Edinburgh Research Explorer

The genetic basis and evolution of red blood cell sickling in deer

1   **The genetic basis and evolution of red blood cell sickling in deer**

2

3   Alexander Esin[1,2], L. Therese Bergendahl[3], Vincent Savolainen[4,5], Joseph A. Marsh[3],

4   Tobias Warnecke[1,2*]

5

6   [1]Molecular Systems Group, MRC London Institute of Medical Sciences (LMS), Du

7   Cane Road, London W12 0NN, United Kingdom

8   [2]Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London,

9   Du Cane Road, London W12 0NN, United Kingdom

10   [3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine,

11   University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United

12   Kingdom

13   [4]Department of Life Sciences, Silwood Park Campus, Imperial College London,

14   Ascot SL5 7PY, United Kingdom

15   [5]University of Johannesburg, PO Box 524, Auckland Park, 2006, South Africa

16

17   [*]Corresponding author

18   Corresponding author contact information:
19
20   Tobias Warnecke (tobias.warnecke@lms.mrc.ac.uk, Tel: +44 (0) 20 83838232)

21   **Crescent-shaped red blood cells, the hallmark of sickle cell disease, present a**

22   **striking departure from the biconcave disc shape normally found in mammals.**

23   **Characterized by increased mechanical fragility, sickled cells promote**

24   **haemolytic anaemia and vaso-occlusions and contribute directly to disease in**

25   **humans. Remarkably, a similar sickle-shaped morphology has been observed in**

26   **erythrocytes from several deer species, without obvious pathological**

27   **consequences. The genetic basis of erythrocyte sickling in deer, however, remains**

28   **unknown. Here, we determine the sequences of human β-globin orthologs in 15**

29   **deer species and use protein structural modelling to identify a sickling**

30   **mechanism distinct from the human disease, coordinated by a derived valine**

31   **(E22V) that is unique to sickling deer. Evidence for long-term maintenance of a**

32   **trans-species sickling/non-sickling polymorphism suggests that sickling in deer is**

33   **adaptive. Our results have implications for understanding the ecological regimes**

34   **and molecular architectures that have promoted convergent evolution of sickling**

35   **erythrocytes across vertebrates.**

36

37

38 Human sickling is caused by a single amino acid change (E6V) in the adult β-globin

39 (HBB) protein[1]. Upon deoxygenation, steric changes in the haemoglobin tetramer

40 enable an interaction between 6V and a hydrophobic acceptor pocket (known as the

41 EF pocket) on the β-surface of a second tetramer[2,3]. This interaction promotes

42 polymerization of mutant haemoglobin (HbS) molecules, which ultimately coerces

43 red blood cells into the characteristic sickle shape. Heterozygote carriers of the HbS

44 allele are typically asymptomatic[4] whereas HbS homozygosity has severe

45 pathological consequences and is linked to shortened lifespan[5]. Despite this, the HbS

46 allele has been maintained in sub-Saharan Africa by balancing selection because it

47 confers – by incompletely understood means – a degree of protection against the

48 effects of *Plasmodium* infection and malaria[6].

49

50 Sickling red blood cells were first described in 1840 – seventy years prior to their

51 discovery in humans[7] – when Gulliver[8] reported unusual erythrocyte shapes in blood

52 from white-tailed deer (*Odocoileus virginianus*). Subsequent research revealed that

53 sickling is widespread amongst deer species worldwide[8-11] (Fig. 1, Supplementary

54 Table 1). It is not, however, universal: red blood cells from reindeer (*Rangifer*

55 *tarandus*) and European elk (*Alces alces*, known as moose in North America) do not

56 sickle; neither do erythrocytes from most North American wapiti (*Cervus*

57 *canadensis*)[11,12].

58

59 Sickling deer erythrocytes are similar to human HbS cells with regard to their gross

60 morphology and the tubular ultrastructure of haemoglobin polymers[13-16]. Moreover, as

61 in humans, sickling is reversible through modulation of oxygen supply or pH[9,17] and

62 mediated by specific β-globin alleles[18,19], with both sickling and non-sickling alleles

63   segregating in wild populations of white-tailed deer[20]. As in humans, α-globin – two

64   copies of which join two β-globin proteins to form the haemoglobin tetramer – is not

65   directly implicated in sickling etiology[18,21]. Also as in humans, foetal haemoglobin

66   molecules, which incorporate distinct β-globin paralogs in human and deer (see

67   below), do not promote sickling under the same conditions[19]. But whereas HbS

68   sickling occurs when oxygen tension is low, deer erythrocytes sickle under high pO_2

69   and at alkaline pH[17]. Consequently, prime conditions for sickling in deer are likely

70   found in lung capillaries (rather than in systemic capillaries were oxygen is unloaded),

71   although *in vivo* sickling can also be observed in peripheral venous blood[22], especially

72   following exercise regimes that induce transient respiratory alkalosis[23]. Further,

73   unlike in humans, sickled deer erythrocytes do not exhibit increased mechanical

74   fragility *in vitro*[17,18] and the sickling allele in white-tailed deer (previously labelled

75   β[III]) is the major allele, with ≥60% of individuals homozygous for β[III] (REF. 20,24).

76   Remarkably, β[III] homozygotes do not display aberrant haematological values or

77   obvious pathological traits[25]. Together, these observations are consistent with reduced

78   physiological costs of sickling in deer. However, it is unknown whether sickling is

79   simply innocuous, as previously suggested[23], or plays an HbS-like adaptive role. In

80   addition, partial peptide digests of sickling white-tailed deer β-globins did not recover

81   the E6V mutation that causes sickling in humans[24], leaving the genetic basis of

82   sickling in deer unresolved.

83

84

85   **Results**

86

87   *The molecular basis of sickling in deer is distinct from that in the human disease*

4

88

89   To dissect the molecular basis of sickling in deer and elucidate its evolutionary

90   history and potential adaptive significance, we used a combination of whole-genome

91   sequencing, locus-specific assembly and targeted amplification to determine the

92   sequence of the HBB_A gene, which encodes the adult β-globin chain, in a

93   phylogenetically broad sample of 15 deer species, including both sickling and non-

94   sickling taxa (Fig. 1, Supplementary Table 1). Globin genes in mammals are located

95   in paralog clusters, which – despite a broadly conserved architecture – constitute

96   hotbeds of pseudogenization, gene duplication, conversion, and loss[26,27]. In ruminants,

97   the entire β-globin cluster is triplicated in goat (*Capra hircus*)[28] and duplicated in

98   cattle (*Bos taurus*)[29], where two copies of the ancestral β-globin gene sub-

99   functionalized to become specifically expressed in adult (HBB_A) and foetal (HBB_F)

100   blood. Based on a recent draft assembly of a white-tailed deer (*O. virginianus*

101   *texanus*) genome, the architecture of the β-globin cluster mirrors that seen in cattle,

102   consistent with the duplication event pre-dating the Bovidae-Cervidae split

103   (Supplementary Figure 1). Primers designed before this assembly became available

104   frequently co-amplified HBB_A and HBB_F (see Methods and Supplementary Figure 2).

105   In the first instance, we therefore assigned foetal and adult status based on residues

106   specifically shared with either HBB_A or HBB_F in cattle, which results in independent

107   clustering of putative HBB_A and HBB_F genes on an $HBB_{A/F}$ gene tree (Supplementary

108   Figure 3). To confirm these assignments, we sequenced mRNA from the red cell

109   component of blood from an adult Père David's deer (*Elaphurus davidianus*) and

110   assembled the erythrocyte transcriptome *de novo* (see Methods). We identified a

111   highly abundant β-globin transcript (>200,000 transcripts/million) corresponding

112   precisely to the putative adult β-globin gene amplified from genomic DNA of the

113   same individual (Supplementary Figure 4). Reads that uniquely matched the putative

114   HBB$_A$ gene were >2000-fold more abundant than reads uniquely matching the

115   putative HBB$_F$ gene, which is expressed at low levels. This is similar to the situation

116   in humans, where transcripts of HBG, a distinct paralog that convergently evolved

117   foetal expression, are found at low abundance in adult blood[30]. Finally, our

118   assignments are consistent with partial peptide sequences for white-tailed deer[24],

119   fallow deer (*Dama dama*)[31] and reindeer[32] that were previously obtained from the

120   blood of adult individuals.

121

122   We then considered deer HBB$_A$ orthologs in a wider mammalian context, restricting

123   analysis to species with high-confidence HBB assignments (see Methods). Treating

124   wapiti as non-sickling, and four species as indeterminate (no or insufficient

125   phenotyping of sickling; see Supplementary Table 1), we find three residues (Fig. 1)

126   that discriminate sickling from non-sickling species: 22 (non-sickling: E, sickling:

127   V/I), 56 (*n-s*: H, *s*: G), and 87 (*n-s*: K, *s*: Q/H). The change at residue 22, from an

128   ancestral glutamic acid to a derived valine (isoleucine in *Pudu puda*) is reminiscent of

129   the human HbS mutation and occurs at a site that is otherwise highly conserved

130   throughout mammalian evolution.

131

132   *Structural modelling supports an interaction between 22V and the EF pocket*

133

134   To understand how sickling-associated amino acids promote polymerization, we

135   examined these residues in their protein structural context. Residue 22 lies on the

136   surface of the haemoglobin tetramer, at the start of the second alpha helix (Fig. 2a).

137   Close to residue 22 are residue 56 and two other residues that differ between non-

6

138   sickling reindeer and moose (but not wapiti) and established sickling species: 19 (*n-s*:

139   K, *s*: N) and 120 (*n-s*: K, *s*: G/S). Together these residues form part of a surface of

140   increased hydrophobicity in sickling species (Fig. 2b). Distal to this surface, residue

141   87 is situated at the perimeter of the EF pocket, which in humans interacts with 6V to

142   laterally link two β-globin molecules in different haemoglobin tetramers and stabilize

143   the parallel strand architecture of the HbS fibre[2,3,33,34]. Mutation of residue 87 in

144   humans can have marked effects on sickling dynamics[35]. For example, erythrocytes

145   derived from HbS/Hb Quebec-Chori (T87I) compound heterozygotes sickle like HbS

146   homozygotes[36] while Hb D-Ibadan (T87K) inhibits sickling[37].

