

1 **TITLE: Alpha globin variation in the long-tailed macaque suggests malaria selection**

2

3 **AUTHORS:**

4 C.L. Faust<sup>1,2,3</sup>, F. Rangkuti<sup>1</sup>, S. G. Preston<sup>1</sup>, A. Boyd<sup>1</sup>, P. Flammer<sup>1</sup>, B. Bia<sup>1</sup>, N. J. Rose<sup>4</sup>, F.

5 B. Piel<sup>1,5</sup>, A. L. Smith<sup>1\*</sup>, A.P. Dobson<sup>3</sup>, S. Gupta<sup>1\*</sup>, and B. S. Penman<sup>1,6\*</sup>

6 \*Corresponding authors

7

8 **AFFILIATIONS:**

9 <sup>1</sup>Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK

10 <sup>2</sup>Institute of Biodiversity, Animal Health and Comparative Medicine; Wellcome Centre for

11 Molecular Parasitology, University of Glasgow, Glasgow, G12 8QQ, UK

12 <sup>3</sup>Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New  
13 Jersey, 08544, USA

14 <sup>4</sup>National Institute for Biological Standards and Control, Blanche Lane, South Mimms,  
15 Potters Bar, Hertfordshire, EN6 3QG, UK

16 <sup>5</sup>School of Public Health, Faculty of Medicine, Imperial College London, London, UK

17 <sup>6</sup>Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, School  
18 of Life Sciences, University of Warwick, Coventry CV4 7AL, UK

19

20 CORRESPONDANCE TO: B.Penman@warwick.ac.uk (BSP), sunetra.gupta@zoo.ox.ac.uk

21 (SG) and adrian.smith@zoo.ox.ac.uk (ALS)

22 **Abstract**

23 Human haemoglobin variants, such as sickle, confer protection against death from malaria;  
24 consequently, frequencies of such variants are often greatly elevated in humans from malaria  
25 endemic regions. Among non-human primates, the long-tailed macaque, *Macaca*  
26 *fascicularis*, also displays substantial haemoglobin variation. Almost all *M.*  
27 *fascicularis* haemoglobin variation is in the alpha globin chain, encoded by two linked  
28 genes: *HBA1* and *HBA2*. We demonstrate that alpha globin variation in *M.*  
29 *fascicularis* correlates with the strength of malaria selection. We identify a range of missense  
30 mutations in *M. fascicularis* alpha globin and demonstrate that some of these exhibit a  
31 striking *HBA1* or *HBA2* specificity, a pattern consistent with computational simulations of  
32 selection on genes exhibiting copy number variation. We propose that *M.*  
33 *fascicularis* accumulated amino acid substitutions in its alpha globin genes under malaria  
34 selection, in a process that closely mirrors, but does not entirely converge with, human  
35 malaria adaptation.

36

37 **Main text**

38 **Introduction**

39 It is well established that certain human haemoglobin mutations reach high frequencies due  
40 to the protection they offer against death from malaria (Haldane 1949; Allison 1954a; Allison  
41 1954b; Taylor, et al. 2012). Non-human primate species also host malaria parasites (Coatney,  
42 et al. 1971), but the question of whether non-human primate haemoglobins are under malaria  
43 selection is unresolved. In typical adult humans, over 97% of haemoglobin is made up of beta  
44 globin transcribed from the *HBB* gene, and alpha globin transcribed from two genes in the  
45 alpha globin cluster (*HBA1* and *HBA2*) that encode identical proteins (Weatherall and Clegg  
46 2001b). Haemoglobin in other primates appears to be broadly similar, although gene  
47 duplications and deletions of *HBA* occur frequently as a result of unequal crossing over  
48 (Hoffman, et al. 2008). The common ancestor of old world primates, a group that includes  
49 macaques, is likely to have had three *HBA* genes in its alpha globin cluster (Hoffman, et al.  
50 2008). 587 amino acid changes have been reported in human *HBB*; 89 in human *HBA1* and  
51 150 in *HBA2* (Patrinos, et al. 2004). Only three of these amino acid substitutions reach  
52 significant frequencies in human populations: haemoglobin S (that leads to sickle cell  
53 anaemia when inherited in the homozygous state), haemoglobin C and haemoglobin E  
54 (Weatherall and Clegg 2001a), and all are caused by mutations in *HBB*. The former two offer  
55 significant malaria protection (Taylor, et al. 2012); clinical studies to prove the same for

56 haemoglobin E are lacking. In addition to amino acid substitutions, thalassaemic mutations  
57 affecting the rate of alpha or beta globin subunit production exist in humans. There is strong  
58 evidence that alpha thalassaemia protects against severe malaria (Taylor, et al. 2012); beta  
59 thalassaemia likely also offers protection, but far fewer studies have been carried out to  
60 confirm this.

61

62 Long-tailed macaques (*Macaca fascicularis*) display the highest level of haemoglobin  
63 variation reported in a non-human primate species (supplementary table 1). Extensive  
64 surveys of haemolysates during the 1960s-80s revealed that *M. fascicularis* populations  
65 possess a variety of adult haemoglobins that can be distinguished using starch-gel  
66 electrophoresis (Barnicot, et al. 1966) (fig. 1, supplementary table 2). Band “A” haemoglobin  
67 migrates similarly to human adult haemoglobin and is likely the wild-type haemoglobin as it  
68 is present in all populations of *M. fascicularis* and sister species of long-tailed macaques.  
69 Band “Q” haemoglobin migrates anodally to A at an alkaline pH. Changes in the alpha globin  
70 subunit are responsible for the difference between A and Q (Barnicot, et al. 1966),  
71 specifically Q differs from A by the substitution of aspartic acid for glycine at alpha globin  
72 position 71 (p. Gly71Asp) (Takenaka, et al. 1988). A third electrophoretic variant, “P”, has  
73 only been found in mainland Southeast Asian populations of *M. fascicularis*. P results from  
74 an as-yet-undefined amino acid change at position 15 or 16 of alpha globin (Barnicot, et al.  
75 1966) and P has been observed to polymerize *in vitro*. Finally, a “minor” haemoglobin (“X”)  
76 that migrates more slowly than A at an alkaline pH has been observed in some studies of *M.*  
77 *fascicularis* (Barnicot, et al. 1966; Ishimoto, et al. 1970). X is distinguished from A by four  
78 variant sites, all in alpha globin (p. Asn9Lys, p. Trp14Leu, p. Gly19Arg and p. Gly71Arg)  
79 (Wade, et al. 1970). In contrast to the widespread variation in alpha globin, only one beta  
80 globin variant has been found in a single isolated population of *M. fascicularis* (Kawamoto,  
81 et al. 1984).

82

83 The first evidence for malaria selection acting on humans was the geographical association  
84 JBS Haldane observed between the presence of malaria disease and the presence of inherited  
85 blood disorders (Haldane 1949). Here, we demonstrate a geographical association between  
86 macaque haemoglobin variation and the presence of virulent macaque malarias, which  
87 provides a compelling rationale to extend Haldane’s malaria hypothesis to non-human  
88 primates. We further show, using parallel amplicon sequencing and population genetic

89 simulations, that the *M. fascicularis* alpha globin cluster exhibits patterns which are  
90 consistent with selection.

91

92

### 93 **Results**

#### 94 ***M. fascicularis* alpha globin variation is more likely to be observed in the presence of** 95 **virulent macaque malarias.**

96 For six locations where electrophoretic alpha globin phenotypes have been surveyed (fig. 1,  
97 supplementary table 2), macaque malaria surveys have also been carried out (Faust and  
98 Dobson 2015). These are Cambodia, Singapore, Thailand, Peninsular Malaysia and the  
99 Indonesian islands of Java and Bali (supplementary table 3, supplementary fig. 1). The **A**  
100 band occurs in every macaque population surveyed; thus we consider it to be ancestral.

101 Unlike **P** and **Q**, **X** was not always examined in every study. This leads us to define **A** in the  
102 absence of **P** or **Q** to be a *non-variant alpha globin phenotype*, and any phenotype containing  
103 **Q**, **P** or both to be a *variant alpha globin phenotype*. Bayesian inference allows us to  
104 calculate the probability of observing at least one virulent macaque malaria species  
105 (*Plasmodium coatneyi* or *P. knowlesi*) in a region, and this estimate is used as a proxy for  
106 malaria selection (see Methods for virulent malaria justification). This Bayesian hierarchical  
107 model allows us to define the probability of observing a macaque with a variant alpha globin  
108 phenotype in regions with low or high malaria selection (Eq. 1, Methods, supplementary  
109 information section 1.2). In regions with low malaria selection, the probability of observing a  
110 long-tailed macaque with a variant alpha globin phenotype is 0.034 (95% credible interval  
111 (CI): 0.012, 0.053). In regions with high malaria selection, the probability of observing a  
112 long-tailed macaque with a variant alpha globin phenotype is much higher, 0.686 (95%CI:  
113 0.660, 0.715). Sensitivity analyses demonstrate that lower probabilities of variant phenotypes  
114 are always found in areas that are unlikely to have virulent malarias, regardless of the specific  
115 proxy for malaria selection used (supplementary information section 1.2; supplementary figs.  
116 2-4). There is, therefore, a geographical association between the presence of virulent malarias  
117 and variant alpha globin phenotypes in long-tailed macaque populations. The distribution of  
118 macaque haemoglobin variants is not simply explained by the heterozygosity of the macaque  
119 populations in these locations (supplementary information section 1.3, supplementary fig. 5);  
120 however we must acknowledge that it is not possible to reliably assess the heterozygosity of  
121 all the macaque populations in all the geographical locations of interest.

