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

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

[1]Molecular Systems Group, MRC London Institute of Medical Sciences (LMS), Du Cane Road, London W12 0NN, United Kingdom

[2]Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, Du Cane Road, London W12 0NN, United Kingdom

[3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom

[4]Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot SL5 7PY, United Kingdom

[*]Corresponding author

Crescent-shaped red blood cells, the hallmark of sickle cell disease, present a striking departure from the biconcave disc shape normally found in mammals. Characterized by increased mechanical fragility, sickled cells promote haemolytic anaemia and vaso-occlusions and contribute directly to disease in humans. Remarkably, a similar sickle-shaped morphology has been observed in erythrocytes from several deer species, without pathological consequences. The genetic basis of erythrocyte sickling in deer, however, remains unknown, limiting the utility of deer as comparative models for sickling. Here, we determine the sequences of human β-globin orthologs in 15 deer species and identify a set of co-evolving, structurally related residues that distinguish sickling from non-sickling deer. Protein structural modelling indicates a sickling mechanism distinct from human sickle cell disease, coordinated by a derived valine (E22V) in the second alpha helix of the β-globin protein. The evolutionary history of deer β-globins is characterized by incomplete lineage sorting, episodes of gene conversion between adult and foetal β-globin paralogs, and the presence of a trans-species polymorphism that is best explained by long-term balancing selection, suggesting that sickling in deer is adaptive. Our results reveal structural and evolutionary parallels and differences in erythrocyte sickling between human and deer, with implications for understanding the ecological regimes and molecular architectures that favour the evolution of this dramatic change in erythrocyte shape.

Introduction

Human sickling is caused by a single amino acid change (E6V) in the adult β-globin (HBB) protein (*1*). Upon deoxygenation, steric changes in the haemoglobin tetramer enable an interaction between 6V and a hydrophobic acceptor pocket (known as the EF pocket) on the β-surface of a second tetramer (*2*, *3*). This interaction promotes polymerization of mutant haemoglobin (HbS) molecules, which ultimately coerces red blood cells into the characteristic sickle shape. Heterozygote carriers of the HbS allele are typically asymptomatic (*4*) whereas HbS homozygosity has severe pathological consequences and is linked to shortened lifespan (*5*). Despite this, the HbS allele has been maintained in sub-Saharan Africa by balancing selection because it confers – by incompletely understood means – a degree of protection against the effects of *Plasmodium* infection and malaria (*6*).

Sickling red blood cells were first described in 1840 – seventy years prior to their discovery in humans (*7*) – when Gulliver (*8*) reported unusual erythrocyte shapes in blood from white-tailed deer (*Odocoileus virginianus*). Subsequent research spanning more than a century revealed that sickling, at least as an *in vitro* phenotype, is widespread amongst deer species worldwide (*8*-*11*) (Fig. 1, Table S1). It is not, however, universal: red blood cells from reindeer (*Rangifer tarandus*) and European elk (*Alces alces*, known as moose in North America) do not sickle; neither do erythrocytes from most North American wapiti [*Cervus canadensis*, 25 out of 27 sampled in (*12*); 5 out of 5 sampled in (*11*)]. Below, we will use the term moose for *A. alces* to avoid confusion, as wapiti are also commonly referred to as elk in North America.

Sickling deer erythrocytes are similar to human HbS cells with regard to their gross morphology and the tubular ultrastructure of haemoglobin polymers (*13-16*). Moreover, as in humans, sickling is reversible through modulation of oxygen supply or pH (*9*, *17*). As in humans, deer sickling is mediated by specific β-globin alleles (*18*, *19*), with both sickling and non-sickling alleles segregating in wild populations of white-tailed deer (*20*). As in humans, foetal haemoglobin does not sickle under the same conditions (*19*) and α-globin – two copies of which join two β-globin proteins to form the haemoglobin tetramer - is not directly implicated in sickling etiology (*18*, *21*).

At the same time, a suit of striking differences emerged: whereas human sickling occurs when oxygen tension is low, deer erythrocytes sickle under high pO_2 and at alkaline pH (*17*). Unlike in humans, the sickling allele (previously labelled β^{III}) in white-tailed deer – the only species where allelic diversity has been explored systematically – is the major allele, with a large fraction of individuals (≥60%) homozygous for β^{III} (*20*, *22*). Finally, partial peptide digests suggested that sickling white-tailed deer did not carry the E6V mutation that causes sickling in humans (*22*), leaving the genetic basis of sickling in deer unresolved.

In addition, whereas sickling in humans is unequivocally harmful, it is not known whether sickling in deer is similarly costly, inconsequential or adaptive. Sickling is evident *in vitro* in most deer species, has been induced *in vivo* in sika deer (*Cervus nippon*) by intravenous administration of sodium bicarbonate (*23*) and can be triggered by exercise regimes that lead to transient respiratory alkalosis (*24*).

However, whether sickling occurs intravascularly at appreciable frequency under physiological conditions remains uncertain (*23*). Animals homozygous for the sickling allele do not display aberrant haematological values or other pathological traits (*25*), in line with findings that – unlike human sickle cells – sickled deer erythrocytes do not exhibit increased mechanical fragility *in vitro* (*17*, *18*).

Results and Discussion

The molecular basis of sickling in deer is distinct from the human disease state

To dissect the molecular basis of sickling in deer and elucidate its evolutionary history and potential adaptive significance, we used a combination of whole-genome sequencing, locus-specific assembly and targeted amplification to determine the sequences of β-globin genes in 15 deer species, including sickling and non-sickling taxa (Fig. 1, Table S1). Globin genes in mammals are located in paralog clusters, which – despite a broadly conserved architecture – constitute hotbeds of pseudogenization, gene duplication, conversion, and loss (*26*, *27*). In ruminants in particular, globin cluster evolution is highly dynamic. The entire β-globin cluster is triplicated in goat (*Capra hircus*) (*28*) and duplicated in cattle (*Bos taurus*) (*29*), where the two copies of the ancestral β-globin gene sub-functionalized to become specifically expressed in adult (HBB_A) and foetal (HBB_F) blood. Consistent with this duplication event pre-dating the Bovidae-Cervidae split, primers designed to amplify HBB_A in both these families frequently co-amplified HBB_F (see Materials and Methods, Fig. S1c,d). In the first instance, we assigned foetal and adult status based on residues specifically shared with either HBB_A or HBB_F in cattle, resulting in independent clustering of putative $\mathrm{HBB_A}$ and $\mathrm{HBB_F}$ genes on an $\mathrm{HBB_{A/F}}$ gene tree (Fig. S2). To confirm these assignments, we sequenced mRNA from the blood (red cell component) of an adult Père David's deer (*Elaphurus davidianus*) and assembled the erythrocyte transcriptome *de novo* (see Materials and Methods). We identified a highly abundant β-globin transcript (>200,000 transcripts/million, Fig. S3a) corresponding precisely to the putative adult β-globin gene amplified from genomic DNA of the same individual (Fig. S3b). Reads that uniquely matched the putative $\mathrm{HBB_A}$ gene were >2000-fold more abundant than reads uniquely matching the putative $\mathrm{HBB_F}$ gene (see Materials and Methods), which is expressed at low levels, as is the case in human adults (*30*). Our assignments are further consistent with peptide sequences for white-tailed deer (*22*), fallow deer (*Dama dama*) (*31*) and reindeer (*32*) that were previously obtained from the blood of adult individuals. We then considered deer $\mathrm{HBB_A}$ orthologs in a wider mammalian context, restricting analysis to species with high-confidence HBB assignments (see Materials and Methods). Treating wapiti as non-sickling, and four species as indeterminate (no or insufficient phenotyping of sickling; Table S1), we find three residues (Fig. 1) that discriminate sickling from non-sickling species: 22 (non-sickling: E, sickling: V/I), 56 (*n-s*: H, *s*: G), and 87 (*n-s*: K, *s*: Q/H). The change at residue 22, from an ancestral glutamic acid to a derived valine (isoleucine in *Pudu puda*) is reminiscent of the human HbS mutation and occurs at a site that is otherwise highly conserved throughout mammalian evolution. The only other amino acid state at residue 22 in our alignment is a biochemically conservative change to aspartic acid (D) in the brown rat (*Rattus norvegicus*).

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

To understand how sickling-associated amino acids promote polymerization, we examined these residues in their protein structural context. Residue 22 lies on the surface of the haemoglobin tetramer, at the start of the second alpha helix (Fig. 2a). Close to residue 22 are residue 56 and two other residues that differ between non-sickling reindeer and moose (but not wapiti) and established sickling species: 19 (*n-s*: K, *s*: N) and 120 (*n-s*: K, *s*: G/S). Together these residues form part of a surface of increased hydrophobicity in sickling species (Fig. 2b). Distal to this surface, residue 87 is situated at the perimeter of the EF pocket, which in humans interacts with 6V to laterally link two β-globin molecules in different haemoglobin tetramers and stabilize the parallel strand architecture of the HbS fibre (*2*, *3*, *33*, *34*). Mutation of residue 87 in humans can have marked effects on sickling dynamics (*35*). For example, erythrocytes derived from HbS/Hb Quebec-Chori (T87I) compound heterozygotes sickle like HbS homozygotes (*36*) while Hb D-Ibadan (T87K) inhibits sickling (*37*).

Given the similarity between the human E6V mutation and E22V in sickling deer, we hypothesized that sickling occurs through an interaction in *trans* between residue 22 and the EF pocket. To test whether such an interaction is compatible with fibre formation, we carried out directed docking simulations centred on these two residues using a homology model of oxy β-globin from white-tailed deer (see Materials and Methods). We then used the homodimeric interactions from docking to build polymeric haemoglobin structures, analogous to how the 6V-EF interaction leads to extended fibres in HbS homozygotes. Strikingly, nearly half of our docking models resulted in HbS-like straight, parallel strand fibres (Fig. 2c). In contrast, when we performed similar docking simulations centred on residues other than 22V, nearly all were incompatible or much less compatible with fibre formation (Fig. 2d). Out of all 145 β-globin residues, only 19N, which forms a contiguous surface with 22V, has a higher propensity to form HbS-like fibres. By contrast, when docking is carried out using the deoxy β-globin structure 22V is incompatible with fibre formation, consistent with the observation that sickling in deer occurs under oxygenated conditions. Importantly, when this methodology is applied to human HbS, we find that 6V has the highest fibre formation propensity out of all residues under deoxy conditions (Fig. S4a), providing validation for the approach.

Next, we used a force field model to compare the energetics of fibre formation across deer species. We find that known non-sickling species and species suspected to be non-sickling based on their β-globin primary sequence (Chinese water deer, roe deer) exhibit energy terms less favourable to fibre formation than sickling species (Fig. 2e). To elucidate the relative contribution of 22V and other residues to fibre formation, we introduced all single amino acid differences found amongst adult deer β-globin individually into a sickling (*O. virginianus*) and non-sickling (*R. tarandus*) background *in silico* and considered the change in fibre interaction energy. Changes at residue 22 have the strongest predicted effect on fibre formation, along with two residues – 19 and 21 – in its immediate vicinity (Fig. S4b). Smaller effects of amino acid substitutions at residue 87, as well as residues 117 (N in *P. puda* and *O. virginianus*) and 118 (Y in *D. dama*) hint at species-specific modulation of sickling propensity. Taken together, the results support the formation of HbS-like fibres in sickling deer erythrocytes via surface interactions centred on residues 22V and 87Q in β-globin molecules of different haemoglobin tetramers.