147

148   Given the similarity between the human E6V mutation and E22V in sickling deer, we

149   hypothesized that sickling occurs through an interaction in *trans* between residue 22

150   and the EF pocket. To test whether such an interaction is compatible with fibre

151   formation, we carried out directed docking simulations centred on these two residues

152   using a homology model of oxy β-globin from white-tailed deer (see Methods). We

153   then used the homodimeric interactions from docking to build polymeric haemoglobin

154   structures, analogous to how the 6V-EF interaction leads to extended fibres in HbS

155   homozygotes. Strikingly, nearly half of our docking models resulted in HbS-like

156   straight, parallel strand fibres (Fig. 2c). In contrast, when we performed similar

157   docking simulations centred on residues other than 22V, nearly all were incompatible

158   or much less compatible with fibre formation (Fig. 2d). Out of all 145 β-globin

159   residues, only 19N, which forms a contiguous surface with 22V, has a higher

160   propensity to form HbS-like fibres. By contrast, when docking is carried out using the

161   deoxy β-globin structure, 22V is incompatible with fibre formation, consistent with

162   the observation that sickling in deer occurs under oxygenated conditions. Importantly,

163   when this methodology is applied to human HbS, we find that 6V has the highest

164   fibre formation propensity out of all residues under deoxy conditions (Supplementary

165   Figure 5), providing validation for the approach.

166

167   Next, we used a force field model to compare the energetics of fibre formation across

168   deer species. We find that known non-sickling species (Fig. 1) and two species

169   suspected to be non-sickling based on their β-globin primary sequence – Chinese

170   water deer (*Hydropotes inermis*) and roe deer (*Capreolus capreolus*) – exhibit energy

171   terms less favourable to fibre formation than sickling species (Fig. 2e). To elucidate

172   the relative contribution of 22V and other residues to fibre formation, we introduced

173   all single amino acid differences found amongst adult deer β-globins individually into

174   a sickling (*O. virginianus*) and non-sickling (*R. tarandus*) background *in silico* and

175   considered the change in fibre interaction energy. Changes at residue 22 have the

176   strongest predicted effect on fibre formation, along with two residues – 19 and 21 – in

177   its immediate vicinity (Supplementary Figure 6). Smaller effects of amino acid

178   substitutions at residue 87, as well as residues 117 (N in *P. puda* and *O. virginianus*)

179   and 118 (Y in *D. dama*) hint at species-specific modulation of sickling propensity. *In*

180   *silico* residue swaps at a shorter evolutionary time-scale, between non-sickling *C.*

181   *canadensis* and sickling sika deer (*Cervus nippon*)*,* similarly implicate 22V as a key

182   determinant of sickling (Supplementary Figure 6).

183

184   Taken together, the results support the formation of HbS-like fibres in sickling deer

185   erythrocytes via surface interactions centred on residues 22V and 87Q in β-globin

186   molecules of different haemoglobin tetramers. In contrast, previous attempts to model

187   interactions in the deer haemoglobin fibre, based on preliminary crystallographic data

8

188   for white-tailed deer haemoglobin[24,38,39], either incorrectly assumed a hexagonal fibre

189   architecture or proposed different relative orientations and contacts that fail to predict

190   differences between sickling and non-sickling chains.

191

192   *Evidence for incomplete lineage sorting during the evolution of* HBB$_A$

193

194   To shed light on the evolutionary history of sickling and elucidate its potential

195   adaptive significance, we considered sickling and non-sickling genotypes in

196   phylogenetic context. First, we note that the HBB$_A$ gene tree and the species tree

197   (derived from 20 mitochondrial and nuclear genes) are significantly discordant

198   (Approximately Unbiased test p<1e-61, see Methods). Notably, sickling and non-

199   sickling genotypes are polyphyletic on the species tree but monophyletic on the

200   HBB$_A$ tree where wapiti, an Old World deer, clusters with moose and reindeer, two

201   New World deer (Fig. 3a). Gene tree-species tree discordance can result from a

202   number of evolutionary processes, including incomplete lineage sorting, gene

203   conversion, introgression, and classic convergent evolution, where point mutations

204   arise and fix independently in different lineages. In our case, the convergent evolution

205   scenario fits the data poorly. Discordant amino acid states are found throughout the

206   HBB$_A$ sequence and are not limited to sickling-related residues. Furthermore, in

207   many instances, amino acids shared between phylogenetically distant species are

208   encoded by the same underlying codons. Conspicuously, this includes the case of

209   residue 120 where all three codon positions differ between sickling species

210   (GGT/AGT) and non-sickling relatives (AAG in reindeer, moose, and cattle;

211   Supplementary Figure 7, Supplementary Data File 1). Even if convergence were

212   driven by selection on a narrow adaptive path through genotype space, precise

9

213   coincidence of mutational paths at multiple non-synonymous and synonymous sites

214   must be considered unlikely. Rather, these patterns are *prima facie* consistent with

215   incomplete lineage sorting, a process that might have prominently accompanied the

216   rapid divergence of Old World from New World deer during the Miocene[40].

217

218   *Gene conversion affects HBB$_A$ evolution but does not explain the phyletic pattern of*

219   *sickling*

220

221   To shore up this conclusion and rule out alternative evolutionary scenarios, we next

222   asked whether identical genotypes, rather than originating from a common ancestor,

223   might have been independently reconstituted from genetic diversity present in other

224   species (via introgression) or in other parts of the genome (via gene conversion). To

225   evaluate the likelihood of introgression and particularly gene conversion, which has

226   been attributed a major role in the evolution of mammalian globin genes[26], we first

227   searched for evidence of recombination in an alignment of deer HBB$_A$ and HBB$_F$

228   genes. HBB$_F$ is the principal candidate to donate non-sickling residues to HBB$_A$ in a

229   conversion event given that it is itself refractory to sickling[19] and – as a recent

230   duplicate of the ancestral HBB$_A$ gene – retains high levels of sequence similarity.

231   Using a combination of phylogeny-based and probabilistic detection methods and

232   applying permissive criteria that allow inference of shorter recombinant tracts (see

233   Methods), we identify eight candidate HBB$_F$-to-HBB$_A$ events, two of which, in

234   Chinese water deer and wapiti, are strongly supported by different methods (Fig. 3b).

235   Importantly, however, we find no evidence for gene conversion involving residue 22

236   (Fig. 3b, Supplementary Figure 8) even when considering poorly supported candidate

237   events. Recombination between HBB$_F$ and/or HBB$_A$ genes therefore does not explain

10

238   the distribution of glutamic acids and valines at residue 22 across Old World and New

239   World deer. Consistent with this, removal of putative recombinant regions does not

240   affect the HBB_F/HBB_A gene tree, with wapiti robustly clustered with other non-

241   sickling species whereas white-tailed deer and pudu cluster with Old World sickling

242   species (Supplementary Figure 8). We further screened raw genome sequencing data

243   from white-tailed deer and wapiti for potential donor sequences beyond HBB_F, such

244   as HBE or pseudogenized HBD sequences, but did not find additional candidate

245   donors. Thus, although gene conversion is a frequent phenomenon in the history of

246   mammalian globins[26] and contributes to HBB evolution in deer, it does not by itself

247   explain the phylogenetic distribution of key sickling/non-sickling residues. Rather,

248   gene conversion introduces additional complexity on a background of incomplete

249   lineage sorting.

250

251   *Balancing selection has maintained ancestral variation in HBB_A*

252

253   The presence of incomplete lineage sorting and gene conversion confounds

254   straightforward application of rate-based (dN/dS-type) tests for selection, making it

255   harder to establish whether the sickling genotype is simply tolerated or has been under

256   selection. We therefore examined earlier protein-level data on HBB_A allelic diversity.

257   This allows us to include additional alleles previously identified from partial peptide

258   digests, for which we have no nucleotide-level data. For white-tailed deer, this

259   includes β^{II}, which is associated with a different flavour of polymerization that results

260   in matchstick-shaped erythrocytes[41], and two rarer non-sickling alleles, β^V and β^{VII}. β^{II}

261   encodes 22V and expectedly clusters with other sickling HBB_A sequences (Fig. 3c).

262   More importantly, the non-sickling white-tailed deer alleles cluster with non-sickling

11

263   HBB$_A$ orthologs rather than with the conspecific β^{II} and β^{III} alleles (Fig. 3c), as does

264   the HBB$_A$ sequence from *O. v. texanus*, for which we can also demonstrate clustering

265   at the nucleotide level (Supplementary Figure 3). Similarly, an alternate adult β-

266   globin chain previously observed in fallow deer[31], a predominantly sickling Old

267   World deer, clusters with non-sickling sequences (Fig. 3c). Finally, phenotypic

268   heterogeneity in wapiti[12] and sika deer[42] sickling indicates that rare sickling and non-

269   sickling variants, respectively, also segregate in these two species. Taken together,

270   these findings point to the long-term maintenance of ancestral variation through

271   successive speciation events dating back to the most common ancestor of Old World

272   and New World deer, an estimated ~13.6 million years ago (mya) [CI: 9.84-

273   17.33mya][43].

274

275   Might this polymorphism have been maintained simply by chance or must balancing

276   selection be evoked to account for its survival? We currently lack information on

277   broader patterns of genetic diversity at deer HBB$_A$ loci and surrounding regions that

278   would allow us to search for footprints of balancing selection explicitly. However, we

279   can estimate the probability *P* that a trans-species polymorphism has been maintained

280   along two independent lineages by neutral processes alone as

281     $$P = (e^{-T/2N_e}) \text{ x } (e^{-T/2N_e})$$

282   where *T* is the number of generations since the two lineages split and N_e is the

283   effective population size[44,45]. For simplicity, we assume N_e to be constant over *T* and

284   the same for both lineages. In the absence of reliable species-wide estimates for N_e,

285   we can nonetheless ask what N_e would be required to meet a given threshold

286   probability. Conservatively assuming an average generation time of 1 year[46,47] and a

287   split time of 7.2mya (the lowest divergence time estimate in the literature[43]), N_e would

12

288   have to be 2,403,419 to reach a threshold probability of 0.05. Although deer can have

289   large census population sizes, an N_e >2,000,000 for both fallow and white-tailed deer

290   is comfortably outside what we would expect for large-bodied mammals, >4-fold

291   higher than estimates for wild mice[48] and >2-fold higher even than estimates for

292   African populations of *Drosophila melanogaster*[49]. Consequently, we argue that the

293   HBB_A trans-species polymorphism is inconsistent with neutral evolution and instead

294   reflects the action of balancing selection.