122

123 ***Multiple unique HBA1 and HBA2 globin sequences are present in Indonesian long-tailed***  
124 ***macaques***

125 The two alpha globin genes (HBA1 and HBA2) in the *Macaca\_fascicularis\_5.0* genome  
126 (NC\_022291.1) can be distinguished based on differences in their downstream sequences (for  
127 specific primers see Methods and SI). Out of a sample of 78 Indonesian *M. fascicularis* we  
128 successfully amplified and sequenced a 334 nucleotide region of both HBA1 and HBA2 from  
129 77 animals. This 334 nucleotide region included exon 2 and parts of its flanking introns. We  
130 identified 13 unique HBA1 and 12 unique HBA2 globin sequences based on 24 variable sites  
131 (table 1). The majority of animals possessed 2 unique *HBA1* sequences (fig. 2A,  
132 supplementary fig. 6). For most animals we observed 3 unique *HBA2* sequences, but one  
133 macaque had 5 unique *HBA2* sequences (fig. 2A, supplementary fig. 7). These results are  
134 consistent with a single copy of *HBA1* and up to 3 copies of *HBA2* within the alpha globin  
135 gene cluster of these animals. From the proportion of reads each sequence contributed to the  
136 total (supplementary figs. 6-8), it would appear that some animals may possess more than 3  
137 copies of *HBA1* or *HBA2* in their alpha globin clusters. We are reluctant to over-interpret the  
138 relative proportions of reads found, since it is possible that primers may have been biased  
139 towards amplifying certain sequences, but overall it seems likely that gene duplication of at  
140 least *HBA2*, and likely both *HBA1* and *HBA2* occurs within this population.

141

142 Given our uncertainty over the exact number of copies of each of *HBA1* and *HBA2* present in  
143 each animal, it was not possible to predict the exact haplotypic combinations of *HBA1* and  
144 *HBA2* sequences present in each alpha globin cluster. However, some patterns are apparent,  
145 e.g. two unique *HBA2* globin sequences: HBA2.1 and HBA2.2 are always found together,  
146 and at similar proportions for each macaque (supplementary fig. 7) suggesting they may be  
147 linked on the same haplotype. In supplementary tables 6 and 7 we propose a potential set of  
148 haplotypes which could account for most of the genotypes in our sample. The most frequent  
149 haplotypes under this scheme are HBA1.1-HBA2.1-HBA2.2; HBA1.2-HBA2.3; HBA1.1-  
150 HBA2.5 and HBA1.3-HBA2.1-HBA2.2-HBA2.4.

151

152 ***Amino acid position 71 displays specificity of SNPs between HBA1 and HBA2.***

153 We recorded seven nonsynonymous mutations across both *HBA1* and *HBA2* from this  
154 population of long-tailed macaques (table 1), of which Gly57Asp; Val73Arg; Gly71Arg and  
155 Gly71Glu had not previously been reported (fig. 2B,C). We observed two different  
156 substitutions at amino acid position 71, which has been shown to be the key amino acid site

157 differentiating allozymes in electrophoretic studies (Takenaka, et al. 1988). All position 71  
158 SNPs in *HBA1* sequences caused a change from glycine to arginine (Gly71Arg), and all  
159 position 71 SNPs in *HBA2* sequences caused a change from glycine to glutamic acid  
160 (Gly71Glu). Previous work identified a Gly71Asp substitution in **Q** bands in haemolysate  
161 from a single southern Sumatran *M. fascicularis* (Takenaka, et al. 1988). Both the previously  
162 observed amino acid change and the different *HBA2* substitution observed in our study  
163 involve a large negatively charged amino acid (glutamic acid or aspartic acid) replacing a  
164 small non-charged amino acid (glycine). We propose that these changes will give rise to  
165 similar phenotypic consequences, and that the Gly71Glu of *HBA2* is extremely likely to  
166 generate the **Q** band of *M. fascicularis* haemolysate (fig. 2D). The Gly71Arg change we  
167 observe in *HBA1* is one of the changes identified as characteristic of the **X** band (a so-called  
168 “minor haemoglobin” identified in some *M. fascicularis* – see Introduction). Other changes  
169 that are found in the **X** band occur in exon 1 (Wade, et al. 1970), and are beyond the scope of  
170 this analysis. If all the sequences we have identified are expressed, we predict the distribution  
171 of electrophoretic types within our sample to be as follows: **A**:8, **AQ**: 40, **AQX**:27, **AX**:2  
172 (fig. 2D). Such a distribution of electrophoretic phenotypes is not unprecedented: populations  
173 with a high frequency of the **AQ** phenotype are known to exist in Indonesian populations of  
174 *M. fascicularis* (see supplementary table 2). **X** can be observed alongside the **A** and **Q** bands  
175 in *M. fascicularis* in other parts of its range (Barnicot, et al. 1966; Ishimoto, et al. 1970).  
176 Although **X** has not been reported in electrophoretic surveys of *M. fascicularis* from  
177 Indonesia, it is likely previous studies did not use protocols capable of detecting this variant  
178 (Kawamoto and Ischak 1981; Kawamoto, et al. 1984; Perwitasari-Farajallah, et al. 1999).  
179

180 ***Multiple peptide sequences are possible in HBA1 and HBA2.***

181 Of the nonsynonymous substitutions we found at amino acid positions other than 71,  
182 His78Gln and Thr67Ile have been previously reported to occur within *M. fascicularis* **A** band  
183 (Takenaka, et al. 1988). Thr67Ile is only found in *HBA1* sequences in our sample; His78Gln  
184 is found in both *HBA1* and *HBA2*. We identified two additional substitutions unique to  
185 *HBA1*: Gly57Asp and Val73Arg. Gly57Asp is likely to alter the electrophoretic properties of  
186 haemoglobin, by analogy with the same change in human alpha globin where it gives rise to  
187 the fast migrating Hb J-Norfolk (Baglioni 1962). Likewise, Val73Arg may give rise to  
188 similar electrophoretic properties as Gly71Arg change. Interestingly all 16 individuals  
189 displaying Val73Arg also had Gly71Arg, but how Val73Arg may relate to the X band of *M.*  
190 *fascicularis* haemoglobin is unclear. We also observed a novel Val55Ile substitution in *HBA2*

191 in 20 individuals. The biochemical similarity of valine and isoleucine means it is possible  
192 that this change would not significantly alter the properties of haemoglobin.

193

194 Despite the range of substitutions observed among our samples, it was not possible to  
195 phylogenetically determine evolutionary relationships between these different *M. fascicularis*  
196 alpha globin exon 2 sequences (supplementary fig. 9). This is likely due to the fact that we  
197 had only a relatively short sequence of 334 nucleotides. The length of our sequence and the  
198 fact that only 10/68 (15%) of codons in exon 2 are variable means dn/ds ratios are unlikely to  
199 be able to provide reliable insights into whether positive selection is evident among these  
200 sequences (Anisimova, et al. 2002).

201

202 ***Natural selection can maintain high frequency HBA1 or HBA2 specific polymorphisms.***

203 The HBA1 and HBA2 specificity of certain amino acid substitutions in our sample is a  
204 surprising finding given HBA paralogues typically have highly similar coding regions in  
205 primates (almost certainly a consequence of gene conversion)(Hoffman, et al. 2008). To  
206 understand whether natural selection can drive the *HBA1* or *HBA2* specificity of amino acid  
207 substitutions in *M. fascicularis*, we simulated *HBA1* and *HBA2* in a finite diploid population  
208 using an individual based model (see Methods). Mutations generating two different amino  
209 acid substitutions at the same site in alpha globin, one neutral and one potentially under  
210 selection, were able to enter the population via migration. Gene conversion (*c*) and reciprocal  
211 crossing over (*r*) could occur between *HBA1* and *HBA2* and the probabilities of each were  
212 varied (fig. 3A-C). Our model thus assumed the simulated population to be connected to a  
213 wider global population of *M. fascicularis*, acting as the source of haemoglobin variation.

214

215 The simultaneous maintenance of both mutations as *HBA1* or *HBA2* specific polymorphisms  
216 in the population is most likely when there is (i) an advantage to the state where some, but  
217 not all, alpha globin genes in a genotype encode the selected variant, and (ii) a cost to the  
218 state where more than two of the alpha globin genes in a genotype encode the selected variant  
219 (fig. 3A-C). *HBA1* or *HBA2* specificity of both mutations simultaneously is often (but not  
220 exclusively) achieved when both the neutral and the selected mutations are present on the  
221 same chromosome, and that chromosome is elevated to a high frequency (fig. 3D). Selection  
222 is also likely to elevate the selected variant alone in an *HBA1* or *HBA2* specific manner (fig.  
223 3D). Without selection, we see fewer scenarios in which any mutations are present at high

224 frequencies. Among those where mutations do reach higher frequencies there is no bias  
225 towards *HBA1* or *HBA2* specificity (fig. 3E).

226

227 Figure 3A-E assume a high level of gene flow with a large global *M. fascicularis* population  
228 continually providing new genetic diversity. In the absence of such a process, populations  
229 will, eventually, become fixed for a single alpha globin cluster (bearing the selected variant if  
230 selection is present). Figure 3F and 3G illustrate how long multiple alpha globin clusters  
231 bearing *HBA1* and *HBA2* specific mutations persist in the absence of gene flow (each mutant  
232 chromosome starting at a frequency of 5%). Selection extends the average time that multiple  
233 mutant chromosomes can coexist, because chromosomes bearing the selected mutation (some  
234 of which may also carry a neutral mutation) are preferentially maintained in the population.

235

236 We cannot be certain of the historical population size of Indonesian *M. fascicularis* , and  
237 whether or not the mutations we have found in our sample arose *de novo* in Indonesia or were  
238 imported from elsewhere. It is therefore not possible to calculate exactly how likely it is that  
239 the pattern in our sample arose under an entirely neutral model – but we can say that under  
240 two possibilities: frequent challenge with diverse mutations (fig. 3A-E) or an entirely closed  
241 system (fig. 3F-G), selection acting on at least some of the mutations makes the maintenance  
242 of multiple HBA1 or HBA2 specific mutations far more likely.