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

To shed light on the evolutionary history of sickling and elucidate its potential adaptive significance, we considered sickling and non-sickling genotypes in phylogenetic context. First, we note that the HBB$_A$ gene tree and the species tree (derived from 20 mitochondrial and nuclear genes, see Materials and Methods) are highly discordant (Fig. 3a). Notably, non-sickling and sickling genotypes are polyphyletic on the species tree but monophyletic on the HBB$_A$ tree where wapiti, an Old World deer, clusters with moose and reindeer, two New World deer. Gene tree-species tree discordance can result from a number of evolutionary processes, including incomplete lineage sorting, gene conversion, introgression, and classic convergent evolution, where point mutations arise and fix independently in different lineages. In our case, the convergent evolution scenario fits the data poorly. Discordant amino acid states are found throughout the HBB$_A$ sequence and are not limited to sickling-related residues (see, for example, the tract of amino acids between residue 44 and 66 shared by *R. tarandus* and *C. capreolus* in Fig. 1). Furthermore, in many instances, amino acids shared between phylogenetically distant species are encoded by the same underlying codons. Conspicuously, this includes the case of residue 120 where all three codon positions differ between sickling species (GGT/AGT) and non-sickling relatives (AAG in reindeer, moose, and the non-sickling ancestor; Fig. S5, Supplementary Data File 1). Even if convergence were driven by selection on a narrow adaptive path through genotype space, precise coincidence of mutational paths at multiple non-synonymous and synonymous sites must be considered unlikely. Rather, these patterns are *prima facie* consistent with incomplete lineage sorting.

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

To shore up this conclusion and rule out alternative evolutionary scenarios, we next asked whether identical genotypes, rather than originating from *de novo* mutations, might have been independently reconstituted from genetic diversity already present in other species (via introgression) or in other parts of the genome (via gene conversion). To evaluate the likelihood of introgression and particularly gene conversion, which has been attributed a prominent role in the evolution of mammalian globin genes (*26*), we first searched for evidence of recombination in an alignment of deer HBB$_A$ and HBB$_F$ genes. HBB$_F$, the most recently diverged paralog of HBB$_A$ and itself refractory to sickling (*19*), is the prime candidate to donate non-sickling residues to HBB$_A$ in a conversion event. Using a combination of phylogeny-based and probabilistic detection methods and applying permissive criteria that allow inference of shorter recombinant tracts (see Materials and Methods), we identify eight candidate HBB$_F$-to-HBB$_A$ events, two of which, in Chinese water deer and wapiti, are strongly supported by different methods and likely account for hybrid genotypes in exon 2 (Fig. 3b). In addition, we find evidence for recombination in wapiti exon 3. Unlike regions further upstream, the segment affected is 100% identical to other *Cervus spp*., which might be explained by allelic recombination during incomplete lineage sorting (see below). Note, however, that we find no evidence for gene conversion at residue 22 (Fig. 3b, Fig. S6) even when considering poorly supported candidate events. Recombination between HBB_F and/or HBB_A genes therefore does not explain the re-appearance of 22V in white-tailed deer and pudu (or 22E in wapiti). Consistent with this, removal of putative recombinant regions does not affect the HBB_F/HBB_A gene tree, with wapiti robustly clustered with other non-sickling species whereas white-tailed deer and pudu cluster with Old World sickling species (Fig. S6c). We further screened raw genome sequencing data from white-tailed deer and wapiti for potential donor sequences beyond HBB_F, such as HBE or pseudogenized HBD sequences, but did not find additional candidate donors. Thus, although gene conversion is a frequent phenomenon in the history of mammalian globins (*26*) and contributes to evolution of HBB loci in deer, it does not by itself explain the recurrence of key sickling/non-sickling residues. Rather, gene conversion introduces additional complexity on a background of incomplete lineage sorting.

Balancing selection has maintained ancestral variation in HBB_A

The presence of incomplete lineage sorting and gene conversion confounds straightforward application of rate-based (dN/dS-type) tests for selection, making it harder to establish whether the sickling genotype is simply tolerated or has been under selection. We therefore examined earlier protein-level data on HBB_A allelic diversity in extant deer populations. Intriguingly, we find evidence for long-term maintenance of ancestral variation. Two rare non-sickling β-globin alleles in white-tailed deer (previously identified from partial peptide digests) cluster with the non-sickling β-globin of reindeer rather than with the white-tailed deer sickling allele (Fig. 3c), albeit with modest bootstrap support. In addition, protein-level allelic diversity has also been observed in fallow deer, a sickling Old World deer, where the alternate β-chain

(*31*) is closely related to the moose sequence but substantially different from the sickling allele that we recovered in our sample (Fig. 3c). Finally, phenotypic heterogeneity in wapiti (*12*) and sika deer (*38*) sickling indicates that rare sickling and non-sickling variants, respectively, also segregate in these two species. Taken together, these findings point to the long-term maintenance of ancestral variation through successive speciation events dating back to the most common ancestor of Old World and New World deer, an estimated ~13.6 million years ago (mya) [CI: 9.84-17.33mya, (*39*)].

Might this polymorphism have been maintained simply by chance or must balancing selection be evoked to account for its survival? We currently lack information on broader patterns of genetic diversity at deer HBB_A loci and surrounding regions that would allow us to search for footprints of balancing selection explicitly. However, we can estimate the probability *P* that a trans-species polymorphism has been maintained along two independent lineages by neutral processes alone as

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

where *T* is the number of generations since the two lineages split and N_e is the effective population size (*40*, *41*). For simplicity, N_e is assumed to be constant over *T* and the same for both lineages. In the absence of reliable species-wide estimates for N_e, we can nonetheless ask what N_e would be required to meet a given threshold probability. Conservatively assuming an average generation time of 1 year (*42*, *43*) and a split time of 7.2mya [the lowest divergence time estimate in the literature (*39*)], N_e would have to be 2,403,419 to reach a threshold probability of 0.05 (1,563,460 for *P*=0.01). Although deer populations can have a large census population sizes, an N_e >2,000,000 for both fallow and white-tailed deer is comfortably outside what we would expect for large-bodied mammals, >4-fold higher than estimates for wild mice (*44*) and >2-fold higher even than estimates for African populations of *Drosophila melanogaster* (*45*). Consequently, we posit that the HBB_A trans-species polymorphism is inconsistent with neutral evolution and instead reflects the action of balancing selection.

Balanced polymorphisms shared between human and chimp are principally related to immune function and parasite pressure (*46*, *47*), so it is tempting to speculate that similar selection pressures might operate in deer, which host a number of intra-erythrocytic parasites including *Babesia* (*48*) and *Plasmodium* (*49*). Unlike in humans, however, the maintenance of two distinct alleles cannot be attributed to heterozygote advantage: Sickling homozygotes are the norm in white-tailed deer (*20*, *22*, *50*) and not associated with an outward clinical phenotype, indicating that the cost of bearing two sickling alleles must be comparatively low. Why, then, is the non-sickling allele being maintained? One possibility is that the sickling allele is cost-free most of the time, as previously suggested (*24*), but carries a burden in a particular environment, so that the sickling or non-sickling allele might be favoured in different sub-populations depending on local ecology. In this scenario, allelic diversity might be maintained by migration-selection balance. However, what ecological features might define such environments remains an open question. Considering only species for which phenotypic sickling information is available, there is a marked geographic asymmetry in sickling status, where non-sickling species are restricted to arctic and subarctic (elk, reindeer) or mountainous (wapiti) habitat. This might indicate that the sickling allele loses its adaptive value in colder climates (perhaps linked to the lower prevalence of blood-born parasites). However, based on genotype, we would also predict Chinese water deer and roe deer to be amongst the non-sicklers yet these species are widespread in temperate regions and their ranges overlap extensively with sickling species, challenging the hypothesis that ambient temperature is a primary driver of sickling. Future epidemiological studies coupled to ecological and population genetic investigations will be required to unravel the evolutionary ecology of sickling in deer, establish whether parasites are indeed ecological drivers of between- and within-species differences in HBB_A genotype and, ultimately, whether deer might serve as a useful comparative system to elucidate the link between sickling and protection from the effects of *Plasmodium* infection, which remains poorly understood in humans.

Materials and Methods

Sample collection and processing. Blood, muscle tissue, and DNA samples were acquired for 15 species of deer from a range of sources (Table S1).

The white-tailed deer blood sample was heat-treated on import to the United Kingdom in accordance with import standards for ungulate samples from non-EU countries (IMP/GEN/2010/07). Fresh blood was collected into PAXgene Blood DNA tubes (PreAnalytix) and DNA extracted using the PAXgene Blood DNA kit (PreAnalytix). DNA from previously frozen blood samples was extracted using the QIAamp DNA Blood Mini kit (Qiagen). DNA from tissue samples was extracted with the QIAamp DNA Mini kit (Qiagen) using 25mg of tissue. Total RNA was isolated from an *E. davidianus* blood sample using the PAXgene Blood RNA kit (PreAnalytix) three days after collection into a PAXgene Blood RNA tube

(PreAnalytix). All extractions were performed according to manufacturers' protocols. For each sample, we validated species identity by amplifying and sequencing the cytochrome b (*CytB*) gene. With the exception of *Cervus albirostris*, we successfully amplified *CytB* from all samples using primers MTCB_F/R (Fig. S1) and conditions as described in (*51*). Phusion High-Fidelity PCR Master Mix (ThermoFisher) was used for all amplifications. PCR products were purified using the MinElute PCR Purification Kit (Qiagen) and Sanger-sequenced with the amplification primers. The *CytB* sequences obtained were compared to all available deer *CytB* sequences in the 10kTrees Project (*52*) using the *ape* package (function *dist.dna* with default arguments) in R (*53*). In all cases, the presumed species identity of the sample was confirmed (Table S2).

Whole genome sequencing. *O. virginianus* genomic DNA was prepared for sequencing using the NEB DNA library prep kit (New England Biolabs) and sequenced on the Illumina HiSeq platform. The resulting 229 million paired-end reads were filtered for adapters and quality using Trimmomatic (*54*) with the following parameters: *ILLUMINACLIP:adapters/TruSeq3-PE-2.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:50*. Inspection of the remaining 163.5M read pairs with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) suggested that overrepresented sequences had been successfully removed.

Mapping and partial assembly of the *O. virginianus* β-globin locus. To seed a local assembly of the *O. virginianus* β-globin locus we first mapped *O. virginianus* trimmed paired-end reads to the duplicated β-globin locus in the hard-masked *B.*

taurus genome (UMD 3.1.1; chr15: 48973631-49098735). The β-globin locus is defined here as the region including all *B. taurus* β-globin genes (HBE1, HBE4, HBB, HBE2, HBG), the intervening sequences and 24kb either side of the two outer β-globins (HBE1, HBG). The mapping was performed using bowtie2 (*55*) with default settings and the optional *--no-mixed* and *--no-discordant* parameters. 110 reads mapped without gaps and a maximum of one nucleotide mismatch. These reads, broadly dispersed across the *B. taurus* β-globin locus (Fig. S7), were used as seeds for local assembly using a customised aTRAM (*56*) pipeline (see below). Prior to assembly, the remainder of the reads were filtered for repeat sequences by mapping against Cetartiodactyla repeats in Repbase (*57*). The aTRAM.pl wrapper script was modified to accept two new arguments: *max_target_seqs <int>* limited the number of reads found by BLAST from each database shard; *cov_cutoff <int>* passed a minimum coverage cut-off to the underlying Velvet 1.2.10 assembler (*58*). The former modification prevents stalling when the assembly encounters a repeat region, the latter discards low coverage contigs at the assembler level. aTRAM was run with the following arguments: *-kmer 31 -max_target_seqs 2000 -ins_length 270 -exp_coverage 8 -cov_cutoff 2 -iterations 5*. After local assembly on each of the 110 seed reads, the resulting contigs were combined using Minimo (*59*) with a required minimum nucleotide identity of 99%. To focus specifically on assembling the adult β-globin gene, only contigs that mapped against the *B. taurus* adult β-globin gene ±500bp (chr15: 49022500-49025000) were retained and served as seeds for another round of assembly. This procedure was repeated twice. The final 59 contigs were compared to the UMD 3.1.1 genome using BLAT and mapped exclusively to either the adult or foetal *B. taurus* β-globin gene. From the BLAT alignment, we identified short sequences that were perfectly conserved between the assembled deer contigs and the *B. taurus* as well as the sheep (*Ovis aries*) assembly (Oar_v3.1). Initial forward and reverse primers (Ovirg_F1/Ovirg_R1, Fig. S1a) for β-globin amplification were designed from these conserved regions located 270bp upstream (chr15:49022762-49022786) and 170bp downstream (chr15:49024637-49024661) of the *B. taurus* adult β-globin gene, respectively.

