

# Haemoglobin S and haemoglobin C: ‘quick but costly’ versus ‘slow but gratis’ genetic adaptations to *Plasmodium falciparum* malaria

David Modiano<sup>1,†</sup>, Germana Bancone<sup>1,†</sup>, Bianca Maria Ciminelli<sup>2</sup>, Fiorenza Pompei<sup>2</sup>, Isa Blot<sup>3</sup>, Jacques Simporé<sup>4</sup> and Guido Modiano<sup>2,\*</sup>

<sup>1</sup>Dipartimento di Scienze di Sanità Pubblica, University of Rome ‘La Sapienza’, Rome, Italy, <sup>2</sup>Dipartimento di Biologia, University of Rome ‘Tor Vergata’, Rome, Italy, <sup>3</sup>Centre National de Transfusion Sanguine, Ouagadougou, Burkina Faso and <sup>4</sup>Centre de Recherche Biomoléculaire Pietro Annigoni (CERBA), Ouagadougou, Burkina Faso

Received September 20, 2007; Revised and Accepted November 27, 2007

**Haemoglobin S (HbS;  $\beta 6\text{Glu} \rightarrow \text{Val}$ ) and HbC ( $\beta 6\text{Glu} \rightarrow \text{Lys}$ ) strongly protect against clinical *Plasmodium falciparum* malaria. HbS, which is lethal in homozygosity, has a multi-foci origin and a widespread geographic distribution in sub-Saharan Africa and Asia whereas HbC, which has no obvious CC segregational load, occurs only in a small area of central West-Africa. To address this apparent paradox, we adopted two partially independent haplotypic approaches in the Mossi population of Burkina Faso where both the local S ( $S^{\text{Benin}}$ ) and the C alleles are common (0.05 and 0.13). Here we show that: both C and  $S^{\text{Benin}}$  are monophyletic; C has accumulated a 4-fold higher recombinational and DNA slippage haplotypic variability than the  $S^{\text{Benin}}$  allele ( $P = 0.003$ ) implying higher antiquity; for a long initial lag period, the C alleles did apparently remain very few. These results, consistent with epidemiological evidences, imply that the C allele has been accumulated mainly through a recessive rather than a semidominant mechanism of selection. This evidence explains the apparent paradox of the uni-epicentric geographic distribution of HbC, representing a ‘slow but gratis’ genetic adaptation to malaria through a transient polymorphism, compared to the polycentric ‘quick but costly’ adaptation through balanced polymorphism of HbS.**

## INTRODUCTION

Haemoglobin S (HBB E6V) and Haemoglobin C (HBB E6K) provide considerable protection from severe *Plasmodium falciparum* malaria (1–3 for Hb S; 3–4 for Hb C) and from mild malaria attacks (1,3,5). The S allele has become polymorphic independently in different locations (6), it is common all over tropical and equatorial Africa, in Arabia and in India and its large diffusion is explained by the relationships between the fitness  $w$  of its three genotypes ( $w_{SS} \approx 0 \ll w_{AA} < w_{AS}$ ) under a strong *P. falciparum* malaria selective pressure (7). This makes the A/S polymorphism the best example of balanced polymorphism of human biology, namely of a class of genetic adaptations intrinsically ‘bad’ because at the equilibrium the frequency of the advantaged

heterozygous genotype can be at most 50% and the segregational load may be very high (as in the case of the A/S polymorphism where one of the two homozygous genotypes is even lethal). The C allele, instead, occurs in a single and quite restricted area of central West Africa (unicentricity and epicentricity) and even in this area its frequency is not dramatically higher than that of the S allele (3). Since the CC homozygosity provides a full (or very nearly so) protection against *P. falciparum* malaria, two selective models can be figured out: a strong protection also of the AC heterozygotes (semidominant model) or a mild protection of the AC heterozygotes (recessive model). Both these models would expect at the long run the C allele fixation, hence a full protection for the whole population, but they dramatically differ for three

\*To whom correspondence should be addressed at: Facoltà di Scienze Matematiche, Fisiche e Naturali, Dipartimento di Biologia, Università di Roma ‘Tor Vergata’, Via della Ricerca Scientifica, 00133 Rome, Italy. Tel: +39 0672594341; +39 0672594330; Fax: +39 062023500; Email: modiano@uniroma2.it

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint Authors.

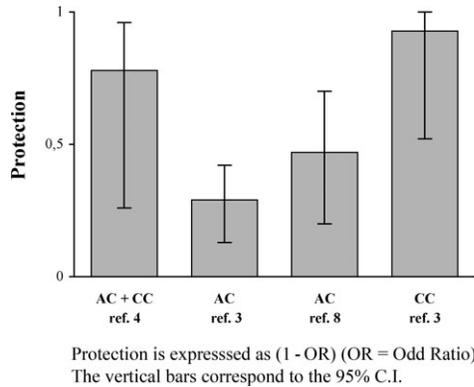


Figure 1. Protection against *P. falciparum* clinical malaria.

fundamental aspects: (i) the probability that the C allele starts to be accumulated, instead of disappearing by pure chance; (ii) its exportability to neighbouring populations by demic diffusion and (iii) the time required to attain a common frequency and, eventually, fixation. With the semidominant model, the probability of accumulation and the exportability are high and the time required to become polymorphic short; on the contrary, with the recessive model, the probability of accumulation and the exportability are small and the time required to become polymorphic long. The explanation of the contradiction between the extremely large diffusion of the costly S allele, on one side, and the very restricted diffusion of the costless C allele on the other side, must be based on a correct choice between the two models. Three sources of information can be utilized to choose between the two models:

- direct, epidemiological data (Fig. 1). They strongly suggest that the AC protection is much lower than the CC protection, but, owing to the large confidence intervals, do not rule out a partially semidominant model (3);
- the C and S frequencies in the only African area where they coexist. The fact that the frequency of the C allele is not dramatically higher than that of the S allele favours the recessive model of selection for the C allele, but is far from proving it because this finding could have been the consequence of a delayed appearance of the C allele with respect to that of the S allele.
- the uni-epi-centricity mentioned above. They would be neatly explained by the recessive model, but almost incompatible with the semidominant one.

On these bases, the recessive model has been proposed as the most likely one (3), though further evidence on (a) and/or (b) issues was needed to consider it conclusively proved.

In the present investigation we focused on the (b) issue by studying, through a haplotypic approach, the evolutionary histories of the C and S alleles in a single population. The survey was performed in the Mossi of Burkina Faso, central West Africa, where the C and S<sup>Benin</sup> (the haplotype where the S mutation of Mossi is found) alleles are both common ( $0.128 \pm 0.004$  and  $0.051 \pm 0.003$ ) (3), thus providing an ideal opportunity to study the evolutionary stories of these two malaria-protective alleles within the same epidemiologic

context and genetic background. The ages of the two alleles have been estimated through the classical approach based on the linkage disequilibrium (LD) decay, namely on the extent of accumulation of initially absent C and S<sup>Benin</sup> haplotypes (hereafter designated 'new' haplotypes) produced by recombination and/or by DNA slippage events. Moreover, we devised a novel semiquantitative approach to gather information on the time-course of the C allele accumulation.