295

296   *A distinct genetic basis for sickling in sheep*

297

298   While sickling in deer is particularly well-documented, the capacity for reversible

299   haemoglobin polymerization has also been observed in a small coterie of other

300   vertebrates[10,50], including some species of fish[50], mongoose[51], and notably also goat

301   and sheep (*Ovis aries*)[11,52]. For most of these species, we have no information on

302   sickling-associated genotypes and allelic diversity. Sheep, where sickling has been

303   found in a variety of domestic breeds[52,53], are an exception in this regard. Two HBB_A

304   alleles, HbA and HbB, were previously identified[54]. HbA homozygotes and HbA/HbB

305   heterozygotes sickle whereas HbB homozygotes do not[52]. We first compared the

306   sheep reference sequence (Texel breed) included in Fig. 1 with partial peptide

307   information for both alleles[54] and found it to be fully consistent with the non-sickling

308   HbB allele. We then surveyed amino acid variation at the β-globin gene across 75

309   breeds of sheep, selected to cover global sheep genetic diversity[55]. We observed all

310   seven amino acids known to discriminate HbA from HbB but found no variation at

311   residues 6 or 22 (Supplementary Figure 9), suggesting, first, that the genetic diversity

312   panel captures HbA and, second, that HbA, lacking 6V and 22V, promotes

313   polymerization by yet another mechanism. This conclusion is consistent with

314   phenomenological differences in sickling dynamics between deer and sheep,

315   including a) the finding that sickling in the latter only occurs when cells are

316   suspended in hypertonic saline and incubated at 37ºC (REF. 11), making it less likely

317   that sickling frequently takes place *in vivo* under physiological conditions, and b) the

318   observation that the sickling allele is dominant in sheep but recessive in deer[19].

319   Importantly, these results also indicate that sickling evolved independently in deer

320   (Cervidae) and their sister clade (Caprinae).

321

322   **Discussion**

323

324   Given the dramatic change in erythrocyte shape brought about by haemoglobin

325   polymerization it is conspicuous that multiple vertebrate lineages have independently

326   converged on this phenotype. In principle, recurrent emergence could be the result of

327   non-adaptive forces. Recent findings suggest that symmetric protein complexes like

328   haemoglobin exist at the edge of supramolecular self-assembly, often being a short

329   mutational distance away from the propensity to form polymers[56]. However, in deer,

330   the long-term maintenance of a trans-species polymorphism is inconsistent with

331   selective neutrality and instead argues for fitness effects along multiple lineages. By

332   direct implication, even though sickling is remarkably well tolerated *in vivo*, perhaps

333   owing to unique properties of deer erythrocytes (Supplementary Discussion), it cannot

334   be perfectly innocuous[23]. Rather, it must exert a physiological effect that is strong and

335   frequent enough to be targeted by selection.

336

337   What are the ecological driving factors behind the maintenance of sickling (and non-

338   sickling) alleles over evolutionary time? It has previously been suggested that

339   haemoglobin polymerization might, by radically altering the intracellular environment

340   of red blood cells, provide a generic defense mechanism against red blood cell

341   parasites[50]. Deer certainly harbour a number of intra-erythrocytic parasites, including

342   *Babesia*[57] and *Plasmodium*[58,59]. The latter was recently found to be widespread in

343   white-tailed deer, but, interestingly, associated with very low levels of parasitaemia[59].

344   Also worth noting in this regard is the marked geographic asymmetry in sickling

345   status, where established non-sickling species are restricted to arctic and subarctic

346   (elk, reindeer) or mountainous (wapiti) habitat. Might this indicate that the sickling

347   allele loses its adaptive value in colder climates, perhaps linked to the lower

348   prevalence of blood-born parasites? Although a general cross-species link between

349   sickling and parasite burden is tantalizing, it is important to highlight that there is

350   currently no concrete evidence for such a connection and alternative hypotheses

351   should be considered. For example, with no evidence for heterozygote advantage,

352   might it be that allelic diversity has been maintained by migration-selection balance?

353   Or do the timescales involved render such a scenario improbable? Exploring

354   geographic structure in the distribution of sickling and non-sickling alleles will be

355   important in this regard and might point to the ecological factors involved in

356   maintaining either allele. More generally, future epidemiological studies coupled to

357   population genetic investigations will be required to unravel the evolutionary ecology

358   of sickling in deer and establish whether parasites are indeed ecological drivers of

359   between- and within-species differences in HBB_A genotype. Ultimately, such analyses

360   will determine whether deer constitute a useful comparative system to elucidate the

15

361 link between sickling and protection from the effects of *Plasmodium* infection, which

362 remains poorly understood in humans.

363

364

365 **Methods**

366
367 <u>Sample collection and processing.</u> Blood, muscle tissue, and DNA samples were

368 acquired for 15 species of deer from a range of sources (Supplementary Table 1).

369 The white-tailed deer blood sample was heat-treated on import to the United

370 Kingdom in accordance with import standards for ungulate samples from non-EU

371 countries (IMP/GEN/2010/07). Fresh blood was collected into PAXgene Blood DNA

372 tubes (PreAnalytix) and DNA extracted using the PAXgene Blood DNA kit

373 (PreAnalytix). DNA from previously frozen blood samples was extracted using the

374 QIAamp DNA Blood Mini kit (Qiagen). DNA from tissue samples was extracted with

375 the QIAamp DNA Mini kit (Qiagen) using 25mg of tissue. Total RNA was isolated

376 from an *E. davidianus* blood sample using the PAXgene Blood RNA kit

377 (PreAnalytix) three days after collection into a PAXgene Blood RNA tube

378 (PreAnalytix). All extractions were performed according to manufacturers' protocols.

379 For each sample, we validated species identity by amplifying and sequencing the

380 cytochrome b (*CytB*) gene. With the exception of *Cervus albirostris*, we successfully

381 amplified *CytB* from all samples using primers MTCB_F/R (Supplementary Figure 2)

382 and conditions as described in REF. 60. Phusion High-Fidelity PCR Master Mix

383 (ThermoFisher) was used for all amplifications. PCR products were purified using the

384 MinElute PCR Purification Kit (Qiagen) and Sanger-sequenced with the amplification

385 primers. The *CytB* sequences obtained were compared to all available deer *CytB*

386 sequences in the 10kTrees Project[61] using the *ape* package (function *dist.dna* with

387 default arguments) in R[62]. In all cases, the presumed species identity of the sample

388 was confirmed (Supplementary Table 2).

389

390 <u>Whole genome sequencing.</u> *O. virginianus* genomic DNA was prepared for

391 sequencing using the NEB DNA library prep kit (New England Biolabs) and

392 sequenced on the Illumina HiSeq platform. The resulting 229 million 100bp paired-

393 end reads were filtered for adapters and quality using Trimmomatic[63] with the

394 following parameters: *ILLUMINACLIP:adapters/TruSeq3-PE-2.fa:2:30:10*

395 *LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50*. Inspection of the

396 remaining 163.5M read pairs with FastQC

397 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) suggested that

398 overrepresented sequences had been successfully removed.

399

400 <u>Mapping and partial assembly of the *O. virginianus* β-globin locus.</u> To seed a local

401 assembly of the *O. virginianus* β-globin locus we first mapped *O. virginianus*

402 trimmed paired-end reads to the duplicated β-globin locus in the hard-masked *B.*

403 *taurus* genome (UMD 3.1.1; chr15: 48973631-49098735). The β-globin locus is

404 defined here as the region including all *B. taurus* β-globin genes [HBE1, HBE4,

405 HBB_A (ENSBTAG00000038748), HBE2, HBB_F (ENSBTAG00000037644)], the

406 intervening sequences and 24kb either side of the two outer β-globins (HBE1, HBB_F).

407 The mapping was performed using bowtie2 (REF. 64) with default settings and the

408 optional *--no-mixed* and *--no-discordant* parameters. 110 reads mapped without gaps

409 and a maximum of one nucleotide mismatch. These reads, broadly dispersed across

410 the *B. taurus* β-globin locus (Supplementary Figure 10), were used as seeds for local

17

411   assembly using a customised aTRAM[65] pipeline (see below). Prior to assembly, the

412   remainder of the reads were filtered for repeat sequences by mapping against

413   Cetartiodactyla repeats in Repbase[66]. The aTRAM.pl wrapper script was modified to

414   accept two new arguments: *max_target_seqs \<int\>* limited the number of reads found

415   by BLAST from each database shard; *cov_cutoff \<int\>* passed a minimum coverage

416   cut-off to the underlying Velvet 1.2.10 assembler[67]. The former modification prevents

417   stalling when the assembly encounters a repeat region, the latter discards low

418   coverage contigs at the assembler level. aTRAM was run with the following

419   arguments: *-kmer 31 -max_target_seqs 2000 -ins_length 270 -exp_coverage 8 -*

420   *cov_cutoff 2 -iterations 5*. After local assembly on each of the 110 seed reads, the

421   resulting contigs were combined using Minimo[68] with a required minimum nucleotide

422   identity of 99%. To focus specifically on assembling the adult β-globin gene, only

423   contigs that mapped against the *B. taurus* adult β-globin gene ±500bp (chr15:

424   49022500-49025000) were retained and served as seeds for another round of

425   assembly. This procedure was repeated twice. The final 59 contigs were compared to

426   the UMD 3.1.1 genome using BLAT and mapped exclusively to either the adult or

427   foetal *B. taurus* β-globin gene. From the BLAT alignment, we identified short

428   sequences that were perfectly conserved between the assembled deer contigs and the

429   *B. taurus* as well as the sheep assembly (Oar_v3.1). Initial forward and reverse

430   primers (Ovirg_F1/Ovirg_R1, Supplementary Figure 2) for β-globin amplification

431   were designed from these conserved regions located 270bp upstream

432   (chr15:49022762-49022786) and 170bp downstream (chr15:49024637-49024661) of

433   the *B. taurus* adult β-globin gene, respectively. Our local assembly is consistent with

434   a recent draft genome assembly

18

435   (https://www.ncbi.nlm.nih.gov/assembly/GCF_002102435.1/) from a white-tailed

436   deer from Texas (*O. virginianus texanus*).