Globin gene amplification and sequencing. Amplification of β-globin from *O. virginianus* using primers Ovirg_F1 and Ovirg_R1 yielded two products of different molecular weights (~2000bp and ~1700bp; Fig. S1c), which were isolated by gel extraction and Sanger-sequenced using the amplification primers. The high molecular weight product had higher nucleotide identity to the adult (93%) than to the foetal (90%) *B. taurus* β-globin coding sequence. Note that the discrepancy in size between the adult and foetal β-globin amplicons derives from the presence of two tandem Bov-tA2 SINEs in intron 2 of the adult β-globin gene in cattle, sheep, and *O. virginianus* and is therefore likely ancestral. We designed a second set of primers to anneal immediately up- and downstream, and in the middle of the adult β-globin gene (Ovirg_F2, Ovirg_R2, Ovirg_Fmid2, Fig. S1a,b). Amplification from DNA extracts of other species with Ovirg_F1/Ovirg_R1 produced mixed results, with some species showing a two-band pattern similar to *O. virginianus*, others only a single band – corresponding to the putative adult β-globin (Fig. S1d). Using these primers, no product could be amplified from *R. tarandus*, *H. inermis*, and *C. capreolus*. We identified a 3bp mismatch to the Ovirg_R1 primer in a partial assembly of *C. capreolus* (Genbank accession: GCA_000751575.1; scaffold: CCMK010226507.1) that is likely at fault. A re-designed reverse primer (Ccap_R1) successfully amplified the adult β-globin gene from the three deer species above as well as *C. canadensis*

(Fig. S1). All amplifications were performed using Phusion High-Fidelity PCR Master Mix (ThermoFisher), with primers as listed in Fig. S1a, and 50-100ng of genomic DNA. Annealing temperature and step timing were chosen according to manufacturer guidelines. Amplifications were run for 35 cycles. Gel extractions were performed on samples resolved on 1% agarose gels for 40 minutes at 90V using the MinElute Gel Extraction Kit (Qiagen) and following the manufacturer's protocol. PCR purifications were performed using the MinElute PCR Purification Kit (Qiagen) following the manufacturer's protocol. All samples were sequenced using the Sanger method with amplification primers and primer Ovirg_Fmid2.

Transcriptome sequencing and assembly. RNA was extracted from the red cell component of a blood sample of an adult Père David's deer using the PAXgene Blood RNA kit (Qiagen). An mRNA library was prepared using a Truseq mRNA library prep kit and sequenced on the MiSeq PE150 platform, yielding 25,406,472 paired-end reads, which were trimmed for adapters and quality-filtered using Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with a base quality threshold of 30. The trimmed reads were used as input for *de novo* transcriptome assembly with Trinity (*60*) using default parameters. A blastn homology search against these transcripts, using the *O. virginianus* adult β-globin CDS as query, identified a highly homologous transcript (E-value = 0; no gaps; 97.5% sequence identity compared with 92.2% identity to the foetal β-globin). The CDS of this putative β-globin transcript was 100% identical to the sequence amplified from Père David's deer genomic DNA (Fig. S3b). We used *emsar* (*61*) with default parameters to assess transcript abundances. The three most abundant reconstructed transcripts correspond to full or partial α- and β-globin transcripts, including one transcript, highlighted above, that encompasses the entire adult β-globin CDS. These transcripts are an order of magnitude more abundant than the fourth most abundant (Fig. S3a), in line with the expected predominance of α- and β-globin transcripts in mature adult red blood cells. To investigate whether the foetal β-globin could be detected in the RNA-seq data and because amplification of the foetal β-globin from Père David's deer genomic DNA was not successful, we mapped reads against the foetal β-globin gene of *C. e. elaphus*, the closest available relative. Given that the CDS of the adult β-globins in these species are 100% identical, we expected that the foetal orthologs would likewise be highly conserved. We therefore removed reads with more than one mismatch and assembled putative transcripts from the remaining 1.3M reads using the Geneious assembler v.10.0.5 (*62*) with default parameters (*fastest* option enabled). We recovered a single contig with high homology to the *C. e. elaphus* foetal β-globin CDS (only a single mismatch across the CDS). We then estimated the relative abundance of adult and the putative foetal transcripts by calculating the proportion of reads that uniquely mapped to either the adult or foetal CDS. 1820532 reads mapped uniquely to the adult sequence whereas 872 mapped uniquely to the foetal CDS, a ratio of 2088:1.

Structural analysis. Homology models were built for *O. virginianus* and *R. tarandus* β-globin sequences using the MODELLER-9v15 program for comparative protein structure modelling (*63*) using both oxy (1HHO) and deoxy (2HHB) human haemoglobin structures as templates. The structures were used for electrostatic calculations using the Adaptive Poisson-Boltzmann Solver (*64*) plugin in the Visual Molecular Dynamics (VMD) program (*65*). The surface potentials were visualised in

VMD with the conventional red and blue colours, for negative and positive potential respectively, set at ±5 kT/e.

Modelling of haemoglobin fibres. We first used the program HADDOCK (*66*) with the standard protein-protein docking protocol to generate ensembles of docking models of β-globin dimers. In each docking run, a different interacting surface centred around a specific residue was defined on each β-globin chain. All residues within 3Å of the central residue were defined as "active" and were thus constrained to be directly involved in the interface, while other residues within 8Å of the central residue were defined as "passive" and were allowed but not strictly constrained to form a part of the interface. We performed docking runs with the interaction centred between residue 87 and all other residues, generating at least 100 water-refined β-globin dimer models for each (although 600 *O. virginianus* oxy β-globin 22V-87Q models were built for use in the interaction energy calculations). The β-globin dimers were then evaluated for their ability to form HbS-like fibres out of full haemoglobin tetramers. Essentially, the contacts from the β-globin dimer models were used to build a chain of five haemoglobin molecules, in the same way that the contacts between 6V and the EF pocket lead to an extended fibre in HbS. HbS-like fibres were defined as those in which a direct contact was formed between the first and third haemoglobin tetramers in a chain (analogous to the axial contacts in HbS fibres, see Fig. 2c), and in which the chain is approximately linear. This linearity was measured as the distance between the first and third plus the distance between the third and the fifth haemoglobin tetramers, divided by the distance between the first and the fifth. A value of 1 would indicate a perfectly linear fibre, while we considered any chains with a value <1.05 to be approximately linear and HbS-like. Finally, chains containing significant steric clashes between haemoglobin tetramers (defined as >3% of Cα atoms being within 2.8Å of another Cα atom) were excluded. Fibre formation propensity was then defined as the fraction of all docking models that led to HbS-like fibres.

Interaction energy analysis. Using the 270 22V-87Q models of *O. virginianus* β-globin dimers that can form HbS-like fibres, we used FoldX (*67*) 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.

Detection of recombination events. We considered two sources of donor sequence for recombination into adult β-globins: adult β-globin orthologs in other deer species and the foetal β-globin paralog within the same genome. *H. inermis* $\mathrm{HBB_F}$ was omitted from this analysis since the sequence of intron 2 was only partially determined. We used the Recombination Detection Program (RDP v.4.83) (*68*) to test for signals of recombination in an alignment of complete adult and foetal deer β-globin genes that were successfully amplified and sequenced, enabling all subtended detection methods (including primary scans for BootScan and SiScan) except LARD, treating the sequences as linear and listing all detectable events. In humans, conversion tracts of lengths as short as 110bp have been detected in the globin genes (*69*) and tracts as short as 50bp in other gene conversion hotspots (*70*, *71*). Given the presence of multiple regions of 100% nucleotide identity across the alignment of adult and foetal deer β-globins (Fig. 3b), we suspected that equally short conversion tracts might also be present. We therefore lowered window and step sizes for all applicable detection methods in RDP (Fig. S6b) at the cost of a lower signal-to-noise ratio. As the objective is to test whether recombination events could have generated the phyletic distribution of sickling/non-sickling genotypes observed empirically, this is conservative.

Phylogenetic analyses. Adult deer β-globin coding sequences were aligned with MUSCLE to a set of non-chimeric mammalian β-globin CDSs (*26*). The mammalian phylogeny (Fig. S8) is principally based on the Timetree of Life (*39*) with the order Carnivora regrafted to branch above the root of the Chiroptera and Artiodactyla to match findings in (*72*). The internal topology of Cervidae was taken from the Cetartiodactyla consensus tree of the 10kTrees Project (*52*). *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 (*73*).

Data availability. HBB_A and HBB_F full gene sequences (coding sequence plus intervening introns) have been submitted to GenBank with accession numbers KY800429-KY800452. Père David's deer RNA sequencing and white-tailed deer whole genome sequencing raw data has been submitted to the European Nucleotide

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

References

1. V. M. Ingram, Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature*. **180**, 326–328 (1957).

2. D. J. Harrington, K. Adachi, W. E. Royer Jr, The high resolution crystal structure of deoxyhemoglobin S. *Journal of Molecular Biology*. **272**, 398–407 (1997).

3. B. C. Wishner, K. B. Ward, E. E. Lattman, W. E. Love, Crystal structure of sickle-cell deoxyhemoglobin at 5 Å resolution. *Journal of Molecular Biology*. **98**, 179–194 (1975).

4. D. A. Sears, The morbidity of sickle cell trait. *The American Journal of Medicine*. **64**, 1021–1036 (1978).

5. O. S. Platt *et al*., Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N. Engl. J. Med*. **330**, 1639–1644 (1994).

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

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

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

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

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

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

12. Y. B. Weber, L. Giacometti, Sickling Phenomenon in the Erythrocytes of

Wapiti (Cervus Canadensis). *Journal of Mammalogy*. **53**, 917–919 (1972).

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

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

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

16. H. Kitchen, C. W. Easley, F. W. Putnam, W. J. Taylor, Structural comparison of polymorphic hemoglobins of deer with those of sheep and other species. *The Journal of Biological Chemistry*. **243**, 1204–1211 (1968).

17. D. Seiffge, Haemorheological studies of the sickle cell phenomenon in european red deer (Cervus elaphus). *Blut*. **47**, 85–92 (1983).

18. H. Kitchen, F. W. Putnam, W. J. Taylor, Hemoglobin Polymorphism: Its Relation to Sickling of Erythrocytes in White-Tailed Deer. *Science*. **144**, 1237–1239 (1964).

19. W. J. Taylor, C. W. Easley, Sickling phenomena of deer. *Annals of the New York Academy of Sciences*. **241**, 594–604 (1974).

20. M. J. Harris, T. H. J. Huisman, F. A. Hayes, Geographic distribution of hemoglobin variants in the white-tailed deer. *Journal of Mammalogy*. **54**, 270–274 (1973).

21. M. J. Harris, J. B. Wilson, T. H. J. Huisman, Structural studies of hemoglobin α chains from Virginia white-tailed deer. *Archives of Biochemistry and Biophysics*. **151**, 540–548 (1972).

22. K. Shimizu *et al.*, The primary sequence of the beta chain of Hb type III of the Virginia white-tailed deer (Odocoilus Virginianus), a comparison with putative sequences of the beta chains from four additional deer hemoglobins, types II, IV, V, and VIII, and relationships between intermolecular contacts, primary sequence and sickling of deer hemoglobins. *Hemoglobin*. **7**, 15–45 (1983).

23. C. J. Parshall, S. J. Vainisi, M. F. Goldberg, E. D. Wolf, In vivo erythrocyte sickling in the Japanese sika deer (Cervus nippon): methodology. *Am J Vet Res*. **36**, 749–752 (1975).

24. C. F. Whitten, Innocuous Nature of the Sickling (Pseudosickling) Phenomenon in Deer. *British Journal of Haematology*. **13**, 650–655 (1967).