## RESULTS

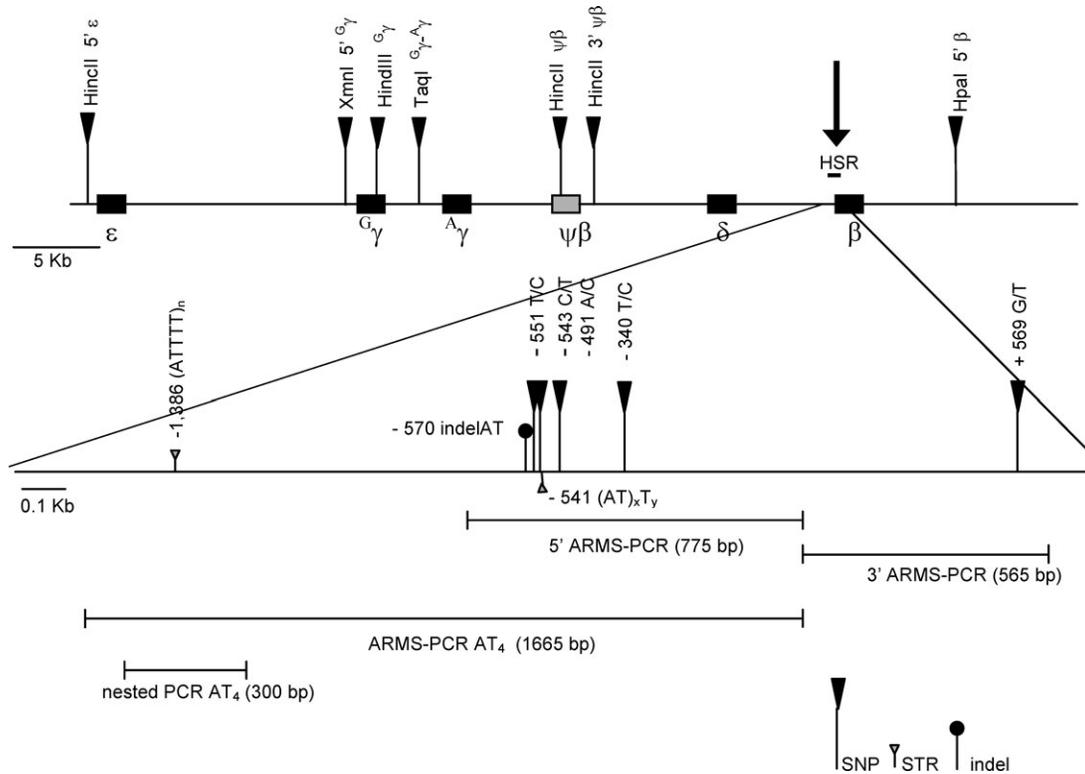
Due to the existence of a Hot Spot of Recombination region (HSR), it is necessary to subdivide the data on the slowly evolving markers (sites which haplotypes with respect to the  $\beta^6$  codon can only change by recombination) into two classes: those lying on the same side of the 6th codon with respect to HSR (downstream or 3' markers) and those lying on the opposite side (upstream or 5' markers) (Fig. 2).

### The haplotype variability for the 3' slowly evolving markers

Table 1 reports the frequencies of the haplotypes made up of the  $\beta^6$  codon and another marker (3' two-loci haplotypes), and Table 2 the frequencies of the 3' multi-loci haplotypes. Only one C and one S<sup>Benin</sup> haplotype (delAT, T, T, C, T) were found among the unambiguously characterized 50 C and 25 S<sup>Benin</sup> chromosomes. This confirms previous reports on Afro-Americans (11,12) and, combined with the present observation that this haplotype is not common among the A haplotypes ( $3/23 = 0.13 \pm 0.07$ ), prove that the C and S<sup>Benin</sup> alleles are both monophyletic in the Mossi.

### The haplotype variability for the 5' slowly evolving markers

The data are presented at the two-loci (Table 3) and multiloci (Table 4) haplotype level. The region of ca. 40 Kb here studied shows a very low (substantially nil) recombination rate (13) so that it can be considered formally as a multiallelic site which 'alleles' correspond to the haplotypes. Out of the 128 theoretically possible haplotypes (i.e.  $2^7$ , where 7 is the number of SNPs studied), 10 have been found or inferred by ML (Maximum Likelihood) with a frequency  $\geq 2$  in the sample of 152 A clusters examined (Table 4). These frequency estimates are compatible with those of several other studies dealing with less numerous samples (6,12,14–17). In contrast to the absence of variability observed for the 3' markers (see above), the C and the S<sup>Benin</sup> clusters show some variability for the 5' markers: the original C and S<sup>Benin</sup> haplotypes are clearly recognizable, but three diverse types of C and one of S<sup>Benin</sup> recombinant ('new' = not ancestral) haplotypes have been found in a sample of 58 C and 42 S<sup>Benin</sup> clusters and their overall frequency among the C is much higher than that among the S<sup>Benin</sup> alleles (7/58 versus 1/42).



**Figure 2.** The  $\beta$ -globin region and the markers studied. The XmnI 5'  $G_\gamma$  is the  $G_\gamma$  -158 C/T site, the + allele (corresponding to T) is known to promote the production of HbF during extra-uterine life, thus mitigating the severity of sickle cell and  $\beta$  thalassaemia diseases (9).

**Table 1.** Absence of 'new' C and  $S^{\text{Benin}}$  haplotypes for the 3' slowly evolving markers (LD decay = 0) (two-loci haplotypes)

Marker	Allele	Chromosomes		$S^{\text{Benin}}$
		A	C	
-570 indelAT	insAT	6/56	0/63	0/58
-551 (T/C)	C	23/71	0/63	0/58
-543 (C/T)	T	5/72	63/63	58/58
-491 (A/C)	C	22/68	62/62	58/58
-340 (T/C)	C	2/56	0/51	0/28
+569 (G/T)	T	25/59	59/59	28/28
+7446 HpaI	(-)	28/170	54/54	34/34

For each marker, its position is referred to the cap site of the  $\beta$ -globin gene (accession no NG\_000007). Estimates were obtained by direct counting (genotyping carried out either by ARMS-PCR or on homozygotes for the  $\beta^0$  site). Hardy-Weinberg equilibrium was verified for all testable sites (-551, -491, +569 and +7446, among the A chromosomes).

### The haplotype variability for the fast evolving (microsatellite) markers

The results concern two different simple tandem repeats (STRs). In both cases, the frequencies have been obtained by direct counting. For the  $(AT)_xT_y$  microsatellite (Table 5) neither C nor  $S^{\text{Benin}}$  clusters show any variation, despite the high variability occurring among the A chromosomes ( $H_A = 0.723$ ), thus confirming the already mentioned monophyletic origin of the two  $\beta^6$  variants and indicating that the two mutations occurred recently in different microsatellite haplotypes [the  $(AT)_7T_7$  for C and the  $(AT)_8T_4$  for  $S^{\text{Benin}}$ ] of the same 3' slowly evolving markers haplotype (delAT, T, T, C,

T, T, Hpa I-) (Table 1). The  $(ATTTT)_n$  STR at ca. -1400 bp is quite variable within the A clusters ( $H_A = 0.621$ ), and some degree of variability is displayed also by C ( $H_C = 0.186$ ) and  $S^{\text{Benin}}$  ( $H_S = 0.054$ ) clusters (Table 6) which are very different from each other for the frequency of the 'new' haplotypes (18/175 versus 3/108,  $P \approx 0.015$ ).

The data on the upstream slowly evolving markers (Table 4), combined with those on the  $(ATTTT)_n$  site (Table 6), subjected to partially independent mechanisms of evolution (recombination versus recombination plus DNA slippage), conclusively show that the journey accomplished by the C allele in the direction of attaining the same haplotype variation of the A allele has been much longer than that of the  $S^{\text{Benin}}$  allele.

### DISCUSSION

The C allele, if fixed, could provide a full protection to all individuals from severe *P. falciparum* malaria, whereas the S allele, even at its best, only protects a minority of the population and even this partial protection is paid with a quite high segregational load. Yet, the C allele occurs only in a single and very limited geographic area of central West Africa and in Thailand (19) while the S allele is distributed all over Africa, Arabia and India. In other words, at the world-wide scale, the protection afforded by the C allele is orders of magnitude smaller than that of the S allele.

The goal of the present investigation was to explain this apparent evolutionary contradiction by studying, through a haplotypic approach, a population where both these alleles

**Table 2.** Absence of 'new' C and S<sup>Benin</sup> haplotypes for the 3' slowly evolving markers (LD decay = 0) [multi-loci haplotypes (direct counting estimates)]

HAPLO ID	Markers <sup>a</sup>						Chromosomes A (n = 23)	C (n = 50)	S <sup>Benin</sup> (n = 25)
		-570	-551	-543	-491	+569			
1	del	C	C	A	G	6			
2	del	T	C	A	G	4			
3	del	T	C	C	T	4			
<b>4</b>	<b>del</b>	<b>T</b>	<b>T</b>	<b>C</b>	<b>T</b>	<b>3</b>	<b>50</b>	<b>25</b>	
5	ins	T	C	C	T	2			
6	del	T	C	A	T	2			
7	del	T	C	C	G	1			
8	del	C	C	A	T	1			

Bold indicates the original  $\beta^C$  and  $\beta^S$  haplotype. ML Arlequin (10) estimates (on AA homozygotes) have been obtained on 17 subjects. They were compatible with the direct counting estimates. Only the latter, being more reliable, will be considered later on. HAPLO ID, haplotype identification number.

<sup>a</sup>Two SNPs have been discarded: the -340 T/C because it was poorly informative and the +7446 HpaI because ARMS-PCR was not feasible.