437

438   <u>Globin gene amplification and sequencing.</u> Amplification of β-globin from *O.*

439   *virginianus* using primers Ovirg_F1 and Ovirg_R1 yielded two products of different

440   molecular weights (~2000bp and ~1700bp; Supplementary Figure 2), which were

441   isolated by gel extraction and Sanger-sequenced using the amplification primers. The

442   high molecular weight product had higher nucleotide identity to the adult (93%) than

443   to the foetal (90%) *B. taurus* β-globin coding sequence. Note that the discrepancy in

444   size between the adult and foetal β-globin amplicons derives from the presence

445   of two tandem Bov-tA2 SINEs in intron 2 of the adult β-globin gene in cattle, sheep,

446   and *O. virginianus* and is therefore likely ancestral. We designed a second set of

447   primers to anneal immediately up- and downstream, and in the middle of the adult β-

448   globin gene (Ovirg_F2, Ovirg_R2, Ovirg_Fmid2, Supplementary Figure 2).

449   Amplification from DNA extracts of other species with Ovirg_F1/Ovirg_R1 produced

450   mixed results, with some species showing a two-band pattern similar to *O.*

451   *virginianus*, others only a single band – corresponding to the putative adult β-globin

452   (Supplementary Figure 2). Using these primers, no product could be amplified from

453   *R. tarandus*, *H. inermis*, and *C. capreolus*. We identified a 3bp mismatch to the

454   Ovirg_R1 primer in a partial assembly of *C. capreolus* (Genbank accession:

455   GCA_000751575.1; scaffold: CCMK010226507.1) that is likely at fault. A re-

456   designed reverse primer (Ccap_R1) successfully amplified the adult β-globin gene

457   from the three deer species above as well as *C. canadensis* (Supplementary Figure 2).

458   All amplifications were performed using Phusion High-Fidelity PCR Master Mix

459   (ThermoFisher), with primers as listed in Supplementary Figure 2, and 50-100ng of

19

460   genomic DNA. Annealing temperature and step timing were chosen according to

461   manufacturer guidelines. Amplifications were run for 35 cycles. Gel extractions were

462   performed on samples resolved on 1% agarose gels for 40 minutes at 90V using the

463   MinElute Gel Extraction Kit (Qiagen) and following the manufacturer's protocol.

464   PCR purifications were performed using the MinElute PCR Purification Kit (Qiagen)

465   following the manufacturer's protocol. All samples were sequenced using the Sanger

466   method with amplification primers and primer Ovirg_Fmid2.

467

468   <u>Transcriptome sequencing and assembly.</u> RNA was extracted from the red cell

469   component of a blood sample of an adult Père David's deer using the PAXgene Blood

470   RNA kit (Qiagen). An mRNA library was prepared using a Truseq mRNA library

471   prep kit and sequenced on the MiSeq platform, yielding 25,406,472 paired-end reads

472   of length 150bp, which were trimmed for adapters and quality-filtered using Trim

473   Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with a

474   base quality threshold of 30. The trimmed reads were used as input for *de novo*

475   transcriptome assembly with Trinity[69] using default parameters. A blastn homology

476   search against these transcripts, using the *O. virginianus* adult β-globin CDS as query,

477   identified a highly homologous transcript (E-value = 0; no gaps; 97.5% sequence

478   identity compared with 92.2% identity to the foetal β-globin). The CDS of this

479   putative β-globin transcript was 100% identical to the sequence amplified from Père

480   David's deer genomic DNA (Supplementary Figure 4). We used *emsar*[70] with default

481   parameters to assess transcript abundances. The three most abundant reconstructed

482   transcripts correspond to full or partial α- and β-globin transcripts, including one

483   transcript, highlighted above, that encompasses the entire adult β-globin CDS. These

484   transcripts are an order of magnitude more abundant than the fourth most abundant

485   (Supplementary Figure 4), in line with the expected predominance of α- and β-globin

486   transcripts in mature adult red blood cells. To investigate whether the foetal β-globin

487   could be detected in the RNA-seq data and because amplification of the foetal β-

488   globin from Père David's deer genomic DNA was not successful, we mapped reads

489   against the foetal β-globin gene of *C. e. elaphus*, the closest available relative. Given

490   that the CDS of the adult β-globins in these species are 100% identical, we expected

491   that the foetal orthologs would likewise be highly conserved. We therefore removed

492   reads with more than one mismatch and assembled putative transcripts from the

493   remaining 1.3M reads using the Geneious assembler v.10.0.5 (REF. 71) with default

494   parameters (*fastest* option enabled). We recovered a single contig with high homology

495   to the *C. e. elaphus* foetal β-globin CDS (only a single mismatch across the CDS). We

496   then estimated the relative abundance of adult and the putative foetal transcripts by

497   calculating the proportion of reads that uniquely mapped to either the adult or foetal

498   CDS. 1820532 reads mapped uniquely to the adult sequence whereas 872 mapped

499   uniquely to the foetal CDS, a ratio of 2088:1.

500

501   <u>Structural analysis.</u> Homology models were built for *O. virginianus* and *R. tarandus*

502   β-globin sequences using the MODELLER-9v15 program for comparative protein

503   structure modelling[72] using both oxy (1HHO) and deoxy (2HHB) human

504   haemoglobin structures as templates. The structures were used for electrostatic

505   calculations using the Adaptive Poisson-Boltzmann Solver[73] plug-in in the Visual

506   Molecular Dynamics (VMD) program[74]. The surface potentials were visualised in

507   VMD with the conventional red and blue colours, for negative and positive potential

508   respectively, set at ±5 kT/e.

509

510   Modelling of haemoglobin fibres. We first used the program HADDOCK[75] with the

511   standard protein-protein docking protocol to generate ensembles of docking models of

512   β-globin dimers. In each docking run, a different interacting surface centred around a

513   specific residue was defined on each β-globin chain. All residues within 3Å of the

514   central residue were defined as "active" and were thus constrained to be directly

515   involved in the interface, while other residues within 8Å of the central residue were

516   defined as "passive" and were allowed but not strictly constrained to form a part of

517   the interface. We performed docking runs with the interaction centred between

518   residue 87 and all other residues, generating at least 100 water-refined β-globin dimer

519   models for each (although 600 *O. virginianus* oxy β-globin 22V-87Q models were

520   built for use in the interaction energy calculations). The β-globin dimers were then

521   evaluated for their ability to form HbS-like fibres out of full haemoglobin tetramers.

522   Essentially, the contacts from the β-globin dimer models were used to build a chain of

523   five haemoglobin molecules, in the same way that the contacts between 6V and the

524   EF pocket lead to an extended fibre in HbS. HbS-like fibres were defined as those in

525   which a direct contact was formed between the first and third haemoglobin tetramers

526   in a chain (analogous to the axial contacts in HbS fibres, see Fig. 2c), and in which

527   the chain is approximately linear. This linearity was measured as the distance between

528   the first and third plus the distance between the third and the fifth haemoglobin

529   tetramers, divided by the distance between the first and the fifth. A value of 1 would

530   indicate a perfectly linear fibre, while we considered any chains with a value <1.05 to

531   be approximately linear and HbS-like. Finally, chains containing significant steric

532   clashes between haemoglobin tetramers (defined as >3% of Cα atoms being within

533   2.8Å of another Cα atom) were excluded. Fibre formation propensity was then

534   defined as the fraction of all docking models that led to HbS-like fibres.

535

Interaction energy analysis. Using the 270 22V-87Q models of *O. virginianus* β-globin dimers that can form HbS-like fibres, we used FoldX[76] and the 'RepairPDB' and 'BuildModel' functions to mutate each dimer to the sequences of all other adult deer species. Note that since *C. e. elaphus, C. e. bactrianus* and *E. davidianus* have identical amino acid sequence, only one of these was included here. The energy of the interaction was then calculated using the 'AnalyseComplex' function of FoldX, and then averaged over all docking models. The same protocol was then used for the analysis of the effects of individual mutations, using all possible single amino acid substitutions observed in the adult deer sequences, except that the interaction energy was presented as the change with respect to the wild-type sequence.

546

Deer species tree and wider mammalian phylogeny. The mammalian phylogeny depicted in Fig. 1 is principally based on the Timetree of Life[43] with the order Carnivora regrafted to branch above the root of the Chiroptera and Artiodactyla to match findings in[77]. The internal topology of Cervidae was taken from the Cetartiodactyla consensus tree of the 10kTrees Project[61]. *C. canadensis* and *Cervus elaphus bactrianus*, not included in the 10kTrees phylogeny, were added as sister branches to *C. nippon* and *C. e. elaphus*, respectively, following REF. 78. Supplementary Figure 11 provides a graphical overview of these changes. To generate Fig. 1, we aligned adult deer β-globin coding sequences to a set of non-chimeric mammalian adult β-globin CDSs[26].

557

Gene tree reconstruction. All trees were built using RAxML v8.2.10 based on alignments made with MUSCLE v3.8.1551. Unless stated otherwise, we used the

23

560   RAxML joint maximum likelihood and bootstrap analysis (option –f a) with random

561   seeds, a single partition, and 100 bootstrap replicates. The GTRGAMMA model was

562   used for nucleotide alignments and the best fitting protein model was automatically

563   chosen by RAxML using the PROTGAMMAAUTO option.

564   As some historical alleles (β^{II}, β^{V}, β^{VII}) are only available at the peptide level, Fig. 3c

565   was built at the protein level. HBB_A ($\pm HBB_F$) trees (Fig. 3a, Supplementary Figures

566   2&6), on the other hand, are nucleotide-level trees build from an alignment of coding

567   exons and intervening introns. Note here, that intron 2, which is comparatively long

568   and less constrained than coding sequence, contributes a comparatively large number

569   of phylogenetically informative sites. In fact, the intron 2 tree re-capitulates the

570   exon+intron tree almost perfectly, with a minor difference in the precise location of *C.*

571   *canadensis* in the non-sickling cluster. Note further, that a large comparative

572   contribution of intron 2 to the overall phylogenetic signal is fortuitous in this context.