25. H. Kitchen, W. J. Taylor, The sickling phenomenon of deer erythrocytes. *Adv. Exp. Med. Biol*. **28**, 325–336 (1972).

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

Repeated evolution of chimeric fusion genes in the β-globin gene family of laurasiatherian mammals. *Genome Biol Evol*. **6**, 1219–1234 (2014).

27. R. C. Hardison, Evolution of Hemoglobin and Its Genes. *Cold Spring Harb Perspect Med*. **2**, a011627–a011627 (2012).

28. T. M. Townes, M. C. Fitzgerald, J. B. Lingrel, Triplication of a four-gene set during evolution of the goat beta-globin locus produced three genes now expressed differentially during development. *Proceedings of the National Academy of Sciences of the United States of America*. **81**, 6589–6593 (1984).

29. J. C. Schimenti, C. H. Duncan, Structure and organization of the bovine beta-globin genes. *Mol Biol Evol*. **2**, 514–525 (1985).

30. J. E. Craig, S. L. Thein, J. Rochette, Fetal hemoglobin levels in adults. *Blood Reviews*. **8**, 213–224 (1994).

31. M. Angeletti *et al*., Different functional modulation by heterotropic ligands (2,3-diphosphoglycerate and chlorides) of the two haemoglobins from fallow-deer (Dama dama). *Eur J Biochem*. **268**, 603–611 (2001).

32. R. Petruzzelli *et al*., The primary structure of hemoglobin from reindeer (Rangifer tarandus tarandus) and its functional implications. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*. **1076**, 221–224 (1991).

33. K. Adachi, L. R. Reddy, S. Surrey, Role of hydrophobicity of phenylalanine beta 85 and leucine beta 88 in the acceptor pocket for valine beta 6 during hemoglobin S polymerization. *The Journal of Biological Chemistry*. **269**, 31563–31566 (1994).

34. R. L. Nagel *et al*., Beta-chain contact sites in the haemoglobin S polymer. *Nature*. **283**, 832–834 (1980).

35. K. Adachi, P. Konitzer, S. Surrey, Role of gamma 87 Gln in the inhibition of hemoglobin S polymerization by hemoglobin F. *The Journal of Biological Chemistry*. **269**, 9562–9567 (1994).

36. H. E. Witkowska *et al*., Sickle cell disease in a patient with sickle cell trait and compound heterozygosity for hemoglobin S and hemoglobin Quebec-Chori. *N. Engl. J. Med*. **325**, 1150–1154 (1991).

37. E. J. Watson-Williams, D. Beale, D. Irvine, H. Lehmann, A new haemoglobin, D Ibadan (beta-87 threonine -- lysine), producing no sickle-cell haemoglobin D disease with haemoglobin S. *Nature*. **205**, 1273–1276 (1965).

38. Butcher, P. D. & Hawkey, C. M. Red blood cell sickling in mammals. In: R. J. Montali, G. Migaki, *The Comparative Pathology of Zoo Animals* (Smithsonian Institute, 1980).

39. S. B. Hedges, J. Marin, M. Suleski, M. Paymer, S. Kumar, Tree of Life Reveals Clock-Like Speciation and Diversification. *Mol Biol Evol*. **32**, 835–845 (2015).

40. C. Wiuf, K. Zhao, H. Innan, M. Nordborg, The Probability and Chromosomal Extent of *trans*-specific Polymorphism. *Genetics*. **168**, 2363–2372 (2004).

41. Z. Gao, M. Przeworski, G. Sella, Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution*. **69**, 431–446 (2015).

42. K. H. Baker *et al.*, Strong population structure in a species manipulated by humans since the Neolithic: the European fallow deer (Dama dama dama). *Heredity*. **119**, 16–26 (2017).

43. N. Ryman, R. Baccus, C. Reuterwall, M. H. Smith, Effective Population Size, Generation Interval, and Potential Loss of Genetic Variability in Game Species under Different Hunting Regimes. *Oikos*. **36**, 257 (1981).

44. D. L. Halligan, F. Oliver, A. Eyre-Walker, B. Harr, P. D. Keightley, Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLoS Genet*. **6**, e1000825 (2010).

45. Adaptive genic evolution in the Drosophila genomes. *Proceedings of the National Academy of Sciences of the United States of America*. **104**, 2271–2276 (2007).

46. E. M. Leffler *et al.*, Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science*. **339**, 1578–1582 (2013).

47. J. C. Teixeira *et al.*, Long-Term Balancing Selection in LAD1 Maintains a Missense Trans-Species Polymorphism in Humans, Chimpanzees, and Bonobos. *Mol Biol Evol*. **32**, 1186–1196 (2015).

48. B. D. Perry, D. K. Nichols, E. S. Cullom, Babesia odocoilei Emerson and Wright, 1970 in white-tailed deer, Odocoileus virginianus (Zimmermann), in Virginia. *Journal of Wildlife Diseases*. **21**, 149–152 (1985).

49. P. C. Garnham, K. L. Kuttler, A malaria parasite of the white-tailed deer (Odocoileus virginianus) and its relation with known species of Plasmodium in other ungulates. *Proc. R. Soc. Lond., B, Biol. Sci*. **206**, 395–402 (1980).

50. P. R. Ramsey, J. C. Avise, M. H. Smith, D. F. Urbston, Biochemical variation and genetic heterogeneity in South Carolina deer populations. *The Journal of Wildlife Management*. **43**, 136 (1979).

51. A. Naidu, R. R. Fitak, A. Munguia Vega, M. Culver, Novel primers for complete mitochondrial cytochrome b gene sequencing in mammals. *Molecular Ecology Resources*. **12**, 191–196 (2012).

52. C. Arnold, L. J. Matthews, C. L. Nunn, The 10kTrees website: A new online resource for primate phylogeny. *Evol. Anthropol*. **19**, 114–118 (2010).

53. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. **20**, 289–290 (2004).

54. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

55. Ben Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Meth*. **9**, 357–359 (2012).

56. J. M. Allen, D. I. Huang, Q. C. Cronk, K. P. Johnson, aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics*. **16**, 98 (2015).

57. W. Bao, K. K. Kojima, O. Kohany, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA 2014 5:1*. **6**, 11 (2015).

58. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*. **18**, 821–829 (2008).

59. T. J. Treangen, D. D. Sommer, F. E. Angly, S. Koren, M. Pop, Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics*. **Chapter 11**, Unit 11.8 (2011).

60. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. **8**, 1494–1512 (2013).

61. S. Lee *et al.*, EMSAR: estimation of transcript abundance from RNA-seq data by mappability-based segmentation and reclustering. *BMC Bioinformatics*. **16**, 278 (2015).

62. M. Kearse *et al.*, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. **28**, 1647–1649 (2012).

63. N. Eswar *et al.*, Comparative Protein Structure Modeling Using Modeller. *Curr Protoc Bioinformatics*. **0 5**, Unit–5.6.30 (2006).

64. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America*. **98**, 10037–10041 (2001).

65. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *Journal of Molecular Graphics*. **14**, 33–38 (1996).

66. Cyril Dominguez, A. Rolf Boelens, A. M. J. J. Bonvin, HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc*. **125**, 1731–1737 (2003).

67. R. Guerois, J. E. Nielsen, L. Serrano, Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*. **320**, 369–387 (2002).

68. D. P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol*. **1**, vev003 (2015).

69. M. N. Papadakis, G. P. Patrinos, Contribution of gene conversion in the evolution of the human beta-like globin gene family. *Human Genetics*. **104**, 117–125 (1999).

70. A. J. Jeffreys, C. A. May, Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*. **36**, 151–156 (2004).

71. E. Bosch, M. E. Hurles, A. Navarro, M. A. Jobling, Dynamics of a human interparalog gene conversion hotspot. *Genome Research*. **14**, 835–844 (2004).

72. R. W. Meredith *et al*., Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science*. **334**, 521–524 (2011).

73. C. J. Ludt, W. Schroeder, O. Rottmann, R. Kuehn, Mitochondrial DNA phylogeography of red deer (Cervus elaphus). *Molecular Phylogenetics and Evolution*. **31**, 1064–1083 (2004).

74. Junge, R. E., Duncan, M. C., Miller, R. E., Gregg, D. & Kombert, M. Clinical presentation and antiviral therapy for poxvirus infection in pudu (Pudu puda). *Journal of Zoo and Wildlife Medicine* **31,** 412–418 (2000).

75. Weisberger, A. S. The Sickling Phenomenon and Heterogeneity of Deer Hemoglobin. *Proceedings of the Society for Experimental Biology and Medicine* **117,** 276–280 (1964).

76. Ogawa, E., Hasegawa, M., Fujise, H. & Kobayashi, K. Erythrocyte Sickling in Sika Deer (Cervus nippon). *J. Vet. Med. Sci*. **53,** 1075–1077 (1991).

Acknowledgments

We thank the Zoological Society of London Whipsnade Zoo (F. Molenaar), Bristol Zoological Society (S. Dow, K. Wyatt), the Royal Zoological Society of Scotland Highland Wildlife Park (J. Morse), the Penn State Deer Research Center (D. Wagner), and the Northeast Wildlife DNA Laboratory (N. Chinnici) for samples, the MRC LMS Genomics Facility for DNA and RNA sequencing, B.N. Sacks, J. Mizzi, and T. Brown for access to Tule elk sequencing data, P.D. Butcher for discussions, and P.

Sarkies, A. Brown, and B. Lehner for comments on the manuscript. This work was supported by an Imperial College Interdisciplinary Cross-Campus Studentship to A.E, an MRC Career Development Award (MR/M02122X/1) to J.A.M., a Leverhulme Trust Fellowship to V.S., and MRC core funding and an Imperial College Junior Research Fellowship to T.W.

Author contributions

A.E. performed laboratory experiments and evolutionary analyses and contributed to experimental design, data analysis and interpretation. L.T.B. and J.A.M. designed and performed structural modelling, and contributed to data analysis and interpretation. V.S. contributed tissue samples. T.W. conceived the study, contributed to experimental design, data analysis, and interpretation and wrote the manuscript with the input from all authors.

Author information

The authors declare no competing financial interests.

