid motifs, two of them identical to those of *falciparum*. Thus, nearly all of the synonymous site differences observed in the CR are between RATs that exist within any single allele. This is a strong indication that while RAT diversity may have an ancient origin, it has been maintained within individual alleles and can therefore withstand even the most constricted bottleneck. For example, all 25 *Csp* CR alleles contain at least one copy of each of the eight most common RATs (A, B, C, D, E, and F, which amount to 96% of all NANP repeats; and M and N, which amount to 84% of all NVDP repeats). If any one of these alleles were the sole survivor following a bottleneck, it alone would possess nearly all the diversity currently known for the species. Intragenic recombination between the RATs originally present in one allele can generate size polymorphisms in the resulting alleles. The process of bottleneck reduction, ensued by generation of new variations through intragenic recombination, may have occurred numerous times in the evolution of the species, and may continue to do so, given the nature of the parasite life style and its propensity for being confronted by population bottlenecks; for example, upon colonization of new geographic regions or during seasonal epidemic relapses (Rich and Ayala, 2000; Rich et al., 2000).

The variation among the *Csp* alleles can be accounted for by duplication and/or deletion of the repeated segments within any one gene, which may come about by the general slipped-strand process for generating length variation in repetitive DNA regions. This process occurs by several mechanisms, each of which is well understood at the molecular level and may involve either intra- or inter-helical exchange of DNA (Levinson and Gutman, 1987). Intragenic recombination is often associated with the evolution of mini- or micro-satellite DNA loci, such as those recently described in *P. falciparum* (Su and Wellems, 1996; Anderson et al., 1999). However, intragenic recombination has also been implicated in generating variability within coding regions in a variety of eukaryotes, including, for example, the *Drosophila* yolk protein gene and the human α_2-globin gene (Ho et al., 1996; Oron-Karni et al., 1997).

The origin of new RATs may be attributed to: (1) repla-

Table 5
Amino acid and nucleotide sequence of the repeat allotypes (RAT) and their incidence. (Adapted from Rich et al., 1997)

RAT	Motif		*falciparum*		*reichenowi*	
	Amino Acid	Nucleotide	Percentage (%)	Number	Percentage (%)	Number
A	NANP	Aatgcaaaccca	55.1	566	38.5	10
B	NANP	.....t..t	16.1	165	30.8	8
C	NANP	.....t...	7.6	78	–	–
D	NANP	.....c..t.a	6.2	64	3.8	1
E	NANP	c	6.2	64	–	–
F	NANP	..c.....c	5.1	52	–	–
G	NANP	.....c.....	3.1	32	7.7	2
H	NANP	.....t	0.3	3	–	–
I	NANP	..c.....	0.2	2	3.8	1
J	NANP	.....c.....c	0.1	1	–	–
Z	NANP	.....t..c	–	–	15.4	4
M	NVDP	....t.g.t...	52.3	46	20.0	1
N	NVDP	....t.g.t..c	31.8	28	40.0	2
O	NVDP	..c.t.g.t..t	14.8	13	–	–
P	NVDP	....t.g.t..t	1.1	1	20.0	2
X	NVNP	....t...t..c	–	–	100.0	4

Table 6
Nucleotide diversity (π) within and between Group I and II alleles of the *P. falciparum Msp*-1 genes. Blocks are as defined by Tanabe et al. (1987). Block length varies between Group I and II alleles; the value given is the average of the two. (From Rich and Ayala, 2000; Rich et al., 2000)