**Table 3.** The 5' SNPs haplotypes: the C LD decay is much larger than the S<sup>Benin</sup> LD decay (two-loci haplotypes)

Marker	Allele <sup>a</sup>	Chromosomes		C (n = 58)		S <sup>Benin</sup> (n = 42)	
		A n	q	n	q	n	q
HincII 5' $\epsilon$	+	32/160	0.200	2	0.034	0	0.0
XmnI 5' $G_{\gamma}^b$	+	18/188	0.096	3	0.052	0	0.0 <sup>b</sup>
HindIII $G_{\gamma}$	+	68/184	0.370	54	0.931	1	0.024
TaqI $G_{\gamma}-A_{\gamma}$	-	78/184	0.424	2	0.034	41	0.976
HincII $\psi\beta$	+	18/174	0.103	3	0.052	0	0.0
HincII 3' $\psi\beta$	-	32/190	0.168	2	0.034	0	0.0

For each  $\beta^6$  allele, the estimates have been obtained by direct gene counting (analyses carried out on homozygous subjects).

Hardy-Weinberg equilibrium was verified for all testable sites.

<sup>a</sup>The minor allele among the A chromosomes.

<sup>b</sup>The (+) allele at this site is known to confer an HPFH phenotype in erythropoietic stress conditions, so that SS homozygosity results in a milder phenotype (9). However, S<sup>Benin</sup> is apparently never associated with this allele.

are common. The approach consisted in comparing the C and S<sup>Benin</sup> LD decays and is based on the assumption that all the C (or the S<sup>Benin</sup>) haplotypes have the same selective value (irrespective of whether they are in the ancestral or in a 'new' haplotype), therefore their LD decays proceeded at a constant rate, equal for C and S<sup>Benin</sup>, through the whole process from their birth up to now. In other words, it is here assumed that these alleles are the only determinants of the selective value of the cluster they belong to. The approach here adopted can be successful only if (i) both the C and S<sup>Benin</sup> alleles are monophyletic and (ii) their relative LD decays are neither both just started nor both almost completed. Monophyleticism is necessary because only in this case it can be assumed that any C and S<sup>Benin</sup> 'new' haplotype has been produced by recombination or DNA slippage thus allowing one to infer the allele age from the observed LD decay: our findings demonstrated that both the C and the S<sup>Benin</sup> alleles are monophyletic (see Results). Also the second condition was fulfilled: the C LD decay turned out to be far from both the extremes (0.165: see Table 4), allowing one to compare the LD decays of the two alleles, hence their relative ages.

### The comparison between the C and S<sup>Benin</sup> haplotype variabilities and its implications: C is more ancient than S<sup>Benin</sup>

This comparison has been carried out through two partially independent sets of data, those on the upstream recombinant haplotypes and those on the STR haplotypes (produced by recombination and/or DNA slippage). In both cases, the result of such comparison can be expressed in terms either of number of distinct types of events which produced the 'new' haplotypes or of the LD decays of the ancestral haplotypes.

With the present sample sizes, the maximum number of distinguishable types of recombination events was 3, both for the C and S<sup>Benin</sup> haplotypes (for C with the haplotype ID no. 1, or with the ID no. 3 or with anyone of the remaining pooled 'uncommon' A haplotypes; for S<sup>Benin</sup> with the ID no. 2, or with the ID no. 3 or with anyone of the remaining pooled A haplotypes: see Table 4). For the C allele, all the three types of recombinants have been found, whereas only one was found for the S<sup>Benin</sup> allele. The 4-fold difference observed between the C and the S<sup>Benin</sup> relative overall LD decays, though large, is not significant (0.165 versus 0.04, see footnotes of Table 4;  $P \approx 0.2$ , a value calculated by taking into account, also the C and the S<sup>Benin</sup> sample sizes). As to the (ATTTT)<sub>n</sub> STR (Table 6) both the possible one-step slippage haplotypes (ATTTT<sub>5→6</sub> and ATTTT<sub>5→4</sub>) were found among the C alleles, but only one (ATTTT<sub>5→4</sub>) among the S<sup>Benin</sup> alleles. The cumulative frequencies of the 'new' haplotypes were 18/175 (= 0.103) versus 3/108 (= 0.028), a 3.7-fold difference ( $P \approx 0.015$ ).

On the whole, present data show that the C haplotype variability is about 4-fold greater than that of the S<sup>Benin</sup> and the combined statistical significance of this difference is high ( $P \approx 0.003$ ). In spite of recent reports on biallelic HSRs with alleles showing different efficiencies in promoting recombination (20,21), in the present case higher haplotype variability implies greater antiquity for at least two reasons: (i) a study on the HSR  $\beta$  globin cluster showed 'no evidence of polymorphism in recombination rate' (20) and (ii) the A → C and the A → S<sup>Benin</sup> mutations occurred in the same 3' haplotype (Table 2) making even more unlikely that they are associated with different HSR alleles (if any).

**Table 4.** The 5' SNPs haplotypes: the C LD decay is much larger than the S<sup>Benin</sup> LD decay (ML frequency estimates of the multi-loci haplotypes)

Haplotype ID	Markers							Chromosomes		
	HincII 5' ε	XmnI 5' G <sub>γ</sub>	HindIII G <sub>γ</sub>	TaqI G <sub>γ</sub> -A <sub>γ</sub>	HincII ψβ	HincII 3'ψβ	A (n=152) n q ± SE	C (n=58) n q ± SE	S <sup>Benin</sup> (n=42) n q ± SE	
1	–	–	–	–	–	+	<b>61</b> <b>0.401 ± 0.040</b>	2 0.034 ± 0.023	<b>41</b> <b>0.976 ± 0.023</b>	
2	–	–	+	+	–	+	<b>40</b> <b>0.266 ± 0.038</b>	<b>51</b> <b>0.879 ± 0.046</b>	1 0.024 ± 0.023	
3	+	–	–	+	–	–	20 0.131 ± 0.030	2 0.034 ± 0.021	0	
4	–	+	+	+	+	+	7 0.047 ± 0.018	3 0.051 ± 0.029	0	
5	+	–	+	+	–	–	7 0.047 ± 0.016	0	0	
6	–	–	+	+	–	–	3 0.016 ± 0.012	0	0	
7	–	+	+	+	–	+	2 0.015 ± 0.010	0	0	
8	–	–	+	+	+	+	2 0.014 ± 0.012	0	0	
9	–	–	–	–	+	+	2 0.014 ± 0.013	0	0	
10	–	+	+	+	+	–	2 0.011 ± 0.009	0	0	
all the others							6 0.038 ± 0.017	0	0	

Bold indicate the original C and S<sup>Benin</sup> haplotypes. The C allele LD decay =  $(1-0.879)/(1-0.266) = 0.165$  (see Materials and Methods). The S<sup>Benin</sup> allele LD decay =  $(1-0.976)/(1-0.401) = 0.040$  (see Materials and Methods).

**Table 5.** The (AT)<sub>x</sub>T<sub>y</sub> STR at –541: all the C and the S<sup>Benin</sup> alleles are still associated with the original allele

Allele x	y	Chromosomes		C	S <sup>Benin</sup>
		A (n=66) n	q		
7	7	28	0.424	63	
7	8	18	0.273		
6	9	6	0.091		
8	5	6	0.091		
8	7	4	0.061		
8	4	3	0.045		58
9	4	1	0.015		

Heterozygosity:  $H_A = 0.723$ ;  $H_C = 0$ ;  $H_S = 0$ .

### The absolute age of the C and S<sup>Benin</sup> alleles

To estimate the absolute age of these two alleles from the extent of accumulation of 'new' C and S<sup>Benin</sup> haplotypes, different types of markers have been studied and partially independent estimates have been obtained. Two sets consisted of upstream or downstream slowly evolving markers; the remaining were fast evolving markers, but for the (AT)<sub>x</sub>T<sub>y</sub> STR 'new' haplotypes could only have been formed through DNA slippage, whereas for the (ATTTT)<sub>n</sub> STR they could have been produced also by recombination. It is worth to point out that the ultimate, overall accumulation of 'new' haplotypes should not be affected either by the demographic or by the selective history of the population and not even by its time-course (see Statistical Methods).