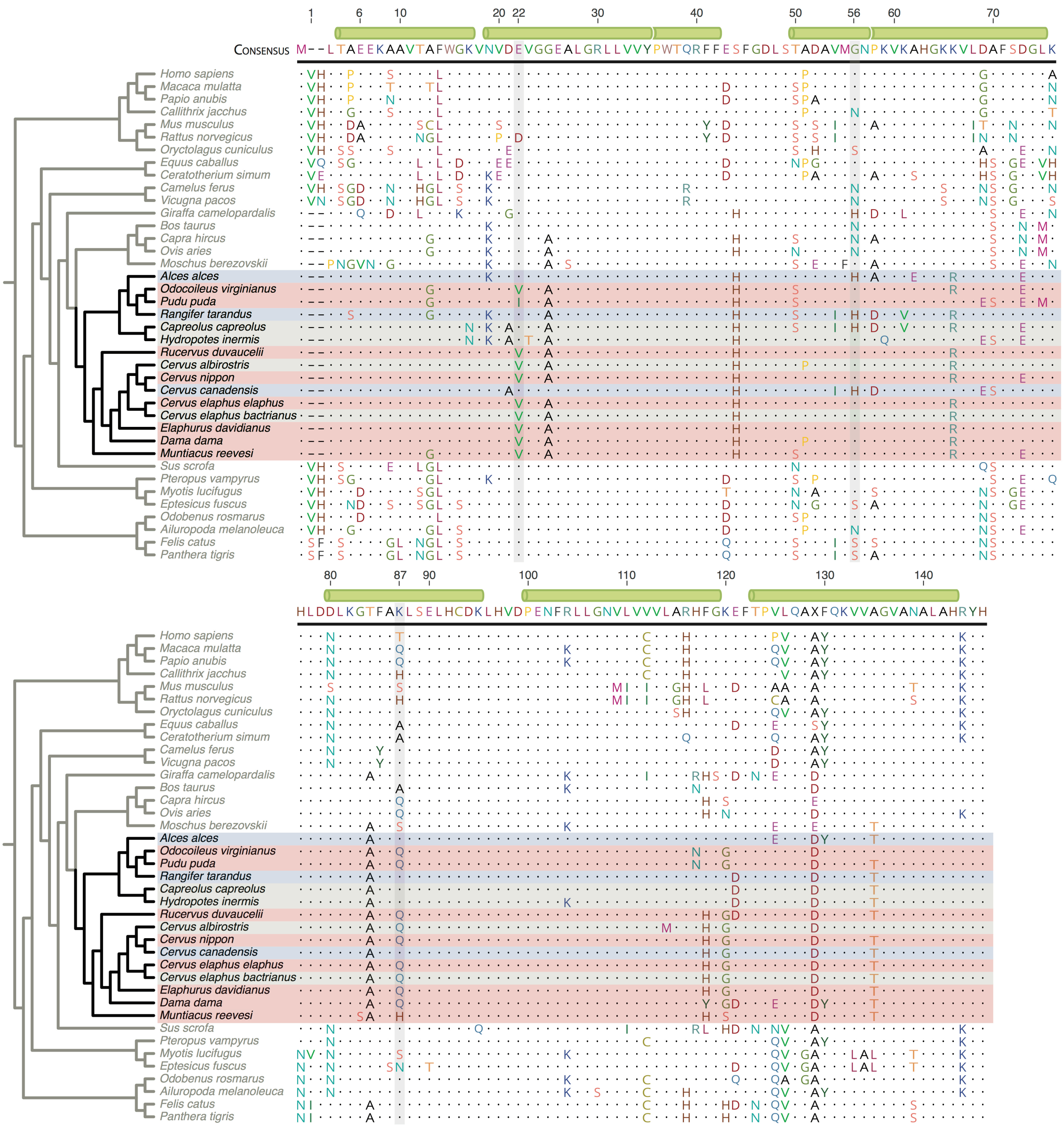


Fig. 1. Mammalian adult β-globin peptide sequences in phylogenetic context. To facilitate comparisons with prior classic literature, residues here and in the main text are numbered according to human HBB, skipping the leading methionine. Dots represent residues identical to the consensus sequences, defined by the most common amino acid (X indicates a tie). Key residues discussed in the text are highlighted. β-globin sequences from deer are coloured according to documented sickling state: red = sickling, blue = non-sickling, grey = indeterminate (Table S1). Green cylinders highlight the position of α-helices in the secondary structure of human HBB.

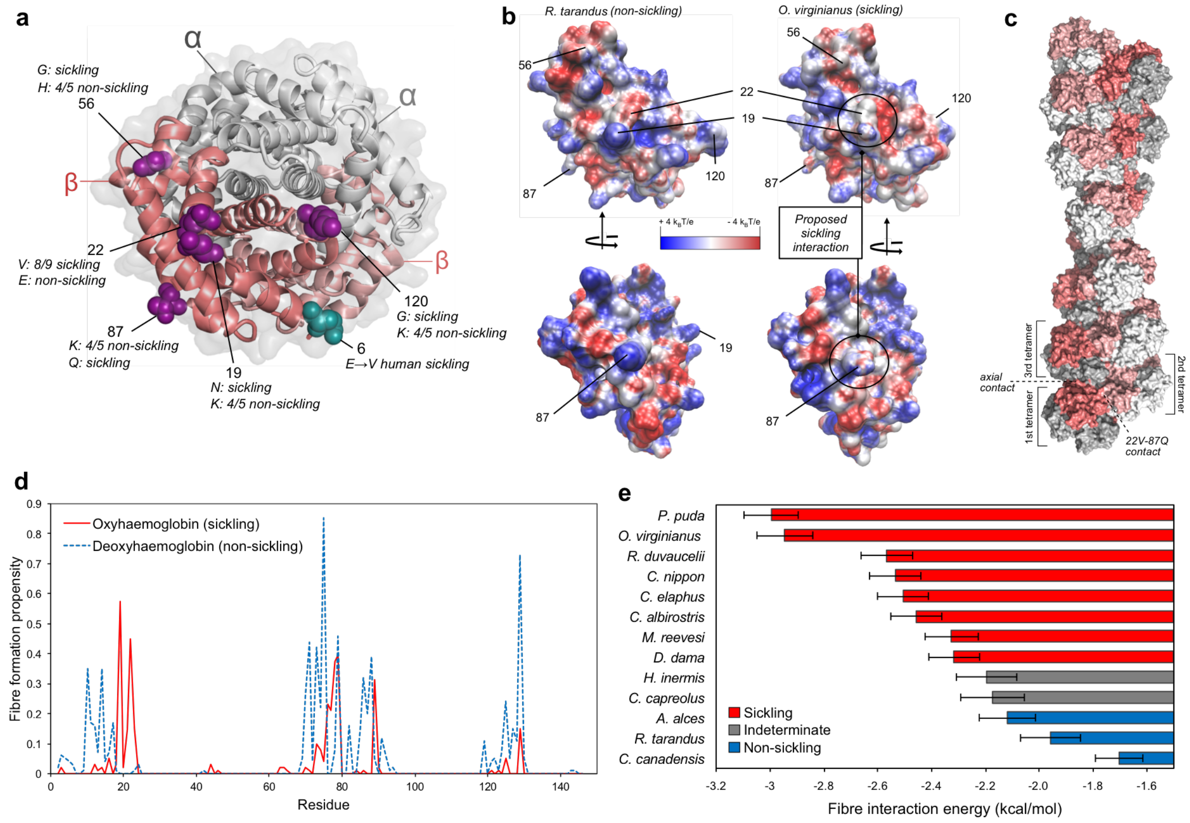


Fig. 2. Structural basis for sicking of deer haemoglobin. a, Structure of oxyhaemoglobin (PDB ID: 1HHO), with the key residues associated with sickling highlighted in one of the β-globin chains. **b**, Comparison of the electrostatic surfaces of oxy β-globin from a non-sickling (*R. tarandus*) and sickling (*O. virginianus*) species. **c**, Example of a haemoglobin fibre formed via directed docking between residues 22V and 87Q of *O. virginianus* oxy β-globin. **d**, Fibre formation propensity derived from docking simulations centred at a given focal residue in *O. virginianus* oxy and deoxy β-globin. 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.

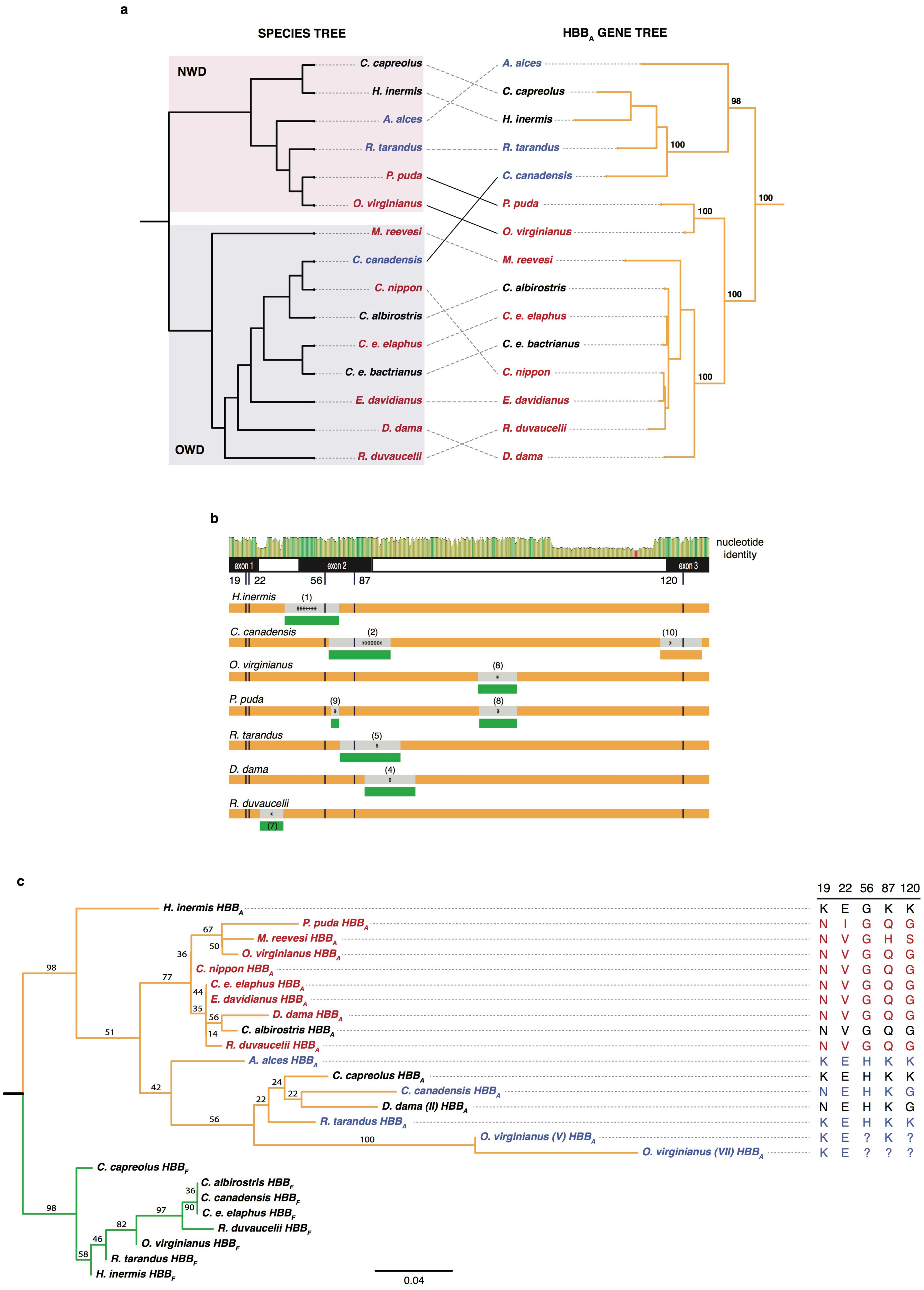

a
SPECIES TREE
HBB_A GENE TREE
NWD
OWD
C. capreolus
H. inermis
A. alces
R. tarandus
P. puda
O. virginianus
M. reevesi
C. canadensis
C. nippon
C. albirostris
C. e. elaphus
C. e. bactrianus
E. davidianus
D. dama
R. duvaucelii
98
100
100
100
100
100
100
b
nucleotide identity
exon 1
exon 2
exon 3
19
22
56
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
P. puda HBB_A N I G Q G
M. reevesi HBB_A N V G H S
O. virginianus HBB_A N V G Q G
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
C. capreolus HBB_A K E H K K
C. canadensis HBB_A N E H K G
D. dama (II) HBB_A N E H K G
R. tarandus HBB_A K E H K K
O. virginianus (V) HBB_A K E ? K ?
O. virginianus (VII) HBB_A K E ? ? ?
C. capreolus HBB_F
C. albirostris HBB_F
C. canadensis HBB_F
C. e. elaphus HBB_F
R. duvaucelii HBB_F
O. virginianus HBB_F
R. tarandus HBB_F
H. inermis HBB_F
98
67
50
36
77
44
35
56
14
51
42
24
22
22
56
100
98
36
90
97
82
46
58
0.04

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 a 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 Materials and Methods). Details for individual events (numbered in parentheses) are given in Fig. S6a. **c,** Maximum likelihood tree of adult (orange) and foetal (green) β-globin proteins. 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.