Block	Length (codons)	Synonymous			Non-synonymous		
		Group I	Group II	Group I + Group II	Group I	Group II	Group I + Group II
1	55	0.019	0.021	0.017	0.017	0.010	0.013
2	55	0.106	0.185	0.150	0.449	0.497	0.553
3	202	0.038	0.006	0.042	0.018	0.000	0.023
4	31	0.031	0.000	0.020	0.307	0.000	0.215
5	35	0.000	0.000	0.070	0.000	0.000	0.026
6	227	0.000	0.000	0.282	0.004	0.001	0.300
7	73	0.000	0.000	0.361	0.003	0.000	0.072
8	95	0.000	0.000	0.338	0.000	0.003	0.711
9	107	0.000	0.023	0.409	0.005	0.043	0.126
10	126	0.008	0.000	0.448	0.011	0.000	0.394
11	35	0.000	0.000	0.128	0.000	0.000	0.068
12	79	0.000	0.000	0.000	0.000	0.000	0.000
13	84	0.000	0.042	0.040	0.005	0.007	0.052
14	60	0.000	0.018	0.212	0.002	0.005	0.371
15	89	0.000	0.000	0.216	0.001	0.003	0.089
16	217	0.002	0.032	0.277	0.005	0.027	0.185
17	99	0.002	0.019	0.007	0.010	0.027	0.016

cement or silent substitutions in a codon; and (2) the slippage mechanism that leads to RAT proliferation. The two amino acid motifs and the different RAT types have arisen by the first process. The variation in the number of RATs arises by the second process. Process (2) occurs with a frequency several orders of magnitude greater than process (1) (Schug et al., 1998).

Notice that only two amino acid motifs are present in the whole set of 25 *Csp* sequences and that both motifs are present in every one of the sequences (Tables 4 and 5). Thus, there is no evidence that any replacement substitution has occurred in the recent evolution of *P. falciparum*. With respect to RAT variation, process 1 may have actually occurred in RATs H, I, J, and P, but it is not unreasonable to assume that most of the variation observed in *P. falciparum* may have recently arisen by process (2) from a single ancestral allele, if this were as heterogeneous as any one of the extant *falciparum* alleles (or, for that matter, as heterogeneous as the only known *P. reichenowi* sequence, which, as noted, includes ten different RATs, encoding three different amino acid motifs; Table 5).

5. The *Msp*-1 and *Msp*-2 antigenic polymorphisms

We have examined two other antigenic proteins, MSP-1 and MSP-2, to ascertain whether their extensive polymorphisms are consistent with a recent origin of the global populations of *P. falciparum* (Ayala et al., 1999; Rich and Ayala, 2000; Rich et al., 2000).

The *Msp*-1 gene codes for the merozoite surface antigen protein-1, which is a large 185–215 kDa protein precursor that is proteolytically cleaved into several membrane protein constituents. The known alleles of *Msp*-1 belong to one or the other of two allelic classes (Group I and Group II). The two classes are commonly designated by the strains in which they were originally identified: K1 (Group I) and MAD20 (Group II). There is considerable nucleotide substitution and length variation between the two classes but much less variation within each class (Tanabe et al., 1987). Based on this nucleotide divergence, Hughes (1992) has estimated the age of the *Msp*-1 cenancestor at ~35 million years.

Tanabe et al. (1987) have partitioned the MSP-1 protein into 17 blocks, and classified seven of these blocks as highly variable, five blocks as semi-conserved, and five blocks (#1, 3, 5, 12, and 17, which include the two terminal segments) as conserved (Table 6). Amino acid polymorphisms appear for the most part when comparisons are made between the two allele groups, whereas amino acid, as well as synonymous, polymorphisms are very low within each allele group. An exception is block 2, which encodes a set of repetitive tripeptides and thus is subject to the intragenic recombination process described above for *Csp,* as a mechanism for generating polymorphism. We have identified repeats also within other polymorphic *Msp*-1 blocks (in particular, blocks 4, 8 and 14), which were heretofore assumed to be non-repetitive. We'll focus on the repeats detected within block 8, which exhibits the highest amino acid polymorphism between the two allele groups ($\pi = 0.711$ in Table 6). Three repeats occur within this block (Fig. 2), two in Group I alleles (R1a and R1b) and one in Group II alleles (R2a) (Rich and Ayala, 2000; Rich et al., 2000). R2a is a 9-bp repeat tandemly replicated five times in all Group II alleles (the five uppermost alleles in Fig. 2). R1a is a 7-bp repeat replicated five times, and R1b is a 6-bp replicated four times in all Group I alleles. The occurrence of these repeats cannot statistically be attributed to chance (Ayala et al., 1999). Moreover, in the recently