Table 7 presents all the age estimates. They range between 38 and 120 generations for the C allele, and 10 and 28 generations for the S<sup>Benin</sup> allele, depending on the value of the recombination rate ( $R$ ) and/or on the type of haplotype (SNPs or STRs) considered. For the C allele, present age estimate is in agreement with literature data (75–150 generations with an upper limit <275 generations) (24). The uncertainty about the haplotypes frequency estimates is relatively small when compared with the very large one concerning  $R$ . These sources of uncertainty make the present absolute age estimates

no more than tentative; however, reasonable values seem to be 100 generations for the C and 25 generations for the S<sup>Benin</sup> allele.

### The time-course of the C accumulation: a major lag followed by a short phase of rapid frequency increase

The present approach has consisted in evaluating the extent of the deviations, if any, from the theoretical expectation that the LD decays of different 'new' haplotypes are all equal. By studying two types of 'new' haplotypes [the 5' SNP and the (ATTTT)<sub>n</sub> haplotypes], large and highly significant differences among their LD decays have been found for both of them ( $P \approx 0.0005$ , see Table 8; and  $P \approx 0.007$ ; see Table 9), thus making the combined likelihood vanishingly small ( $P \approx 4 \times 10^{-6}$ ). This finding shows that for a long initial period, the C alleles remained very few, thus ruling out both a rapid self-sustained (not due to immigration) increase of the population size and a strong selective advantage of the C allele during that period even though *P. falciparum* was already there (25,26).

*Implications of the above findings.* The fact that the selective advantage of the C allele towards *P. falciparum* malaria had been very poor when this allele was still rare implies that such protection was very mild when it was brought about only through the AC heterozygosity. This rules out the semidominant model and proves by exclusion the recessive model. A further evidence is the finding that C was born much earlier than the S<sup>Benin</sup> allele. In fact, if the AC genotype afforded a protection comparable with that of AS, the C allele—having existed for a much longer time than the S<sup>Benin</sup> and being at the homozygous state highly advantageous (instead than lethal)—would have attained a frequency exceeding that of the S<sup>Benin</sup> allele to a much larger extent than it actually did (0.13 versus 0.05). Indeed it should have approached fixation.

The slow increase of the number of C alleles during the lag phase was probably due to a combined effect of a population expansion and a selective advantage of the AC heterozygotes both of mild degree. As to their rapid post-lag expansion, since

**Table 6.** The (ATTTT)<sub>n</sub> STR at ca. -1400 bp: the C 'new' haplotypes are much more frequent than the S<sup>Benin</sup> 'new' haplotypes

Allele	Chromosomes A		C		S <sup>Benin</sup> Previous data <sup>a</sup>		Present data		Combined data	
	n	q	n	q	n	q	n	q	n	q
4	65	0.337	17	0.097	0	0	3	0.037	3	0.028
5	87	0.451	157	0.897	26	1	79	0.963	105	0.972
6	39	0.202	1	0.006	0	0	0	0	0	0
7 <sup>b</sup>	2	0.010	0	0	0	0	0	0	0	0
Total	193	1.000	175	1.000	26	1.00	82	1.000	108	1.000

Comparing C with S<sup>Benin</sup> for the frequency of 'new' haplotypes: 18/175 versus 3/108 ( $P \approx 0.015$ ). Heterozygosity:  $H_A = 0.621$ ,  $H_C = 0.186$ ,  $H_S = 0.054$ .  
<sup>a</sup>Ref.18.

<sup>b</sup>The (ATTTT)<sub>7</sub> allele has never been reported.

**Table 7.** C and S<sup>Benin</sup> absolute allele ages

Recombination rate in the HSR region; $R \times 10^{-3}$	Estimates based on 'new' SNPs haplotypes				Estimates based on 'new' STR haplotypes <sup>a</sup>			
	Upstream C	S <sup>Benin</sup>	Downstream C	S <sup>Benin</sup>	(ATTTT) <sub>n</sub> C	S <sup>Benin</sup>	(AT) <sub>x</sub> T <sub>y</sub> C	S <sup>Benin</sup>
2.9 <sup>b</sup>	120	28	<500 <sup>f</sup>		64	17	<100 <sup>g</sup>	
3.8 <sup>c</sup>	92	22				56		15
5.7 <sup>d</sup>	61	14				44		12
7.1 <sup>e</sup>	49	11				38		10

The estimates (no. of generations) have been obtained by the procedures described in Materials and Methods ('Statistical Methods' section).

<sup>a</sup>Rate of production 0.0009 (see Materials and Methods).

<sup>b-c</sup>Recombination rate estimates according to refs 20,13,22; and 23, respectively.

<sup>d</sup>On the hypothesis of 500 generations and considering the C and S<sup>Benin</sup> sample size, 54 and 34 respectively, one would have expected 3.8 C and 2.4 S<sup>Benin</sup> 'new' haplotypes (see Materials and Methods), whereas none was found (Table 1).

<sup>e</sup>On the hypothesis of 100 generations and considering the C and S<sup>Benin</sup> sample size, 63 and 58 respectively, one would have expected 5.7 C and 5.2 S<sup>Benin</sup> 'new' haplotypes (see Materials and Methods), whereas none was found (Table 3).

**Table 8.** The LD decays estimated by studying various C haplotypes differ significantly from each other, instead of being all equal (recombinant haplotypes)

Haplotype ID (numbers as in Table 4)	A ( $n = 152$ )		C ( $n = 58$ )		relative LD decay <sup>c</sup>
	q		Absolute freq obs <sup>a</sup>	exp <sup>b</sup>	
Ancestral C	2	0.266	51		0.165
Recombinant C	1	0.401	2	3.8	0.085
	3	0.131	2	1.3	0.259
	4	0.047	3	0.4	1.085
	Others <sup>d</sup>	0.155	0	1.5	0
Total		0.734	7		

Comparing ID no. 4 versus all other haplotypes:  $\chi^2_{(Yates)} = 12.05$ , 1 d.f. ( $P \approx 0.0005$ ).

<sup>a</sup>Arlequin estimates (10).

<sup>b</sup>For each 'new' haplotype its expected value is the product ( $f_A \times 0.165 \times 58$ ) where  $f_A$  is its frequency in the A cluster, 0.165 is the overall observed LD decay and 58 is the sample size.

<sup>c</sup>See 'Statistical Methods'.

<sup>d</sup>All haplotypes other than ID no. 1-4.

so tremendous an increase in so short a time cannot be accounted for by selection only, it is mandatory to postulate that a huge Mossi population expansion accompanied by a strong mating structure (spanning in the whole range from

**Table 9.** The proportion of "new" microsatellite haplotypes within the C chromosomes differs from that within the A chromosomes ( $P \approx 0.007$ ) instead of being equal

	(ATTTT) <sub>4</sub>	(ATTTT) <sub>6</sub>
C	17	1
A	65	39

inbreeding to village and territory isolation) favouring the production of homozygotes, played a costarring, rather than a marginal, role in the process of the C allele accumulation. This state of affairs, specific of this particular system, makes unfeasible (because too arbitrary and speculative) any simulation approach not based on a reliable knowledge of the demographic, mating structure and malarial histories of the Mossi during the last 100-200 generations.

In summary, present findings, by ruling out a delayed birth of the C allele as the reason why its frequency is not dramatically higher than that of S<sup>Benin</sup> and by showing that the C allele accumulation process initiated with a long lag, conclusively prove the recessive model. This conclusion is supported by recent *in vitro* studies, which showed that CC parasitized Red Blood Cells are very different from AA parasitized RBCs for three features relevant for the severity of the

disease (cytoadherence, rosetting and agglutination by immune sera), whereas AC parasitized RBCs are much less modified (27).

The fact that the AC heterozygosity protection is much lower than the CC protection may be helpful to figure out hypotheses on the molecular basis of the C protection. For example, considering that the expected approximate proportion of  $\alpha_2\beta_2^C$  haemoglobin is ca. 100% in the CC homozygotes and only 25% in the AC heterozygotes, a reasonable (perhaps too naïf and simplistic) hypothesis is that the extent of the protection depends on the percentage of  $\alpha_2\beta_2^C$  Hb.

