Demography and adaptation promoting evolutionary transitions in a mammalian genus that diversified during the Pleistocene

Menno Jong, Zhipeng Li, Yanli Qin, Erwan Quéméré, Karis Baker, Wen Wang, A. Rus Hoelzel

Received: 12 September 2019 | Revised: 25 February 2020 | Accepted: 3 April 2020

DOI: 10.1111/mec.15450

FROM THE COVER

MOLECULAR ECOLOGY WILEY

Demography and adaptation promoting evolutionary transitions in a mammalian genus that diversified during the Pleistocene

Menno J. de Jong[1] | Zhipeng Li[2] | Yanli Qin[3] | Erwan Quéméré[4,5] | Karis Baker[1] | Wen Wang[3] | A. Rus Hoelzel[1]

[1]Molecular Ecology Group, Department of Biosciences, Durham University, Durham, UK

[2]Department of Special Animal nutrition and Feed Science, Institute of Special Animal and Plant Sciences, Chinese Academy of Agricultural Sciences, Changchun City, China

[3]Center for Ecological and Environmental Sciences, Northwestern Polytechnical University, Xi'an, China

[4]Comportement et Ecologie de la Faune Sauvage (CEFS), INRA, Université de Toulouse, Castanet-Tolosan, France

[5]Ecology and Ecosystems Health, Ouest, INRAE, Rennes, France

Correspondence
A. Rus Hoelzel, Molecular Ecology Group, Department of Biosciences, Durham University, South Road, Durham DH1 3LE, UK.
Email: a.r.hoelzel@dur.ac.uk

Wen Wang, Center for Ecological and Environmental Sciences, Northwestern Polytechnical University, Xi'an, China.
Email: wwang@mail.kiz.ac.cn

Funding information
Whitehead Trust; British Deer Society, Grant/Award Number: 201811; Northwestern Polytechnical University

Abstract

Species that evolved in temperate regions during the Pleistocene experienced periods of extreme climatic transitions. Consequent population fragmentation and dynamics had the potential to generate small, isolated populations where the influence of genetic drift would be expected to be strong. We use comparative genomics to assess the evolutionary influence of historical demographics and natural selection through a series of transitions associated with the formation of the genus *Capreolus*, speciation within this genus during the Quaternary and during divergence among European roe deer (*C. capreolus*) populations. Our analyses were facilitated by the generation of a new high-coverage reference genome for the Siberian roe deer (*C. pygargus*). We find progressive reductions in effective population size (N_e), despite very large census sizes in modern *C. capreolus* populations and show that low N_e has impacted the *C. capreolus* genome, reducing diversity and increasing linkage disequilibrium. Even so, we find evidence for natural selection shared among *C. capreolus* populations, including a historically documented founder population that has been through a severe bottleneck. During each phylogenetic transition there is evidence for selection (from dN/dS and nucleotide diversity tests), including at loci associated with diapause (delayed embryonic development), a phenotype restricted to this genus among the even-toed ungulates. Together these data allow us to assess expectations for the origin and diversification of a mammalian genus during a period of extreme environmental change.

KEYWORDS

adaptation, demography, genetic drift, population genomics, roe deer

Menno de Jong, Zhipeng Li and Yanli Qin made equal contributions.

1 | INTRODUCTION

Much of our understanding of evolution from empirical studies is based on the interpretation of neutral genetic diversity and the effects of demographic processes involving population expansions, contractions, divisions and connectivity. We have usefully interpreted such data to advance our understanding of evolutionary process in the context of environmental change, and to promote more effective strategies for the conservation of biodiversity. At the same time, it has been difficult to fully address the key objective of conserving adaptive potential beyond an expectation that greater diversity at neutral loci will indirectly reflect the rest of the genome and therefore mean greater potential for adaptive change (see Hoelzel, Bruford, & Fleischer, 2019). Genomic methodologies widen our opportunities, but it remains difficult to interpret the importance of apparent adaptive change, and to understand the relative importance of genetic drift and natural selection during the evolution of a species. Evolutionary histories are likely to be affected by the interaction of three forces: changing environments, stochastic processes affecting the fate of alleles, and natural selection. In this study we gain insight into this interaction by considering the history of a genus for which the dominant environmental impact was governed by Quaternary environmental change, and by using genomic data to consider both adaptive and stochastic processes through a series of evolutionary transitions.

We chose the genus *Capreolus* because it evolved relatively recently, it speciated during the Pleistocene, and demographic fluctuations have been prominent for species in this genus throughout the Quaternary, including both natural and anthropogenic founder events (Baker & Hoelzel, 2014). The genus *Capreolus* consists of two extant sister species within the subfamily Capreolinea. Although Capreolinea are also known as New World deer, the European (*Capreolus capreolus*) and the Siberian roe deer (*Capreolus pygargus*) are distributed parapatrically (Matosiuk et al., 2014) in the temperate zone of Eurasia (Figure S1). They are phenotypically very similar and are difficult to distinguish based on morphology alone (Plakhina, Zvychanaya, Kholodova, & Danilkin, 2014). Mitochondrial DNA (mtDNA) studies show reciprocally monophyletic divergence between these species (Randi, Pierpaoli, & Danilkin, 1998), although they may hybridize where their ranges overlap (Matosiuk et al., 2014;Plakhina et al., 2014). Phylogenetic evidence indicates a split from the sister lineage *Hydropotes* (Chinese and Korean water deer) 7–11 million years ago (Ma; Mennecart et al., 2017). The earliest known fossil record for the genus *Capreolus* in Eurasia is from the late Miocene in Udunga, eastern Russia (*Capreolus constantini*), and from Europe in Germany 1.02 Ma (*Capreolus cusanoides*; Croitor, 2014). *Capreolus capreolus* is known in Europe from at least 600 thousand years ago (Lister, Grubb, & Sumner, 1998). Data from mtDNA suggested a split between *C. capreolus* and *C. pygargus* 2–4 Ma (Randi et al., 1998). As for other deer species in Europe, there is evidence for a retreat into southern refugia during the last glacial period (Sommer, Fahlke, Schmolcke, Benecke, & Zachos, 2009) and a subsequent re-expansion during the Holocene. Two subspecies have been proposed for the Siberian species (*C. p. tianschanicus* and *C. p. pygargus*; Lee, 2016).

We track the demographic history of the two species in this genus and consider the effects of both drift and selection during a series of evolutionary transitions. We begin with an investigation of the divergence of the genus *Capreolus* from other cervid lineages, followed by the divergence of the two species within the genus (using genomic data), and then focus on the differentiation of populations of *C. capreolus* in Europe (based on RADseq [restriction site associated DNA sequencing] data). We look for loci under selection that distinguish *Capreolus* spp. from other Old and New World deer lineages, including loci that may be associated with distinguishing characteristics such as obligatory embryonic diapause, found only in *Capreolus* among the Artiodactyla (Beyes, Nause, Bleyer, Kaup, & Neumann, 2017). We consider genomic differences between the two *Capreolus* species that may reflect either genetic drift or selection (or both). At the population level we investigate populations of known and empirically assessed demographic histories while looking for signals of selection. Because the potential for selection to drive evolution despite strong drift in small populations remains poorly understood, we include an anthropogenic founder population (a 19th century introduction in East Anglia, UK) where we know the timing and size of the founder group, and that the population has remained mostly isolated (Baker & Hoelzel, 2014). A comparative analysis with natural roe deer populations allows us to consider the shared or differential impact of drift and selection on this founder population. Our broader objective is to better understand the interaction between environment, selection and drift during the evolutionary history of a lineage.

2 | MATERIALS AND METHODS

2.1 | DNA extraction through genome sequencing for *C. pygargus*

Liver tissue was collected from the Er He wild animal farm at Shulan, Jilin city, Jilin province, in northeastern China (126.58°N, 43.855°E) from a 7-month-old male roe deer (*Capreolus pygargus tianschanicus*), from which a high-molecular-weight DNA sample (≥50 kb) was isolated and DNA size was selected for sample improvement following the 10X Genomics HWM DNA sample preparation protocol (https://support.10xgenomics.com/genome-exome/sample-prep). The source population was large enough that inbreeding should not be an important factor and although we cannot rule out admixture, we have no reason to suspect that this is an issue. All animal-specific procedures were approved and authorized by the Chinese Academy of Agricultural Sciences Animal Care and Use Committee, and the Institute of Special Animal and Plant Sciences Wild Animal and Plant Subcommittee.

A library with a 600-bp insert-size was constructed by applying the 10X Genomic Chromium system. The Chromium Controller instrument (10X Genomics Inc.) was used for barcoding long stretches of DNA in Genome Gel Beads. Subsequently, these DNA fragments were reduced to smaller sizes for short read sequencing on an Illumina Hiseq platform (Illumina), generating 2 × 150

paired-end sequences. Sequencing was performed by Novagene. After removing low-quality reads, 285.67 Gb clean sequences were retained for further genome analyses. Quality control was undertaken using the TRIMMOMATIC program with default parameters (java -jar trimmomatic-0.36.jar PE ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50).

SUPERNOVA software (version 1.2.2) was applied to construct the scaffolds using the barcode and reads data generated from 10X Genomic libraries. We used the following parameters: supernova run --maxread = 1,500,000,000; supernova mkoutput --style = pseudohap2 --minsize = 500 (other parameters were set as defaults). To verify the quality of the assembly, we used Benchmarking Universal Single-Copy Orthologs (BUSCO; version 2.0; RRID:SCR_015008; Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) software to assess the genome completeness. Our assembly covered 94% of the core genes, with 3,858 genes being complete (single: 91.9%, duplicated: 2.1%, fragmented: 2.9%, missing: 3.1%; for further detail see Table S1).

2.2 | Acquisition of raw reads and genome assembly of *C. capreolus*

The raw reads and genome assembly of *C. capreolus* were generated for a previous study (Kropatsch et al., 2013), and kindly provided to us by the authors. The authors collected a blood sample from a male roe deer from Hohenstein-Born (Germany; 50′09″N, 8′05″E) and prepared this sample for paired-end sequencing on an Illumina 1.9 platform. Full details of the sequencing protocol are described in Kropatsch et al., 2013. Of the 422,979,622 + 422,818,638 *C. capreolus* forward and reverse reads, we dropped respectively 142,876 and 124,768 reads (dropping the entire read) which were contaminated with adapter sequences, retaining 422,836,746 forward and 422,693,870 reverse reads. We used the software TRIMMOMATIC (Bolger, Lohse, & Usadel, 2014) to discard all *C. capreolus* reads with an average PHRED33-quality score below 20, retaining 412,716,619 read pairs. All reads were trimmed to a length of 101 bp. Average read depth was ~32× after quality control and trimming.