573   In order to understand patterns of lineage sorting and introgression, it is desirable to

574   eliminate spurious phylogenetic signals introduced by gene conversion, which

575   strongly affects exonic sequence (as evident in Fig. 3b) but is much less prevalent in

576   intron 2. Since we explicitly demonstrate (in Supplementary Figure 6) that including

577   sequence affected by gene conversion does not affect the overall tree topology, we

578   present exon+intron (i.e. gene) trees throughout for simplicity.

579

580   <u>Topology testing.</u> To test for significant phylogenetic discordance between the HBB_A

581   gene tree and the species tree as depicted in Fig. 3a we compared both topologies

582   using the Approximately Unbiased (AU) test[79] implemented in CONSEL[80]. The

583   unconstrained maximum likelihood (ML) HBB_A gene tree was tested against an

584   alternative ML tree (derived from 200 maximum likelihood starting trees) built under

585   a single constraint: to recover the well-established monophyletic groups of Old World

586   and New World deer. Branching patterns *within* these major clades were allowed to

587   vary. With this approach, we conservatively test the significance of the incongruent

588   placement of *O. virginianus* and *P. pudu* sickling alleles with the Old World deer (and

589   *C. canadensis* with New World deer) without considering confounding signals from

590   within-clade branching that might arise, for example, due to gene conversion. Both

591   the constrained and unconstrained ML trees were calculated with RAxML as

592   described above. Per site log-likelihoods were computed for the unconstrained and

593   constrained ML trees with RAxML (option –f G).

594

595   <u>Detection of recombination events.</u> We considered two sources of donor sequence for

596   recombination into adult β-globins: adult β-globin orthologs in other deer species and

597   the foetal β-globin paralog within the same genome. *H. inermis* HBB$_F$ was omitted

598   from this analysis since the sequence of intron 2 was only partially determined. We

599   used the Recombination Detection Program (RDP v.4.83)[81] to test for signals of

600   recombination in an alignment of complete adult and foetal deer β-globin genes that

601   were successfully amplified and sequenced, enabling all subtended detection methods

602   (including primary scans for BootScan and SiScan) except LARD, treating the

603   sequences as linear and listing all detectable events. In humans, conversion tracts of

604   lengths as short as 110bp have been detected in the globin genes[82] and tracts as short

605   as 50bp in other gene conversion hotspots[83,84]. Given the presence of multiple regions

606   of 100% nucleotide identity across the alignment of adult and foetal deer β-globins

607   (Fig. 3b), we suspected that equally short conversion tracts might also be present. We

608   therefore lowered window and step sizes for all applicable detection methods in RDP

609   (Supplementary Figure 8) at the cost of a lower signal-to-noise ratio. As the objective

25

610 is to test whether recombination events could have generated the phyletic distribution

611 of sickling/non-sickling genotypes observed empirically, this is conservative.

612

613 <u>Data availability.</u> HBB_A and HBB_F full gene sequences (coding sequence plus

614 intervening introns) have been submitted to GenBank with accession numbers

615 KY800429-KY800452. An alignment of these sequences is also available as

616 Supplementary Data. Père David's deer RNA sequencing and white-tailed deer whole

617 genome sequencing raw data has been submitted to the European Nucleotide Archive

618 (ENA) with the accession numbers PRJEB20046 and PRJEB20034, respectively.

619

620

621 **References**

622

623 1.   Ingram, V. M. Gene mutations in human haemoglobin: the chemical difference

624   between normal and sickle cell haemoglobin. *Nature* **180,** 326–328 (1957).

625 2.   Harrington, D. J., Adachi, K. & Royer, W. E., Jr. The high resolution crystal

626   structure of deoxyhemoglobin S. *Journal of Molecular Biology* **272,** 398–407

627   (1997).

628 3.   Wishner, B. C., Ward, K. B., Lattman, E. E. & Love, W. E. Crystal structure of

629   sickle-cell deoxyhemoglobin at 5 Å resolution. *Journal of Molecular Biology*

630   **98,** 179–194 (1975).

631 4.   Sears, D. A. The morbidity of sickle cell trait. *The American Journal of*

632   *Medicine* **64,** 1021–1036 (1978).

633 5.   Platt, O. S. *et al.* Mortality in sickle cell disease. Life expectancy and risk

634   factors for early death. *N. Engl. J. Med.* **330,** 1639–1644 (1994).

635   6.   Piel, F. B. *et al.* Global distribution of the sickle cell gene and geographical
636   confirmation of the malaria hypothesis. *Nature Communications* **1,** 104 (2010).

637   7.   Herrick, J. B. Peculiar elongated and sickle-shaped red blood corpuscles in a
638   case of severe anemia. *Arch. Int. Med.* **5,** 517 (1910).

639   8.   Gulliver, G. Observations on certain peculiarities of form in the blood
640   corpuscles of the mammiferous animals. *Lond. Edinb. Dubl. Phil. Mag* **17,**
641   325–327 (1840).

642   9.   Undritz, E., Betke, K. & Lehmann, H. Sickling phenomenon in deer. *Nature*
643   **187,** 333–334 (1960).

644   10.   Hawkey, C. M. *Comparative Mammalian Haematology*. (Heinemann
645   Educational Books, 1975).

646   11.   Butcher, P. D. & Hawkey, C. M. Haemoglobins and erythrocyte sickling in the
647   artiodactyla: A survey. *Comparative Biochemistry and Physiology Part A:*
648   *Physiology* **57,** 391–398 (1977).

649   12.   Weber, Y. B. & Giacometti, L. Sickling Phenomenon in the Erythrocytes of
650   Wapiti (Cervus Canadensis). *Journal of Mammalogy* **53,** 917–919 (1972).

651   13.   Simpson, C. F. & Taylor, W. J. Ultrastructure of sickled deer erythrocytes. I.
652   The typical crescent and holly leaf forms. *Blood* **43,** 899–906 (1974).

653   14.   Schmidt, W. C. *et al.* The structure of sickling deer type III hemoglobin by
654   molecular replacement. *Acta Crystallogr Sect B Struct Crystallogr Cryst Chem*
655   **33,** 335–343 (1977).

656   15.   Pritchard, W. R., Malewitz, T. D. & Kitchen, H. Studies on the mechanism of
657   sickling of deer erythrocytes. *Experimental and Molecular Pathology* **2,** 173–
658   182 (1963).

659   16.   Kitchen, H., Easley, C. W., Putnam, F. W. & Taylor, W. J. Structural

660   comparison of polymorphic hemoglobins of deer with those of sheep and other

661   species. *The Journal of Biological Chemistry* **243,** 1204–1211 (1968).

662   17.   Seiffge, D. Haemorheological studies of the sickle cell phenomenon in

663   european red deer (Cervus elaphus). *Blut* **47,** 85–92 (1983).

664   18.   Kitchen, H., Putnam, F. W. & Taylor, W. J. Hemoglobin Polymorphism: Its

665   Relation to Sickling of Erythrocytes in White-Tailed Deer. *Science* **144,** 1237–

666   1239 (1964).

667   19.   Taylor, W. J. & Easley, C. W. Sickling phenomena of deer. *Annals of the New*

668   *York Academy of Sciences* **241,** 594–604 (1974).

669   20.   Harris, M. J., Huisman, T. H. J. & Hayes, F. A. Geographic distribution of

670   hemoglobin variants in the white-tailed deer. *Journal of Mammalogy* **54,** 270–

671   274 (1973).

672   21.   Harris, M. J., Wilson, J. B. & Huisman, T. H. J. Structural studies of

673   hemoglobin α chains from Virginia white-tailed deer. *Archives of Biochemistry*

674   *and Biophysics* **151,** 540–548 (1972).

675   22.   Parshall, C. J., Vainisi, S. J., Goldberg, M. F. & Wolf, E. D. In vivo erythrocyte

676   sickling in the Japanese sika deer (Cervus nippon): methodology. *Am J Vet Res*

677   **36,** 749–752 (1975).

678   23.   Whitten, C. F. Innocuous Nature of the Sickling (Pseudosickling) Phenomenon

679   in Deer. *British Journal of Haematology* **13,** 650–655 (1967).

680   24.   Shimizu, K. *et al.* The primary sequence of the beta chain of Hb type III of the

681   Virginia white-tailed deer (Odocoilus Virginianus), a comparison with putative

682   sequences of the beta chains from four additional deer hemoglobins, types II,

683   IV, V, and VIII, and relationships between intermolecular contacts, primary

684   sequence and sickling of deer hemoglobins. *Hemoglobin* **7,** 15–45 (1983).

685 25. Kitchen, H. & Taylor, W. J. The sickling phenomenon of deer erythrocytes.

686 *Adv. Exp. Med. Biol.* **28,** 325–336 (1972).

687 26. Gaudry, M. J., Storz, J. F., Butts, G. T., Campbell, K. L. & Hoffmann, F. G.

688 Repeated evolution of chimeric fusion genes in the β-globin gene family of

689 laurasiatherian mammals. *Genome Biol Evol* **6,** 1219–1234 (2014).

690 27. Hardison, R. C. Evolution of Hemoglobin and Its Genes. *Cold Spring Harb*

691 *Perspect Med* **2,** a011627–a011627 (2012).

692 28. Townes, T. M., Fitzgerald, M. C. & Lingrel, J. B. Triplication of a four-gene

693 set during evolution of the goat beta-globin locus produced three genes now

694 expressed differentially during development. *Proceedings of the National*

695 *Academy of Sciences of the United States of America* **81,** 6589–6593 (1984).

696 29. Schimenti, J. C. & Duncan, C. H. Structure and organization of the bovine

697 beta-globin genes. *Mol Biol Evol* **2,** 514–525 (1985).

698 30. Craig, J. E., Thein, S. L. & Rochette, J. Fetal hemoglobin levels in adults.