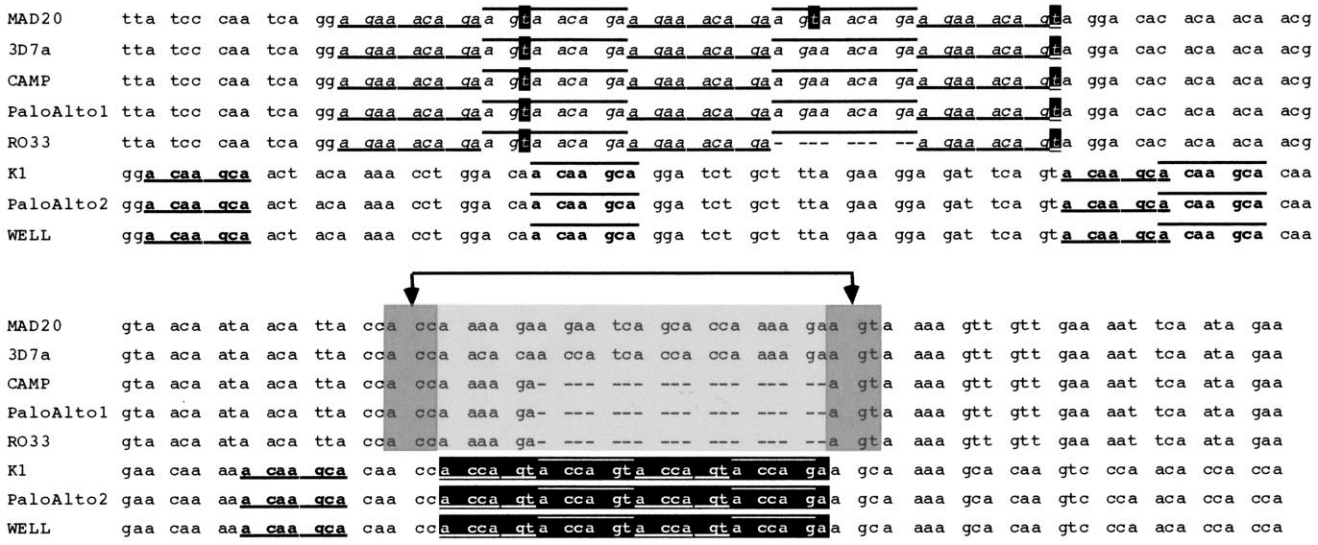

Fig. 2. Partial alignment of *Msp*1 (Block 8) Group I and II alleles (from Rich and Ayala, 2000 and Rich et al., 2000). Alternating odd and even occurrence of a repeat is indicated by underline and overbar, respectively. Region R2a consists of five tandem repeats of a 9-bp sequence (agaaacaga, in italics) highlighted in the five Group II alleles (top); one copy is missing in the RO33 allele. Regions R1a and R1b consist of two repeats, measuring 7-bp (acaagca, in boldface; repeated five times) and 6-bp (accagt, shown in inverted text; repeated four times) found in Group I alleles. The five 7-bp repeats (except for two) are separated by several codons, while the 6-bp repeats occur in tandem. There are no repeat sequences shared between Group I and II; however, the 6-bp repeat in Group I alleles clearly derives from a deletion of the intervening lightly shaded portion of Group II alleles, followed by duplication of the resulting accagt motif (junction indicated by arrows).

completed genomic sequences of *P. falciparum* chromosomes 2 and 3, the nucleotide sequences of repeats R1a, R1b, and R2a appear 25, 116, and 11 times, respectively, within the 947 kb of chromosome 2; 39, 52, and 7 times, respectively, within the 1,060 kb of chromosome 3. None of the three-nucleotide repeats ever appears in tandem on