243

## 244 Discussion

245 We have conducted the first DNA analysis of the alpha globin cluster of *M. fascicularis*. The  
246 most striking feature of our results is that several amino acid changes are limited to *HBA1* or  
247 *HBA2*, with a stark separation of two possible substitutions at alpha globin amino acid  
248 position 71. Gly71Arg and Gly71Glu both occur as part of more than one unique *HBA1*  
249 (Gly71Arg) or *HBA2* (Gly71Glu) sequence. This indicates that their *HBA1* and *HBA2*  
250 specificity is widespread in Indonesian *M. fascicularis*, and not a result of inbreeding in the  
251 colony. Our population genetic simulations show that such *HBA1* or *HBA2* specificity is  
252 more likely to be maintained over longer periods of time if at least one of these amino acid  
253 substitutions is under selection. Amino acid substitutions at alpha globin position 71 are  
254 associated with the electrophoretic phenotypes **A**, **AQ** and **AQX** explored in historical  
255 studies. Our geographical analysis showed that there is an association between variant  
256 haemoglobin electrophoretic phenotypes in *M. fascicularis* and the presence of virulent  
257 macaque malarias. We contend, therefore, that the most likely selective pressure to account

258 for the *HBA1* and *HBA2* specificity of alpha globin variants in *M. fascicularis* is malaria  
259 selection.

260

261 The closest evolutionary relative of *M. fascicularis* is the rhesus macaque, *Macaca mulatta*.  
262 Data from a recently published whole genome sequencing study of *M. mulatta* (Xue, et al.  
263 2016) allows us to analyse whether the polymorphisms we identified in *M. fascicularis* occur  
264 in its sister species. We were able to obtain partial alpha globin exon 2 sequences for 98 *M.*  
265 *mulatta* (supplementary information section 1.5). The only non-synonymous change detected  
266 in the *M. mulatta* samples was the His78Gln substitution also found in *M. fascicularis*, where  
267 it occurs in both *HBA1* and *HBA2* sequences (supplementary table 8). Unfortunately, the  
268 short reads used for whole genome sequencing make it impossible to distinguish *HBA1* from  
269 *HBA2* sequences in the *M. mulatta* samples. The reasons for the maintenance of His78Gln as  
270 a (possible) trans-species polymorphism are unclear. However, the fact that we observed  
271 none of the *HBA1* or *HBA2* specific amino acid changes belonging to *M. fascicularis* in the  
272 *M. mulatta* data shows that *HBA1* or *HBA2* specific amino acid substitutions are not  
273 necessarily a feature of macaque alpha globin generally. It has been noted that *M. mulatta*  
274 and *M. fascicularis* differ in their susceptibility to malaria, and that malaria itself may have  
275 driven their speciation (Wheatley 1980). Higher admixture of *M. fascicularis* with *M. mulatta*  
276 has been suggested to increase susceptibility to *P. cynomolgi* in breeding colonies (Zhang, et  
277 al. 2017). Our observations of *M. mulatta* and *M. fascicularis* alpha globin are consistent with  
278 these hypotheses.

279

280 Five other macaque species: *Macaca nemestrina*, *Macaca arctoides*, *Macaca assamensis*,  
281 *Macaca radiata*, and *Macaca sinica* possess variant haemoglobins which may be similar to  
282 *M. fascicularis* Q haemoglobin (fig. 4, supplementary table 9). *Macaca fascicularis* and *M.*  
283 *nemestrina* were the only macaques naturally found infected with the virulent parasites *P.*  
284 *knowlesi* and *P. coatneyi* (Eyles, et al. 1962) (supplementary table 10). However, recent  
285 surveys of *M. arctoides* from Thailand present evidence that these species are also naturally  
286 infected with *P. knowlesi* and *P. coatneyi* (Fungfuang, et al. 2020). *M. radiata*, the Bonnet  
287 macaque, is native to southwest India and is infected with three malaria species, including *P.*  
288 *fragile* (Ramakrishnan and Mohan 1962; Dissanaike, et al. 1965). *Plasmodium fragile*  
289 undergoes deep vascular schizogony like *P. coatneyi* and *P. knowlesi* and causes 33%  
290 mortality in intact *M. mulatta* (Eyles 1963; Coatney, et al. 1971). A range of different amino  
291 acid substitutions may be responsible for these different variant haemoglobins (see fig. 4

292 legend), but it is striking that a correlation between the presence of virulent malaria and  
293 variant macaque haemoglobins may extend beyond *M. fascicularis*.

294

295 A protective effect of haemoglobin electrophoretic variant phenotypes could help explain  
296 puzzling results from experimental infection trials. *P. knowlesi* infection has a consistently  
297 mild course in *M. fascicularis* exported from the Philippines, whilst ‘Malayan’ animals  
298 directly exported from Singapore suffered fatal infection (Schmidt, et al. 1977). The **AQ**  
299 phenotype appears universal among *M. fascicularis* in the Philippines (supplementary table  
300 2). Long-tailed macaques from Singapore, by contrast, are more likely to display an **A** band  
301 alone (proportion A band alone = 0.7, n = 10) (Barnicot, et al. 1966) (fig. 1, supplementary  
302 table 2). A further study showed that *P. coatneyi* was fatal to splenectomized *M. fascicularis*  
303 from Mauritius, but not to splenectomized Philippine *M. fascicularis* (Migot-Nabias, et al.  
304 1999). Mauritian animals are less likely to display variant alpha globin phenotypes  
305 (proportion variants = 0.09; n = 201) (Kondo, et al. 1993) than Philippine animals (proportion  
306 variants = 1.0; n = 118) (Ishimoto, et al. 1970; Ishimoto 1972), further suggesting that variant  
307 alpha globin phenotypes predict survival when infected with virulent malarias.

308

309 The infection study of *M. fascicularis* exported from the Philippines and Singapore also  
310 suggests a potential mechanism of protection by alpha globin variants. Only ring stage  
311 parasites were observed in blood of the lethally infected Singaporean animals, suggesting  
312 parasites were sequestering outside of peripheral circulation. However, all asexual  
313 development stages of the parasite were observed in the blood of the non-lethally-infected  
314 Philippine animals, suggesting parasite sequestration was less successful (Schmidt, et al.  
315 1977). In humans, there is an association between sequestered parasite biomass and severe  
316 disease (Dondorp, et al. 2005). The parasite antigen PfEMP1 (*Plasmodium falciparum*  
317 erythrocyte membrane protein 1) is an important regulator of cytoadherence in the human  
318 parasite *Plasmodium falciparum*. *P. knowlesi* possesses a similar antigen called SICA  
319 (Schizont-infected cell agglutination antigen) (Brown and Brown 1965; Korir and Galinski  
320 2006) although its role in cytoadherence is not well characterized. Human  
321 haemoglobinopathies can affect the expression of PfEMP1 (Fairhurst, et al. 2005; Cholera, et  
322 al. 2008), reducing cytoadherence. It is possible that macaque haemoglobin variants can  
323 affect the expression of *P. knowlesi* cytoadherence molecules, and that macaque red blood  
324 cells containing more than one major adult haemoglobin (as we expect to occur in all  
325 Philippine animals) are associated with reduced sequestration and improved health outcomes.

326

327 The extensive variation of *M. fascicularis* alpha globin is contrasted by just one beta globin  
328 amino acid substitution reported in *M. fascicularis* from Bali (Kawamoto, et al. 1984).  
329 *Macaca fascicularis* may therefore present an inversion of the situation in *Homo sapiens*,  
330 whose major malaria protective amino acid substitutions occur in beta globin, not alpha.  
331 Primate beta globin is encoded by a single gene in the beta globin cluster, whilst primate  
332 alpha globin is encoded by two or more linked genes in the alpha globin cluster. An alpha  
333 globin cluster can therefore contain two genes encoding structurally different alpha globin  
334 proteins. If such an alpha globin cluster becomes fixed in a population, the entire population  
335 might be able to express two or more different types of haemoglobin. It may be that this state  
336 is particularly advantageous against malaria. Since beta globin is typically encoded by just  
337 one gene in vertebrates, the equivalent situation is extremely unlikely to emerge for beta  
338 globin.

339

340 No human population has fully fixed a malaria protective haemoglobinopathy mutation,  
341 although the Tharu population of the Terai region of Nepal (a holoendemic malaria region)  
342 has come close, with alpha thalassaemia frequencies > 80%. The alpha thalassaemic mutation  
343 in question is a deletion of one of the two alpha globin genes in the alpha globin cluster.  
344 Deleting just one alpha globin gene in the cluster allows the remaining alpha globin gene to  
345 continue to support the function of the cell and the entire population can, theoretically, enjoy  
346 the same malaria protective phenotype if everyone is homozygous for the deletion. It may be,  
347 although the mutations are very different, that high frequencies of alpha thalassaemic  
348 deletions in the Tharu population and the fixation of the **AQ** phenotype among Philippine  
349 *Macaca fascicularis*, represent similar states of population adaptation to malaria.

350

351 An additional observation from our sequencing was that many or all of our studied  
352 population possessed more than two copies of alpha globin per chromosome, specifically at  
353 least two copies of *HBA2* in addition to *HBA1*. The sheer diversity of patterns observed when  
354 proportions of sequences are considered (supplementary figs. 6 –8) is also suggestive of  
355 variation in the number of copies of alpha globin. Previous studies have shown that  
356 triplication or quadruplication of alpha globin reaches high frequencies in certain *M.*  
357 *fascicularis* populations (Takenaka, et al. 1991; Takenaka, et al. 1993), and have noted  
358 varying proportions of variant haemoglobins in different samples (Barnicot, et al. 1966). If  
359 we allow for the possibility that an alpha globin cluster containing three alpha globin genes

360 generates an excess of alpha globin chains, and that this has some malaria protective effect,  
361 then it is possible that *M. fascicularis* alpha globin copy number variation arose through its  
362 malaria protective advantage, and this advantage was subsequently enhanced by the  
363 incorporation of amino acid changes into some of the duplicated alpha globin genes. The  
364 contrasting routes that humans and *M. fascicularis* appear to have taken to achieve malaria  
365 protection may follow from the higher frequency of alpha globin deletions in humans (alpha  
366 thalassaemia), as opposed to alpha globin duplications in *M. fascicularis*.