699 *Blood Reviews* **8,** 213–224 (1994).

700 31. Angeletti, M. *et al.* Different functional modulation by heterotropic ligands

701 (2,3-diphosphoglycerate and chlorides) of the two haemoglobins from fallow-

702 deer (Dama dama). *Eur J Biochem* **268,** 603–611 (2001).

703 32. Petruzzelli, R. *et al.* The primary structure of hemoglobin from reindeer

704 (Rangifer tarandus tarandus) and its functional implications. *Biochimica et*

705 *Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1076,**

706 221–224 (1991).

707 33. Adachi, K., Reddy, L. R. & Surrey, S. Role of hydrophobicity of phenylalanine

708 beta 85 and leucine beta 88 in the acceptor pocket for valine beta 6 during

709 hemoglobin S polymerization. *The Journal of Biological Chemistry* **269,**

710   31563–31566 (1994).

711   34.   Nagel, R. L. *et al.* Beta-chain contact sites in the haemoglobin S polymer.

712   *Nature* **283,** 832–834 (1980).

713   35.   Adachi, K., Konitzer, P. & Surrey, S. Role of gamma 87 Gln in the inhibition

714   of hemoglobin S polymerization by hemoglobin F. *The Journal of Biological*

715   *Chemistry* **269,** 9562–9567 (1994).

716   36.   Witkowska, H. E. *et al.* Sickle cell disease in a patient with sickle cell trait and

717   compound heterozygosity for hemoglobin S and hemoglobin Quebec-Chori. *N.*

718   *Engl. J. Med.* **325,** 1150–1154 (1991).

719   37.   Watson-Williams, E. J., Beale, D., Irvine, D. & Lehmann, H. A new

720   haemoglobin, D Ibadan (beta-87 threonine -- lysine), producing no sickle-cell

721   haemoglobin D disease with haemoglobin S. *Nature* **205,** 1273–1276 (1965).

722   38.   Amma, E. L., Sproul, G. D., Wong, S. & Huisman, T. H. J. Mechanism of

723   sickling in deer erythrocytes. *Annals of the New York Academy of Sciences*

724   **241,** 605–613 (1974).

725   39.   Girling, R. L., Schmidt, W. C., Jr, Houston, T. E., Amma, E. L. & Huisman, T.

726   H. J. Molecular packing and intermolecular contacts of sickling deer type III

727   hemoglobin. *Journal of Molecular Biology* **131,** 417–433 (1979).

728   40.   Fernández, M. H. & Vrba, E. S. A complete estimate of the phylogenetic

729   relationships in Ruminantia: a dated species-level supertree of the extant

730   ruminants. *Biological Reviews* **80,** 269–302 (2005).

731   41.   Taylor, W. J. & Simpson, C. F. Ultrastructure of sickled deer erythrocytes. II.

732   The matchstick cell. *Blood* **43,** 907–914 (1974).

733   42.   Butcher, P. D. & Hawkey, C. M. Red blood cell sickling in mammals. In: R. J.

734   Montali, G. Migaki, *The Comparative Pathology of Zoo Animals* (Smithsonian

735   Institute, 1980).

736 43. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of Life

737   Reveals Clock-Like Speciation and Diversification. *Mol Biol Evol* **32,** 835–845

738   (2015).

739 44. Wiuf, C., Zhao, K., Innan, H. & Nordborg, M. The Probability and

740   Chromosomal Extent of trans-specific Polymorphism. *Genetics* **168,** 2363–

741   2372 (2004).

742 45. Gao, Z., Przeworski, M. & Sella, G. Footprints of ancient-balanced

743   polymorphisms in genetic variation data from closely related species. *Evolution*

744   **69,** 431–446 (2015).

745 46. Baker, K. H. *et al.* Strong population structure in a species manipulated by

746   humans since the Neolithic: the European fallow deer (Dama dama dama).

747   *Heredity* **119,** 16–26 (2017).

748 47. Ryman, N., Baccus, R., Reuterwall, C. & Smith, M. H. Effective Population

749   Size, Generation Interval, and Potential Loss of Genetic Variability in Game

750   Species under Different Hunting Regimes. *Oikos* **36,** 257 (1981).

751 48. Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B. & Keightley, P. D.

752   Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLoS Genet.*

753   **6,** e1000825 (2010).

754 49. Shapiro, J. A. *et al.* Adaptive genic evolution in the Drosophila genomes.

755   *Proceedings of the National Academy of Sciences of the United States of*

756   *America* **104,** 2271–2276 (2007).

757 50. Koldkjær, P., McDonald, M. D., Prior, I. & Berenbrink, M. Pronounced in vivo

758   hemoglobin polymerization in red blood cells of Gulf toadfish: a general role

759   for hemoglobin aggregation in vertebrate hemoparasite defense? *American*

31

760 *Journal of Physiology - Regulatory, Integrative and Comparative Physiology*

761 **305,** R1190–R1199 (2013).

762 51. Hawkey, C. M. & Jordan, P. Sickle-cell erythrocytes in the mongoose

763 Herpestes sanguineus. *Transactions of the Royal Society of Tropical Medicine*

764 *and Hygiene* **61,** 180–181 (1967).

765 52. Butcher, P. D. & Hawkey, C. M. The nature of erythrocyte sickling in sheep.

766 *Comparative Biochemistry and Physiology Part A: Physiology* **64,** 411–418

767 (1979).

768 53. Evans, E. T. R. Sickling Phenomenon in Sheep. *Nature* **217,** 74–75 (1968).

769 54. Tucker, E. M. Genetic variation in the sheep red blood cell. *Biol Rev Camb*

770 *Philos Soc* **46,** 341–386 (1971).

771 55. Kijas, J. W. *et al.* Genome-wide analysis of the world's sheep breeds reveals

772 high levels of historic mixture and strong recent selection. *Plos Biol* **10,**

773 e1001258 (2012).

774 56. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins

775 evolve on the edge of supramolecular self-assembly. *Nature* **365,** 1596 (2017).

776 57. Perry, B. D., Nichols, D. K. & Cullom, E. S. Babesia odocoilei Emerson and

777 Wright, 1970 in white-tailed deer, Odocoileus virginianus (Zimmermann), in

778 Virginia. *Journal of Wildlife Diseases* **21,** 149–152 (1985).

779 58. Garnham, P. C. & Kuttler, K. L. A malaria parasite of the white-tailed deer

780 (Odocoileus virginianus) and its relation with known species of Plasmodium in

781 other ungulates. *Proc. R. Soc. Lond., B, Biol. Sci.* **206,** 395–402 (1980).

782 59. Martinsen, E. S. *et al.* Hidden in plain sight: Cryptic and endemic malaria

783 parasites in North American white-tailed deer (Odocoileus virginianus).

784 *Science Advances* **2,** e1501486 (2016).

785   60.   Naidu, A., Fitak, R. R., Munguia Vega, A. & Culver, M. Novel primers for

786   complete mitochondrial cytochrome b gene sequencing in mammals.

787   *Molecular Ecology Resources* **12,** 191–196 (2012).

788   61.   Arnold, C., Matthews, L. J. & Nunn, C. L. The 10kTrees website: A new

789   online resource for primate phylogeny. *Evol. Anthropol.* **19,** 114–118 (2010).

790   62.   Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and

791   Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).

792   63.   Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for

793   Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

794   64.   Ben Langmead & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.

795   *Nat Meth* **9,** 357–359 (2012).

796   65.   Allen, J. M., Huang, D. I., Cronk, Q. C. & Johnson, K. P. aTRAM - automated

797   target restricted assembly method: a fast method for assembling loci across

798   divergent taxa from next-generation sequencing data. *BMC Bioinformatics* **16,**

799   98 (2015).

800   66.   Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of

801   repetitive elements in eukaryotic genomes. *Mobile DNA 2014 5:1* **6,** 11 (2015).

802   67.   Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read

803   assembly using de Bruijn graphs. *Genome Research* **18,** 821–829 (2008).

804   68.   Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next

805   generation sequence assembly with AMOS. *Curr Protoc Bioinformatics*

806   **Chapter 11,** Unit 11.8 (2011).

807   69.   Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq

808   using the Trinity platform for reference generation and analysis. *Nat Protoc* **8,**

809   1494–1512 (2013).

810 70. Lee, S. *et al.* EMSAR: estimation of transcript abundance from RNA-seq data

811   by mappability-based segmentation and reclustering. *BMC Bioinformatics* **16,**

812   278 (2015).

813 71. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop

814   software platform for the organization and analysis of sequence data.

815   *Bioinformatics* **28,** 1647–1649 (2012).

816 72. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Curr*

817   *Protoc Bioinformatics* **0 5,** Unit–5.6.30 (2006).

818 73. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A.

819   Electrostatics of nanosystems: application to microtubules and the ribosome.

820   *Proceedings of the National Academy of Sciences of the United States of*

821   *America* **98,** 10037–10041 (2001).

822 74. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics.

823   *Journal of Molecular Graphics* **14,** 33–38 (1996).

824 75. Cyril Dominguez, Rolf Boelens, A. & Bonvin, A. M. J. J. HADDOCK: A

825   Protein−Protein Docking Approach Based on Biochemical or Biophysical

826   Information. *J. Am. Chem. Soc.* **125,** 1731–1737 (2003).

827 76. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting Changes in the Stability of

828   Proteins and Protein Complexes: A Study of More Than 1000 Mutations.

829   *Journal of Molecular Biology* **320,** 369–387 (2002).

830 77. Meredith, R. W. *et al.* Impacts of the Cretaceous Terrestrial Revolution and

831   KPg Extinction on Mammal Diversification. *Science* **334,** 521–524 (2011).

832 78. Ludt, C. J., Schroeder, W., Rottmann, O. & Kuehn, R. Mitochondrial DNA

833   phylogeography of red deer (Cervus elaphus). *Molecular Phylogenetics and*

834   *Evolution* **31,** 1064–1083 (2004).

835 79. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree

836 Selection. *Systematic Biology* **51,** 492–508 (2002).

837 80. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of

838 phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).

839 81. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4:

840 Detection and analysis of recombination patterns in virus genomes. *Virus Evol*

841 **1,** vev003 (2015).