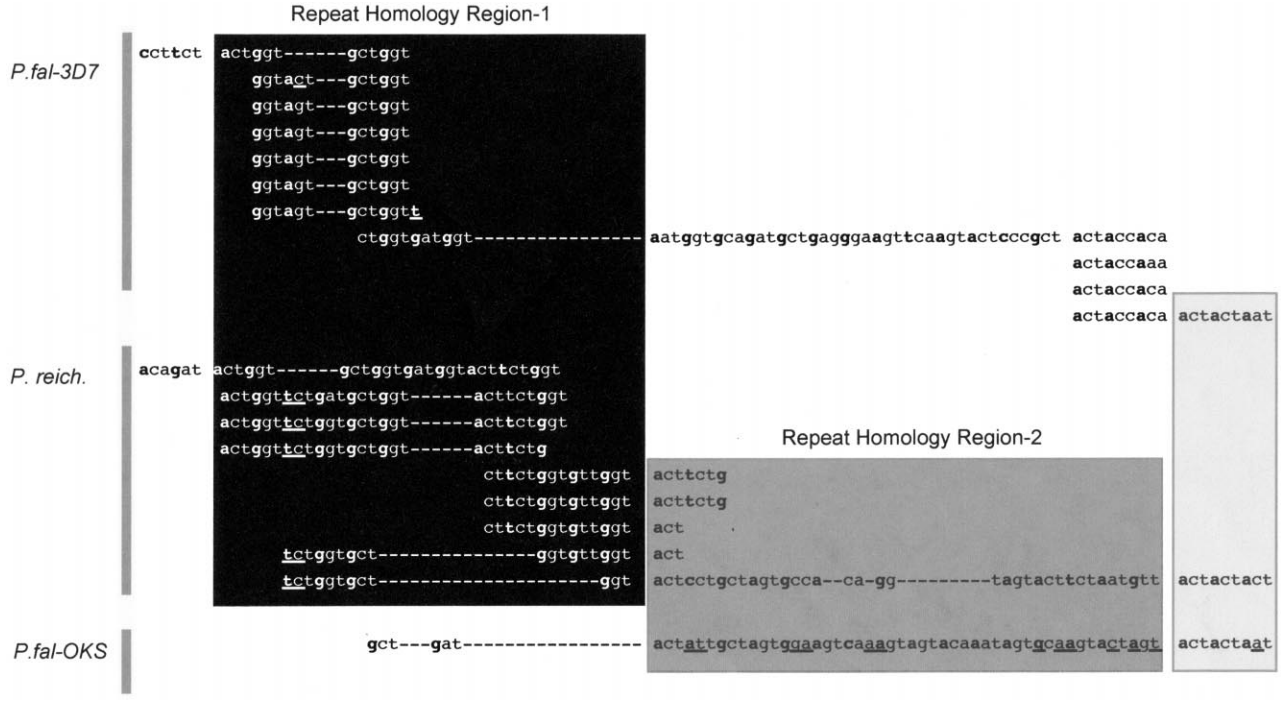

Fig. 3. Partial nucleotide alignments of three *Msp*-2 gene sequences to manifest the repeats and the homologies between *P. falciparum* 3D7 and *P. reichenowi* (RHR1, black shading) and between *P. falciparum* OKS and *P. reichenowi* (RHR2, dark shading). Sequences should be read left to right and down where homologous repeats are present. The open box at the 3′ end of RHR2 shows a region of high similarity among all three alleles. Bold letters indicate the first nucleotide of each codon. Differences between aligned sequences are highlighted by underline. Repeats within and between sequences are aligned to show their homology (from Rich and Ayala, 2000 and Rich et al., 2000).