367

368 There are alternative explanations for the observed patterns of genetic variation and fixation  
369 in alpha and beta globin genes in long-tailed macaques. Mitochondrial and low-coverage  
370 whole genome sequences demonstrate there is significant population structure between  
371 mainland and insular populations of long-tailed macaques (Tosi and Coke 2007;  
372 Kanthaswamy, et al. 2013; Yao, et al. 2020). While we find a compelling correlation between  
373 alpha globin phenotypic variation and malaria selection, it is possible that genetic drift on  
374 insular populations may explain some fixation of the **AQ** or **A** states, although we have  
375 checked for evidence of bottlenecks with available genetic data (supplementary figure 5).  
376 Our ability to test for positive selection (i.e. dn/ds ratios) within alpha globin is currently  
377 limited by only having short read sequences from a single population (Kryazhimskiy and  
378 Plotkin 2008). Long-read sequencing of globin genes across the population structure of long-  
379 tailed macaques would allow the application of dn/ds ratio tests for positive selection on *M.*  
380 *fascicularis* alpha globin.

381

382 Haldane's malaria hypothesis was developed to explain the geographical association between  
383 human malaria and heritable blood disorders. The malaria hypothesis has been validated with  
384 clinical evidence for a malaria protective effect of haemoglobinopathies in human  
385 populations. We find a geographical association between *M. fascicularis* alpha globin variant  
386 phenotypes and malaria selection, measured as the presence of virulent malaria species.  
387 Furthermore, we find that the specificity of amino acid variants to particular copies of alpha  
388 globin may be a signature of natural selection. Further research is required to prove that this  
389 selection is from malaria. Long read sequencing would provide higher quality data in order  
390 to correctly phase and assign haplotypes within this gene complex. This, combined with SNP  
391 panels to control for population structure and test for signatures of positive selection would  
392 provide more confidence in these findings. A significant challenge is demonstrating a benefit

393 of variant alpha globins at the individual level in *M. fascicularis*. This could be done with  
394 experimental infections, but such experimentation carries significant ethical concerns.

395

396 It is becoming clear that there are many parallels between malaria resistance mechanisms  
397 among different vertebrate species. There are higher rates of adaptation in mammalian  
398 proteins that interact with *Plasmodium* species versus matched controls, and domains of  
399 alpha spectrin have been identified as potential sites for primate evolution in response to  
400 malaria (Ebel, et al. 2017). Examples of human malaria resistance traits with parallels in  
401 other species include sickle haemoglobin (a convergent form of sickle haemoglobin exists in  
402 deer (Esin, et al. 2017)) and FY variation (Duffy negativity confers human resistance to  
403 *Plasmodium vivax*; yellow baboon FY variation affects their susceptibility to malaria-like  
404 *Hepatocystis* parasites (Tung, et al. 2009)). As we add to the list of ways that different hosts  
405 have adapted the same proteins to combat the problem of malaria, we increase our potential  
406 to uncover biochemical similarities that advance our understanding of how each protective  
407 mechanism operates at the molecular level.

408

409

## 410 **Materials and Methods**

### 411 *Geographical analyses*

412 We identified twelve electrophoretic population surveys of *M. fascicularis* alpha globin from  
413 the literature (Barnicot, et al. 1966; Barnicot, et al. 1970; Ishimoto, et al. 1970; Ishimoto  
414 1972; Nozawa, et al. 1977; Smith and Ferrell 1980; Kawamoto and Ischak 1981; Kawamoto,  
415 et al. 1984; Kawamoto, et al. 1989; Tomiuk 1989; Kondo, et al. 1993; Perwitasari-Farajallah,  
416 et al. 1999) (supplementary table 2). One study did not include sufficient geographical  
417 information so was excluded (Tomiuk 1989). Another survey was based on samples derived  
418 from an introduced population of Mauritian long-tailed macaques (Kondo, et al. 1993), where  
419 there is no active malaria transmission. We used the remaining surveys to conduct the  
420 geographic analyses. Specificity of sample origins ranged from troop-level latitude and  
421 longitude (Kawamoto, et al. 1984) to whole countries. Since our malaria selection likelihood  
422 calculation (detailed below) was at the regional level, we aggregated the alpha globin data by  
423 region (i.e. Sumatra Utara).

424

425 We sought to analyse a possible link between long-tailed macaque alpha globin phenotypic  
426 variation and malaria selection. This required us to develop a proxy for malaria selection

427 across the range of long-tailed macaques. We used the likely presence of virulent malaria (*P.*  
428 *coatneyi* or *P. knowlesi*) in a region as our malaria selection proxy. We consider *P. coatneyi*  
429 and *P. knowlesi* to be the most virulent malarias infecting *M. fascicularis*, for the following  
430 reasons: (i) *P. coatneyi* and *P. knowlesi* have been shown to be capable of killing some,  
431 though not all, experimentally infected *M. fascicularis* (Schmidt, et al. 1977; Migot-Nabias,  
432 et al. 1999); (ii) unlike *P. cynomolgi*, *P. inui*, or *P. fieldi*, *P. coatneyi* and *P. knowlesi* cause  
433 lethal infections in the sister species of *M. fascicularis*, *M. mulatta* (Coatney, et al. 1971), and  
434 (iii) like *P. falciparum*, but unlike other macaque malarias, *P. knowlesi* and *P. coatneyi*  
435 undergo deep vascular schizogony – attachment of infected RBCs to the vascular  
436 endothelium– which may be associated with increased pathology (Desowitz, et al. 1969;  
437 Miller, et al. 1971). To assess the presence of *P. coatneyi* or *P. knowlesi* and malaria  
438 sampling effort across the range of *M. fascicularis*, we used a systematic review of  
439 publications that reported surveys of malaria in primates (Faust and Dobson 2015).

440

441 Using the number of long-tailed macaques surveyed for malaria and the number of macaques  
442 infected with virulent malarias (supplementary table 3), we fitted a probability density  
443 function (PDF; equation 1) for the presence of virulent malarias at each location using a beta  
444 distribution with a uniform prior:

445

$$\text{Beta}(\tau, \alpha, \rho) = \frac{\tau^{\alpha-1}(1-\tau)^{\rho-1}}{B(\alpha, \rho)} \quad \text{Equation 1}$$

446 where  $\alpha$  is the number of individuals with virulent malarias,  $\rho$  is the number of individuals  
447 without virulent malarias and  $0 \leq \tau \leq 1$ . The cumulative distribution function (CDF) for the  
448 probability density function (PDF) defined by equation 1 can then be used to determine a  
449 likelihood that virulent malaria is present or not present in a given locality by using a specific  
450 cutoff. We used a cutoff of 0.02 for the results reported in the main text, meaning we took a  
451 “likelihood of malaria being present” equal to the area under the PDF for a given region  
452 (equation 1) where the probability of observing a virulent malaria was  $>2\%$  (and vice versa,  
453 the “likelihood of malaria not being present” was 1- the aforementioned area). A sensitivity  
454 analysis adjusting this cutoff is described in the supplementary methods (supplementary  
455 information section 1.2). We used a Metropolis-Hastings sampler with 100000 estimates,  
456 20000 burn in and thinning every 100 to estimate the proportion of long-tailed macaques that  
457 had variant alpha globin phenotypes in areas with high malaria selection (high likelihood of

458 virulent malarias) compared to low malaria selection (low likelihood of virulent malarias).  
459 The MCMC sampler was implemented in the *sampyl* package in Python 2.7 (Python  
460 Software Foundation 2010).

461

#### 462 ***Genetic characterization of HBA1 and HBA2 exon 2***

463 Genomic DNA for 78 *Macaca fascicularis* from the UK National Institute for Biological  
464 Standards and Control (NIBSC) breeding colony was obtained from archived samples held  
465 by NIBSC. These samples were taken during historical health screening of the long-tailed  
466 macaques as part of standard colony management. The ancestors of these *M. fascicularis*  
467 were from Indonesia (further geographical specificity is unknown), and they have been bred  
468 in the UK for ~14 generations but still retain high MHC diversity (Mitchell, et al. 2012).  
469 *HBA1* and *HBA2* were amplified separately using gene-specific primers (supplementary table  
470 5). For the initial PCR, reactions were run with 30ng template gDNA and initially heated to  
471 98°C for 30sec, followed by 20 cycles of denaturation (98°C, 15 sec), annealing (70°C or  
472 68°C, 20 sec), and extension (72°C, 30 sec), and then a final extension at 72°C for 2 minutes  
473 using Q5 High-Fidelity Polymerase (New England BioLabs). We conducted a nested PCR to  
474 sequence exon 2 (204bp; 68 amino acids) and flanking introns (5' end: 9 bp; 3' end: 121 bp)  
475 for both HBA amplicons, while adding a unique barcode to each sample: a 7 nucleotide  
476 sequence at the 5' end of the primers formed a (7,4) Hamming barcode (Bystrykh 2012). 78  
477 samples were multiplexed in a single MiSeq library in this way. The nested PCR was  
478 performed in a 50 µl solution containing polymerase master mix, 300 nM each of barcoded  
479 forward primer and barcoded reverse primer, and 1 µl of template DNA (supplementary table  
480 2). Reactions were initially heated to 98 °C for 30 sec, followed by 10 cycles of 98 °C for 15  
481 s, 68 °C for 20 s, 72 °C for 30 s. Reactions were completed at 72 °C for 2 min. Barcoded  
482 PCR products were pooled and purified using a QIAquick PCR Purification Kit (Qiagen,  
483 Hilden, Germany) according to manufacturer's protocol. Illumina library preparation was  
484 performed with this pool using the NEBNext Ultra kit (NEB, Ipswich, MA) according to the  
485 manufacturer's instructions, with size selection. Final concentration was measured by qPCR  
486 using the NEBNext Library Quant Kit for Illumina (NEB) according to the manufacturer's  
487 instructions. Libraries were diluted from 26.2 nM (*HBA1*) and 33.4 nM (*HBA2*) to a  
488 concentration of 4 nM and pooled for sequencing on a MiSeq (Illumina, San Diego, CA).

489

490 Illumina output was demultiplexed using custom scripts in Python 2.7 (Python Software  
491 Foundation 2010). Output was denoised and dereplicated using dada2 v1.2.1 (Callahan, et al.