842 82. Papadakis, M. N. & Patrinos, G. P. Contribution of gene conversion in the

843 evolution of the human beta-like globin gene family. *Human Genetics* **104,**

844 117–125 (1999).

845 83. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion

846 activity in human meiotic crossover hot spots. *Nat Genet* **36,** 151–156 (2004).

847 84. Bosch, E., Hurles, M. E., Navarro, A. & Jobling, M. A. Dynamics of a human

848 interparalog gene conversion hotspot. *Genome Research* **14,** 835–844 (2004).

849

850

851 **Acknowledgments**

863

864 **Author contributions**

865 A.E. performed laboratory experiments and evolutionary analyses and contributed to

866 experimental design, data analysis and interpretation. L.T.B. and J.A.M. designed and

867 performed structural modelling, and contributed to data analysis and interpretation.

868 V.S. contributed tissue samples. T.W. conceived the study, contributed to

869 experimental design, data analysis, and interpretation and wrote the manuscript with

870 input from all authors.

871

872 **Competing financial interests**

873 The authors declare no competing financial interests.

874

Figure Legends

Fig. 1. Mammalian adult β-globin peptide sequences in phylogenetic context. To

facilitate comparisons with prior classic literature, residues here and in the main text

are numbered according to human HBB, skipping the leading methionine. Dots

represent residues identical to the consensus sequences, defined by the most common

amino acid (X indicates a tie). Key residues discussed in the text are highlighted.

HBB$_A$ sequences from deer are coloured according to documented sickling state: red

= sickling, blue = non-sickling, grey = indeterminate (Supplementary Table 1). Green

cylinders highlight the position of α-helices in the secondary structure of human

HBB. See Methods and Supplementary Figure 11 for derivation of the accompanying

cladogram.

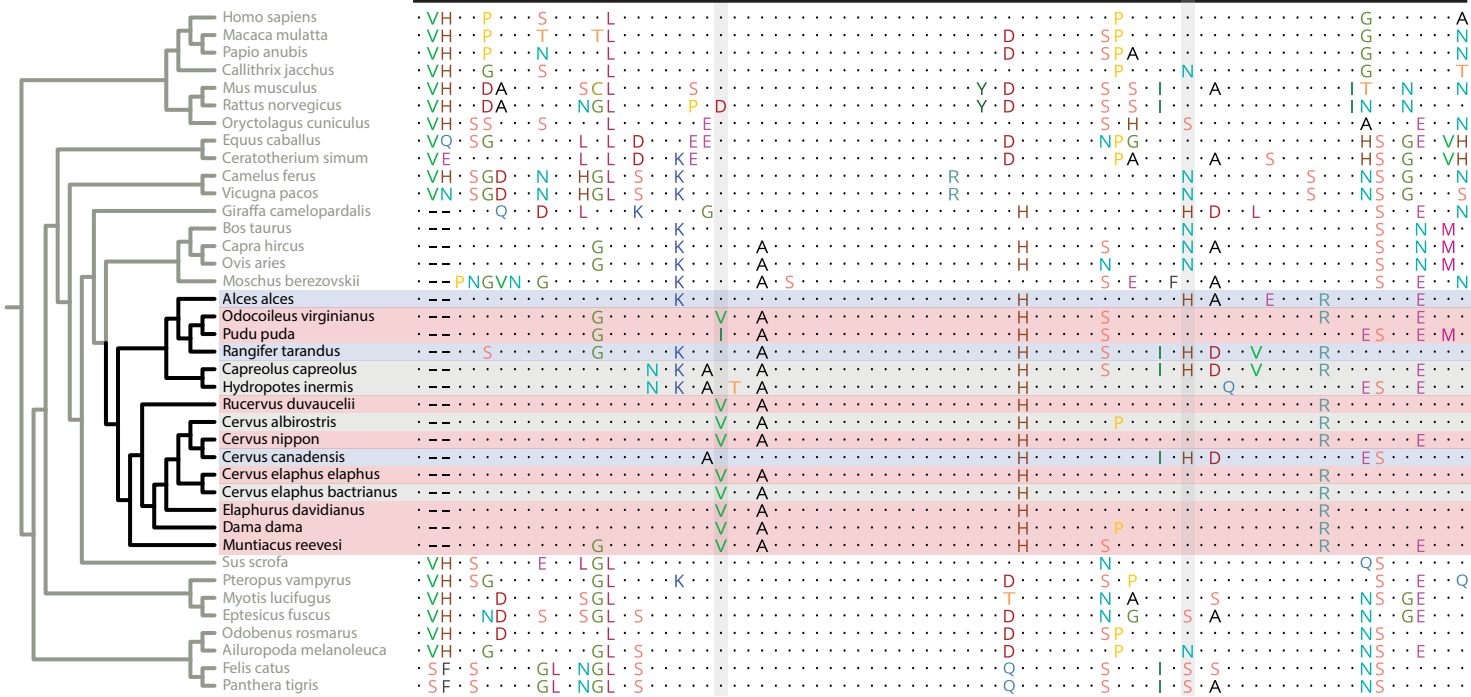**Fig. 2. Structural basis for sicking of deer haemoglobin. a**, Structure of

oxyhaemoglobin (PDB ID: 1HHO), with the key residues associated with sickling

highlighted in one of the β-globin chains. **b**, Comparison of the electrostatic surfaces

of oxy HBB$_A$ from a non-sickling (*R. tarandus*) and sickling (*O. virginianus*) species.

c, Example of a haemoglobin fibre formed via directed docking between residues 22V

and 87Q of *O. virginianus* oxy HBB$_A$. **d**, Fibre formation propensity derived from

docking simulations centred at a given focal residue in *O. virginianus* oxy and deoxy

HBB$_A$. These values represent the fraction of docking models that result in HbS-like

haemoglobin fibre structures. **e**, Fibre interaction energy for different deer species,