a

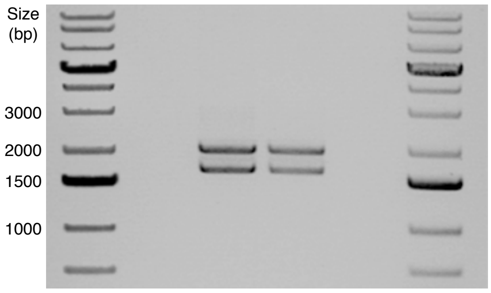
Name	Sequence (5' to 3')	Tm (°C)
Ovirg_F1	ATTTGTATTGCTGGAATGACTGAGA	64.5
Ovirg_R1	CCATCTGAGTAAGAGACAGTGAAAT	61.5
Ovirg_F2	CTTACACTTGCTTCTGACACAACCG	67.7
Ovirg_R2	GGAAAGCAGGAAAGGGAGCT	65.7
Ccap_R1	CCATCTGAGTAAGAGATCATGAAAT	61.3
Ovirg_Fmid2	GTGATAAGCTGCACGTGGAT	62.7
MTCB_F	CCHCCATAAATAGGNGAAGG	57.6
MTCB_R	WAGAAYTTCAGCTTTGGG	55.3

b

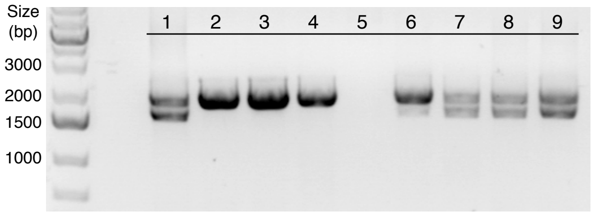
Sample	Adult β-globin	Foetal β-globin
WTD	Ovirg_F1 + Ovirg_R1	GE
SD	Ovirg_F1 + Ovirg_R1	GE
RED	Ovirg_F1 + Ovirg_R1	GE
BACT	Ovirg_F1 + Ovirg_R1	GE
WLD	Ovirg_F1 + Ovirg_R1	GE
BFD	Ovirg_F1 + Ovirg_R1	Not amplified
PDD	Ovirg_F1 + Ovirg_R1	Not amplified
SIKA	Ovirg_F1 + Ovirg_R1	Not amplified
ELK	Ovirg_F1 + Ovirg_R1	Not amplified
MUNT	Ovirg_F1 + Ovirg_R1	Not amplified
PUDU	Ovirg_F1 + Ovirg_R1	Not amplified
REIN	Ovirg_F1 + Ccap_R1	Ovirg_F2 + Ovirg_R1
ROE	Ovirg_F1 + Ccap_R1	Ovirg_F2 + Ovirg_R1
CWD	Ovirg_F1 + Ccap_R1	Ovirg_F2 + Ovirg_R1
WAP	Ovirg_F1 + Ccap_R1	Ovirg_F2 + Ovirg_R1

GE, Gel Excision.

c

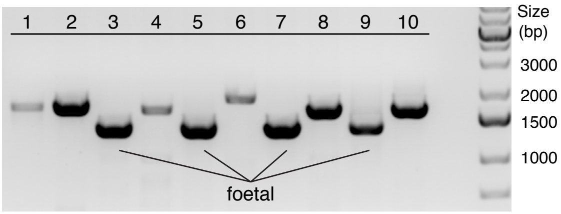


d



1-SD; 2-SIKA; 3-PDD; 4-BFD; 5-REIN; 6-ELK; 7-RED; 8-WLD; 9-BACT

e



1-WLD; 2-SD; 3,4-WAP; 5,6-REIN; 7,8-ROE; 9,10-CWD

Fig. S1.

Amplification of deer β-globin genes. a, Primers used in this study. **b**, Primer combinations used to amplify adult and foetal β-globin genes in different species. Where possible, gel excision was used to isolate the co-amplified foetal β-globin band. In certain cases, the adult gene could also be selectively amplified using primers Ovirg_F1/Ovirg_R2 (see panel e, lanes 1,2). **c**, Agarose gel showing co-amplification of adult and foetal β-globin genes in white-tailed deer with Ovirg_F1/Ovirg_R1; the two lanes show two different individuals. **d**, Agarose gel showing heterogeneity of amplification products in different deer using Ovirg_F1/Ovirg_R1. **e**, Agarose gel showing selective amplification of adult and foetal β-globin genes. Lanes 1 & 2: adult β-globin using Ovirg_F1/Ovirg_R2; all other lanes with primer combinations described in panel b.

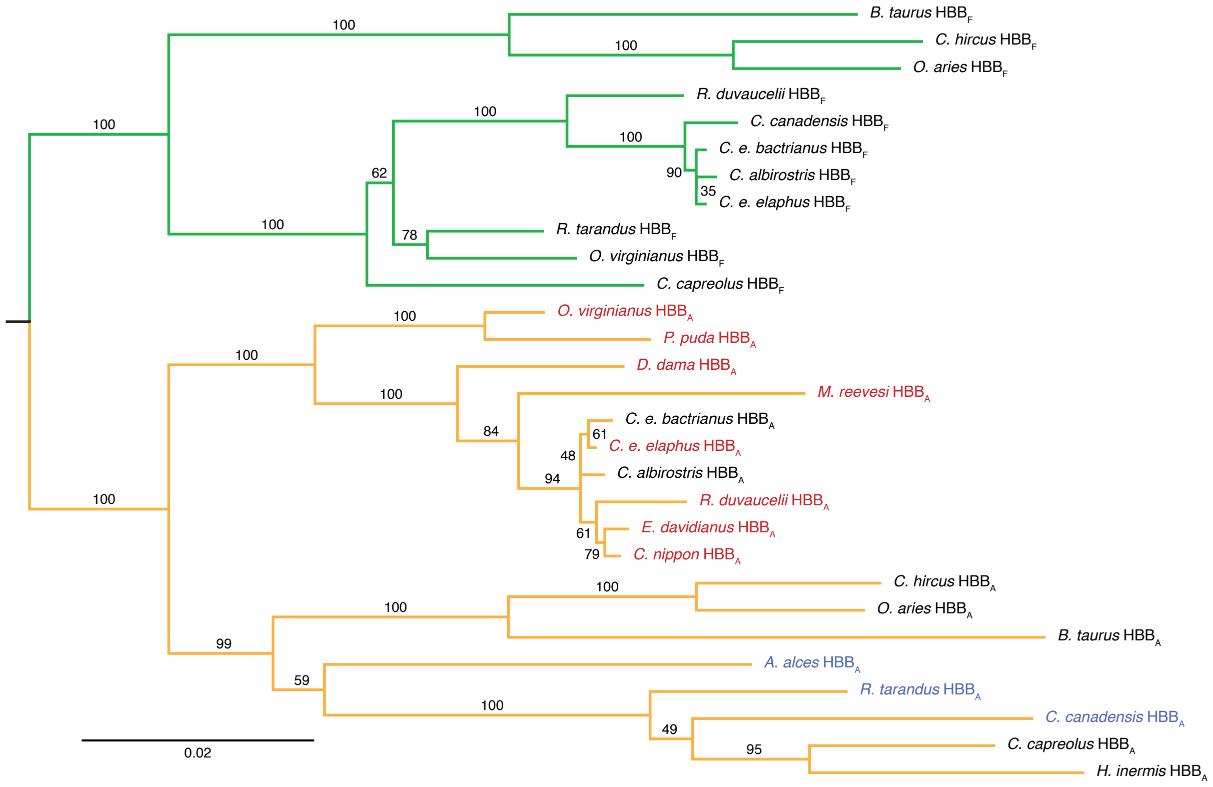


Fig. S2

Putative HBB$_A$ and HBB$_F$ genes cluster separately on an HBB$_A$ /HBB$_F$ gene tree. The tree is a maximum likelihood reconstruction based on a nucleotide alignment of complete exonic and intronic sequences (see Materials and Methods). The sequence of *H. inermis* HBB$_F$ was only partially resolved and is hence omitted. Branches are coloured as adult (orange) or foetal (green). Tip labels for HBB$_A$ are coloured according to the species' propensity to sickle where known (red = sickling, blue = non-sickling). Bootstrap values are derived from 100 bootstrap replicates. The scale bar shows the number of nucleotide substitutions per site.

a

ID of reconstructed transcript	Transcript length	Predicted product	Genbank ID (species)*	Abundance (TPM)
DN5482_c3_g1_i1	731	α-globin chain	JF811751.1 (*Pantholops hodgsonii*)	325146
DN1442_c2_g1_i1	294	β-globin mRNA	NM_001014902.3 (*Bos taurus*)	285287
DN27361_c1_g1_i1	747	β-globin mRNA	XM_006061581.1 (*Bubalus bubalis*)	224345
DN10562_c0_g1_i1	651	ubiquitin A-52 residue ribosomal protein fusion product 1 (UBA52)	XM_019963827.1 (*Bos indicus*)	2255
DN9617_c0_g5_i1	322	16S bacterial ribosomal RNA gene	CP019213.1, nucleotides 680747-681068 (*Escherichia coli*)	2219
DN27542_c0_g1_i1	2033	5'-aminolevulinate synthase 2 (ALAS2)	XM_005894452.1 (*Bos mutus*)	2010
DN20291_c0_g1_i1	621	S100 calcium binding protein A12 (S100A12)	XM_005909570.2 (*Bos mutus*)	1850
DN52287_c4_g1_i1	2893	18S ribosomal RNAgene	JN412502.1 (*Bubalus bubalis*)	1478
DN5641_c0_g1_i1	979	ferritin heavy chain 1 (FTH1)	NM_174062.3 (*Bos taurus*)	1404
DN9617_c0_g2_i1	1480	16S bacterial ribosomal RNA gene	CP018801.1 (4464861-4466340) (*Escherichia coli*)	1348

*Best match in the non-redundant nucleotide database (queried using MegaBLAST); TPM, transcripts per million.

b

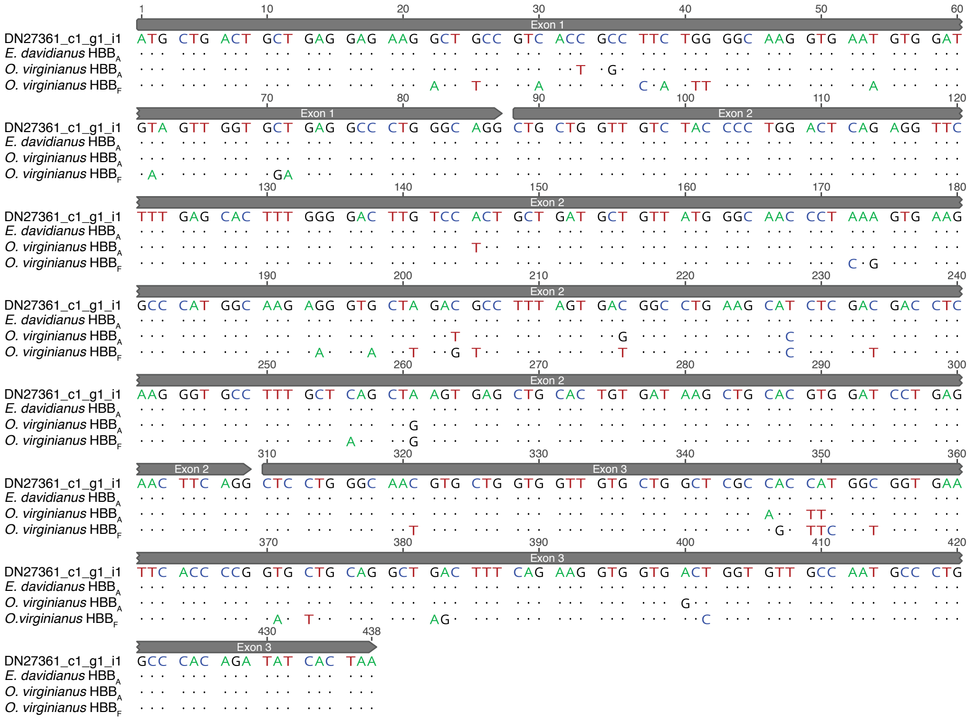

Exon 1
Exon 2
Exon 3
DN27361_c1_g1_i1
E. davidianus HBB_A
O. virginianus HBB_A
O. virginianus HBB_F
ATG CTG ACT GCT GAG GAG AAG GCT GCC GTC ACC GCC TTC TGG GGC AAG GTG AAT GTG GAT
GTA GTT GGT GCT GAG GCC CTG GGC AGG CTG CTG GTT GTC TAC CCC TGG ACT CAG AGG TTC
TTT GAG CAC TTT GGG GAC TTG TCC ACT GCT GAT GCT GTT ATG GGC AAC CCT AAA GTG AAG
GCC CAT GGC AAG AGG GTG CTA GAC GCC TTT AGT GAC GGC CTG AAG CAT CTC GAC GAC CTC
AAG GGT GCC TTT GCT CAG CTA AGT GAG CTG CAC TGT GAT AAG CTG CAC GTG GAT CCT GAG
AAC TTC AGG CTC CTG GGC AAC GTG CTG GTG GTT GTG CTG GCT CGC CAC CAT GGC GGT GAA
TTC ACC CCG GTG CTG CAG GCT GAC TTT CAG AAG GTG GTG ACT GGT GTT GCC AAT GCC CTG
GCC CAC AGA TAT CAC TAA
438

Fig. S3

Reconstructing adult β-globin sequence from RNA sequencing data. a, the ten most abundant transcripts in the *de novo* assembled *E. davidianus* red blood cell transcriptome. **b**, Nucleotide alignment of the *E. davidianus* β-globin CDS derived from the *de novo* transcriptome assembly, the putative *E. davidianus* adult β-globin CDS derived from amplification, and the foetal and adult β-globin CDSs from *O. virginianus*.