492 2016), and further manipulated using data.table v1.10.4 (Dowle and Srinivasan 2017) in R  
493 v3.2 (R Core Team 2015). Unique sequences that were found at a frequency less than 2% of  
494 an individual's total reads were assumed to be PCR errors and were excluded from this  
495 analysis.

496

497 ***Population genetic model***

498 **Model structure**

499 We set up an individual based model of a diploid population of constant size  $N$ . The alpha  
500 globin cluster consists of two linked alpha globin genes (*HBA1* and *HBA2*). Each individual  
501 in the model thus possesses 4 globin genes, 2 in a cluster inherited from their mother and 2 in  
502 a cluster inherited from their father. The ancestral alpha globin type is designated type  $\alpha^A$ .  
503 Two further alpha globin types,  $\alpha^B$  and type  $\alpha^C$ , represent possible (and mutually exclusive)  
504 amino acid changes in alpha globin. Every generation there is probability  $m$  that a single new  
505 migrant of a randomly generated genotype replaces an existing member of the population.  
506 The aforementioned randomly generated genotype consists of four globin types, randomly  
507 sampled with replacement from  $\alpha^A$ ,  $\alpha^B$  or  $\alpha^C$ , linked into two alpha globin clusters to create a  
508 diploid genotype.  $\alpha^B$  is always a neutral variant, no more or less fit than  $\alpha^A$ .  $\alpha^C$  may be under  
509 selection, generated by two processes. Firstly,  $d$  is the probability of any individual dying  
510 before reproducing in a given generation as a consequence of an environmental hazard (e.g.  $d$   
511 could represent the burden of malaria on the population). Individuals with 1 or more genes  
512 encoding  $\alpha^C$  in their genotype have a reduced probability of dying due to this hazard, such  
513 that their probability of dying from the hazard is equal to  $(1-p)d$ . Secondly,  $\alpha^C$  may be  
514 associated with a blood disorder. Individuals with 3 or more genes encoding  $\alpha^C$  in their  
515 genotype have probability  $k$  of dying before reproducing. Parameters  $p$  and  $k$  thus tune the  
516 advantages and disadvantages of genotypes containing globin type  $\alpha^C$ . If  $p=0$  and  $k=0$  then  
517  $\alpha^C$  becomes a neutral variant. Every generation, those individuals which survive the  
518 environmental hazard and the potential blood disorder cost go on to form the parents of the  
519 next generation. During the reproduction step, randomly chosen pairs of parents each produce  
520 a single offspring genotype generated according to Mendelian inheritance until the required  
521 population size of  $N$  is reached. Reciprocal crossing over between *HBA1* and *HBA2* (i.e. the  
522 swapping of a maternal *HBA1* with a paternal *HBA1* between the two clusters so that each  
523 *HBA2* ends up linked to an alternate *HBA1*) takes place in each individual with probability  $r$ .  
524 Gene conversion, defined here as the conversion of one randomly chosen sequence within a  
525 genotype to match a randomly chosen sequence from the other three alpha globin sequences

526 present in that genotype, takes place in each individual with probability  $c$ . The population  
527 evolves for  $t$  generations in each simulation.

528

529 In figure 3, a population genetic outcome with *HBA1* or *HBA2* specific mutations is defined  
530 as one in which:

531 •  $\alpha^B$  and  $\alpha^C$  are present in the population, each accounting for at least 5% of alpha  
532 globin sequences overall. This is to ensure that any *HBA1* or *HBA2* specificity was  
533 not a function of a mutation only being present at a very low frequency.

534 • At least 98% of  $\alpha^B$  sequences in the population occur at *HBA1* or *HBA2* only, and  
535 likewise at least 98% of  $\alpha^C$  sequences occur at *HBA1* or *HBA2* only.

536

537 Code used to implement this model is provided in 'AlphaGlobinPopulationGeneticModel.c'  
538 at [github.com/cfaustus/macaque\\_workspace](https://github.com/cfaustus/macaque_workspace).

539

#### 540 Choice of parameters

541 We simulated a population size ( $N$ ) of 10000. We simulated 10000 generations of evolution  
542 ( $t=10000$ ). Longer and larger simulations were not possible due to computational limitations,  
543 so it was not practical to simulate the mutations arising entirely *de novo* within the population  
544 at a realistic rate. To get around this limitation we considered scenarios in which mutations  
545 arrive in the population at random, assumed to be generated in a wider global population of  
546 *M. fascicularis* (fig. 3A-E), and simulations in which genetic diversity is introduced at the  
547 beginning of the simulation, and the stability of that diversity considered over time, (fig. 3D-  
548 E).

549

550 We tested three possible values for the probability of gene conversion ( $c$ ) and reciprocal  
551 crossing over ( $r$ ) in our simulations: 0,  $10^{-5}$  and  $5 \times 10^{-5}$ . These were chosen based on the rate  
552 of unequal crossing in the human alpha globin cluster during meiosis (a study of human  
553 sperm observed  $10^{-5}$  unequal crossing over events per sperm (Lam and Jeffreys 2007), thus a  
554 probability of  $10^{-5}$  per meiosis event). Although our model is not simulating unequal crossing  
555 over, this is the best estimate we have for a reasonable alpha globin recombination rate.

556

#### 557 **Supplementary Material**

558 **Supplementary** Information includes supplementary text sections 1.1 – 1.5, supplementary  
559 tables 1-10, and supplementary figures 1-9.  
560 The raw amplicon sequencing data of *Macaca fascicularis* generated in this study have been  
561 deposited in the Sequence Read Archive (BioProject ID: [PRJNA639946](#)) under the accession  
562 numbers: SRR12404495- SRR12404572 (HBA1) and SRR12404678- SRR12404755  
563 (HBA2)). Code associated with this research is available at  
564 [github.com/cfaustus/macaque\\_workspace](https://github.com/cfaustus/macaque_workspace).

565

566 **Acknowledgements.**

567 This work was supported by a Princeton-Oxford Collaborative Grant; the Wellcome Trust  
568 (BSP grant 096063/Z/11/Z); the Truman Foundation (CLF); National Defense Science and  
569 Engineering (CLF); the Schlumberger Foundation (FR); the Oxford Centre for Islamic  
570 Studies (FR); the Royal Society (SG) and the European Research Council (SG, grant  
571 DIVERSITY). The funders had no role in study design, data collection and analysis, decision  
572 to publish, or preparation of the manuscript.

573

574 **Competing Interests.**

575 The authors declare that there are no competing interests.

576

577 **References:**

578 Allison AC. 1954a. The distribution of the sickle-cell trait in East Africa and elsewhere, and  
579 its apparent relationship to the incidence of subtertian malaria. *Trans R Soc Trop Med Hyg*  
580 48:312-318.  
581 Allison AC. 1954b. Protection afforded by sickle-cell trait against subtertian malarial  
582 infection. *Brit Med J* 1:290.  
583 Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and Power of Bayes Prediction of  
584 Amino Acid Sites Under Positive Selection. *Mol Biol Evol* 19:950-958.  
585 Baglioni C. 1962. A Chemical Study of Hemoglobin Norfolk. *J Biol Chem* 237:69-74.  
586 Barnicot N, Wade P, Cohen P. 1970. Evidence for a Second Haemoglobin  $\alpha$ -Locus  
587 Duplication in *Macaca irus*. *Nature* 228:379-381.  
588 Barnicot NA, Huehns ER, Jolly CJ. 1966. Biochemical Studies on Haemoglobin Variants of  
589 the Irus Macaque. *Proc R Soc Lond B Biol Sci* 165:224-244.  
590 Bininda-Emonds ORP, Cardillo M, Jones KE, Macphee RDE, Beck RMD, Grenyer R, Price  
591 SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals.  
592 *Nature* 446:507-512.  
593 Brown KN, Brown IN. 1965. Immunity to Malaria: Antigenic Variation in Chronic Infections  
594 of *Plasmodium knowlesi*. *Nature* 208:1286-1288.  
595 Bystrykh LV. 2012. Generalized DNA Barcode Design Based on Hamming Codes. *PLoS*  
596 One 7:e36852.