determined by mutating the 270 22V-87Q docking models compatible with fibre

formation and calculating the energy of the interaction. Error bars represent standard
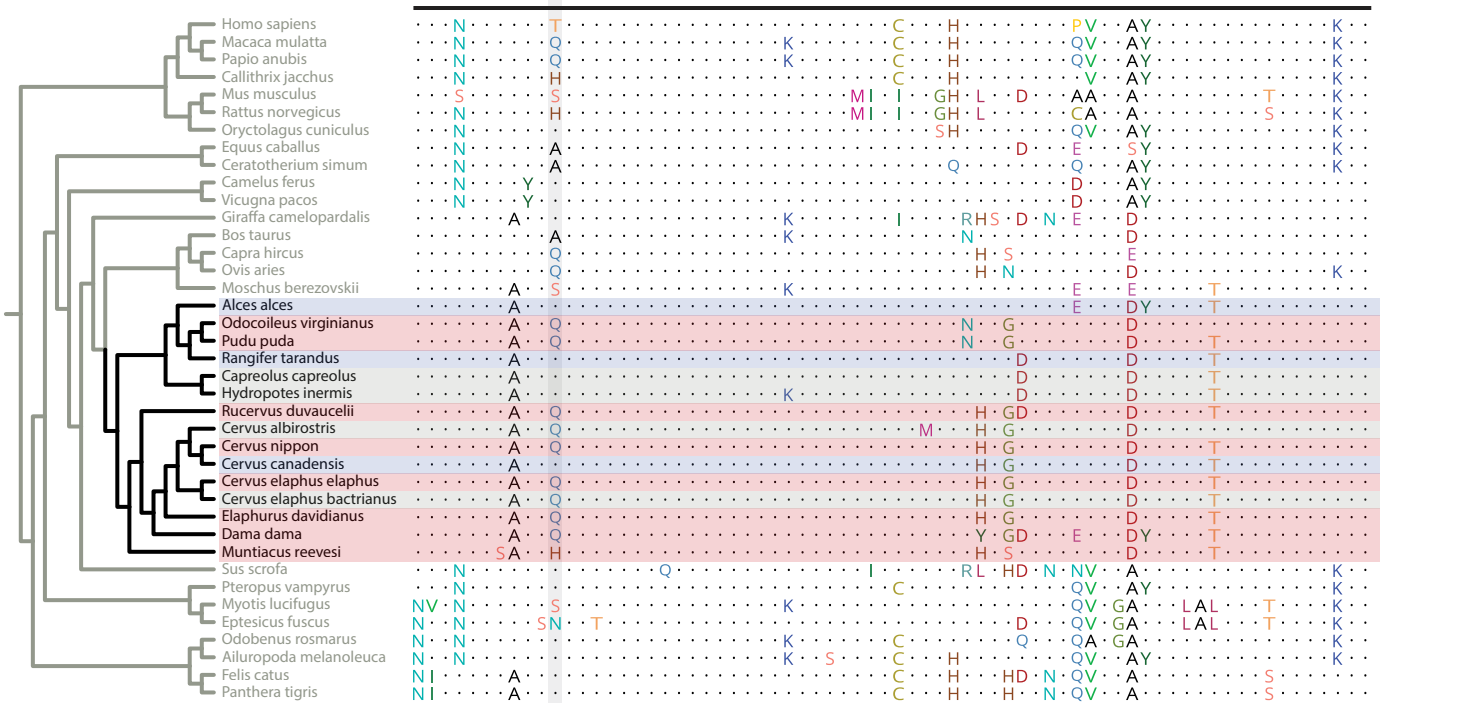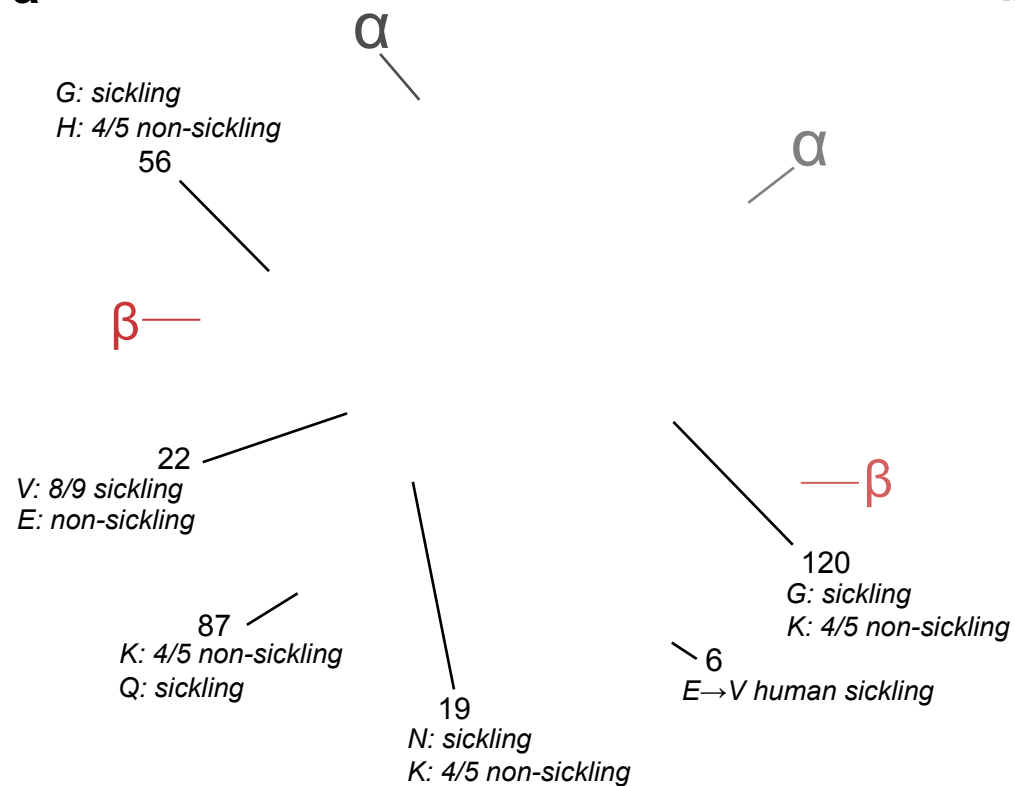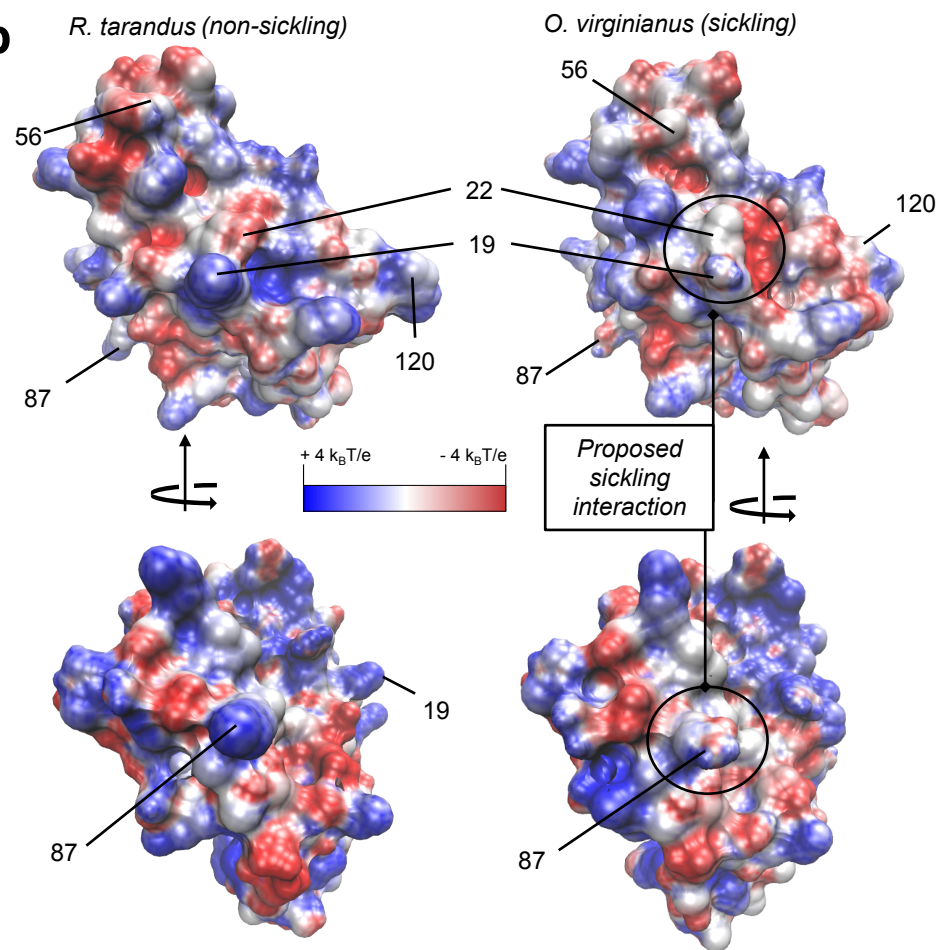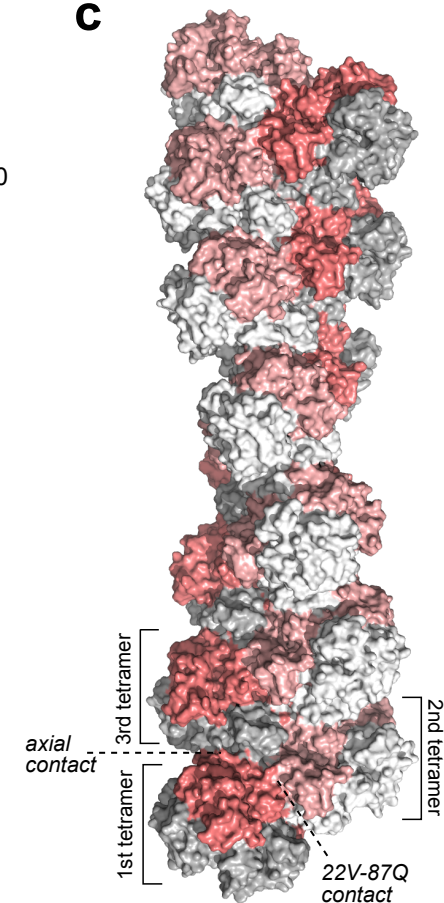
error of the mean.

900

Fig. 3. Evidence for incomplete lineage sorting, gene conversion, and a trans-species polymorphism in the evolutionary history of deer HBB$_A$. a, Discordances between the maximum likelihood HBB$_A$ gene tree and the species tree. Topological differences that violate the principal division into New World deer (NWD, Capreolinae) and Old World deer (OWD, Cervinae) are highlighted by solid black lines. Bootstrap values (% out of 1000 bootstrap replicates) are highlighted for salient nodes. **b**, Gene conversion and/or introgression. The top panel illustrates nucleotide identity between HBB$_A$ and HBB$_F$ orthologs (green: 100%, yellow: 30-100%, red: <30% identity). The low-identity segment towards the end of intron 2 marks repeat elements present in all adult but absent from all foetal sequences. Below, predicted recombination events affecting HBB$_A$ genes (orange), with either an adult ortholog (orange) or a foetal HBB$_F$ paralog (green) as the predicted source, suggestive of introgression or gene conversion, respectively. The number of asterisks indicates how many detection methods (out of a maximum of seven) predicted a given event (see Methods). Details for individual events (numbered in parentheses) are given in Supplementary Figure 8. **c,** Maximum likelihood protein tree of adult (orange) and foetal (green) β-globin. Alternate non-sickling *D. dama* (II) and *O. virginianus* (V, VII) alleles group with non-sickling species (coloured as in Fig. 1). Amino acid identity at key sites is shown on the right. ?: amino acid unresolved in primary source.

920

a

α

α

β

β

G: sickling
H: 4/5 non-sickling

56

22
V: 8/9 sickling
E: non-sickling

87
K: 4/5 non-sickling
Q: sickling

19
N: sickling
K: 4/5 non-sickling

6
E→V human sickling

120
G: sickling
K: 4/5 non-sickling

b

R. tarandus (non-sickling)

O. virginianus (sickling)

56

56

22

19

120

120

87

87

+ 4 k$_B$T/e   − 4 k$_B$T/e

Proposed sickling interaction

87

19

87

c

3rd tetramer

2nd tetramer

axial contact

1st tetramer

22V-87Q contact

d

Fibre formation propensity

Oxyhaemoglobin (sickling)
Deoxyhaemoglobin (non-sickling)

Residue

e

P. puda
O. virginianus
R. duvaucelii
C. nippon
C. elaphus
C. albirostris
M. reevesi
D. dama
H. inermis
C. capreolus
A. alces
R. tarandus
C. canadensis

Sickling
Indeterminate
Non-sickling

Fibre interaction energy (kcal/mol)

a

SPECIES TREE   HBB$_A$ GENE TREE

b

nucleotide identity

exon 1   exon 2   exon 3

19 22   56   87   120

H.inermis   (1)

C. canadensis   (2)   (10)

O. virginianus   (8)

P. puda   (9)   (8)

R. tarandus   (5)

D. dama   (4)

R. duvaucelii   (7)

c

	19	22	56	87	120
H. inermis HBB$_A$	K	E	G	K	K
O. virginianus (β^{II}) HBB$_A$	N	V	?	K	G
P. puda HBB$_A$	N	I	G	Q	G
O. virginianus (β^{III}) HBB$_A$	N	V	G	Q	G
M. reevesi HBB$_A$	N	V	G	H	S
C. nippon HBB$_A$	N	V	G	Q	G
C. e. elaphus HBB$_A$	N	V	G	Q	G
E. davidianus HBB$_A$	N	V	G	Q	G
D. dama HBB$_A$	N	V	G	Q	G
C. albirostris HBB$_A$	N	V	G	Q	G
R. duvaucelii HBB$_A$	N	V	G	Q	G
A. alces HBB$_A$	K	E	H	K	K
O. v. texanus HBB$_A$	K	E	H	K	K
C. capreolus HBB$_A$	K	E	H	K	K
C. canadensis HBB$_A$	N	E	H	K	G
D. dama (β^{II}) HBB$_A$	N	E	H	K	G
R. tarandus HBB$_A$	K	E	H	K	K
O. virginianus (β^V) HBB$_A$	K	E	?	K	?
O. virginianus (β^{VII}) HBB$_A$	K	E	?	?	?

0.04