2.3 | Genome-wide genetic diversity

We used BOWTIE (Langmead & Salzberg, 2012) version 2.2.5 to map the retained sequence reads of both species to their reference genomes. We used SAMTOOLS (Harris, 2007;Li, 2011) version 1.3.3 to filter out reads with a mapping quality below 20 and applied the command "grep –v 'XS:i'" to filter out reads which mapped to multiple locations, retaining 676,985,798 *C. pygargus* and 522,854,805 *C. capreolus* reads. We subsequently called single nucleotide polymorphisms (SNPs) using SAMTOOLS, BCFTOOLS-1.9 (Narasimhan et al., 2016) and VCFTOOLS-0.1.13 (Danecek et al., 2011), and used linux command line tools to count on a per-contig basis the number of heterozygous sites, the total number of sites with genotype information and the spacing between adjacent heterozygous sites. Average read depths after filtering, calculated using the SAMTOOLS depth tool, equalled 22.1 for *C. capreolus* and 39.7 for *C. pygargus* (see the Supporting methods [p. 1] for details about potential impact on error rate).

2.4 | PSMC analyses

Prior to Pairwise Sequentially Markovian Coalescent (PSMC) analyses (Li & Durbin, 2011), we down-sampled *C. pygargus* data to match the depth available for *C. capreolus* (but also include the PSMC illustration for the *C. pygargus* genome for data that was not down-sampled, Fig. S4). We generated diploid fasta files of both genomes by mapping the reads of each species to their own reference genomes, and by subsequently using the vcfutils.pl executable of BCFTOOLS. We feel that due to variation in genome quality, this is the appropriate approach, but also show the results when *C. capreolus* is mapped against *C. pygargus* (Fig. S4). This demonstrates the point, showing an improbable increase in population size prior to the likely time of splitting between the two species. We used the default settings of 64 time intervals defined by 28 parameters (-p "4+25*2+4+6"), and a maximum number of 25 iterations (-N25). We also used the default settings of -t15 and -r5. For *C. capreolus* we set the minimum read depth (*d*) to 7 and the maximum read depth (*D*) to 42; for *C. pygargus* we set *d* to 13 and *D* to 78. We set the generation time parameter to 4–6 years (Nilsen et al., 2009). We assumed a mammalian mutation rate per site per year (Kumar & Subramanian, 2002) of 0.22×10^{-8} and 1.1×10^{-8} per generation assuming a generation time of 5 years and a linear relationship. Although we cannot know the correct mutation rate precisely, this estimate is consistent with a broader survey of rates (Lynch, 2010), and we illustrate the effect of varying the mutation rate (Figure S4). Bootstrapping was undertaken with 100 replicates. For comparison we generated a PSMC analysis of red deer (*Cervus elaphus*) using GenBank data for the genome published by Bana et al. (2017), using the same parameters as for the *Capreolus* spp. analyses, but changing the read depth filters to a minimum of 16 and a maximum of 100, consistent with the average higher read depth of the red deer assembly. Although assembly quality is different for the compared genomes, Patton et al. (2019) have shown that PSMC demographic reconstructions are robust to these differences, generally consistent with our further analyses (see Supporting Information).

2.5 | Identification of genomic regions with low genetic variation

To estimate the spacing between heterozygous sites, which we define here as "runs of homozygosity" (ROHs), we calculated the distance between adjacent SNPs. Note that this definition of ROH deviates from other definitions found in the literature. Because of the low contig sizes for the *C. capreolus* genome, we used the *C. pygargus*

genome as the reference genome for both species (i.e., we mapped the sequencing reads of both species to the *C. pygargus* genome, as described in Sections 2.3 and 2.4), assuming highly conserved synteny. The assumption of synteny among cervids was verified by visual inspection of dot plots, which we generated by mapping all *C. pygargus* contigs of 10 Mb or longer to *Cervus elaphus* chromosomes using LASTZ (Harris, 2007) version 1.02.00. We used a sliding window approach to screen the *C. capreolus* and *C. pygargus* genomes for regions with above average spacing between heterozygous sites (i.e., genomic regions with depleted genetic variation), using the rollapply function of the R-package "zoo" version 1.8.6 (Zeileis & Grothendieck, 2005). We set both stepsize and windowsize to 100 datapoints (i.e., 100 adjacent ROHs). A window was marked as an outlier window based on two criteria: (a) the window should contain at least one ROH in the top 0.05% (= 5%/windowsize) of all ROHs; and (b) the window should contain at least *n* ROHs in the top 5% of all ROHs located on the respective contig, with *n* averaging 15, the exact number being dependent on the number of windows (nwin) on the contig, as described by the R command: n = qpois((1−0.05/nwin), (windowsize/20)). We excluded from our analysis all ROHs with 10% or more missing data points (i.e., all sites of 1 bp or longer with a read depth below 8). ROHs which were truncated at the start or end of a contig but otherwise not exceeding our limit for missing data were not excluded from the analyses. Contigs with a size below 5 Mb were not considered, retaining 164 contigs with a median and mean length of respectively 8.0 and 9.3 Mb (*SD* = 4.4 Mb) and a combined length of 1,519.4 Mb, spanning roughly half of the *C. pygargus* genome.

2.6 | Genome-wide genetic divergence

We calculated the genetic divergence between *Capreolus* species by mapping their raw reads to the reference genome of either sister species using BOWTIE2, and by subsequently calling SNPs and indels using SAMTOOLS, BCFTOOLS and VCFTOOLS. We filtered out reads with a mapping quality below 20 and applied the command "grep −v 'XS:i'" to filter out reads which mapped to more than one location, as well as sites with a read depth below 8. We counted the total number of sites and single nucleotide variations (SNVs) on a per-contig basis using BASH command line tools. We calculated the time to the most recent common ancestor (TMRCA) by constructing and applying a random walk Markov chain model in which we simulated the probability of the occurrence of single nucleotide differences between two populations after these populations split from a shared ancestral population (see Supporting Methods, p. 2). To further consider divergence time, we used an alternative method whereby the third codon positions of exomes were compared between the two species. From a two-way alignment of *C. pygargus* versus *C. capreolus* exomes we selected third codon positions, and subsequently removed all sites which had missing data in one or both species. We counted the total length of the retained alignment, as well the number of sites which differed between the species. All steps were performed using linux command line tools (i.e., grep, awk, tr and sed).

2.7 | *Capreolus pygargus* genome annotation

First, we used LTR _ FINDER (Xu & Wang, 2007; RRID:SCR_015247) and REPEATMODELLER (Smit & Hubley, 2018; version 1.0.4; REPEATMODELER, RRID:SCR_015027) to find repeats. REPEATMASKER (Tarailo-Graovac & Chen, 2009; version 4.0.5, with parameters -nolow -no_is -norna -parallel 1) was then used to search for known and novel transposable elements (TEs) by mapping against the de novo repeat library and Repbase TE library (version 16.02). Subsequently, tandem repeats were annotated using TANDEM REPEAT FINDER (version 4.07b; 53; with parameters 2 7 7 80 10 50 2000 -d -h). In addition, we used REPEATPROTEINMASK software (Tarailo-Graovac & Chen, 2009; with parameters -no LowSimple -*p* value .0001) to identify TE-relevant proteins. In total, the repeat sequences cover about 1.16 Gb and account for 44.6% of the genome. The remaining sequences were annotated using both de novo and homology-based gene prediction approaches. For de novo gene prediction, we used SNAP (Korf, 2004; version 2006-07-28), GENSCAN (Burge & Karlin, 1997; RRID:SCR_012902), GLIMMERHMM (Majoros, Pertea, & Salzberg, 2004; RRID:SCR_002654), and AUGUSTUS (Stanke et al., 2006; version 2.5.5; Augustus: Gene Prediction, RRID:SCR_008417) to analyse the repeat-masked genome. For homology-based predictions, sequences encoding homologous proteins of *Bos taurus* (Ensembl 87 release), *Ovis aries* (Ensembl 87 release) and *Homo sapiens* (Ensembl 87 release) were aligned to the genome using TBLASTN (Birney, Clamp, & Durbin, 2004; version 2.2.26; TBLASTN, RRID:SCR_011822) with an E-value cutoff of 1 e−5. GENWISE (version wise 2.2.0) was then used to annotate structures of the genes. To aid in the subsequent annotation of the genome, we used transcript sequences (kindly provided by Dr Wang from Wang et al., 2019) to predict gene structure and update the EVIDENCEMODELER (EVM) consensus predictions, adding untranslated region annotations and models for alternatively spliced isoforms by using PASA software (Haas, 2003; version 2.1.0; RRID:SCR_014656). Thereafter, we used EVM software (Haas et al., 2003; version 1.1.1; EVIDENCEMODELER, RRID:SCR_014659) to merge the de novo and homology gene sets and form a comprehensive, nonredundant gene set, resulting in 21,777 protein-coding genes. To infer gene functions, we searched the GO, KEGG, TrEMBL and SwissProt databases for best matches to the protein sequences from the previous step by using BLASTP (version 2.2.26) with an E-value cutoff of 1 e−5, and searched the Pfam, PRINTS, ProDom and SMART databases to identify known motifs and domains using INTERPROSCAN software (Jones et al., 2014; version 5.18-57.0; INTERPROSCAN, RRID:SCR_005829).

2.8 | dN/dS tests

We blasted the exons of the 21,777 *C. pygargus* genes (excluding untranslated regions) to the *C. capreolus* genome as well as to *Bos taurus* (Elsik, 2009; GCA_002263795.2) and other published cervid genomes, namely: *Rangifer tarandus* (Li et al., 2017; PRJNA391754), *Odocoileus virginianus* (Seabury et al., 2011; GCA_002102435.1), *Elaphurus*

davidianus (Zhang et al., 2018; GCA_002443075.1), *Cervus elaphus* (Bana et al., 2018; GCA_002197005.1), *Cervus albirostris* (Chen et al., 2019; GCA_006408465.1), *Hydropotes inermus* (Chen et al., 2019; GCA_006459105.1), *Odocoileus hemonius* (mule deer), *Muntiacus muntjak* (Chen et al., 2019; GCA_006409035.1), *Muntiacus crinifrons* (Chen et al., 2019; GCA_006408485.1), *Hyelaphus porcinus* (Chen et al., 2019) and *Muntiacus reevesi* (Chen et al., 2019; GCA_006408525.1).