597 Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2:  
598 High-resolution sample inference from Illumina amplicon data. *Nat Meth* 13:581-583.  
599 Cholera R, Brittain NJ, Gillrie MR, Lopera-Mesa TM, Diakité SAS, Arie T, Krause MA,  
600 Guindo A, Tubman A, Fujioka H, et al. 2008. Impaired cytoadherence of *Plasmodium*  
601 *falciparum*-infected erythrocytes containing sickle hemoglobin. *Proc Natl Acad Sci* 105:991-  
602 996.  
603 Coatney GR, Collins WE, Warren M, Contacos PG. 1971. *The Primate Malaria*. Washington  
604 DC: US Government Printing Office.  
605 Desowitz R, Miller L, Buchanan R, Permpanich B. 1969. The sites of deep vascular  
606 schizogony in *Plasmodium coatneyi* malaria. *Trans R Soc Trop Med Hyg* 63:198-202.  
607 Dissanaike AS, Nelson P, Garnham PCC. 1965. Two new malaria parasites, *Plasmodium*  
608 *cynomolgi ceylonensis* subsp. nov. and *Plasmodium fragile* sp. nov., from monkeys in  
609 Ceylon. *Ceylon Med J* 14:1-14.  
610 Dondorp AM, Desakorn V, Pongtavornpinyo W, Sahassananda D, Silamut K, Chotivanich K,  
611 Newton PN, Pitisuttithum P, Smithyman AM, White NJ, et al. 2005. Estimation of the Total  
612 Parasite Biomass in Acute Falciparum Malaria from Plasma PfHRP2. *PLoS Med* 2:e204.  
613 Dowle M, Srinivasan A. 2017. *data.table*: Extension of ‘*data.frame*’. Version 1.10.4.  
614 Ebel ER, Telis N, Venkataram S, Petrov DA, Enard D. 2017. High rate of adaptation of  
615 mammalian proteins that interact with *Plasmodium* and related parasites. *PLoS Genet*  
616 13:e1007023.  
617 Esin A, Bergendahl LT, Savolainen V, Marsh JA, Warnecke T. 2017. The genetic basis and  
618 evolution of red blood cell sickling in deer. *Nat Ecol Evol* 2:367-376.  
619 Eyles DE. 1963. The species of simian malaria: Taxonomy, morphology, life cycle, and  
620 geographical distribution of the monkey species. *J Parasitol* 49:866-887.  
621 Eyles DE, Fong YL, Warren M, Guinn EG, Sandosham AA, Wharton RH. 1962.  
622 *Plasmodium coatneyi*, a new species of primate malaria from Malaya. *American Journal for*  
623 *Tropical Medicine and Hygiene* 11:597-604.  
624 Fairhurst RM, Baruch DI, Brittain NJ, Ostera GR, Wallach JS, Hoang HL, Hayton K, Guindo  
625 A, Makobongo MO, Schwartz OM, et al. 2005. Abnormal display of PfEMP-1 on  
626 erythrocytes carrying haemoglobin C may protect against malaria. *Nature* 435:1117-1121.  
627 Faust C, Dobson AP. 2015. Primate malarias: diversity, distribution and insights for zoonotic  
628 *Plasmodium*. *One Health* 1:66-75.  
629 Fungfuang W, Udom C, Tongthainan D, Kadir KA, Singh B. 2020. Stump-Tailed Macaques  
630 (*Macaca Arctoides*) are New Natural Hosts for *Plasmodium Knowlesi*, *P. Inui*, *P. Coatneyi*  
631 and *P. Fieldi*. *Malaria J Under Review*.  
632 Haldane J. 1949. Disease and evolution. *Ric Sci Suppl A* 19.  
633 Hoffman F, Opazo J, Storz J. 2008. Rapid rates of lineage-specific gene duplication and  
634 deletion in the alpha globin family. *Mol Biol Evol* 25:591-602.  
635 International Union for the Conservation of Nature. 2019. The IUCN Red List of Threatened  
636 Species. In. [www.iucnredlist.org](http://www.iucnredlist.org).  
637 Ishimoto G. 1972. マカク属サルの血液蛋白変異に関する研究. *J Anthropol Soc Nippon*  
638 80:250-274.  
639 Ishimoto G, Tanaka T, Nigi H, Prychodko W. 1970. Hemoglobin Variation in Macaques.  
640 *Primates* 11:229-241.  
641 Kanthaswamy S, Ng J, Satkoski Trask J, George DA, Kou AJ, Hoffman LN, Doherty TB,  
642 Houghton P, Smith DG. 2013. The genetic composition of populations of cynomolgus  
643 macaques (*Macaca fascicularis*) used in biomedical research. *J Med Primatol* 42:120-131.  
644 Kawamoto Y, Ischak TM. 1981. Genetic Differentiation of the Indonesian Crab-eating  
645 Macaque (*Macaca fascicularis*): I. Preliminary Report on Blood Protein Polymorphism.  
646 *Primates* 22:237-252.

647 Kawamoto Y, Ishida T, Suzuki J, Tanenaka O, Varavudhi P. 1989. A preliminary report on  
648 the genetic variations of crab-eating macaques in Thailand. Kyoto University Overseas  
649 Research Report of Studies on Asian Non-human Primates 7:94-103.

650 Kawamoto Y, Tschak TM, Supriatna J. 1984. Genetic Variations Within and Between Troops  
651 of the Crab-eating Macaque (*Macaca fascicularis*) on Sumatra, Java, Bali, Lombok and  
652 Sumbawa, Indonesia. *Primates* 25:131-159.

653 Kondo M, Kawamoto Y, Nozawa K, Matsubayashi K, Watanabe T, Griffiths O, Stanley M.  
654 1993. Population genetics of crab-eating macaques (*Macaca fascicularis*) on the island of  
655 Mauritius. *Am J Primatol* 29:167-182.

656 Korir C, Galinski M. 2006. Proteomic studies of *Plasmodium knowlesi* SICA variant antigens  
657 demonstrate their relationship with *P. falciparum* EMP1. *Infect Genet Evol* 6:75-79.

658 Kryazhimskiy S, Plotkin JB. 2008. The Population Genetics of dN/dS. *PLoS Genet*  
659 4:e1000304.

660 Lam K-WG, Jeffreys AJ. 2007. Processes of *de novo* duplication of human  $\alpha$ -globin genes.  
661 *Proc Natl Acad Sci* 104:10950-10955.

662 Mahoney WC, Nute PE. 1979. Hemoglobin  $\alpha$ -chain variation in macaques: Primary  
663 structures of the  $\alpha$ I and  $\alpha$ II chains from the adult hemoglobins of Malaysian *Macaca*  
664 *nemestrina*. *Arch Biochem Biophys* 196:64-72.

665 Maita T, Tanioka Y, Nakayama S, Matsuda G. 1985. Amino-acid sequences of the two major  
666 components of adult hemoglobins from the stump-tail monkey, *Macaca speciosa*. *Biol Chem*  
667 *Hoppe Seyler* 366:1149-1154.

668 Migot-Nabias F, Ollomo B, Dubreuil G, Morelli A, Domarle O, Nabias R, Georges AJ,  
669 Millet P. 1999. *Plasmodium coatneyi*: Differential Clinical and Immune Responses of Two  
670 Populations of *Macaca fascicularis* from Different Origins. *Exp Parasitol* 91:30-39.

671 Miller LH, Fremont HN, Luse SA. 1971. Deep vascular schizogony of *Plasmodium*  
672 *knowlesi* in *Macaca mulatta*: distribution in organs and ultrastructure of parasitized red cells.  
673 *American Journal for Tropical Medicine and Hygiene* 20:816.

674 Mitchell JL, Mee ET, Almond NM, Cutler K, Rose NJ. 2012. Characterisation of MHC  
675 haplotypes in a breeding colony of Indonesian cynomolgus macaques reveals a high level of  
676 diversity. *Immunogenetics* 64:123-129.

677 Nozawa K, Shotake T, Ohkura Y, Tanabe Y. 1977. Genetic Variations Within and Between  
678 Species of Asian Macaques. *Jpn J Genet* 52:15-30.

679 Oliver E, Kitchen H. 1968. Hemoglobins of adult *Macaca speciosa*: an amino acid  
680 interchange (alpha 15 (gly--asp)). *Biochem Biophys Res Commun* 31:749.

681 Patrinos GP, Giardine B, Riemer C, Miller W, Chui DHK, Anagnou NP, Wajcman H,  
682 Hardison RC. 2004. Improvements in the HbVar database of human hemoglobin variants and  
683 thalassemia mutations for population and sequence variation studies. *Nucleic Acids Res*  
684 32:D537-D541.

685 Perwitasari-Farajallah D, Kawamoto Y, Suryobroto B. 1999. Variation in blood proteins and  
686 mitochondrial DNA within and between local populations of longtail macaques, on the Island  
687 of Java, Indonesia. *Primates* 40:581-595.

688 Python Software Foundation. 2010. Python v 2.7. Version 2.7.

689 R Core Team. 2015. R: A language and environment for statistical computing. Version v 3.2.  
690 Vienna, Austria: R Foundation for Statistical Computing.

691 Ramakrishnan S, Mohan B. 1962. An enzootic focus of simian malaria in *Macaca radiata*  
692 *radiata* Geoffroy of Nilgiris, Madras State, India. *Indian J Malariaol* 16:87-94.

693 Schmidt L, Fradkin R, Harrison J, Rossan R. 1977. Differences in the virulence of  
694 *Plasmodium knowlesi* for *Macaca irus* (*fascicularis*) of Philippine and Malayan origins.  
695 *American Journal for Tropical Medicine and Hygiene* 26:612.

696 Smith D, Ferrell R. 1980. A family study of the hemoglobin polymorphism in *Macaca*  
697 *fascicularis*. *J Hum Evol* 9:557-563.

698 Takenaka A, Takahashi K, Takenaka O. 1988. Novel Hemoglobin Components and Their  
699 Amino Acid Sequences from the Crab-Eating Macaque (*Macaca fascicularis*). *J Mol Evol*  
700 28:136-144.

701 Takenaka A, Udono T, Miwa N, Varavudhi P, Takenaka O. 1993. High frequency of  
702 triplicated  $\alpha$ -globin genes in tropical primates, crab-eating macaques (*Macaca fascicularis*),  
703 chimpanzees (*Pan troglodytes*), and orang-utans (*Pongo pygmaeus*). *Primates* 34:55-60.

704 Takenaka A, Ueda S, Terao K, Takenaka O. 1991. Multiple alpha-globin genes in crab-eating  
705 macaques (*Macaca fascicularis*). *Mol Biol Evol* 8:320.

706 Taylor SM, Parobek SM, Fairhurst RM. 2012. Haemoglobinopathies and the clinical  
707 epidemiology of malaria: a systematic review and meta-analysis. *Lancet Infect Dis* 12:457-  
708 468.

709 Tomiuk J. 1989. Hemoglobin Polymorphism in Macaques with Reference to the Evolution of  
710 *Macaca fascicularis* and *Macaca mulatta*. *Primates* 30:95-102.

711 Tosi AJ, Coke CS. 2007. Comparative phylogenetics offer new insights into the  
712 biogeographic history of *Macaca fascicularis* and the origin of the Mauritian macaques. *Mol*  
713 *Phylogenet Evol* 42 2:498-504.

714 Tung J, Primus A, Bouley AJ, Severson TF, Alberts SC, Wray GA. 2009. Evolution of a  
715 malaria resistance gene in wild primates. *Nature* 460:388-391.

716 Wade PT, Barnicot NA, Huehns ER. 1970. Structural studies on the major and minor  
717 haemoglobin of the monkey *Macaca irus*. *Biochim Biophys Acta* 221:450-466.

718 Weatherall DJ, Clegg JB. 2001a. Inherited haemoglobin disorders: an increasing global  
719 health problem. *Bull World Health Org* 79:704-712.