It was known since long that the SC genotype is severely disadvantaged (sickle-cell-haemoglobin C disease), but the role of the S and the C alleles in shaping each other's evolutionary fate could not be inferred because the fitness  $w_{AC}$  of the AC heterozygotes was not known. The present result that  $(w_{AC} - w_{AA}) \ll (w_{AS} - w_{AA})$  (because  $\text{Protection}_{AC} \ll \text{Prot}_{AS}$ ) allows one to assign a clear-cut, important and substantially unidirectional role to the S allele. In fact, the weighted mean fitness of the C heterozygotes (AC plus SC) with respect to the AA homozygotes, depends, by definition, on the fitness of the two types of heterozygotes and on their frequencies, according to the following equation:

$$(w_{C \text{ Heteroz}} - w_{AA}) = 2f_A f_C (w_{AC} - w_{AA}) + 2f_S f_C (w_{SC} - w_{AA})$$

Since

$$2f_A f_C (w_{AC} - w_{AA}) > 0 \quad \text{and} \quad 2f_S f_C (w_{SC} - w_{AA}) < 0,$$

$$(w_{C \text{ Heteroz}} - w_{AA}) = 0 \quad \text{when}$$

$$2f_A f_C (w_{AC} - w_{AA}) = 2f_S f_C (w_{SC} - w_{AA}),$$

namely when

$$\frac{f_A}{f_S} = \frac{w_{SC} - w_{AA}}{w_{AC} - w_{AA}}$$

which shows that the overall fitness of the C heterozygotes is equal to that of the AA homozygotes if the frequency of the A allele exceeds that of the S allele to the same extent as the large disadvantage of the SC genotype exceeds the small advantage of the AC genotype. Thus, the interplay between the three alleles may create a kind of pseudo-balanced polymorphism within the C alleles carried by the heterozygotes, where the balance would take place between the advantage of the C alleles carried by the AC heterozygotes and the disadvantage of those of the SC heterozygotes. In other words, even a modest S allele frequency may be sufficient to turn the C cumulative heterozygotes advantage into a disadvantage. The same effect on the S allele may be brought about by the C allele, but it is very unlikely because, owing to the large AS advantage, the C allele frequency required to cause such an effect would be very high. In summary, because of the large SC disadvantage coupled with the small AC advantage, the S allele has been potentially able to make even more unlikely for the C allele to increase its diffusion, particularly so if such coexistence occurred when the C allele was still in its lag phase.

On the whole, two adverse odds had been overcome by the C allele while attaining its present polymorphic status: the low rate of production (as any SNS) and the high probability of

disappearance by pure chance through the whole lag period. The chance of attaining a polymorphic frequency has been even more adverse in populations with a common S allele because of the strong disadvantage of the SC genotype. Therefore, the fact that all this took place in one occasion is perhaps more surprising than the fact that it occurred only once.

A considerable part of this scenario had been already suggested long time ago in a pioneer book (28) where, on the basis of the deviations from the HW equilibrium observed in a pooled sample of 72 African populations, the following fitness were assigned to the six genotypes for the A, S and C alleles:  $w_{AA} = 0.89 \pm 0.03$ ,  $w_{SS} = 0.20 \pm 0.11$ ,  $w_{CC} = 1.31 \pm 0.29$ ,  $w_{AS} = 1$ ,  $w_{AC} = 0.89 \pm 0.035$ ,  $w_{SC} = 0.70 \pm 0.07$ .

In conclusion, the genetic response of the Mossi to *P. falciparum* malaria has been brought about by a 'quick but costly' balanced polymorphism and a 'slow but gratis' transient polymorphism. This, which is the main conclusion of the present investigation, primes two types of general evolutionary implications.

### General implications

*Alleles which protect from malaria in a mainly (or exclusively) recessive fashion.* Three of such alleles,  $\alpha^{-3.7}$  thal (29), *fy* (30) and HbC, are presently known (though for the *fy* allele its past selective value, if any, is still debated (see, for example, the discussion in 31)). For each of them the success, in terms of geographic distribution, has largely been a function of the rate of production (very high for the  $\alpha^{-3.7}$  thal, which is produced by a displaced but homologous crossing over and extremely low for *fy* and C, which require a single, specific SNS) and of the time elapsed since the appearance of the selective factor (very long for the *fy* allele which protects from the very ancient *P. vivax* malaria and relatively short for the  $\alpha^{-3.7}$  and C which protect from the more recent *P. falciparum* malaria). The C allele, being strongly disfavoured for both these aspects, has been much less successful than the  $\alpha^{-3.7}$  thal allele, which is polycentric, and the *fy* allele, which is unicentric (Western sub-saharan Africa) but fixed in a large area.

*General aspects of the pathways of genetic adaptation.* The biological impact (pattern and velocity of evolution, and ultimate fate) of a major adaptive allele  $x$  towards a stringent, continuous and long-lasting selective factor essentially depends on: (i) the rate at which  $x$  is produced (it may range from the very low value of a specific SNS as  $\beta^S$  and  $\beta^C$ , to the much higher value of the loss of function alleles, as the thal alleles); (ii) the selection model and the genetic load of the adaptation: the A/S polymorphism is a balanced polymorphism with a high genetic load ('quick but costly' genetic adaptation), whereas  $\beta^C$  is an almost recessive selective polymorphism with, obviously, no CC segregational load ('slow but gratis' genetic adaptation). Duration, stringency and continuity of the selective pressure appear to be the main extra-genetic factors relevant for the phase achieved by an adaptive process (besides the demographic and mating structure histories of the exposed population).

One can figure out a genetic adaptation as a four-phase (*i-iv*) process usually—but not necessarily—starting with the appearance or even the pre-existence of 'quick but

costly' emergency alleles and ending with the fixation of one 'slow but gratis' allele with no 'emergency' alleles left. This simplistic scheme may be convenient to frame the single actual adaptive scenarios so far known or hypothesized. (i) A possible example of adaptive scenario still in the first phase is the ensemble of the silent, lethal alleles of four structural genes for lysosomal enzymes in the Ashkenazi Jews (32,33), apparently a set of 'quick but costly' alleles providing an 'emergency' adaptation towards an unidentified stringent but recent selective factor and not accompanied by any known 'slow but gratis' adaptive allele. (ii) The second phase consists in the coexistence of 'emergency' alleles with one 'slow but gratis' allele theoretically travelling towards fixation, but still far from it. This scenario is possibly represented by the coexistence, in central West Africa, of the  $\beta^S$  and  $\beta^C$  alleles. (iii) Conclusively demonstrated examples of third-phase scenarios in which one 'slow but gratis' adaptive allele is fixed (or very nearly so), while relics of 'quick but costly' alleles are still present, are not available. However, it has been hypothesized (34) that the Cystic Fibrosis-causing lethal alleles, which are common (cumulative frequency = 0.02) in Northern European populations where a Lactase-Persistence allele ('slow but gratis') is almost fixed, are relics of 'emergency' alleles. They would have provided a partial—and costly—adaptation by mitigating, in the heterozygotes, the severity of the diarrhoea due to the dairy milk diet adopted by these populations when they were (as all mammals) still lactose-intolerant. (iv) The fourth phase, that of a 'slow but gratis' adaptive allele fixed and apparently no longer accompanied by relics of 'quick but costly' alleles, may be represented by the  $f_y$  allele in Central-West Africa. The long time elapsed since the appearance of the supposed selective agent (*P. vivax malaria*) is likely to be the reason why this adaptive allele has attained fixation probably long time ago. Further possible examples are the  $\alpha^{-3.7}$  alleles almost fixed among Tharus of Southern Nepal (35) and in Papua New Guinea (36).

## MATERIALS AND METHODS

### The sample

It consists of 390 unrelated Mossi of Burkina Faso with the following genotypes: 120 AA, 180 AC, 35 AS, 31 CC and 24 SS. All subjects gave informed consent. Not all the specimens have been studied for all the markers. The total number of A, C and S alleles examined for each marker is specified in the Tables.