We used BEDTOOLS (Quinlan & Hall, 2010) version 2.19.1 GETFASTAFROMBED to extract the BLAST hits from the reference genomes, and subsequently MUSCLE (Edgar, 2004) version 3.8.31 for multiple alignments. For each gene we selected the hits with the highest bitscores, giving preference to exons residing on the same contig or chromosome. The credibility of the gene alignments was verified by visual inspection of a subsample of the genes, as well as by using the concatenated alignments to generate a maximum likelihood species tree with 100 bootstrap replicates using the software RAXML (Stamatakis, 2014) with *Bos taurus* as the outgroup (Elsik et al., 2009; GCA_002263795.2). Genes were partitioned for first and second codon positions versus third codon positions. (The actual RAXML command: raxml -fa -m GTRGAMMA -p 12345 -o B_tau -x 12345 -# 100 -s input.phy -n bootstrap100.)

Prior to dN/dS analyses we filtered out alignments that due to indels were evidently out of the reading frame (by removing all alignments for which lengths were not multiples of 3), retaining 14,512 genes. We calculated gene-specific dN/dS rates using PAML4.4 (Yang, 2007) yn00. For each gene, a species was excluded from the analysis if the gene contained a stop codon or at least 50% missing data. We used PAML's CodeML to test for evidence of positive selection by comparing for each gene the performance of two branch-site models: an alternative model allowing for purifying, neutral and positive selection ($\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$), with the corresponding null model only allowing for purifying and neutral selection ($\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 = 1$), by setting model = 2 and Nssites = 2 for both models, and fix_omega = 1 and omega = 1 when running the null model. Significance was evaluated using the Chi-square test implemented in R, applying a Bonferonni correction for multiple testing. In contrast to the procedure used for the dN/dS (ratio of nonsynonomous and synonomous substitution rates) analyses, we did not filter the gene sets prior to CODEML analyses. Instead, we tested all genes, and afterwards filtered the outlier genes based on visual examination of their alignments. Genes were filtered based on levels of missing data, the presence of paralogues and signs of misalignment.

The relative magnitude of the dN/dS ratio for comparisons outside compared to within the genus *Capreolus* was assessed as a fold value (B/A) where A = comparison between *C. pygargus* and *C. capreolus*, and B = the average of comparisons between *C. pygargus* and the included non-*Capreolus* species), presented in Table S2.

Evidence for selection at a "candidate" locus (locus with a relevant putative function based on earlier studies; see Results), *pard3* (see Boroviak et al., 2015; Table S2), was further assessed based on amino acid properties using the program TREESAAP (Woolley, Johnson, Smith, Crandall, & McClellan, 2003). We included all test properties available in the program in the initial run and tested screening window sizes of 20 and 100 codons. Alignments for illustration were generated in GENEIOUS, which was also used to generate a neighbour joining phylogeny based on 100 bootstrap replicates and the Tamura–Nei model for locus ZDHHC23 (g09868).

2.9 | Gene ontology analysis

We searched for gene ontology (GO) terms using FATIGO (Al-Shahrour, Diaz-Uriarte, & Dopazo, 2004) on the Babelomics 5 platform (http://babelomics.bioinfo.cipf.es/). GO term functions were identified with reference to the human database (chosen due to the relatively complete level of annotation and functional analysis, and the lack of a suitable reference from a closely related species). A Fisher exact test for 2 × 2 contingency tables (testing for over-representation in list 1) was used to check for significance. Significance was corrected for multiple testing using the Benjamini and Hochberg's false discovery rate (FDR)-controlling procedure. GO biological processes, molecular function, cellular component, GOSlim GOA, Interpro and Genome-scale metabolic network databases were searched, filtering terms by five to 500 annotated IDs in each database.

2.10 | ddRAD sample selection

We generated SNP data sets for *C. capreolus* populations from four localities: Ayrshire (Scotland), East Anglia (England), Wurttemberg (Germany) and Aurignac (France), sampled over geographical areas of comparable size. The East Anglian roe deer are the descendants of an anthropogenic reintroduction of 12 individuals stocked from the Wurttemberg population in 1884, found to have just one mtDNA control region haplotype (744 bp) among 40 sequenced individuals (Baker & Hoelzel, 2013). The Ayrshire population represents the endemic UK population that became isolated from the roe deer population in mainland Europe around 6000–7000 years ago with the flooding of Doggerland (a land bridge that connected Britain to the continent when the Scandinavian and British ice sheets reached their maximum extent). Samples were collected during culls, or were reused from earlier studies (Baker & Hoelzel, 2013). No animals were killed specifically for this study, and all animals were killed by certified deer stalkers. The laboratory work for the Aurignac population was executed at the CEFS whereas the laboratory work for the other three populations (East Anglia, Ayrshire and Wurttemberg) was undertaken at Durham University. DNA was extracted using Omega-Biotek's E.Z.N.A. tissue DNA kit. A subset of samples was selected for sequencing based on Qubit quantification scores and molecular weight of the DNA assessed by gel electrophoresis.

2.11 | RAD library construction

For the East Anglia, Ayrshire and Wurttemberg data sets we constructed libraries for two sequencing lanes following the

double digest (dd)RADseq protocol (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). Based on in silico simulations with the R package SIMRAD-0.96 (Lepais & Weir, 2014), we used a 6-bp cutter (*Hind*III: AAGCTT) and a 4-bp cutter (*Msp*I: CCGG), with a fragment size selection window of 250 bp width (including all fragments with a length of 275–525 bp, excluding the adapters), targeting 120,000 loci with an average read depth of 30. By multiplying this expected number of loci against their average length (250 bp), a conservative estimate for nucleotide diversity (θ = 1/2,000), and an approximation for the harmonic number of Watterson's estimator (Watterson, 1975), we estimated that this size selection window would yield at maximum ~50,000 SNPs with MAF > 0.05. The actual size selection was executed with a Sage Science PippinPrep machine. The Phusion High-Fidelity kit was used for a 13-cycle PCR (denaturation step: 62°C for 20 s; annealing step: 72°C for 45 s; extension step: 72°C for 5 min). The Aurignac data set was generated independently for a separate study (Gervais et al., 2019) and incorporated opportunistically. It was also produced using the ddRAD strategy (Peterson et al., 2012), and with the same frequent cutting enzyme (*Msp*I) but with a different 6-bp cutting enzyme (*Eco*R1) with size-selected libraries of 270–330 bp. All libraries were paired-end sequenced for 125 bp on an Illumina HiSeq_2500 (version 4 chemistry) machine.

2.12 | SNP calling and filtering

The two sequencing lanes for the East Anglia, Ayrshire and Wurttemberg data sets combined producing 602.6 million reads, which were trimmed to 110 bp and demultiplexed using the software STACKS-1.35 (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011). Over 6.1 million reads had to be discarded due to either low quality, an ambiguous radtag, or a missing read mate, resulting in an average number of 7.3 million read pairs per sample (*SD*: 5.8 million, min.: 0.7 million, max.: 32.4 million). Paired reads were aligned against both the newly generated *C. pygargus* genome as well as the *Cervus elaphus* genome (Bana et al., 2018) using the software BOWTIE version 2.2.5. We chose the red deer genome as a second reference genome because at the time red deer was the species closest to *Capreolus* spp. with a genome assembly up to chromosome level. Overall per-sample alignment rates ranged around 98% and 65% for alignment against *C. pygargus* and *Cervus elaphus* respectively, with over 65% and under 50% of the reads aligning concordantly to one location within the genome. SAMTOOLS version 1.3.3 was used to filter out reads which aligned to more than one location in the genome, which aligned disconcordantly, and reads with a mapping quality below 20. The Aurignac data set was treated in the same way except that the reads were trimmed to 117 bp rather than 110 bp. Trimming settings differ because quality control of the Aurignac data set was executed at CEFS.

SNPs were called using the STACKS refmap pipeline with default settings. Loci for which at least 30% of all individuals had a read depth below 8 were removed. We accepted multiple SNPs per read (i.e., we did not set the –write-single-SNPs flag when running the "populations"-command), as we opted to "thin" our data set downstream. For the Ayrshire, East Anglia and Wurttemberg (AEW) data set aligned to the roe deer genome, our settings resulted in 686,859 loci/stacks, of which 74,518 passed the filters ("sample/population constraints"), consisting of 8,196,980 sites, of which 52,364 (0.64%) were bi-allelic. The bi-allelic SNPs were concentrated on 34,250 loci/stacks. For the Aurignac data set we identified 259,987 loci, of which 50,975 passed the filters ("sample/population constraints"), consisting of 5,607,250 sites, of which 29,488 (0.53%) were biallelic. These bi-allelic SNPs were concentrated on 19,772 loci/stacks. Further details are given in the supporting methods.

PGDSPIDER 2.0.9.0 (Lischer & Excoffier, 2012) and PLINK version 1.90 (Purcell et al., 2007) were used to convert the output from genepop format to a genlight object, implemented in the R package ADEGENET-2.1.1 (Jombart, 2008). All samples with more than 25% missing data were removed, retaining 107 samples (i.e., Ayrshire: 25, Aurignac: 29, East Anglia: 23, Wurttemberg: 30). SNPs with more than 10% missing data were filtered out, as well as SNPs with one copy of the minor allele and SNPs which belonged to the 1% class of SNPs with the highest read depths. We also filtered out a small number of SNPs that mapped to the same location of the reference genome even though they belonged to different STACKS loci, and which therefore probably represent unexpected behaviour of the refmap pipeline.

Prior to population structure analyses, we filtered the data sets on mapping distance using R functions, by selecting at maximum one SNP per 500 bp. We retained 15,802 and 10,401 SNPs for the AEW and Aurignac data sets respectively. For genome-wide genetic variation analyses (i.e., Tajima's D analyses), we used all 27,298 SNPs of the red deer aligned data set.