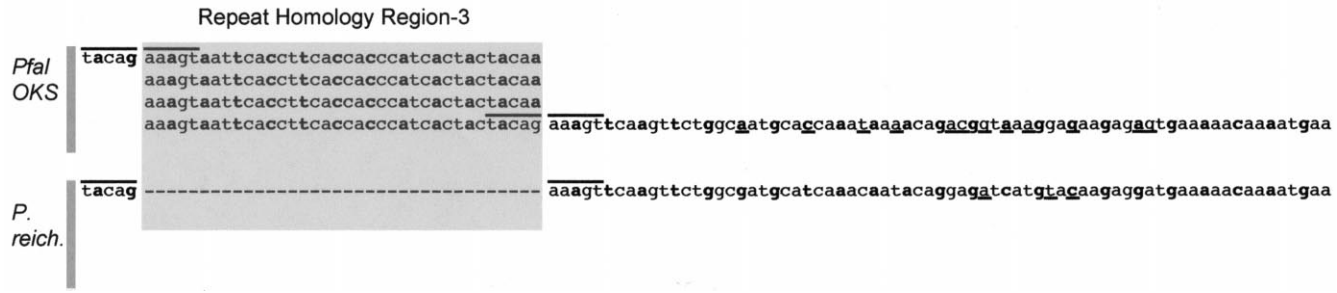

Fig. 4. Nucleotide alignment of two *Msp*-2 gene sequences to manifest the repeats within the RHR3 region of *P. falciparum* OKS (from Rich and Ayala, 2000; Rich et al., 2000). This repeat region is not present in *P. falciparum* 3D7 or in *P. reichenowi*. The repeat region of OKS continues contiguously from first to second to third to fourth row, left to right.

either chromosome 2 or 3. The average distance between each occurrence on these chromosomes is >20 kb, corroborating that their clustered occurrence in the short 147 bp segment of *Msp*-1 block 8 is a strong departure from random expectation. The *Msp*-1 gene is located on chromosome 9, which has not yet been assembled as a complete nucleotide sequence.

The *Msp*-2 gene codes for the merozoite surface protein-2, a glycoprotein anchored, like MSP-1, in the merozoite membrane, but at 45 kDa, much smaller than MSP-1. The *Msp*-2 of *P. falciparum* shows much greater variability in length, amino acid content, and number of repeats than *Csp*, but the pattern of allele polymorphism in *Msp*-2 is consistent with the hypothesis that it has rapidly arisen by intragenic recombination (Rich and Ayala, 2000; Rich et al., 2000).

The MSP-2 protein is characterized by conserved N and C termini (similarly as CSP and MSP-1), with 43 and 74 residues, respectively (Smythe et al., 1991). Bracketed within these segments, is a highly variable repeat region. Two allelic families have been identified and named after the isolates in which they were first identified. The FC27 family is characterized by at least one copy of a 32-amino acid sequence and a variable number of repeats, each 12 amino acids in length. The 3D7/Camp family contains tandem amino acid repeats measuring 4–10 amino acids in length (Felger et al., 1994).

The central portion of the gene manifests the homology of three distinct regions, which Rich and Ayala (2000; Rich et al., 2000) have defined as Repeat Homology Regions (RHR). RHR1 shows common ancestry between the *P. reichenowi Msp*-2 and the 3D7 *Msp*-2 alleles (Fig. 3, black shading). Diversity within this region results from proliferation of a ggtgct hexamer, which is ancestral to both the 3D7/Camp and the *P. reichenowi Msp*-2 allelic repeats within this region. While the conservation of these codons is clear among these two alleles, it appears that they have been lost altogether in the FC27-like alleles (represented by OKS in Fig. 3). However, the region adjacent to RHR1 in the *P. reichenowi Msp*-2 sequence is similar to the first 21 amino acids of the 32-amino acid repeat found within the FC27 family, and this sequence is the basis for the inferred RHR2 (Fig. 3, dark gray shading). The last nine nucleotides of RHR2 also manifest the homology between all three sequences, including the short stretch following the $(actaccaa)_4$ repeat in 3D7.

Fig. 4 manifests a third RHR, located further downstream, and shows the relationship between the 12-amino acid repeats of OKS and *P. reichenowi Msp*-2. The repeat region in OKS is surrounded on either side by a 10-bp sequence (tacagaaagt), which occurs as only a single $5'$ copy in the *P. reichenowi Msp*-2 allele. Despite the lengthy repeat insertion in the OKS sequence, the homology of OKS and the *P. reichenowi Msp*-2 in the region downstream of this repeat is apparent. The comparison of the three RHRs discloses that while the precursor sequences for the various repeats were probably derived from the common *P. falciparum-P. reichenowi* ancestral species, the extant diversity among the *Msp*-2 alleles has been generated since the divergence of the two species. The distinctive dimorphism of the two *P. falciparum* alleles results from proliferation of repeats in two different regions of the molecule. Presumably because the overall MSP-2 molecule is constrained in size, the proliferation of repeats leads to loss of nucleotides along the gene regions; i.e. the 3D7/Camp repeat precursors were lost in FC27 alleles, and the FC27 repeat precursors were lost in the 3D7 alleles.