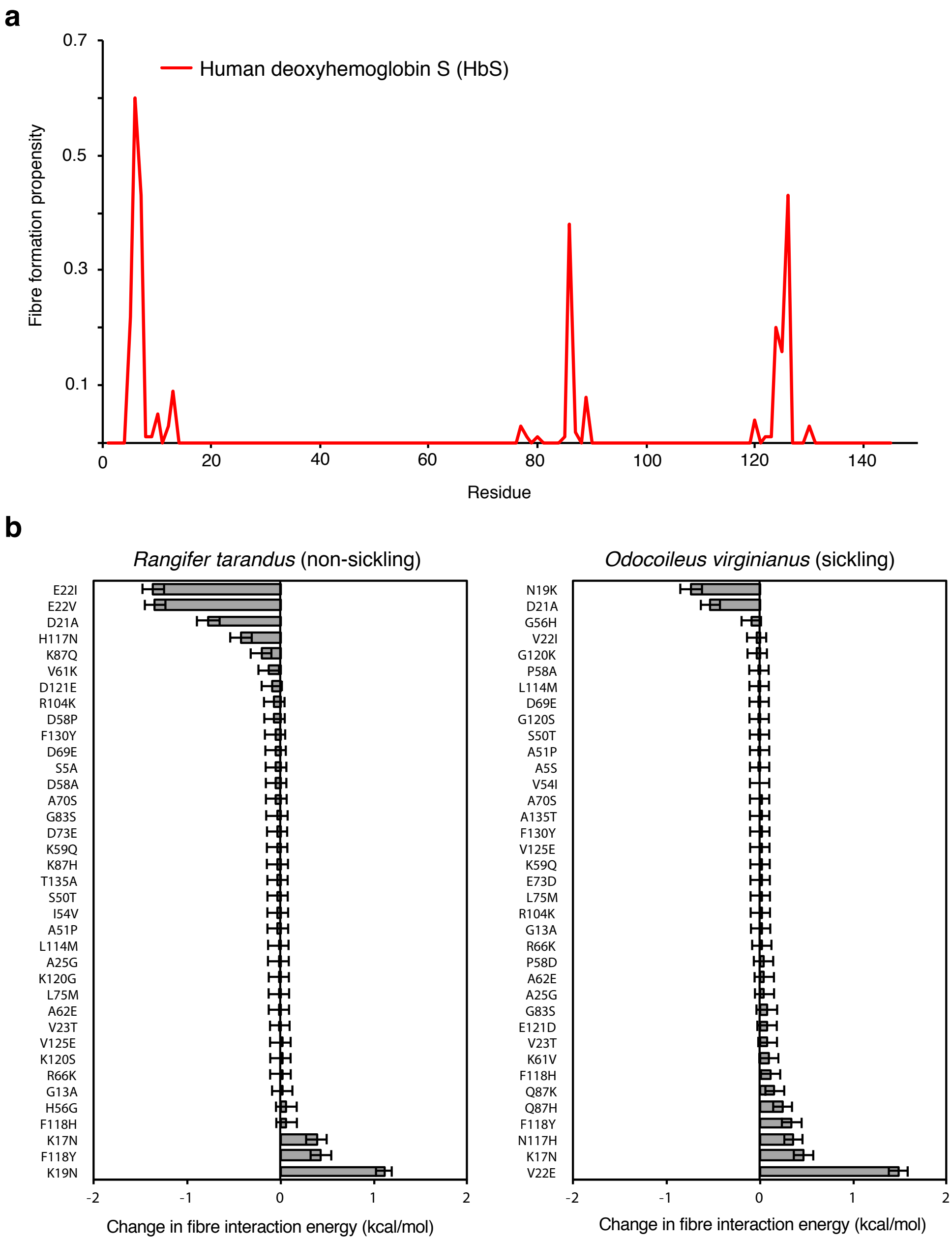
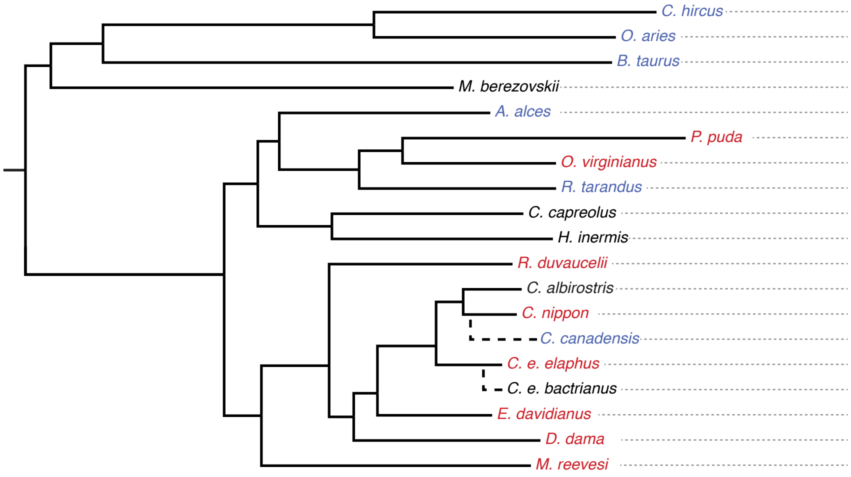


Fig. S4

Analysis of sickling propensity in human and deer. a, Fibre formation propensity assuming an interaction between the EF pocket and a given focal residue on two different β-globin chains, essentially as in Fig. 2d, but using the structure of human deoxyhaemoglobin S (HbS). Fibre formation propensity represents the fraction of the 100 β-globin dimer models built for each position that can form HbS-like fibres. **b**, Effects on fibre interaction energy of replacing defined single amino acids in the primary sequence of either *R. tarandus* or *O. virginianus*. Negative values indicate stronger interactions and thus an increased likelihood of fibre formation. These values show the mean over all 270 22V-87Q docking models compatible with fibre formation, and error bars represent the standard error of the mean.



	19	22	56	87	120
C. hircus	AAA	GAA	AAC	CAG	AGT
O. aries	AAA	GAA	AAC	CAG	AAT
B. taurus	AAA	GAA	AAC	GCG	AAG
M. berezovskii	AAA	GAA	GGC	TCT	AAG
A. alces	AAA	GAA	CAC	AAG	AAG
P. puda	AAT	ATA	GGC	CAG	GGT
O. virginianus	AAT	GTA	GGC	CAG	GGT
R. tarandus	AAA	GAA	CAC	AAG	AAG
C. capreolus	AAA	GAA	CAC	AAG	AAG
H. inermis	AAA	GAA	GGC	AAG	AAG
R. duvaucelii	AAT	GTA	GGC	CAG	GGT
C. albirostris	AAT	GTA	GGC	CAG	GGT
C. nippon	AAT	GTA	GGC	CAG	GGT
C. canadensis	AAC	GAA	CAC	AAG	GGT
C. e. elaphus	AAT	GTA	GGC	CAG	GGT
C. e. bactrianus	AAT	GTA	GGC	CAG	GGT
E. davidianus	AAT	GTA	GGC	CAG	GGT
D. dama	AAT	GTA	GGC	CAA	GGT
M. reevesi	AAC	GTA	GGC	CAT	AGT

Fig. S5

Codons specifying sickling-associated amino acids in HBB_A genes of different deer species. Tip labels and associated codons coloured as in Fig. 3a

a

Recombination event	Sequence(s) with recombination signal	Recombinant source	Breakpoints*		Significance of detection by method						
			Start	End	RDP	GENECONV	BootScan	Maxchi	Chimaera	SiScan	3Seq
1	*H. inermis* (A)	Foetal	172	344	4.25E-10	2.65E-09	2.69E-10	5.26E-04	1.05E-03	1.40E-04	2.12E-07
2	*C. canadensis* (A)	Foetal	308	502	3.96E-06	3.01E-05	3.17E-06	2.60E-03	4.27E-06	1.64E-04	4.29E-06
3	*C. capreolus* (F)	Adult	344	591	4.29E-04	NS	2.04E-03	8.53E-03	1.55E-03	NS	1.34E-03
4	*D. dama* (A)	Foetal	420	580	NS	NS	NS	1.99E-03	NS	NS	NS
5	*R. tarandus* (A)	Foetal	344	534	NS	NS	NS	5.29E-03	NS	NS	NS
6	*R. tarandus* (F) *O. virginianus* (F)	Adult	355	684	NS	NS	NS	1.19E-02	6.44E-03	NS	NS
7	*R. duvaucelii* (A)	Foetal	96	170	NS	1.24E-02	NS	NS	NS	NS	NS
8	*O. virginianus* (A) *P. puda* (A)	Foetal	772	894	NS	NS	NS	2.21E-02	NS	NS	NS
9	*P. puda* (A)	Foetal	317	344	NS	NS	1.44E-02	NS	NS	NS	NS
10	*C. canadensis* (A)	Adult	1336	1466	NS	NS	4.64E-02	NS	NS	NS	NS
11	*O. virginianus* (F) *R. tarandus* (F)	Adult	156	304	NS	NS	4.84E-02	NS	NS	NS	NS

*due to high local conservation, the exact breakpoint position can be uncertain. F: HBB$_F$; A: HBB$_A$; NS: Not significant.

b

RDP option tab	RDP options different from default			
	Window size	Step size	Model	Variable sites per window
BootScan*	20	5	JN90	-
SiScan	20	5	-	-
PhylPro	40	-	-	-
VisRD	100	-	-	-
DSS(TOPAL)	40	5	JN90	-
Distance Plots	40	5	JN90	-
MaxChi	-	-	-	50
Chimaera	-	-	-	50

*In addition for BootScan: Relationship measure = UPGMA, Bootstrap replicates = 300

c

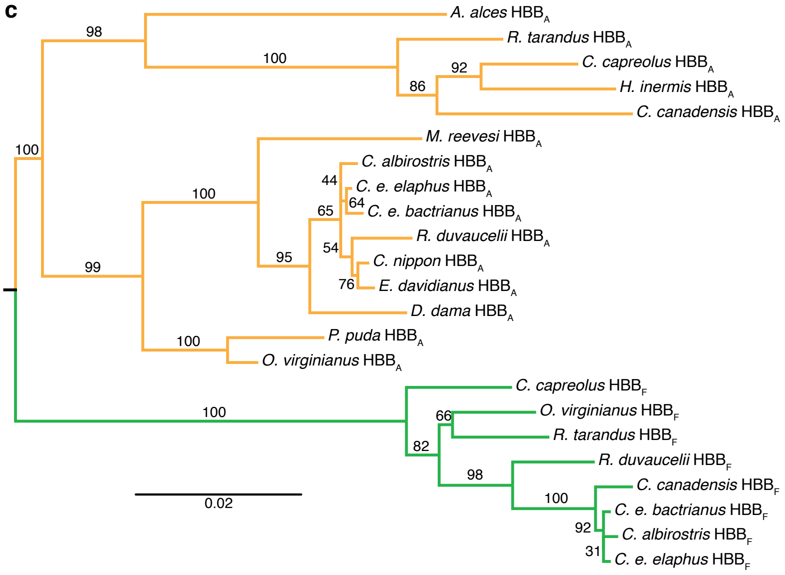


Fig. S6

Detection of gene conversion and introgression events in deer β-globin genes. a, Recombination events predicted from an alignment of deer HBB$_F$ and HBB$_A$ genes by different methods. Breakpoint positions are given relative to each focal sequence. Where two sequences are affected (events 6,8,11) positions refer to the top sequence. **b**, Non-default parameters used for detecting recombination events with RDP. **c**, Maximum likelihood tree derived from the alignment of adult (orange) and foetal (green) β-globin genes after predicted recombinant regions have been removed. Branch support values derived from 100 bootstrap replicates are given.