### The DNA region and the markers

The markers studied (the  $\beta^6$  codon; two STRs; 12 SNPs; 1 dinucleotide insertion/deletion) are distributed along the entire length of the  $\beta$  cluster which contains a hot spot of recombination (HSR) (Fig. 2). Haemoglobin genotypes have been identified either by RFLP analysis (3) or direct sequencing. The region of ca. 1350 bp around the  $\beta^6$  codon (from -700 to +641 bp from the cap site of the  $\beta$  gene) has been

**Table 10.** Primer sequences

Primer name	Position <sup>a</sup>	Sequence
$\beta^6$ FWD <sup>b</sup>	+52/+71	5' TGGTGCACCTGACTCCTGAG 3'
$\beta^C$ FWD	+50/+69	5' CATGGTGCACCTGACTCTTA 3'
$\beta^S$ FWD	+51/+70	5' ATGGTGCACCTGACTCTTGT 3'
$\beta^A$ REV	+89/+69	5' AGTAACGGCAGACTTCTCCTC 3'
$\beta^C$ REV	+90/+68	5' CAGTAACGGCAGACTTCTCATT 3'
$\beta^S$ REV	+90/+67	5' CAGTAACGGCAGACTTCTACA 3'
AT <sub>x</sub> T <sub>y</sub> FWD	-701/-678	5' TCTTGTTCCTCCAAAACCTAATAAG 3'
3'S	+641/+623	5' AAACGATCCTGAGACTTCC 3'
AT <sub>4</sub> FWD	-1526/-1505	5' ATTTAAGAGAATAAAGCAATGG 3'
AT <sub>4</sub> FWD2	-1610/-1590	5' ATGAGGGTTGAGACAGGTAG 3'
AT <sub>4</sub> REV	-1303/-1324	5' TGGAAACCCAGTCGGTTTAG 3'
HincII $\epsilon$ FWD		5' TCACCCAAAGGTACTGTAC 3'
HincII $\epsilon$ REV		5' GATATCCATCTCTCCATTC 3'

<sup>a</sup>Numbering is with respect to the cap site of the  $\beta$ -gene according to the sequence NG\_000007.

<sup>b</sup>This primer amplifies all the three alleles of the sixth codon, thus for the heterozygotes, the A haplotype was determined by subtracting to the diploid genotype, the C (or S) haplotype (amplified through an ARMS-PCR).

**Table 11.** PCR conditions

PCR	FWD primer	REV primer	T <sub>ann</sub> °C	MgCl <sub>2</sub>	Cycle no.
5' ARMS	AT <sub>x</sub> T <sub>y</sub>	$\beta^A$ or $\beta^C$ or $\beta^S$ REV	60–53 <sup>a</sup>	3 mM	14 <sup>a</sup> + 20
3' ARMS	3'S	$\beta^A$ or $\beta^C$ or $\beta^S$ FWD	60–53 <sup>a</sup>	3 mM	14 <sup>a</sup> + 20
AT <sub>4</sub> ARMS	AT <sub>4</sub> FWD2	$\beta^A$ or $\beta^C$ or $\beta^S$ REV	60–53 <sup>a</sup>	3.2 mM	14 <sup>a</sup> + 20
AT <sub>4</sub> nested	AT <sub>4</sub> FWD	AT <sub>4</sub> REV	53	2.5 mM	25

<sup>a</sup>The first 14 cycles were carried out with an annealing temperature starting from 60°C and decreasing of 0.5°C at each cycle; the remaining 20 cycles were carried out with an annealing temperature of 53°C.

studied by sequencing allele-specific PCR (ARMS-PCR) fragments. It comprises five SNPs (-551 T/C, -543 C/T, -491 A/C, -340 T/C and +569 G/T), one STR [-541 (AT)<sub>x</sub>T<sub>y</sub>] and one dinucleotide insertion/deletion (-570 indelAT). Two distinct ARMS-PCR were carried out for each heterozygous subject: the first with the forward primer AT<sub>x</sub>T<sub>y</sub>FWD plus one of the allele specific reverse primers ( $\beta^A$ REV or  $\beta^C$ REV or  $\beta^S$ REV) to amplify a region 750 bp long upstream (5') to the  $\beta^6$  codon; the other with the reverse primer 3'S-REV plus one of the allele specific forward primers ( $\beta^6$  FWD or  $\beta^C$ FWD or  $\beta^S$  FWD) to amplify a region 570 bp long downstream (3') to the  $\beta^6$  codon (for details see Tables 10 and 11). The (ATTTT)<sub>n</sub> STR was studied by acrylamide gel (7%) electrophoresis analysis of a 300 bp PCR fragment obtained through an ARMS-PCR (primers pair: AT<sub>4</sub>FWD2 plus  $\beta^A$  or  $\beta^C$  or  $\beta^S$ REV) followed by a nested PCR (primers pair: AT<sub>4</sub>FWD plus AT<sub>4</sub>REV). All primer sequences and PCR conditions are listed in Tables 10 and 11. All the other SNPs have been studied on homozygous subjects only (100 HbAA, 30 HbCC and 24 HbSS) using primers and PCR conditions already published (37) except than for the HincII site 3' to the  $\epsilon$  gene for which new primers were designed (Table 10).

## STATISTICAL METHODS

Maximum likelihood (ML) haplotype frequency estimates were calculated with the Arlequin 3.0 software (10).

### Estimates of the linkage disequilibrium (LD) decays

They have been calculated as *overall* decays (Table 4), as well as in some cases, referring to *single* haplotypes (Table 8).

The *overall* C LD decay is the *decrease* from 1 [the frequency,  $f_{C, \text{ancC}, 0}$ , among the C haplotypes, of the 'ancestral' C haplotype (the one where the C allele was borne) at the beginning of the process (time 0)] down to the present value after  $n$  generations ( $f_{C, \text{ancC}, n}$ ):

$$\text{Overall LD decay} = \frac{f_{C, \text{ancC}, 0} - f_{C, \text{ancC}, n}}{f_{C, \text{ancC}, 0} - f_{A, \text{ancC}}} = \frac{1 - f_{C, \text{ancC}, n}}{1 - f_{A, \text{ancC}}}$$

where  $f_{A, \text{ancC}}$  is the frequency of the 'ancestral' C haplotype among the A haplotypes (assumed to be equal at time 0 and time  $n$ ), (when  $f_{C, \text{ancC}, n} = f_{A, \text{ancC}}$  the LD decay is completed, i.e. the equilibrium has been reached).

The same applies for the  $S^{\text{Benin}}$  allele.

The LD decay concerning a *single* 'new' haplotype is expressed by the *increase* of its frequency from 0 (its initial frequency) up to its present frequency ( $f_{C, \text{newC}, n}$ ) having as a term of reference the haplotype frequency among the A haplotypes ( $f_{A, \text{newC}, n}$ ):

$$\text{LD decay of a single 'new' haplotype} = \frac{f_{C, \text{newC}, n} - f_{C, \text{newC}, 0}}{f_{A, \text{newC}} - f_{C, \text{newC}, 0}}$$

and, since  $f_{C, \text{newC}, 0}$  is 0, the expression becomes:  $f_{C, \text{newC}, n} / f_{A, \text{newC}}$ .

It is important to point out that both the C versus A and the  $S^{\text{Benin}}$  versus A systems are still in so strong a LD that the ancestral C and  $S^{\text{Benin}}$  haplotypes are clearly identifiable.

### Estimates of the C and $S^{\text{Benin}}$ allele absolute ages

They have been obtained from: (i) the frequency of 'new' (C or  $S^{\text{Benin}}$ ) upstream slowly evolving haplotypes; (ii) the frequency of 'new' downstream slowly evolving haplotypes; (iii) the frequency of 'new' fast evolving haplotypes.

In each case, for the C allele, we applied the general formula:

$$f_{C, \text{ancC}, n} = \{1 - [(f_{C, \text{ancC}, 0} \times f_A \times f_{A, \text{non ancC}} \times R \times 0.5) + (f_{C, \text{ancC}, 0} \times \mu)]\}^n \quad [1]$$

where  $f_A$  is the average frequency of the A allele during the C allele life-span. Since  $f_A$  was 1 at the beginning of the process and is 0.82 now [ $1 - (f_C + f_S) = [1 - (0.13 + 0.05)]$ ], it has been approximated to 1;  $R$  is the recombination rate; 0.5 is the proportion of the recombinant gametes carrying the C allele among those produced by the  $C/A_{\text{non ancestralC}}$  heterozygotes and  $\mu$  is the mutation rate.

For the  $S^{\text{Benin}}$  allele an equivalent formula has been used.

*Estimates based on the frequency of 'new' (C or  $S^{\text{Benin}}$ ) upstream slowly evolving haplotypes.* These 'new' haplotypes

can be produced by recombination only (since  $\mu$  is extremely low).

*The C allele age.* Considering that  $f_{C, \text{ancC}, n} = 0.88$ ,  $f_{A, \text{non ancC}} = (1 - 0.266) = 0.734$  (see Table 4) and  $f_{C, \text{ancC}, 0}$  (the frequency, among the C alleles, of the ancestral C haplotype at the beginning of the process) = 1, the formula [1] becomes:

$$0.88 = (1 - 0.367R)^n$$

so that  $n = \log 0.88 / \log (1 - 0.367R)$ , where  $R$  is the recombination rate in the HSR site.