The paucity of silent substitutions within the non-repetitive regions indicates that intragenic recombination has generated repeat diversity in relatively short periods of time. Empirical estimates of mutation rates among repetitive DNA sequences, such as satellite DNA, are as high as 10^{-2} mutations/generation and therefore several orders of magnitude greater than rates for point mutations (Schug et al., 1998). The high mutation rates, coupled with strong selection for immune evasion, yield an extremely accelerated evolutionary rate for these antigenic genes of *P. falciparum*.

References

Anderson, T.J.C., Su, X.Z., Bockarie, M., Lagog, M., Day, K.P., 1999. Twelve microsatellite markers for characterization of *Plasmodium*

falciparum from finger-prick blood samples. Parasitology 119, 113–125.

Ayala, F.J., Escalante, A.A., Lal, A.A., Rich, S.M., 1998. Evolutionary relationships of human malaria parasites. In: Sherman, I.W. (Ed.). Malaria: Parasite Biology, Pathogenesis, and Protection. ASM Press, Washington, DC, pp. 285–300.

Ayala, F.J., Escalante, A.A., Rich, S.M., 1999. Evolution of *Plasmodium* and the recent origin of the world populations of *Plasmodium falciparum*. Parassitologia 41, 55–68.

Bickle, Q., Anders, R.F., Day, K., Coppel, R.L., 1993. The S-antigen of *Plasmodium falciparum*: repertoire and origin of diversity. Mol. Biochem. Parasitol. 61, 189–196.

Bruce-Chwatt, L.J., de Zulueta, J., 1980. The Rise and Fall of Malaria in Europe. A Historic-Epidemiological Study. Oxford University Press, London.

Coluzzi, M., 1994. Malaria and the afrotropical ecosystems: impact of man-made environmental changes. Parassitologia 36, 223–227.

Coluzzi, M., 1997. Evoluzione Biologica i Grandi Problemi della Biologia. Accademia dei Lincei, Rome, pp. 263–285.

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.

Cowman, A.F., Lew, A.M., 1989. Antifolate drug selection results in duplication and rearrangement of chromosome 7 in *Plasmodium chabaudi*. Mol. Cell Biol. 9, 5182–5188.

de Zulueta, J., 1973. Malaria and Mediterranean history. Parassitologia 15, 1–15.

de Zulueta, J., 1994. Malaria and ecosystems: from prehistory to poster-adication. Parassitologia 36, 7–15.

Escalante, A.A., Ayala, F.J., 1994. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. Proc. Natl. Acad. Sci. USA 91, 11373–11377.

Escalante, A.A., Barrio, E., Ayala, F.J., 1995. Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. Mol. Biol. Evol. 12, 616–626.

Escalante, A.A., Lal, A.A., Ayala, F.J., 1998. Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. Genetics 149, 189–202.

Felger, I., Tavul, L., Kabintik, S., Marshall, V., Genton, B., Alpers, M., Beck, H.P., 1994. *Plasmodium falciparum*: extensive polymorphism in merozoite surface antigen 2 alleles in an area with endemic malaria in Papua. New Guinea. Exp. Parasitol. 79, 106–116.

Garnham, P.C.C., 1996. Malaria Parasites and Other Haemosporidia, Blackwell Scientific, Oxford, pp. 60–84.

Ho, K.F., Craddock, E.M., Piano, F., Kambysellis, M.P., 1996. Phylogenetic analysis of DNA length mutations in a repetitive region of the Hawaiian *Drosophila* yolk protein gene Yp2. J. Mol. Evol. 43, 116–124.

Hudson, R.R., Sáez, A.G., Ayala, F.J., 1997. DNA variation at the *Sod* locus of *Drosophila melanogaster*: an unfolding story of natural selection. Proc. Natl. Acad. Sci. USA 94, 7725–7729.

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, A.L., 1993. Coevolution of immunogenic proteins of *Plasmodium falciparum* and the host's immune system. In: Takahata, N., Clark, A.G. (Eds.). Mechanisms of Molecular Evolution. Sinauer Associates, Sunderland, MA, pp. 109–127.