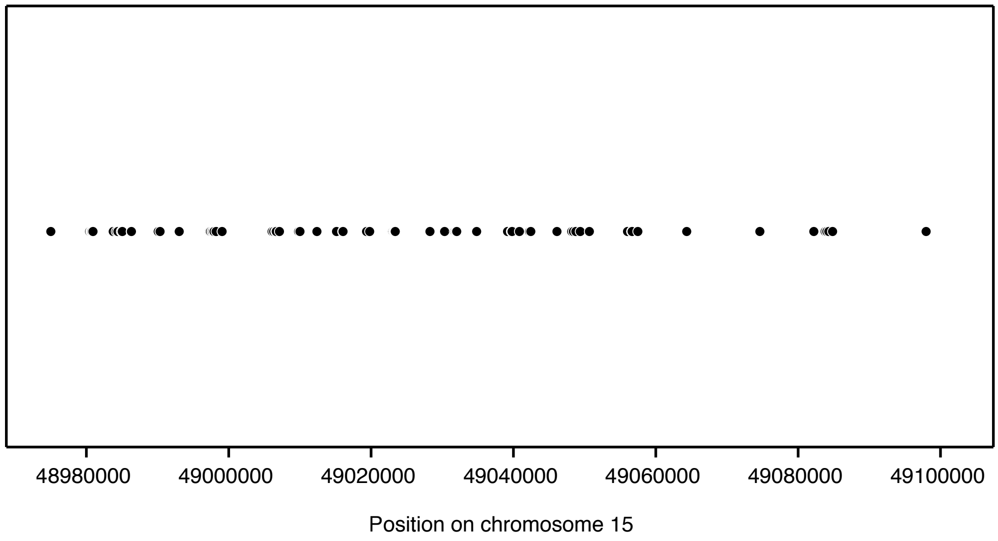
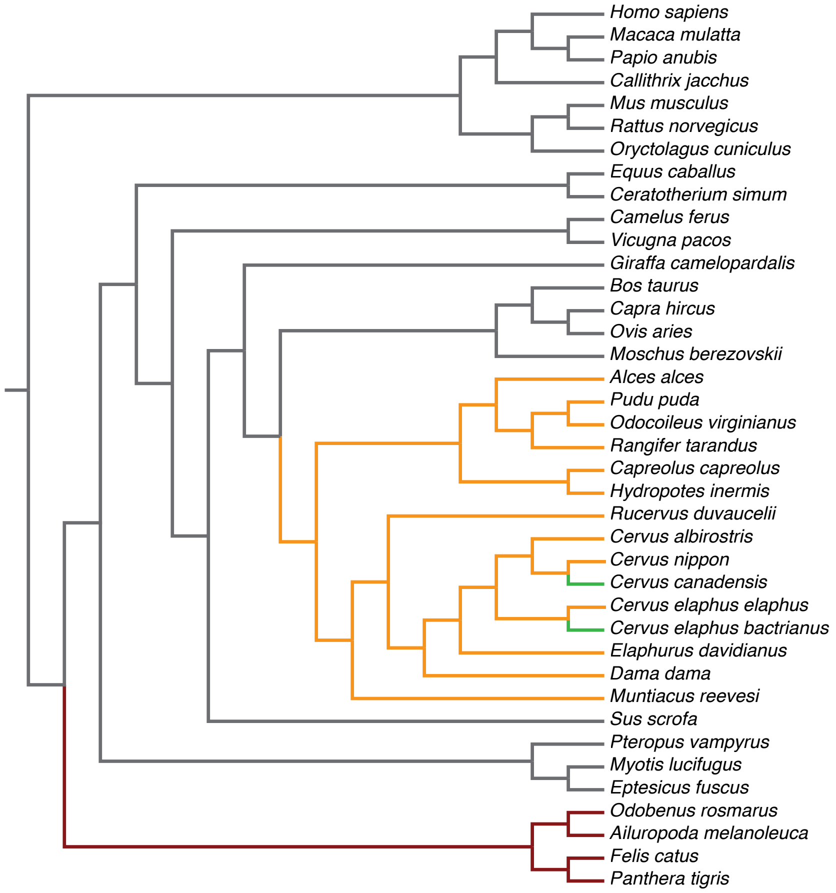


Fig. S7

Mapping locations in the *B. taurus* genome of reads used to seed *O. virginianus* HBB$_A$ local assembly.



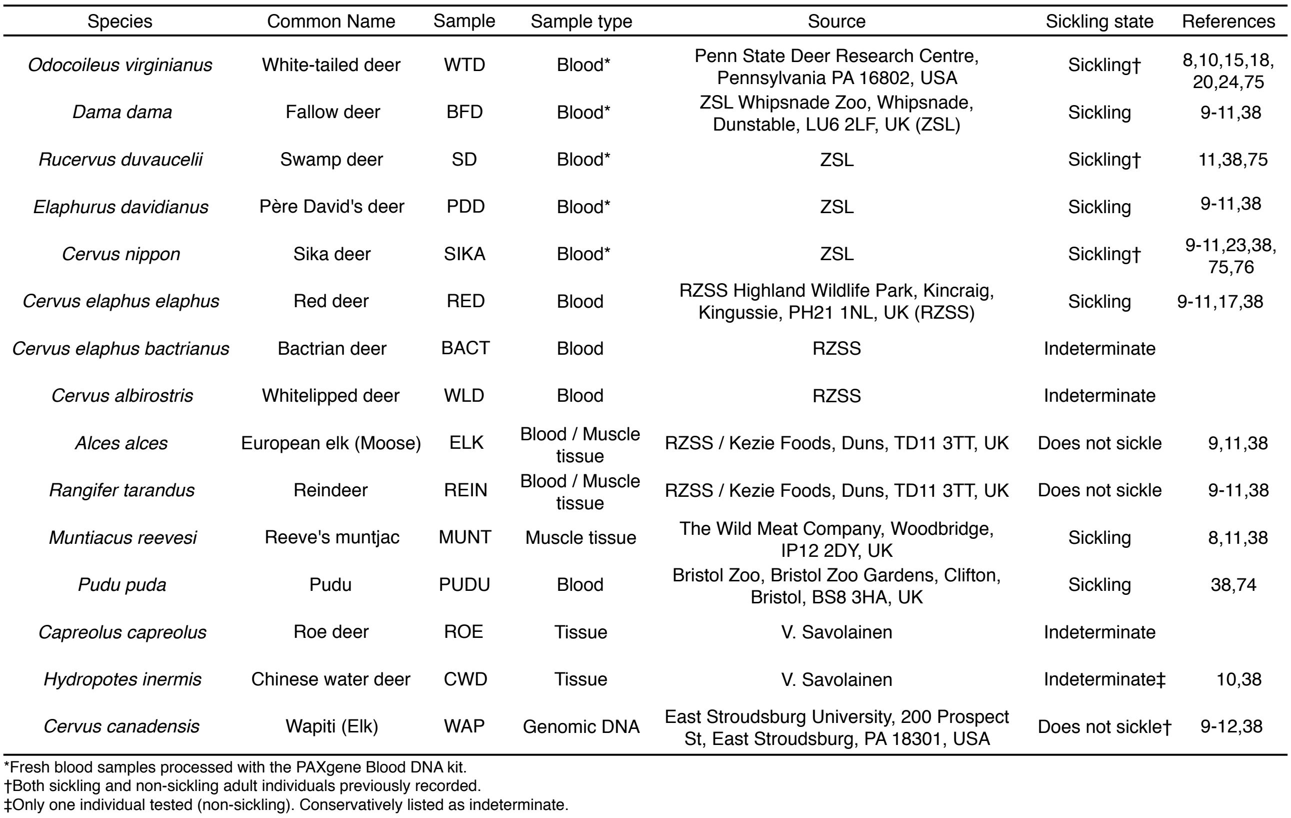
897

898

899 **Fig. S8**

900

901 **Cladogram of the mammalian species phylogeny used in this study**. Coloured
902 branches indicate deviations from and additions to the Timetree of Life phylogeny
903 (see Materials and Methods); red: re-grafted Carnivora; orange: 10kTrees deer
904 phylogeny; green: manually added branches absent from the 10kTrees phylogeny.
905
906



Species	Common Name	Sample	Sample type	Source	Sickling state	References
Odocoileus virginianus	White-tailed deer	WTD	Blood*	Penn State Deer Research Centre, Pennsylvania PA 16802, USA	Sickling†	8,10,15,18, 20,24,75
Dama dama	Fallow deer	BFD	Blood*	ZSL Whipsnade Zoo, Whipsnade, Dunstable, LU6 2LF, UK (ZSL)	Sickling	9-11,38
Rucervus duvaucelii	Swamp deer	SD	Blood*	ZSL	Sickling†	11,38,75
Elaphurus davidianus	Père David's deer	PDD	Blood*	ZSL	Sickling	9-11,38
Cervus nippon	Sika deer	SIKA	Blood*	ZSL	Sickling†	9-11,23,38, 75,76
Cervus elaphus elaphus	Red deer	RED	Blood	RZSS Highland Wildlife Park, Kincraig, Kingussie, PH21 1NL, UK (RZSS)	Sickling	9-11,17,38
Cervus elaphus bactrianus	Bactrian deer	BACT	Blood	RZSS	Indeterminate	
Cervus albirostris	Whitelipped deer	WLD	Blood	RZSS	Indeterminate	
Alces alces	European elk (Moose)	ELK	Blood / Muscle tissue	RZSS / Kezie Foods, Duns, TD11 3TT, UK	Does not sickle	9,11,38
Rangifer tarandus	Reindeer	REIN	Blood / Muscle tissue	RZSS / Kezie Foods, Duns, TD11 3TT, UK	Does not sickle	9-11,38
Muntiacus reevesi	Reeve's muntjac	MUNT	Muscle tissue	The Wild Meat Company, Woodbridge, IP12 2DY, UK	Sickling	8,11,38
Pudu puda	Pudu	PUDU	Blood	Bristol Zoo, Bristol Zoo Gardens, Clifton, Bristol, BS8 3HA, UK	Sickling	38,74
Capreolus capreolus	Roe deer	ROE	Tissue	V. Savolainen	Indeterminate	
Hydropotes inermis	Chinese water deer	CWD	Tissue	V. Savolainen	Indeterminate‡	10,38
Cervus canadensis	Wapiti (Elk)	WAP	Genomic DNA	East Stroudsburg University, 200 Prospect St, East Stroudsburg, PA 18301, USA	Does not sickle†	9-12,38

*Fresh blood samples processed with the PAXgene Blood DNA kit.
†Both sickling and non-sickling adult individuals previously recorded.
‡Only one individual tested (non-sickling). Conservatively listed as indeterminate.

Table S1.

Species considered in this study, previous evidence for sickling and sample origins. For each sample, species identity was confirmed by sequencing the mitochondrial *CytB* gene (see Table S2). For species in which both sickling and non-sickling individuals have been previously identified, the more common phenotype (as found in the associated references) is listed.