2.13 | Population genetic analyses

For population-specific analyses (genetic diversity estimates), we used each population data set independently including the full set of SNPs identified for that population. For comparative analyses (Discriminant Analysis of Principal Components [DAPC], principal components analysis [PCA], F_{ST}) that included Aurignac, we used a reduced data set of up to 301 SNPs which were shared between all data sets. In order to achieve this number of SNPs, we relaxed the mapping quality filter threshold from 20 to 3. The robustness of this less stringently filtered data set was evaluated visually by comparing the PCA output of this data set with the more stringently filtered data set (for the three focal populations), which differed only in the extent of diffusion among sample points for the populations common to each analysis (data not shown). Analyses for heat map dissimilarity, site frequency spectra plots, and admixture analyses were executed in R, using either in-house-built functions or functions implemented in the ADEGENET-2.1.1, APE-5.3 (Jombart & Ahmed, 2011;Paradis, Claude, & Strimmer, 2004), STAMPP-1.5.1 (Pembleton, Cogan, & Forster, 2013) and LEA-2.8.0

(Frichot & Francois, 2015) packages. Pairwise population F_{ST} measures were calculated using an in-house built script following Weir & Cockerham (1984). For DAPC (run in ADEGENET) we set the number of PCs to 1/3 the number of individuals, the number of clusters equal to the number of populations, and the number of discriminant functions set to 3. For population structure estimation (run in LEA) we set *K* (putative number of populations) to 2–6, alpha to 10 and tolerance to 0.00001, and number of iterations to 200. The LD (linkage disequilibrium) calculations were carried out using PLINK version 1.90 (-genome --r2 --ld-window-kb 1,000,000 --ld-window -r2 0), and executed on reduced data sets excluding SNPs with a minor allele frequency count below 5 (with genomic positions based on read mapping against *C. pygargus*). Contemporary gene flow was estimated using BAYESASS (Wilson & Rannala, 2003) version 3 with the reduced data set (300 SNPs) to include all four putative populations. The number of iterations was set to 1,000,000, burn-in to 100,000 and delta values to 0.1. Tajima's D analyses were executed using VCFTOOLS' TajimaD function, using nonoverlapping windows of 1.5 Mb. Watterson's theta was calculated using ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014) with a window of 150 kb (to allow for a comparable scale of comparison to Tajima's D on a shared plot).

2.14 | Estimate of sample-specific genome-wide heterozygosity

We used our SNP data set to calculate for each sample an estimate of the genome-wide heterozygosity (i.e., proportion of heterozygous sites). First, we determined "N_seg," the number of segregating sites. We only considered a site to be segregating if it was segregating within the population to which the individual belonged. Second, we calculated "He_seg," the proportion of heterozygous sites within an individual for those segregating sites. Finally, we calculated genome H_E using the formula: genome H_E = (He_seg * N_seg)/N_total, in which N_total equals the combined length of all polymorphic as well as monomorphic loci/stacks which passed our filter settings.

2.15 | Stairway plots

The demographic histories of our populations were inferred using the STAIRWAY _ PLOT _ v2 software (Liu & Fu, 2015). We chose this method due to its frequent use and demonstrated utility for RADseq data and for considering time frames relevant to our study. Similar to our PSMC analyses, we set the mutation rate to 1.1×10^{-8} and the generation time to 5 years. The population-specific folded site frequency spectrum (SFS) vectors were generated with a custom-built script which binned SNPs in classes based on their number of copies of the minor allele, and which subsequently calculated the size of each bin (i.e., number of SNPs within each bin).

3 | RESULTS

We provide a high-quality reference genome for *Capreolus pygargus tianschanicus* that has an assembly length of 2,607,875,777 bp, distributed over 92,100 scaffolds, with a median (N50) scaffold length of 6,607,211 bp and an average depth of 77× (see Table S1), and compare it to the relatively low-quality (N50 = 10,458 bp) published genome of *Capreolus capreolus* (Kropatsch et al., 2013). Using our assembly, we first consider the characteristics that distinguish the genus *Capreolus* from other cervid species. Our annotation identified 21,777 coding loci in *C. pygargus*. We aligned these genes to 12 available cervid genomes as well as a *Bos taurus* genome, and calculated for each gene pairwise dN/dS rates between *C. pygargus* and each of the other species. As expected, the distribution of dN/dS ratios was dominated by evidence for purifying selection (where dN/dS is <1; Figure 1c). The long-term average was 0.31, suggesting a scaled selection coefficient of roughly -1 (see Mugal, Wolf, & Kaj, 2012).

Using branch-site models in CODEML with the Capreolus genus as the foreground branch and *B. taurus* as outgroup, 18 loci were identified as significant. However, some of these were removed from further consideration due to the evident presence of paralogues, stop codons (possible pseudogenes) or extensive missing data. The remaining eight loci are presented in Table S2 together with data for four candidate loci (chosen based on functions relevant to reproduction identified in earlier studies; see notes in Table S2 that describe the relevant functions). Loci identified by CODEML were associated with development and cellular processes including functions associated with reproduction (Table S2). The candidate genes tested were associated with diapause, which is unique to *Capreolus* among the even-toed ungulates (Beyes et al., 2017). However, note that diapause is thought to be conserved in mammals (present in multiple lineages; Ptak et al., 2012) and so differences in gene expression may in general be more important than changes to the sequence of protein-coding genes. Among the candidate loci, *pard*3 had the highest fold difference for nonsynonymous changes comparing the genus *Capreolus* to other deer species (see Section 2 and Table S2).

PARD3 is an adapter protein involved in cell division and polarization that has been shown to have differential expression in pre- and post-implantation epiblasts early in mouse embryogenesis (Boroviak et al., 2015), and so is potentially relevant to delayed implantation (diapause). Among the loci identified by CODEML as under selection, *adgrb1* interacts with *pard3*, and two others have an association with embryogenesis (*arhgap33* and *nlk*), but none have a high fold difference (see Section 2 and Table S2), and none of the candidate loci were identified by CODEML (Table S2). Further analysis of PARD3 for signatures of selection based on amino acid properties (using TREESAAP; Woolley et al., 2003) supported the interpretation that this locus is under selection in *Capreolus* especially with respect to the activity of the C- or N-terminus of the α-helix (Table S3).

We next considered processes associated with divergence between the two *Capreolus* species. Paralogues to ZDHHC23 in particular, an integral membrane protein that enables transferase activity, were identified from our genome comparisons using

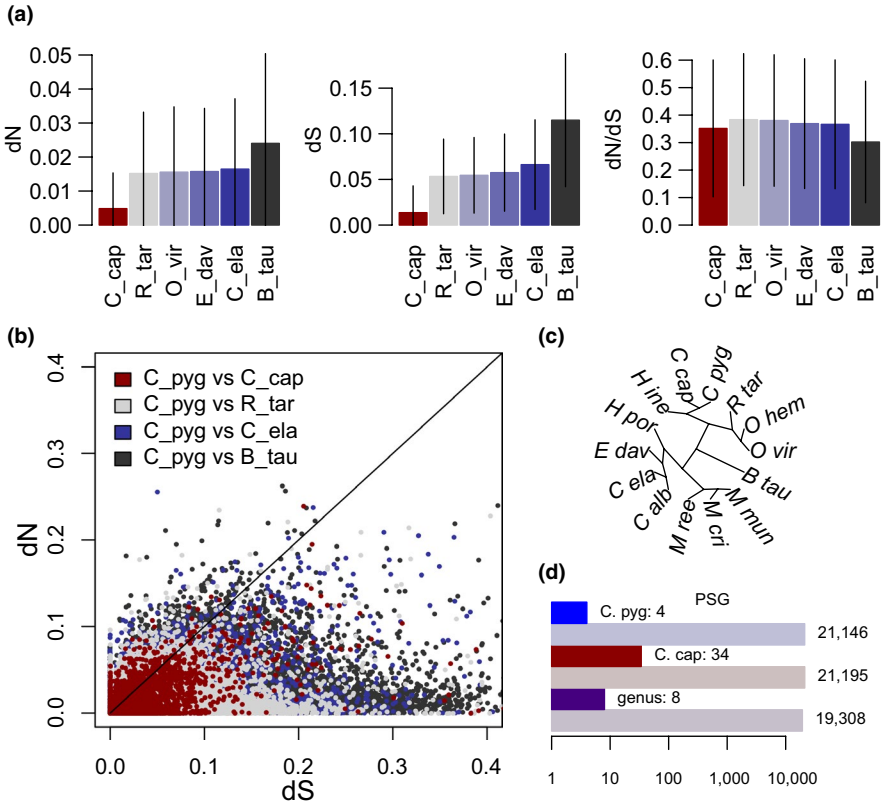

FIGURE 1 dN/dS analyses. B_tau = *Bos taurus* (cattle), C_cap = *Capreolus capreolus* (European roe deer), C_pyg = *Capreolus pygargus* (Siberian roe deer), C_alb = *Cervus albirostris* (Thorold's deer), C_ela = *Cervus elaphus* (red deer), E_dav = *Elaphurus davidianus* (Pere David's deer), H_ine = *Hydropotes inermus* (water deer), H_por = *Hyelaphus porcinus* (Indian hog deer), M_cri = *Muntiacus crinifrons* (black muntjac), M_mun = *Muntiacus muntjak* (common muntjac), M_ree = *Muntiacus reevesi* (Reeves's muntjac), O_hem = *Odocoileus hemonius* (mule deer), O_vir = *Odocoileus virginianus* (white tailed deer), R_tar = *Rangifer tarandus* (reindeer). (a) Barplots showing dN and dS values, calculated using PAML's yn00, for pairwise comparisons between *C. pygargus* and five other cervid species and cattle, for up to 14,512 genes. Bar heights indicate mean gene-specific values. Error bars indicate standard deviation. (b) Scatterplot of dN and dS values, calculated using PAML's yn00, for pairwise comparisons between *C. pygargus* and three other cervid species and cattle. (c) RaxML phylogeny based on full exomes with 100% bootstrap support at all nodes. We generated this phylogenetic tree to verify the gene alignments. (d) Barplots of the number of genes marked by CODEML as neutrally evolving or positively selected genes (PSGs). Light blue/red/purple: neutral genes. Blue: PSGs for *C. pygargus* as the foreground lineage. Red: PSGs for *C. capreolus* as the foreground lineage. Purple: PSGs for the genus *Capreolus* (*C. capreolus* + *C. pygargus*) as the foreground lineage

a CODEML site model. There is a conserved sequence at this locus shared by all species in our alignment with some variation, and divergent paralogues in *C. capreolus*. One of these divergent paralogues shares homology with paralogues found only in Old World deer (Figure 2; Figure S2), suggesting a sympleisiomorphic relationship (cf. exome tree in Figure 1). If the retention of the paralogue is adaptive, it would suggest convergent evolution of *C. capreolus* with the Old World Eurasian cervids, and divergence from the rest of the Capreolinae including *C. pygargus*. An adaptive origin for duplicate coding loci has been suggested for mammalian species based on comparisons between mouse and human genomes, in contrast to other studies that favoured the duplication–degeneration–complementation model (Shiu, Byrnes, Pan, Zhang, & Li, 2006).