720 Weatherall DJ, Clegg JB. 2001b. The Thalassaemia Syndromes. Oxford, UK: Blackwell  
721 Science.

722 Weiss ML, Goodman M, Prychodko W, Moore GW, Tanaka T. 1973. An analysis of  
723 macaque systematics using gene frequency data. *J Hum Evol* 2:213-226.

724 Wheatley BP. 1980. Malaria as a Possible Selective Factor in the Speciation of Macaques. *J*  
725 *Mammal* 61:307-311.

726 Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, Dahdouli M, Deiros DR,  
727 Below JE, Salerno W. 2016. The population genomics of rhesus macaques (*Macaca mulatta*)  
728 based on whole-genome sequences. *Genome Res* 26:1651-1662.

729 Yao L, Witt K, Li H, Rice J, Salinas NR, Martin RD, Huerta-Sánchez E, Malhi RS. 2020.  
730 Population genetics of wild *Macaca fascicularis* with low-coverage shotgun sequencing of  
731 museum specimens. *Am J Phys Anthropol* 173:21-33.

732 Zhang X, Meng Y, Houghton P, Liu M, Kanthaswamy S, Oldt R, Ng J, Trask JS, Huang R,  
733 Singh B, et al. 2017. Ancestry, *Plasmodium cynomolgi* prevalence and rhesus macaque  
734 admixture in cynomolgus macaques (*Macaca fascicularis*) bred for export in Chinese  
735 breeding farms. *J Med Primatol* 46:31-41.

736

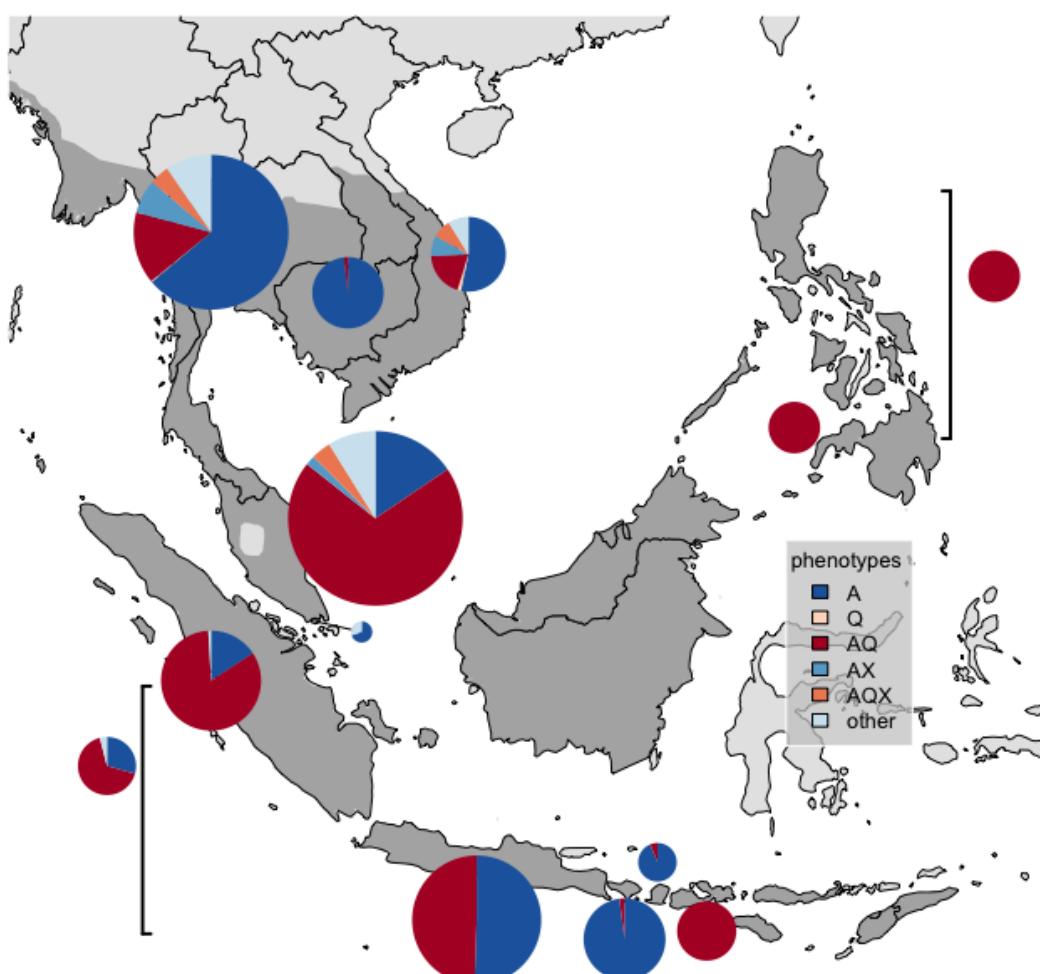
737

738 TABLES

739

740 **Table 1. Unique *HBA1* and *HBA2* sequences in our population of Indonesian long-tailed**  
741 **macaques.** The header row indicates the nucleotide and amino acid positions of each variant  
742 site in *M. fascicularis* *HBA1* (A) and *HBA2* (B). Nonsynonymous substitutions are indicated  
743 at the bottom of the table. The reference alpha globin sequences were taken from  
744 Chromosome 20 of a whole genome shotgun sequencing of *M. fascicularis*:  
745 Macaca\_fascicularis\_5.0 (NC\_022291.1; Indonesian origin). The nucleotide position (nt pos)  
746 is the location along the reference sequence. Unique sequences are named for whether they  
747 were observed in *HBA1* or *HBA2*, and are also assigned a numerical identifier. The “number  
748 of macaques” column gives the frequency of animals in which the sequence is found. A 6  
749 nucleotide sequence (ACGGGA) is present at positions 137834-137839 of the NC\_022291.1  
750 sequence (in the post exon 2 intron of HBA2) which we did not find in any of our samples.  
751 This appears to be a deletion in the intron of all our samples relative to the reference genome.  
752 Since this is not an example of variation within our samples we have not displayed it in table  
753 1B.

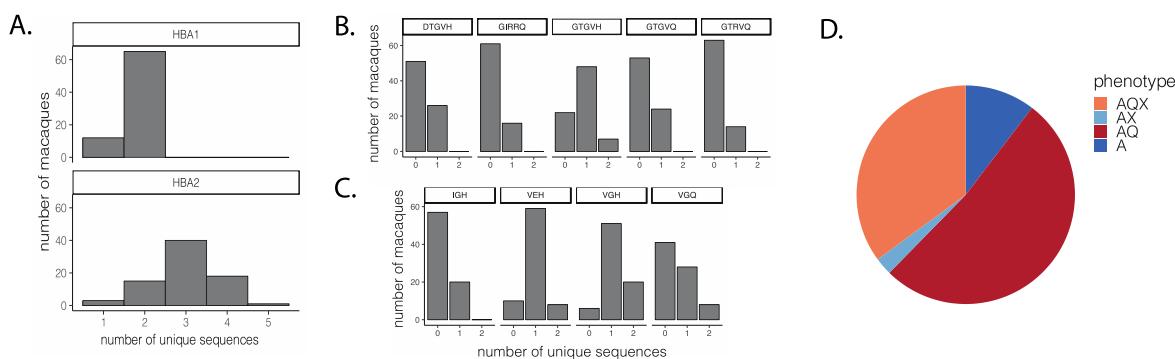
755 **FIGURES**



756

757 **Fig. 1. Map of haemoglobin variant phenotype frequencies.** Surveys of haemoglobin were  
758 collected from published studies of long-tailed macaques, *Macaca fascicularis* (see  
759 supplementary table 2 for full references). The range of *M. fascicularis* is in dark grey  
760 (International Union for the Conservation of Nature 2019). Square brackets indicate surveys  
761 that come from Indonesia or the Philippines without an island specified (for the Philippines  
762 the only island that was specified in any survey was Mindanao). Radii of pie charts are  
763 determined by the sample size of macaques studied at that location (range: n = 10 for  
764 Singapore; n = 677 for Malaysia).