*The  $S^{\text{Benin}}$  allele age.* Since  $f_{S, \text{ancS}, n} = 0.976$  and  $f_{A, \text{non ancS}} = 0.599$  (Table 4), the formula [1] becomes:

$$0.976 = (1 - 0.3R)^n$$

and  $n = \log 0.976 / \log (1 - 0.3R)$ .

The by far most relevant source of uncertainty of these estimates is due to the large range of the different estimates of  $R$ , which span between  $2.9$  and  $7.1 \times 10^{-3}$  (13,20,22,23).

*Estimates based on the frequency of downstream slowly evolving haplotypes.* Also in this case, 'new' haplotypes can be produced by recombination only. Since the observed frequency of recombinants have been 0/54 for the C allele and of 0/34 for the  $S^{\text{Benin}}$  allele (last row of Table 1), these findings are poorly informative allowing one only to identify an upper threshold for these alleles age. The value of this threshold can be inferred from the recombination rate of this region,  $4.5 \times 10^{-2}$  per Mb per generation (13), the distance between the  $\beta^6$  codon and the farthest SNP site (+7446 HpaI), and the frequency of the donor HpaI(+) allele ( $1 - 28/170 = 0.835$ ; see Table 1) which make one to expect that C and  $S^{\text{Benin}}$  'new' haplotypes have been accumulated at a rate of  $1.4 \times 10^{-4}$  per generation.

*Estimates based on the frequency of the 'new' fast evolving haplotypes.* The  $(ATTTT)_n$  site. 'New' one-step C or  $S^{\text{Benin}}$  haplotypes [i.e. those with the  $(ATTTT)_4$  or the  $(ATTTT)_6$  allele] may have been produced by recombination or by DNA slippage. Therefore, by resolving the formula [1] where  $f_{A, \text{non ancC}} = f_{A, \text{non ancS}} = 0.549$  ( $1 - 0.451$ ; see Table 6) the age estimates are:

$$0.897 = [1 - (0.275R + 0.0009)]^n \text{ for the C allele}$$

$$0.972 = [1 - (0.275R + 0.0009)]^n \text{ for the } S^{\text{Benin}} \text{ allele}$$

where 0.897 and 0.972 are the frequency of the ancestral  $(ATTTT)_5$  allele among the present C and, respectively,  $S^{\text{Benin}}$  clusters; and 0.0009 is the here adopted mutation rate of the  $(ATTTT)_n$  site. This figure has been obtained as a weighted mean between the data on Y (38) and autosomal (39) STRs.

*The  $(AT)_xT_y$  site.* Since this site is close to the right end of the HSR region, at a first approximation it will be regarded as located at the same side of the  $\beta^6$  codon with respect to the HSR. In other words, it is here assumed that 'new' C or  $S^{\text{Benin}}$  haplotypes for this site could only have been generated through DNA slippage, namely with a rate equal to the product ( $1 \times 0.0009$ ), where 1 is the frequency of the ancestral

C or S<sup>Benin</sup> haplotype at time 0 and 0.0009 is the DNA mutation rate.

### An approach to ascertain whether the C alleles remained very few for most of the C genealogy life-span (very long lag)

The study of the C LD decay in its wholeness can be informative about the C allele age only, while the LD decay study of the single different 'new' C haplotypes can, in favourable circumstances, provide information also about 'what the C allele genealogy did during its existence'.

The following symbols will be used hereafter:  $N_{C,final}$  is the ultimate absolute number of C alleles in the present five million Mossi. Since  $q_C = 0.13$ ,  $N_{C,final} \approx 0.13 \times 10^7$ ;  $r_{new}$  = combined rate of production of 'new' (recombinant or recombinant + DNA slippage) C haplotypes;  $n_i$  the absolute number of C alleles at the  $i_{th}$  generation;  $e_i$  the expansion factor of the  $n_i$  alleles, which is equal to  $N_{C,final}/n_i$ .

The time course of the C accumulation is expected not to affect the ultimate absolute number of the 'new' C haplotypes ( $N_{C,new}$ ). In fact the expected absolute contribution  $c_i$  of the  $i_{th}$  generation to  $N_{C,new}$  is independent from  $n_i$  being both directly proportional (for the number of recombination events) and inversely proportional (for  $e_i$ ) to it. This is shown by the formula (valid when the LD is still small, as in the case under study)  $c_i = (r_{new} \times n_i \times e_i) = (r_{new} \times n_i \times N_{C,final}/n_i) = (r_{new} \times N_{C,final})$ , which does not contain  $n_i$ . For example, a period of the C genealogy life-span characterized by extremely few C alleles is expected to contribute to the ultimate number of 'new' C alleles as an equally long period with a very high (average) number of C alleles (it may be worth noticing that, if the LD decay were affected by the time-course of the variation of an allele's number, it could not be utilised to infer the allele age as it is usually done). However, their respective contributions differ greatly for the origins: that derived from very few C alleles consists of 'very few recombinant megaclones' (both these features are potentially able to implement the chance deviations), whereas that derived from many C alleles consists of 'very many recombinant microclones' (two features which tend to counteract chance deviations). In other words, a 'very few megaclones pattern' is likely not to comply the theoretical expectation that all the possible LD decays (one per haplotype) be equal the one to the other and to the overall LD decay. On the contrary, a 'very many microclones pattern' should result into a good compliance to theoretical expectations (namely into a subdivision of the 'new' haplotypes mirroring that of the A haplotypes).

The degree of compliance of the C 'new' haplotypes subdivision to the theoretical expectation of being equal to that occurring within the A haplotypes should then ultimately depend on the ratio between the contribution of 'new' C haplotypes provided by the period when the C alleles were very few and that of the period when they were very many. Thus, if, and only if, the first period (the one which may produce 'very few megaclones') had been *much* longer than the second one, the discrepancies between the actual and the expected data accumulated during that initial period are likely not to have been buffered by the later contribution, so

that the final outcome as a rule consists of 'new' C haplotypes subdivided in a way different from that of the A haplotypes, besides of being different from one another, hence also from their mean.

In conclusion, large discrepancies from the expected subdivision of the 'new' C haplotypes would necessarily imply that the C accumulation had been a biphasic process where a very long period with very few C alleles had been followed by a short phase of rapid expansion.

It is worth noting that, if the C alleles are very few, even the pool of 'new' recombinant C haplotypes, and not only its subdivision into the possible recombinant haplotypes, is likely not to comply with its expected frequency. Thus, if the lag had been very long and the C alleles during that lag very few, the confidence intervals of the C allele age estimates inferred from the whole present proportion of 'new' C haplotypes among the C alleles are particularly large.

The just described approach can be utilized, *mutatis mutandis*, also for 'new' DNA slipped C haplotypes and may have general application.

### ACKNOWLEDGEMENTS

We are grateful to the study participants in Burkina Faso for their understanding and cooperation and to the laboratory staff at the Centre Medical Saint Camille of Ouagadougou, Burkina Faso.

*Conflict of Interest statement.* None declared.

### FUNDING

This work was supported by grants from Italian Ministry of Education (MIUR COFIN 2001, 2003), from the University of Rome 'La Sapienza' and from the EU, Sixth Framework Programme, BioMalPar Network of Excellence, N. LSHP-CT-2004-503578.

### REFERENCES

- Allison, A.C. (1954) The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. *Trans. R. Soc. Trop. Med. Hyg.*, **48**, 312–318.
- Hill, A.V., Allsopp, C.E., Kwiatkowski, D., Anstey, N.M., Twumasi, P., Rowe, P.A., Bennett, S., Brewster, D., McMichael, A.J. and Greenwood, B.M. (1991) Common west African HLA antigens are associated with protection from severe malaria. *Nature*, **352**, 595–600.
- Modiano, D., Luoni, G., Sirima, B.S., Simporé, J., Verra, F., Konate, A., Rastrelli, E., Olivieri, A., Calissano, C., Paganotti, G.M. *et al.* (2001) Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature*, **414**, 305–308.
- Agarwal, A., Guindo, A., Cissoko, Y., Taylor, J.G., Coulibaly, D., Kone, A., Kayentao, K., Djimde, A., Plowe, C.V., Doumbo, O. *et al.* (2000) Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood*, **96**, 2358–2363.
- Rihet, P., Flori, L., Tall, F., Traore, A.S. and Fumoux, F. (2004) Hemoglobin C is associated with reduced Plasmodium falciparum parasitemia and low risk of mild malaria attack. *Hum. Mol. Genet.*, **13**, 1–6.
- Pagnier, J., Mears, J.G., Dunda-Belkhdja, O., Schaefer-Rego, K.E., Beldjord, C., Nagel, R.L. and Labie, D. (1984) Evidence for the

- multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc. Natl Acad. Sci. USA*, **81**, 1771–1773.
7. Allison, A.C. (1964) Polymorphism and natural selection in human populations. *Cold Spring Harb. Symp. Quant. Biol.*, **29**, 137–149.
  8. Mockenhaupt, F.P., Ehrhardt, S., Cramer, J.P., Otchwemah, R.N., Anemana, S.D., Goltz, K., Mylius, F., Dietz, E., Eggelte, T.A. and Bienzle, U. (2004) Hemoglobin C and resistance to severe malaria in Ghanaian children. *J. Infect. Dis.*, **190**, 1006–1009.
  9. Labie, D., Dunda-Belkhdja, O., Rouabhi, F., Pagnier, J., Ragusa, A. and Nagel, R.L. (1985) The -158 site 5' to the G gamma gene and G gamma expression. *Blood*, **66**, 1463–1465.
  10. Excoffier, L., Laval, G. and Schneider, S. (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol. Bioinform. Online*.
  11. Trabuchet, G., Elion, J., Dunda, O., Lapoumeroulie, C., Ducrocq, R., Nadiifi, S., Zohoun, I., Chaventre, A., Carnevale, P., Nagel, R.L. *et al.* (1991) Nucleotide sequence evidence of the unicentric origin of the beta C mutation in Africa. *Hum. Genet.*, **87**, 597–601.
  12. Boehm, C.D., Dowling, C.E., Antonarakis, S.E., Honig, G.R. and Kazazian, H.H., Jr. (1985) Evidence supporting a single origin of the beta(C)-globin gene in Blacks. *Am. J. Hum. Genet.*, **37**, 771–777.
  13. Chakravarti, A., Buetow, K.H., Antonarakis, S.E., Waber, P.G., Boehm, C.D. and Kazazian, H.H. (1984) Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.*, **36**, 1239–1258.
  14. Talacki, C.A., Rappaport, E., Schwartz, E., Surrey, S. and Ballas, S.K. (1990) Beta-globin gene cluster haplotypes in Hb C heterozygotes. *Hemoglobin*, **14**, 229–240.
  15. Nagel, R.L., Fabry, W.E., Pagnier, J., Zohoun, I., Wajcman, H., Baudin, V. and Labie, D. (1985) Hematologically and genetically distinct forms of sickle cell anemia in Africa. The Senegal type and the Benin type. *N. Engl. J. Med.*, **312**, 880–884.
  16. Zago, M.A., Silva, W.A., Jr., Dalle, B., Gualandro, S., Hutz, M.H., Lapoumeroulie, C., Tavella, M.H., Araujo, A.G., Krieger, J.E., Elion, J. and Krishnamoorthy, R. (2000) Atypical beta(s) haplotypes are generated by diverse genetic mechanisms. *Am. J. Hematol.*, **63**, 79–84.
  17. Zago, M.A., Silva, W.A., Jr., Gualandro, S., Yokomizu, I.K., Araujo, A.G., Tavela, M.H., Gerard, N., Krishnamoorthy, R. and Elion, J. (2001) Rearrangements of the beta-globin gene cluster in apparently typical beta S haplotypes. *Haematologica*, **86**, 142–145.
  18. Chebloune, Y., Pagnier, J., Trabuchet, G., Faure, C., Verdier, G., Labie, D. and Nigon, V. (1988) Structural analysis of the 5' flanking region of the beta-globin gene in African sickle cell anemia patients: further evidence for three origins of the sickle cell mutation in Africa. *Proc. Natl Acad. Sci. USA*, **85**, 4431–4435.
  19. Sanchaisuriya, K., Fucharoen, G., Sae-ung, N., Siriratmanawong, N., Surapot, S. and Fucharoen, S. (2001) Molecular characterization of haemoglobin C in Thailand. *Am. J. Hematol.*, **67**, 189–193.
  20. Holloway, K., Lawson, V.E. and Jeffreys, A.J. (2006) Allelic recombination and de novo deletions in sperm in the human beta-globin gene region. *Hum. Mol. Genet.*, **15**, 1099–1111.
  21. Jeffreys, A.J. and Neumann, R. (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat. Genet.*, **31**, 267–271.
  22. Wall, J.D., Frisse, L.A., Hudson, R.R. and Di Rienzo, A. (2003) Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am. J. Hum. Genet.*, **73**, 1330–1340.
  23. Schneider, J.A., Peto, T.E., Boone, R.A., Boyce, A.J. and Clegg, J.B. (2002) Direct measurement of the male recombination fraction in the human beta-globin hot spot. *Hum. Mol. Genet.*, **11**, 207–215.
  24. Wood, E.T., Stover, D.A., Slatkin, M., Nachman, M.W. and Hammer, M.F. (2005) The beta-globin recombination hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. *Am. J. Hum. Genet.*, **77**, 637–642.
  25. Rich, S.M., Licht, M.C., Hudson, R.R. and 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.
  26. Joy, D.A., Feng, X., Mu, J., Furuya, T., Chotivanich, K., Krettli, A.U., Ho, M., Wang, A., White, N.J., Suh, E., Beerli, P. and Su, X.Z. (2003) Early origin and recent expansion of *Plasmodium falciparum*. *Science*, **300**, 318–321.
  27. Fairhurst, R.M., Baruch, D.I., Brittain, N.J., Osters, G.R., Wallach, J.S., Hoang, H.L., Hayton, K., Guindo, A., Makobongo, M.O., Schwartz, O.M. *et al.* (2005) Abnormal display of PfEMP-1 on erythrocytes carrying haemoglobin C may protect against malaria. *Nature*, **435**, 1117–1121.
  28. Cavalli-Sforza, L.L. and Bodmer, W.F. (1971) The genetics of human populations. Freeman WH, San Francisco.
  29. Allen, S.J., O'Donnell, A., Alexander, N.D., Alpers, M.P., Peto, T.E., Clegg, J.B. and Weatherall, D.J. (1997) alpha+-Thalassemia protects children against disease caused by other infections as well as malaria. *Proc. Natl Acad. Sci. USA*, **94**, 14736–14741.
  30. Miller, L.H., Mason, S.J., Clyde, D.F. and McGinniss, M.H. (1976) The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.*, **295**, 302–304.
  31. Hamblin, M.T., Thompson, E.E. and Di Rienzo, A. (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.*, **70**, 369–383.
  32. Goodman, R.M. and Motulsky, A.G. (eds) (1979) Genetic diseases among Ashkenazi Jews. Raven Press, New York.
  33. Motulsky, A.G. (1995) Jewish diseases and origins. *Nat. Genet.*, **9**, 99–101.
  34. Modiano, G., Ciminelli, B.M. and Pignatti, P.F. (2007) Cystic fibrosis and lactase persistence: a possible correlation. *Eur. J. Hum. Genet.*, **15**, 255–259.
  35. Modiano, G., Morpurgo, G., Terrenato, L., Novelletto, A., Di Rienzo, A., Colombo, B., Purpura, M., Mariani, M., Santachiara-Benerecetti, S., Brega, A. *et al.* (1991) Protection against malaria morbidity: near-fixation of the alpha-thalassemia gene in a Nepalese population. *Am. J. Hum. Genet.*, **48**, 390–397.
  36. Oppenheimer, S.J., Higgs, D.R., Weatherall, D.J., Barker, J. and Spark, R.A. (1984) Alpha thalassaemia in Papua New Guinea. *Lancet*, **1**, 424–426.
  37. Sutton, M., Bouhassira, E.E. and Nagel, R.L. (1989) Polymerase chain reaction amplification applied to the determination of beta-like globin gene cluster haplotypes. *Am. J. Hematol.*, **32**, 66–69.
  38. Gusmao, L., Sanchez-Diz, P., Calafell, F., Martin, P., Alonso, C.A., Alvarez-Fernandez, F., Alves, C., Borjas-Fajardo, L., Bozso, W.R., Bravo, M.L. *et al.* (2005) Mutation rates at Y chromosome specific microsatellites. *Hum. Mutat.*, **26**, 520–528.
  39. Yan, J., Liu, Y., Tang, H., Zhang, Q., Huo, Z., Hu, S. and Yu, J. (2006) Mutations at 17 STR loci in Chinese population. *Forensic Sci. Int.*, **162**, 53–44.