CODEML analysis with *C. capreolus* as the foreground found 34 loci under positive selection (after eliminating alignments that were compromised by evident paralogues, pseudogenes or extensive missing data) and GO analysis identified enriched terms associated with membrane function and protein interactions (Table S4).

By cross-mapping the raw reads of the *C. pygargus* and *C. capreolus* genome assemblies we observed a proportion of 0.6%–0.7% SNVs when comparing homologous sequences. Using our divergence time estimation approach based on whole genomes (see Supporting Methods), a generation time of 5 years (Nilsen et al. 2009), and a mutation rate of 1.1×10^{-8} sites per generation (Kumar & Subramanian, 2002) we estimated that the two *Capreolus* species diverged 0.9–1.35 Ma (more recent than estimates using mtDNA; Randi et al., 1998). Based on comparing just third codon positions in the exomes of each species we assumed the same generation time and mutation rate, and derived a TMRCA of 1.2 Ma.

The observed percentage of heterozygous sites in the *C. capreolus* genome (0.14%) is low compared to *C. pygargus* (0.32%; Figure S3), but both are within the range seen for other cervid species (Figure 3e). This difference in genetic diversity is consistent with comparative ROH analyses (Figure S3), and reflected in higher estimates of historical N_e for *C. pygargus* compared to *C. capreolus* as inferred from PSMC analyses (Figure 3a). Although

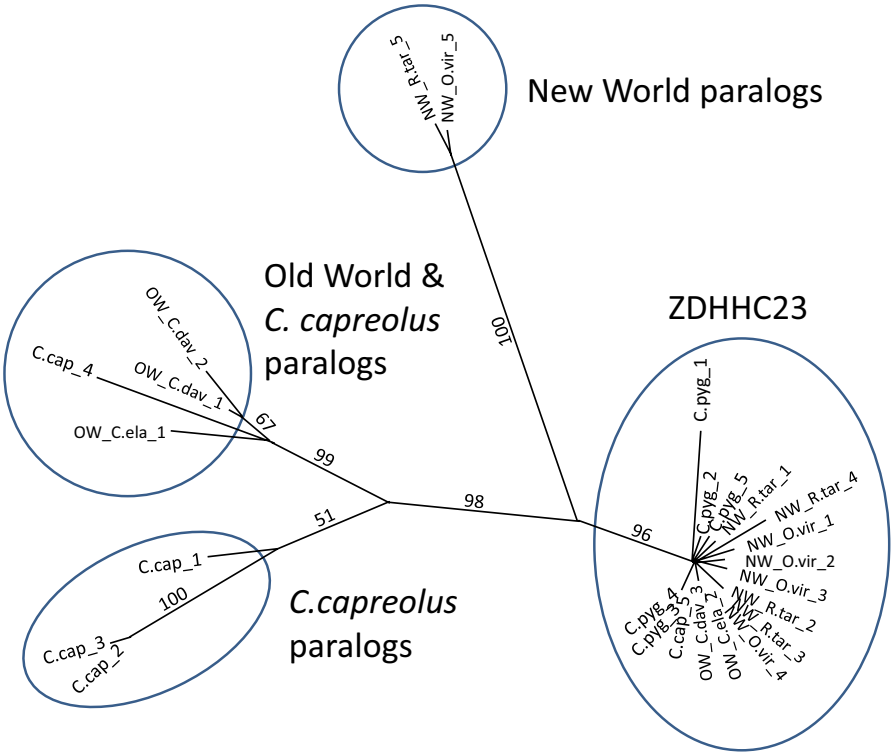

FIGURE 2 Phylogeny of paralogues at ZDHHC23. Unrooted tree of sequences from compared genomes that align closely to ZDHHC23 together with paralogous sequences. OW = Old World: C_cap = *Capreolus capreolus*, C_pyg = *Capreolus pygargus*, E_dav = *Elaphurus davidianus*, C_ela = *Cervus elaphus*; NW = New World: R_tar = *Rangifer tarandus*, O_vir = *Odocoileus virginianus*

differences in genome quality could affect the ROH analyses (Figure S3), all metrics applied consistently show a difference in diversity between the species (e.g., see Figure 3c–e). The N_e of *C. capreolus* has been consistently lower and relatively constant since the origin of the two species, despite modern census numbers in excess of 9 million for *C. capreolus* in Europe (Burbaite, 2009). N_e was low for both species at the apparent time of speciation (Figure 3a), suggesting that drift could have been a factor during the origin of the genus as well as during the subsequent history of each species. Multiple illustrations of the PSMC plots based on different parameters and strategies are presented in Figure S4 (see Section 2).

We then focused on diversity and population structure within *C. capreolus* using the ddRAD (Peterson et al., 2012) genome sampling method. We generated data for populations from four localities (Figure S1). The Ayrshire (Scotland) population represents the surviving native UK population (following the extirpation of roe deer from England and Wales in the Middle Ages; Baker & Hoelzel, 2013). Populations in the UK became isolated from mainland Europe with the flooding of Doggerland, ~5,800 years ago (van der Molen & van Dijck, 2000). The East Anglian (England) roe deer are the descendants of a reintroduction stocked from the Wurttemberg (Germany) population around 1880 (see Baker & Hoelzel, 2013). The inclusion of the East Anglian population serves as an experimental control, and allows us to consider the impact of a bottleneck when the timing and approximate founder size are known.

Stairway plots (Liu & Fu, 2015) track the demographic histories of each population (Figure 4). Population structure can bias inference from these analyses, especially by generating false population declines for the recent period (Heller, Chikhi, & Siegismund, 2013). However, none of these plots show a pronounced recent decline, and earlier simulation analyses supported independent profiles by region for a species (*Equus quagga*) with a similar pattern of population structure (Pedersen et al., 2018; cf. Figure 5). The strong population decline in East Anglia is consistent with the known founder event (although the inferred timing is imprecise for such recent events). A strong recovery is indicated in East Anglia following the bottleneck, but again may not be precise in timing or scale given the reduced power available to resolve recent events. While each profile is distinct, all populations (including East Anglia) show some indication of decline ~8,000–13,000 years ago, coincident with a period of rapid warming following the Younger Dryas climatic event (Lea, Pak, Peterson, & Hughen, 2003). This is most pronounced in Scotland and France, but suggests a shared ancestral event not limited to the founding of regional populations (possibly associated with reduced N_e, or population founding after expansion from glacial refugia). If so, then the stronger signal in Scotland and France may be associated with reduced N_e compared to Germany, probably resulting from fragmentation into regional populations after the last glaciations (Sommer et al., 2009;Templeton, 2008).

Modern populations show genomic signals of diversity that are consistent with these relatively recent demographic histories (Figure 4). For example, there are elevated levels of LD and lower mean measures of overall heterozygosity in all putative populations compared to Germany, and low-frequency alleles are most common in Germany (Figure 5; Figure S5). Population genetic differentiation as assessed using DAPC (Jombart, Devillard, & Balloux, 2010) and PCA is strongest comparing the UK populations to mainland Europe