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.

Kemp, D.J., Coppel, R.L., Anders, R.F., 1987. Repetitive proteins and genes of malaria. Ann. Rev. Microbiol. 41, 181–208.

Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221.

Livingstone, F.B., 1958. Anthropological implications of sickle cell gene distribution in West Africa. Am. Anthropol. 60, 533–562.

Oron-Karni, V., Filon, D., Rund, D., Oppenheim, A., 1997. A novel mechanism generating short deletion/insertions following slippage is suggested by a mutation in the human alpha(2)-globin gene. Hum. Mol. Genet. 6, 881–885.

Pagnier, J., Mears, J.G., Dunda-Belkhodja, O., Shaefer-Regio, K.E., Bedford, C., Nagel, R.L., Labie, D., 1984. Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc. Natl. Acad. Sci. USA 81, 1771–1773.

Qari, S.H., Collins, W.E., Lobel, H.O., Taylor, F., Lal, A.A., 1994. A study of polymorphism in the circumsporozoite protein of human malaria parasites. Am. J. Trop. Med. Hyg. 50, 45–51.

Qari, S.H., Shi, Y.P., Pieniazek, N.J., Collins, W.E., Lal, A.A., 1996. Phylogenetic relationship among the malaria parasites based on small subunit rRNA gene sequences: monophyletic nature of the human malaria parasite. *Plasmodium falciparum*. Mol. Phylogenet. Evol. 6, 157–165.

Rich, S.M., Ayala, F.J., 1998. The recent origin of allelic variation in antigenic determinants of *Plasmodium falciparum*. Genetics 150, 515–517.

Rich, S.M., Ayala, F.J., 1999. Circumsporozoite polymorphism, silent mutations and the evolution of *Plasmodium falciparum*. Parasitology Today 15, 39–40.

Rich, S.M., Ayala, F.J., 2000. Population structure and recent evolution of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA 67, 6994–7001.

Rich, S.M., Hudson, R.R., Ayala, F.J., 1997. *Plasmodium falciparum* antigenic diversity: evidence of clonal population structure. Proc. Natl. Acad. Sci. USA 94, 13040–13045.

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 populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA 95, 4425–4430.

Rich, S.M., Ferreira, M.U., Ayala, F.J., 2000. Origin of antigenic variation in *Plasmodium falciparum*. Parasitol. Today 16, 390–396.

Schug, M.D., Hutter, C.M., Noor, M.A., Aquadro, C.F., 1998. Mutation and evolution of microsatellites in *Drosophila melanogaster*. Genetica 102-103, 359–367.

Sherman, I.W., 1998. A brief history of malaria and discovery of the parasite's life cycle. In: Sherman, I.W. (Ed.), Malaria: Parasite Biology, Pathogenesis, and Protection. ASM Press, Washington, DC, pp. 3–10.

Smythe, J.A., Coppel, R.L., Kay, K.P., Martin, R.K., Oduola, A.M.J., Kemp, D.J., Anders, R.F., 1991. Structural diversity in the *Plasmodium falciparum* merozoite surface antigen 2. Proc. Natl. Acad. Sci. USA 88, 1751–1755.

Stine, O.C., Dover, G.J., Zhu, D., Smith, K.D., 1992. The evolution of two West African populations. J. Mol. Evol. 34, 336–344.

Su, X., Wellems, T.E., 1996. Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats. Genomics 33, 430–444.

Tanabe, K., Mackay, M., Goman, M., Scaife, J.G., 1987. Allelic dimorphism in a surface antigen gene of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. 195, 273–287.

Waters, A.P., Higgins, D.G., McCutchan, T.F., 1991. *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. Proc. Natl. Acad. Sci. USA 88, 3140–3144.

Wiesenfeld, S.L., 1967. Sickle-cell trait in human biological and cultural evolution: development of agriculture causing increased malaria is bound to gene-pool changes causing malaria reduction. Science 157, 1134–1140.

World Health Organization, 1995. Tropical Disease Report, Twelfth Programme Report. World Health Organization, Geneva.