765

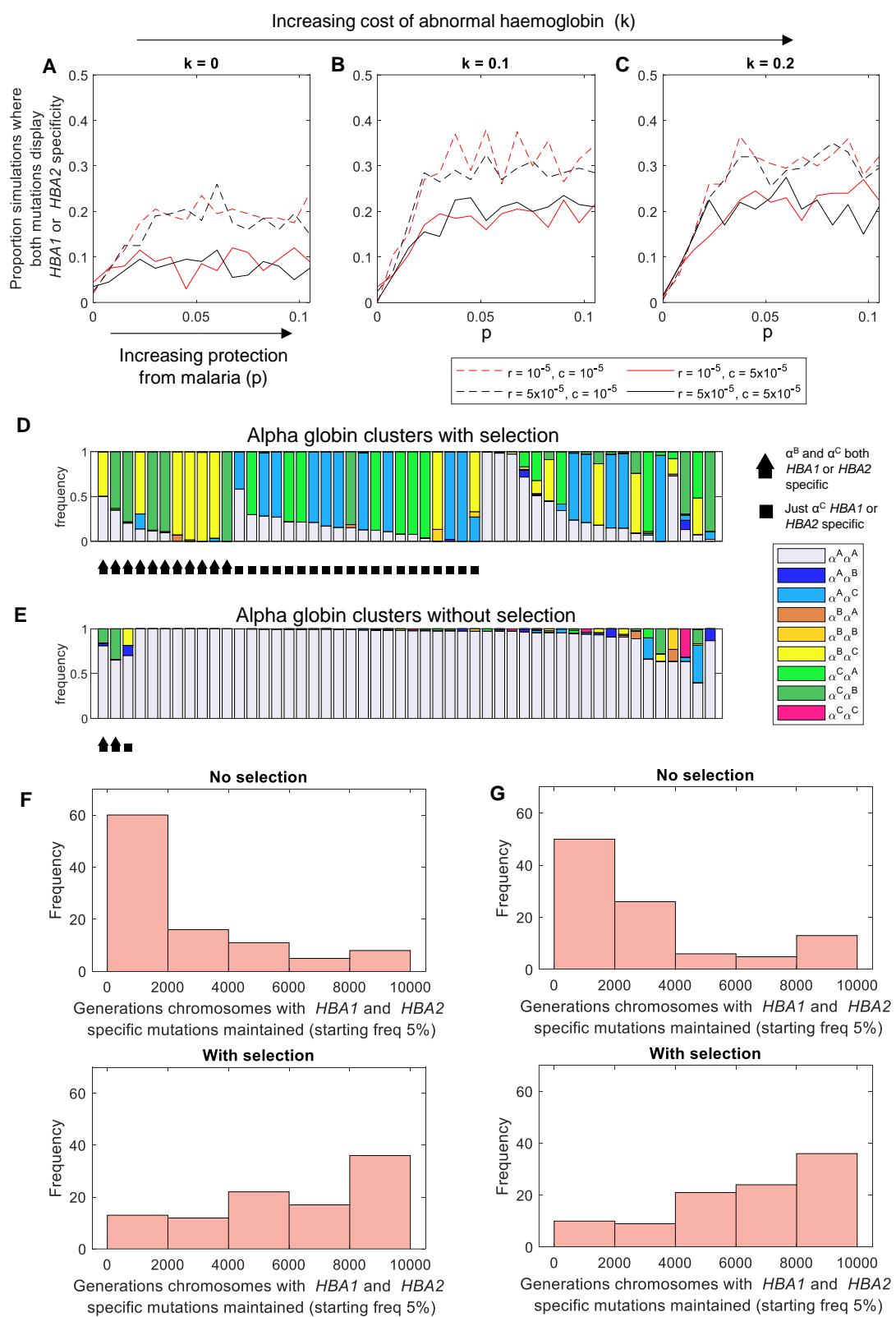


766

767 **Fig. 2. Distribution of *HBA1* and *HBA2* sequences observed in Indonesian *M.***

768 ***fascicularis*.** A) The majority of long-tailed macaques possessed two unique *HBA1* sequences  
769 and three unique *HBA2* sequences. B) Five unique combinations of amino acid acids  
770 (variable sites were found at positions 57, 67, 71, 73, and 78) were observed amongst *HBA1*  
771 in the 77 long-tailed macaques. C) Four unique amino acid combinations (variable amino  
772 acid sites are listed at 55, 71 and 78) were found in *HBA2*. D) Predicted population frequency  
773 of electrophoretic phenotypes on the basis that **Q** can be generated by Gly71Glu and **X** by  
774 Gly71Arg .

775



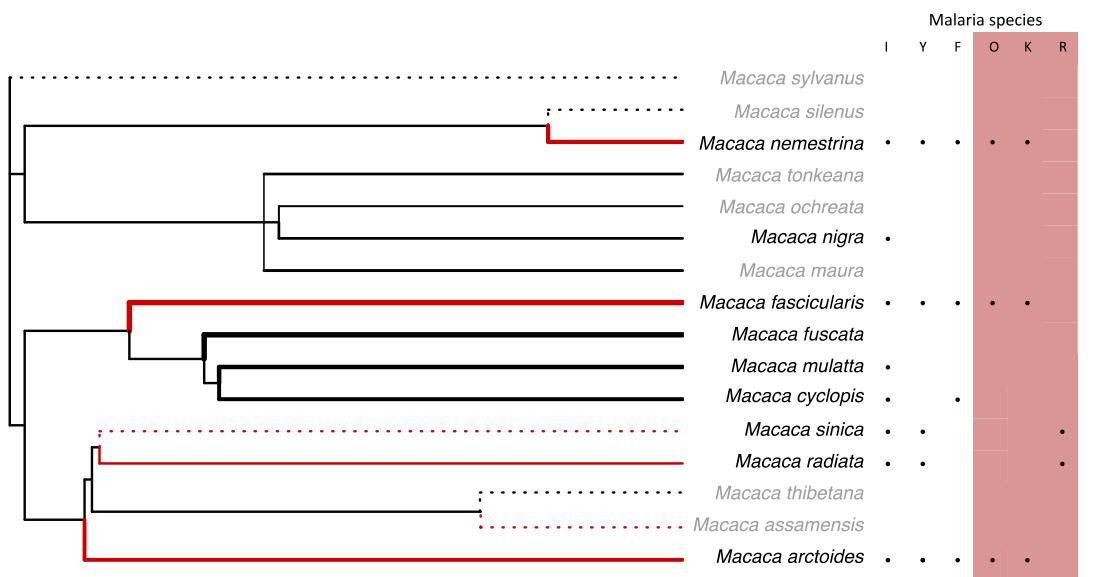
776

777 **Fig. 3. Selection increases the probability of observing *HBA1* or *HBA2* specific amino**  
 778 **acid substitutions.** Three alpha globin types are possible in the model: the ancestral type  
 779 ( $\alpha^A$ ); a neutral variant ( $\alpha^B$ ) and a potentially selected variant ( $\alpha^C$ ). Panels A-C illustrate how  
 780 the properties of  $\alpha^C$  affect the probability of observing a population genetic outcome in which

781 both non ancestral types are present in the population, each associated with only *HBA1* or  
782 only *HBA2* (see Methods for a more detailed description of the thresholds used to define this  
783 state).  $p$  (x axis of each panel) is the protection against an environmental hazard such as  
784 malaria provided by having  $\geq 1$  copies of  $\alpha^C$  in a genotype , and  $k$  (title of each panel) is the  
785 disadvantage (i.e. blood disorder cost) associated with having  $\geq 3$  copies of  $\alpha^C$  in a genotype.  
786 The populations simulated for panels A-C all started with 100%  $\alpha^A$  at both *HBA1* and *HBA2*.  
787 Two different gene conversion (c) and reciprocal crossing over (r) probabilities were used, as  
788 indicated in the legend. Each of these are probability of gene conversion or reciprocal  
789 crossing over affecting the gametes of an individual macaque (see Methods for justification).  
790 Other parameters were:  $N=10000$ ,  $t=10000$ ,  $m= 0.2$  and  $d=0.05$ . 200 repeats were carried out  
791 at each combination of parameters. Panels (D and E) each visualise the distribution of  
792 chromosome types present at the end of 50 individual simulations. Each stacked bar  
793 represents one simulation and the relative proportions of the bands within each stacked bar  
794 indicate different possible chromosomes (see legend). If selection is included,  $k=0.2$  and  $p =$   
795 0.03. If selection is not included  $k=0$  and  $p=0$ . Other parameters in panels D and E were:  $r=$   
796  $10^{-5}$ ,  $c= 10^{-5}$ ,  $N=10000$ ,  $t=10000$ ,  $m= 0.2$  and  $d=0.05$ . Panels (F and G) display histograms of  
797 the number of generations for which at least 2 out of the 3 chromosomes  $\alpha^C\alpha^A$ ,  $\alpha^A\alpha^B$  and  $\alpha^C\alpha^B$   
798 were maintained at frequencies  $>2.5\%$  in simulations of a closed population ( $m=0$ ) where the  
799 starting frequencies of each chromosome were: 5%  $\alpha^A\alpha^B$ ; 5%  $\alpha^C\alpha^A$  ; 5%  $\alpha^C\alpha^B$  and 85%  $\alpha^A\alpha^A$ .  
800 In panel F,  $r= 10^{-5}$  and  $c= 10^{-5}$  and in panel G  $r= 0$  and  $c= 0$ . In both F and G, if selection is  
801 included,  $k=0.2$  and  $p = 0.03$ ; if selection is not included  $k=0$  and  $p=0$ . Other parameters in F  
802 and G were  $N=10000$ ,  $d=0.05$ ,  $t=10000$ . 100 repeats were carried out at each combination of  
803 parameters.

804

805



806

807 **Fig. 4. Major variant alpha globin phenotypes in macaques.** The phylogeny is a species  
808 tree of macaques truncated from the mammalian super tree (Bininda-Emonds, et al. 2007).  
809 The branches of the species tree are annotated with information on the phenotypic variation  
810 in alpha globin from electrophoretic studies (supplementary table 9). The line type of the  
811 branches indicates whether we have (solid) or have not (dotted) been able to identify a  
812 population-level survey of alpha globin phenotypes in the literature (see supplementary table  
813 9 for more details). The width of solid branches reflects the log of the total sample sizes of  
814 haemoglobin surveys (supplementary table 9)- thickest branch (*M. fuscata*) representing 2539  
815 individuals and thinnest branch (*M. ochreata*) representing 17 individuals. Macaque species  
816 with variant alpha globin phenotypes are denoted by red branches and most (not *M. radiata*)  
817 are sympatric over part of their ranges. All variant alpha globin phenotypes for which we  
818 were able to identify the original source papers were reported to migrate similarly to **AQ** in  
819 *M. fascicularis* (original descriptions of *M. sinica* and *M. assamensis* haemoglobin variants  
820 could not be found - see supplementary table 9 for more details). The biochemical origins of  
821 haemoglobin variants in non-*fascicularis* species are as follows: alpha globin Gly71Asp  
822 substitutions have been found in *Macaca nemestrina* (Mahoney and Nute 1979; Takenaka, et  
823 al. 1988); an alpha globin Asp15Gly substitution is responsible for an **A/Q-like**  
824 polymorphism in *Macaca arctoides* (Oliver and Kitchen 1968; Maita, et al. 1985), and the  
825 same polymorphism is also found in *M. assamensis*. *M. sinica* possesses an Ala12Asp  
826 polymorphism. An unknown mutation results in a polymorphism electrophoretically similar  
827 to the **A/Q** polymorphism in *M. radiata* (Weiss, et al. 1973). The adjacent table of malaria  
828 species shows the parasites that have been found in each macaque species: *Plasmodium inui*  
829 (**I**), *P. cynomolgi* (**Y**), *P. fieldi* (**F**), *P. coatneyi* (**O**), *P. knowlesi* (**K**), *P. fragile* (**R**), and *P.*

830 *simovale* (S) (supplementary table 10). The parasite species highlighted in red represent  
831 known virulent malarias, as defined in the Methods and Discussion. Macaque species names  
832 shown in grey are species that have not been sampled for malaria.  
833