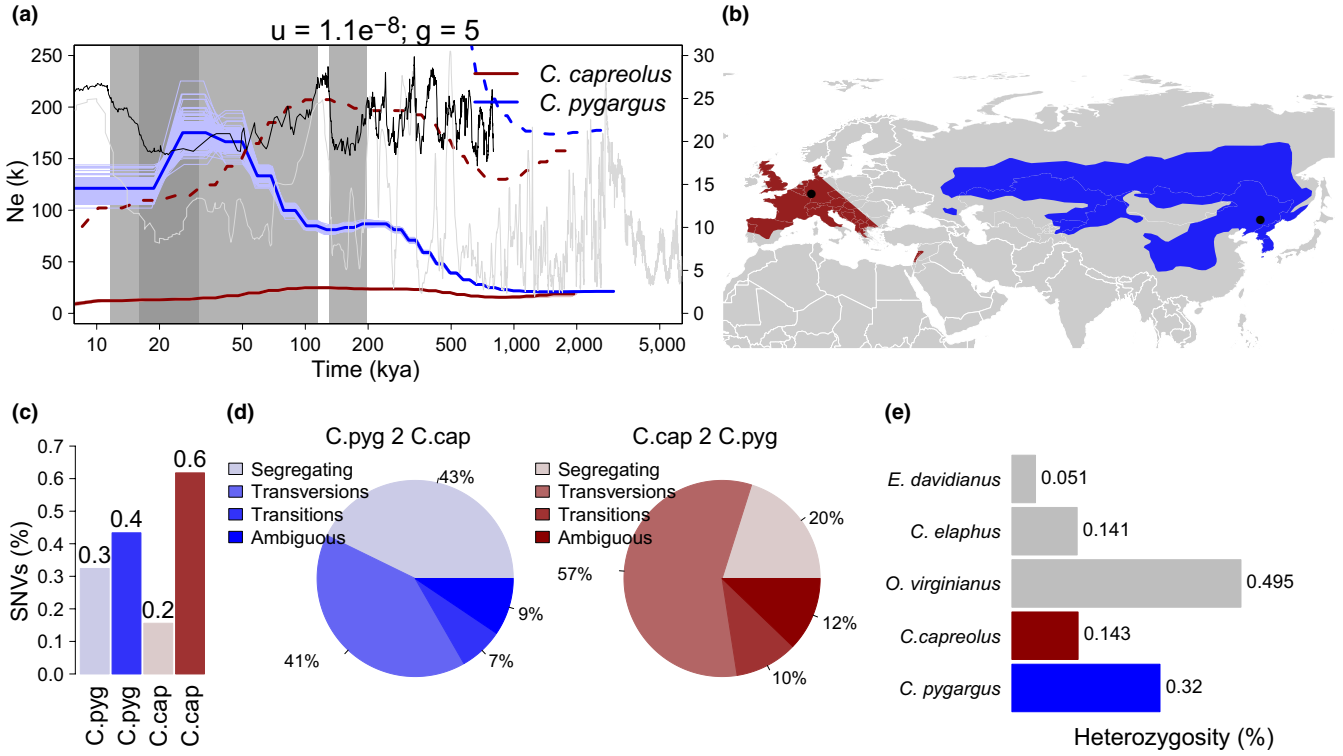

FIGURE 3 Genetic diversity, genetic divergence and demographic history. Red: European roe deer (*Capreolus capreolus*), blue: Siberian roe deer (*Capreolus pygargus*). (a) Changes in effective population size through time estimated from whole genome sequences using PSMC. Dashed coloured lines are relative to the N_e scale depicted on the *y*-axis on the right-hand side of the plot, whereas full lines are relative to the *y*-axis on the left-hand side. Light grey area: Last Glacial Period (11.7–115 thousand years ago [ka]) and Penultimate Glacial Period (130–194 ka). Dark grey area: Last Glacial Maximum (16.3–31 ka). Light grey line: magnetic susceptibility (/100) from Lingtai Loess data (Sun, An, Clemens, Bloemendal, & Vandenberghe, 2010). Black line: atmospheric CO_2/ppm/100 from EPICA Dome C Ice Core 800 ka CO_2 data (Luthi et al., 2008). (b) Geographical distribution of *C. capreolus* and *C. pygargus*. Data from IUCN website. Black dots indicate origins of both whole genome sequences. (c) Barplot of single nucleotide variations (SNVs) in *C. capreolus* relative to *C. pygargus* (right), and conversely *C. pygargus* relative to *C. capreolus* (left). Light blue/red: segregating/heterozygous sites, dark blue/red: fixed sites. (d) Piecharts of composition of SNVs for *C. capreolus* relative to *C. pygargus* (right), and conversely *C. pygargus* relative to *C. capreolus* (left). The categories "transversions," "transitions" and "ambiguous" are subcategories of fixed SNVs, in which "ambiguous" represents sites with multiple alternative alleles calls. (e) Genome-wide heterozygosity (percentage observed number of heterozygous sites of the total length of the genome assembly) of *Capreolus* species compared to other cervids (the latter shown in grey). The estimate for *Elaphurus davidianus* was obtained from Zhu et al. (2018)

(Figure 5c,d). For East Anglia differentiation is probably due to the anthropogenic founder effect, and for Scotland (and France to a lesser extent) due to isolation. F_{ST} values are significant between all four putative populations, but assessment of recent gene flow using BAYESASS indicates gene flow at some level, including between Scotland and East Anglia (Figures S6 and S7, Table S5). A divergence heatmap suggests potential migrants between Scotland and East Anglia (Figure S7) as does a structure analysis implemented in LEA (Figure S8; see Methods). We cannot rule out the possibility that the two samples in Scotland with strong signals for East Anglian origin (possible migrants; Figures S7 and S8) are recent translocations, but we have no supporting evidence for this interpretation. While admixture could theoretically have facilitated the apparently rapid recovery of N_e in East Anglia (Figure 4), diversity and differentiation metrics are more consistent with East Anglia having remained largely isolated (see Figures S5 and S7, Table S5; Baker & Hoelzel, 2014).

We consider evidence for local adaptation in *C. capreolus* by comparing three focal populations (Germany, Scotland and East Anglia sharing the same set of SNPs) each with different demographic histories (Figure 4) and levels of diversity (Figure 5). Summary statistics may show a similar response to selection or demography, but a consistent effect among populations on a particular part of the genome may be more likely to be selection (especially among populations with distinct demographies), either convergent among populations or at the species level. To facilitate a sliding window analysis of Tajima's D (TD), ddRAD sequence data were aligned against the *Cervus elaphus* genome (assembled to chromosomes), made possible by strong synteny between *C. capreolus* and *Cervus elaphus* (see Figure S9). We identified 53 SNPs (out of 27,298 SNPs in total) associated with 11 chromosomal regions where TD was concurrently elevated (suggesting either balancing selection or population contraction) in all focal *C. capreolus* populations (Figure S10). GO analysis for this list of loci identified significant enrichment for terms associated with meiosis in particular, and reproduction in general (Table S4). Assuming independence (but see below), a stochastic pattern driven by genetic drift,

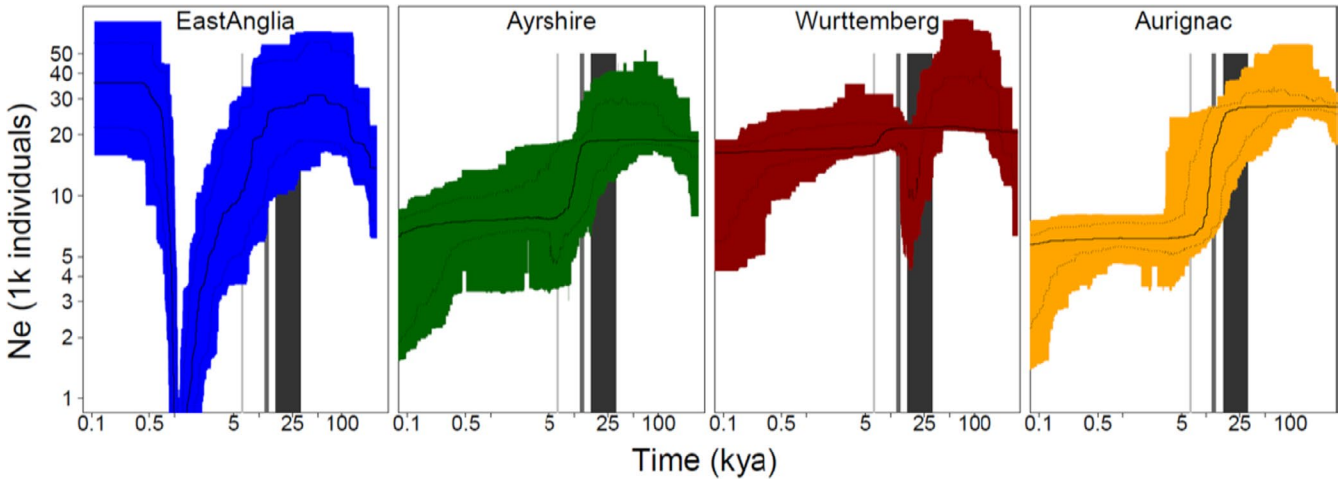

FIGURE 4 Demographic histories of *Capreolus capreolus*. Stairway plots showing demographic histories (mutation rate: 1.1e−8, generation time: 5 years). Grey shaded areas indicate, from left to right, the flooding of Doggerland (6.2–6.5 ka), the Younger Dryas climatic event (11.7–12.9 ka) and the Last Glacial Maximum (16–31 ka). Colour coding reflects regions as in other figures

and counting all instances of positive TD significant at $p \leq .05$, the probability of three convergent, significant values for each occurrence is 3.92×10^{-5} ($N_{EA}/2{,}143 \times N_S/2{,}143 \times N_G/2{,}143$; where N_{EA}, N_S and N_G are the number of incidences in East Anglia, Scotland and Germany, respectively, and 2,143 is the number of windows assessed for TD).

From the observed data we find 30–37 shared significant positive TD windows from pairwise comparisons among populations. Based on independence we would expect 2.3–2.7 events, an approximately 13-fold difference, probably due to shared ancestry. We would expect all three to share a window with significantly elevated TD 0.08 times based on independence, and 13 times this would be just once. There were 11 observed events shared by all three populations, and this is significantly greater than 1 ($\chi^2 = 8.3567$. $p = .0038$), suggesting a shared signal for balancing selection beyond an effect expected from shared ancestry alone.

Negative TD (suggesting either directional selection or population expansion) was measured together with ROH along *C. capreolus* contigs (Figure 6; Figure S11). Convergent ROH from the *C. capreolus* genome and negative TD among all three focal populations were found for two genomic regions which contained a combined total of 15 genes, mostly associated with cellular processes or development (Table S4). Assuming independence (i.e., to test for no correlation and the random occurrence of peaks) using the same criteria as for positive TD ($p < .05$), and given the incidence of significant ROH (11 out of 29,795 possible windows), we can use the multiplicative rule for the probability of independent events and calculate that the probability of each coincidence of significant ROH and negative TD in all three populations would be would be 4.58×10^{-8} if random ($N_{EA}/2{,}143 \times N_S/2{,}143 \times N_G/2{,}143 \times 11/29{,}795$ for ROH). It is possible that shared signals for selection reflect gene surfing (Hallatschek, Hersen, Ramanathan, & Nelson, 2007) during the post-glacial expansion of *C. capreolus* populations out of glacial refugia, but subsequent founder events (as in East Anglia) could disrupt those effects unless they remain strongly adaptive. We also compare TD with Watterson's theta (Figure S12), which shows a similar pattern as seen in Figure S11.

4 | DISCUSSION

In this study we assessed evidence for the impact of genetic drift and natural selection through a series of evolutionary transitions since the origin of the genus *Capreolus* through to the diversification of *C. capreolus* populations. Extensive population structure across the range of both *C. capreolus* (Baker & Hoelzel, 2013;Vernesi et al., 2002) and *C. pygargus* (Lee et al., 2016) probably reflects limited dispersal and rapid demographic change (evident from our genetic analyses, e.g., Figure 3, and from census data; Templeton, 2008, Lee et al., 2016, Burbaite, 2009, Baker & Hoelzel, 2013), which could reduce N_e in both species. At the time of speciation in the middle Pleistocene (estimated here to fall between Marine Isotope Stages 21 and 41; see Ehlers, Gibbard, & Hughes, 2011), N_e was relatively low (Figure 3). Since speciation the two species diverged demographically (Figure 3), and the smaller long-term N_e is reflected in the structure and lower diversity of the *C. capreolus* genome. Earlier data from mtDNA control region sequences showed overlapping values for genetic diversity, but relatively low diversity for *C. capreolus* putative populations (π = 0.25%–0.80%; Baker & Hoelzel, 2014) compared to *C. pygargus* (π = 0.70%–1.23%, apart from an island founder population where π = 0.028%; Sun Lee et al., 2016). Although *C. pygargus* is structured across its range (Lee et al., 2016), the population from Mongolia, the Russian Far East and China (*C. p. tianschanicus*; used for our genome sequence) apparently continued to expand into the last glacial period, while N_e for *C. capreolus* remained small throughout that period (Figure 3).

It is not clear why these demographic trajectories should have been so different, although one possibility could be the differential impact of glaciations in the two regions. For example, in parts of

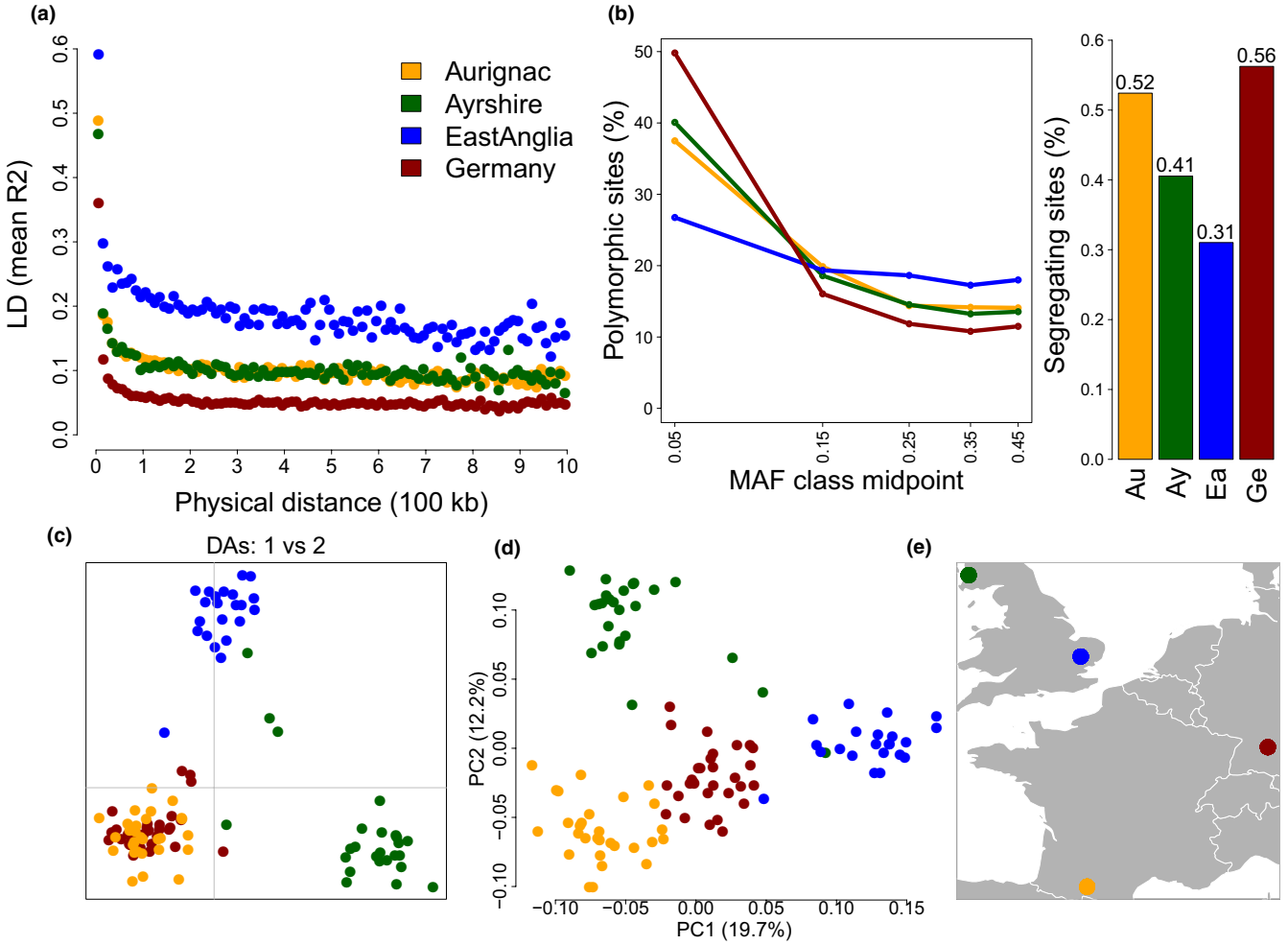

FIGURE 5 Diversity and differentiation of *Capreolus capreolus* populations. Genetic diversity, population structure and demographic history of four *C. capreolus* populations in North Western Europe. (a) LD (measured as the squared correlation coefficient between genotypes [R^2]) versus physical distance (bp). (b) Site frequency spectrum plot and proportion of segregating sites. (c) DAPC analysis based on 250 SNPs shared between data sets (see Section 2). (d) Principal coordinate analyses of Hamming's genetic distance, based on 250 shared SNPs. (e) The geographical locations of the four study populations

Mongolia glaciers apparently retreated during the last glacial maximum due to the dry climate, in contrast to the expanding glaciers in Europe (Batbaatar, Gillespie, Fink, Matmon, & Fujioka, 2018), and ancient DNA from a region nearby (the Denisova cave) confirms the presence of *C. pygargus* 21,000–50,000 years ago (Vorobieva et al., 2011). However, comparison among species suggests that the drivers determining demography over this time frame may be idiosyncratic. For example, the PSMC demographic profile of the milu (*Elaphurus davidianus*) in China (Zhu et al., 2018) is not similar to that of *C. p. tianschanicus* although the geographical distribution is similar (but habitat use distinct, and in recent times milu has become extinct in the wild according to the IUCN). It is also true that other European ungulates show demographic histories quite distinct from *C. capreolis* over the same time frame, for example the bison (*Bison bonasus*), taurine cattle (*Bos taurus*; Gautier et al., 2016) and the red deer (*Cervus elaphus*; Figure S4; after Bana et al. 2018). Although in comparison with *C. pygargus* the PSMC-estimated demographic trajectory of *C. capreolus* is quite flat, there is some indication of a decline following the Eemian interglacial (see Figure 3), and from the stairway plots, again in the Holocene (Figure 4).

Fragmentation and anthropogenic founder events have structured *C. capreolus* populations, affecting regional diversity and linkage equilibrium (Figure 5), and all studied populations show evidence for a similar demographic impact coincident with the major climatic changes at the start of the Holocene (Figure 4). Although this correlation cannot prove causation, evidence for a shared historical shift among now divergent populations would be consistent with a shared response to some environmental driver or shared event (such as coincident population decline, range expansion from glacial refugia, or population splitting at the time of expansion). Although stairway plots are limited in the time frame on which they can provide inference and may provide spurious indications of a bottleneck (see Heller et al., 2013;Patton et al., 2019), we find that the overlapping time frame covered by both PSMC and stairway plots provides similar inference, that all populations show some shared effect at the beginning of the Holocene, and that there is a signal for East Anglia alone that is consistent with

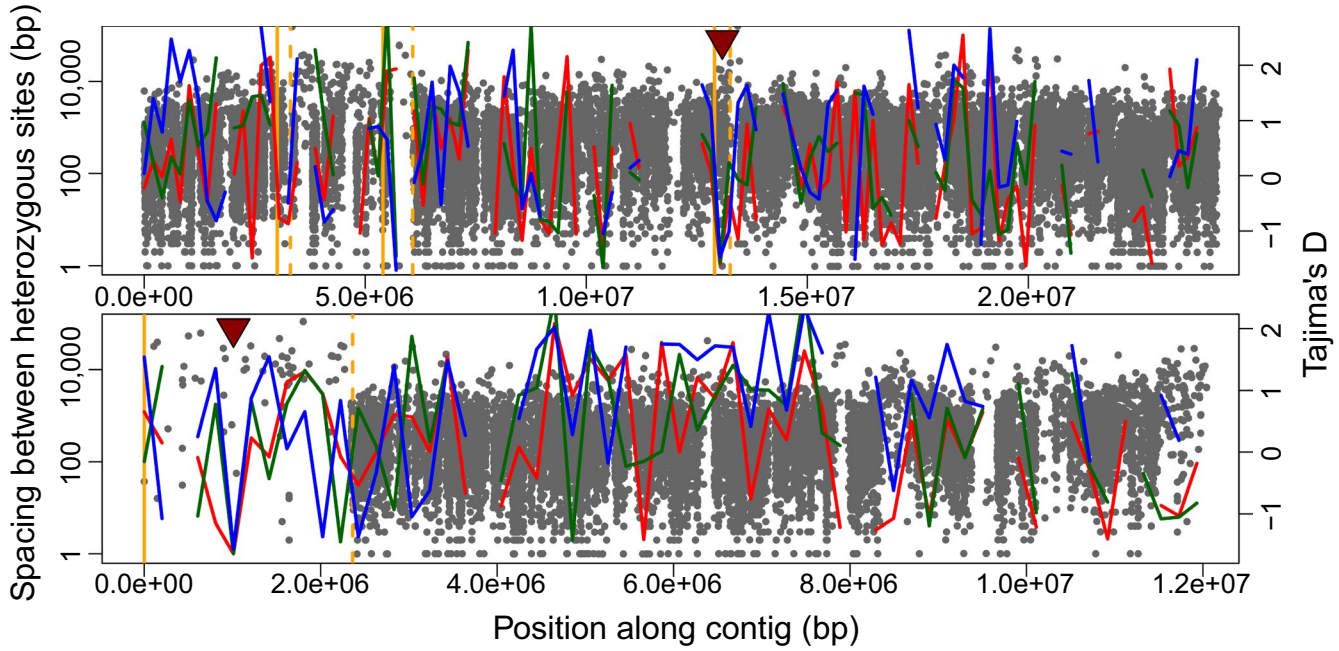

FIGURE 6 Runs of homozygosity (ROH) and Tajima's D. Comparison between ROH in the *Capreolus capreolus* genome for two contigs (see Figure S10 for further comparison), together with Tajima's D data in sliding windows for Scotland (green), East Anglia (blue) and Germany (red). The data set of the Aurignac population was excluded from this analysis because it did not contain enough SNPs shared with the other populations. ROHs are visualized by dark grey dots, which indicate the spacing between heterozygous sites in the *C. capreolus* reference genome. Solid and dashed yellow lines indicate the start and end of regions with unusual long runs of homozygosity. Regions convergent for ROH and low Tajima's D in all populations are indicated with a triangle

the known strong bottleneck there. Together these factors suggest useful inference.

C. capreolus populations have thrived after multiple small, isolated introductions (e.g., in the UK; Baker & Hoelzel, 2013), raising questions about how founder effects may impact local adaptation. In our study we attempt to distinguish signals for selection from drift by comparing several focal populations, including the translocated founder population in East Anglia. Signals for TD were not strongly correlated among these populations (Figure S9), and the populations were significantly differentiated from each other with distinct levels of diversity (Figure 5) and divergent demographic histories (Figure 4). However, they are not fully independent, probably having common ancestors within the Holocene, and therefore even though the likelihood of each concurrent signal by chance alone is very small, it remains possible that shared TD and ROH was driven to some extent by genetic drift or recent demographic events. This is supported by our observation that significant TD coincided among population pairs more often than expected by chance, although even after taking that into account the number of windows with significant positive TD was significantly more than expected by chance, supporting an interpretation of shared balancing selection. Watterson's theta, which reflects diversity, showed a similar relationship with TD (Figure S12).

At the same time, the multiple sites with shared positive TD (suggesting possible balancing selection) also contained genetic loci that were significantly enriched for particular gene functions (Table S4), reinforcing the possibility that adaptive selection was occurring for these loci in all three populations. This includes the East Anglian population, which is known to have been severely bottlenecked. Admixture would have the potential to rescue adaptive potential in this population, although it has apparently remained mostly isolated from other regional populations (Baker & Hoelzel, 2013). Alternatively, rapid growth (Figure 4) may have helped this founder population overcome strong genetic drift in order to be able to respond to selection. The resilience of adaptive selection to drift for populations expanding into novel habitats has been suggested in various studies, for example for a founder population of *Drosophila* in the Hawaiian Islands (see Carson, 1982), for the *Mc1r* locus in old-field mice (*Peromyscus polionotus*; Domingues et al., 2012), and for the origin of "weedy rice" (*Oryza sativa f. spontanea*) by independent de-domestication events from cultivated rice involving strong genetic bottlenecks (Qiu et al., 2017). However, data on the genomic regions preserved by selection in small founder populations are still relatively rare, especially for free-ranging mammals.

In summary, we found some support for adaptive change throughout the history of the genus *Capreolus*, from its origin, through speciation within the genus, and during the differentiation of populations of *C. capreolus*. Gene functions associated with reproductive biology were evidently important, supported by evidence of positive selection during the differentiation of the genus from other cervid lineages (for obligate diapause, unique for the genus among the Artiodactyls), and for balancing selection among populations of *C. capreolus*. Directional selection at loci associated with reproduction was proposed for milu (Zhang et al., 2016), but not at the same loci as we report here for roe deer.

Inference about selection was supported by multiple approaches, including dN/dS ratios (CODEML), amino acid properties (TREESAAP), the phylogeny indicating plesiomorphic selection for the paralogue in *C. capreolus*, neutrality tests and signatures for selective sweeps (at the population level). Although these tests provided consistent inference, greater resolution and the probable identification of a larger number of loci potentially under selection would be possible using resequencing data in future (facilitated by the genome assembly we provide here; see Lowry et al., 2017). For each of these analyses there has been evidence for selection when N_e was relatively low, such as early in the history of the genus (Figure 5), in *C. capreolus* compared to *C. pygargus* throughout most of their history, and in *C. capreolus* populations in the UK compared to the population in Germany. Especially for the anthropogenic founder population in East Anglia, stochastic noise associated with drift probably obscured the signal in some genomic regions, making interpretation at the population level difficult, and may have limited the potential for adaptive change overall. We conclude that both genetic drift and natural selection have been important at each transition stage from the origin of the genus through to the differentiation of populations.

ACKNOWLEDGMENTS

We thank the British Deer Society and associated managers and stalkers and especially Hugh Rose, Annette Kohnen, and Kirstin Waeber for the provision of samples of *C. capreolus*. We thank Regina Kropatsch for providing the raw reads of the *C. capreolus* genome, and Michelle Gaither for advice and assistance. We thank Andreanna Welch and Elena Notarianni for helpful comments and suggestions. This work was supported by the Whitehead Trust and the British Deer Society in support of M.d.J., the Central Public-interest Scientific Institution Basal Research Fund (201811) to Z.P.L., and the Talents Team Construction Fund of Northwestern Polytechnical University (NWPU) to W.W.

AUTHOR CONTRIBUTIONS

A.R.H. and W.W. conceived the study and A.R.H. and M.d.J. wrote the paper with input from all authors. M.d.J., Y.Q., Z.P.L., E.Q. and A.R.H. provided data generation and analyses. K.B. provided DNA samples and field contacts.

DATA AVAILABILITY STATEMENT

Sequence data for the *C. pygargus* genome have been deposited at GenBank under BioProject accession PRJNA526434. Sequences associated with RAD analyses are deposited at GenBank under BioProject accession PRJNA530371. RADseq genotype files are provided on Dryad at: https://doi.org/10.5061/dryad.kkwh70s29. There are no restrictions on data availability.

ORCID

Menno J. de Jong https://orcid.org/0000-0003-2131-9048
Zhipeng Li https://orcid.org/0000-0001-8474-045X
A. Rus Hoelzel https://orcid.org/0000-0002-7265-4180

REFERENCES

Al-Shahrour, F., Diaz-Uriarte, R., & Dopazo, J. (2004). FATIGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, *20*, 578–580. https://doi.org/10.1093/bioinformatics/btg455

Baker, K. H., & Hoelzel, A. R. (2013). Evolution of population genetic structure of the British roe deer by natural and anthropogenic processes (*Capreolus capreolus*). *Ecology and Evolution*, *3*, 89–102.

Baker, K. H., & Hoelzel, A. R. (2014). Influence of Holocene environmental change and anthropogenic impact on the diversity and distribution of roe deer. *Heredity*, *112*, 607–615. https://doi.org/10.1038/hdy.2013.142

Bana, N. A., Nyiri, A., Nagy, J., Frank, K., Nagy, T., Stéger, V., … Orosz, L. (2018). The red deer *Cervus elaphus* genome CerEla1.0: Sequencing, annotating, genes, and chromosomes. *Molecular Genetics and Genomics*, *293*, 665–684. https://doi.org/10.1007/s00438-017-1412-3

Batbaatar, J., Gillespie, A. R., Fink, D., Matmon, A., & Fujioka, T. (2018). Asynchronous glaciations in arid continental climate. *Quaternary Science Reviews*, *182*, 1–19. https://doi.org/10.1016/j.quascirev.2017.12.001

Beyes, M., Nause, N., Bleyer, M., Kaup, F. J., & Neumann, S. (2017). Description of post-implantation embryonic stages in European roe deer (*Capreolus capreolus*) after embryonic diapause. *Anatomia Histologia and Embryologia*, *46*, 582–591.

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research*, *14*, 988–995. https://doi.org/10.1101/gr.1865504

Bolger, A. M., Lohse, M., & Usadel, B. (2014). TRIMMOMATIC: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boroviak, T., Loos, R., Lombard, P., Okahara, J., Behr, R., Sasaki, E., … Bertone, P. (2015). Lineage-specific profiling delineates the emergence and progression of naive pluripotency in mammalian embryogenesis. *Developmental Cell*, *35*, 366–382.

Burbaitė, L., & Csányi, S. (2009). Roe deer population and harvest changes in Europe. *Estonian J. Ecol.*, *58*, 169–180. https://doi.org/10.3176/eco.2009.3.02

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*, 78–94. https://doi.org/10.1006/jmbi.1997.0951

Carson, H. L. (1982). Evolution of *Drosophila* on the newer Hawaiian volcanos. *Heredity*, *48*, 3–25.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). STACKS: Building and Genotyping Loci De Novo from short-read Sequences. *G3*, *1*, 171–182.

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., …Wang, W. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science 364*, 6446.

Croitor, R. (2014). Deer from late Miocene to Pleistocene of western Palearctic: Matching fossil record and molecular phylogeny data. *Zitteliana B*, *32*, 115–153.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Domingues, V. S., Poh, Y.-P., Peterson, B. K., Pennings, P. S., Jensen, J. D., & Hoekstra, H. E. (2012). Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, *66*, 3209–3223. https://doi.org/10.1111/j.1558-5646.2012.01669.x

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340

Ehlers, J., Gibbard, P. L., & Hughes, P. D. (2011). Quaternary glations – extent and chronology. *Developments in Quaternary Science*, *15*, 2–1108.

Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., … Zhao, F. Q. (2009). The genome sequence of

taurine cattle: A window to ruminant biology and evolution. *Science*, *324*, 522–528.

Frichot, E., & Francois, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*, 925–929. https://doi.org/10.1111/2041-210X.12382

Gautier, M., Moazami-Goudarzi, K., Levéziel, H., Parinello, H., Grohs, C., Rialle, S., … Flori, L. (2016). Deciphering the wisent demographic and adaptive histories from individual whole-genome sequences. *Molecular Biology and Evolution*, *33*, 2801–2814. https://doi.org/10.1093/molbev/msw144

Gervais, L., Perrier, C., Bernard, M., Merlet, J., Pemberton, J. M., Pujol, B., & Quéméré, E. (2019). RAD-sequencing for estimating genomic relatedness matrix-based heritability in the wild: A case study in roe deer. *Molecular Ecology Resources*, *19*, 1205–1217. https://doi.org/10.1111/1755-0998.13031

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., … White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*, 5654–5666. https://doi.org/10.1093/nar/gkg770

Hallatschek, O., Hersen, P., Ramanathan, S., & Nelson, D. R. (2007). Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences USA*, *104*, 19926–19930. https://doi.org/10.1073/pnas.0710150104

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. PhD thesis. State College, PA: Pennsylvania State University.

Heller, R., Chikhi, L., & Siegismund, H. R. (2013). The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS ONE*, *8*, e62992. https://doi.org/10.1371/journal.pone.0062992

Hoelzel, A. R., Bruford, M. W., & Fleischer, R. C. (2019). Conservation of adaptive potential and functional diversity. *Cons. Genet.*, *20*, 1–5. https://doi.org/10.1007/s10592-019-01151-x

Jombart, T. (2008). ADEGENET: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). ADEGENET 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*, 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94. https://doi.org/10.1186/1471-2156-11-94

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 35610. https://doi.org/10.1186/s12859-014-0356-4

Kropatsch, R., Dekomien, G., Akkad, D. A., Gerding, W. M., Petrasch-Parwez, E., Young, N. D., … Epplen, J. T. (2013). SOX9 duplication linked to intersex in deer. *PLoS ONE*, *8*, e73734. https://doi.org/10.1371/journal.pone.0073734

Kumar, S., & Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences USA*, *99*, 803–808. https://doi.org/10.1073/pnas.022629899

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–U354. https://doi.org/10.1038/nmeth.1923

Lea, D. W., Pak, D. K., Peterson, L. C., & Hughen, K. A. (2003). Synchroneity of tropical and high-latitude Atlantic temperatures over the last glacial termination. *Science*, *301*, 1361–1364. https://doi.org/10.1126/science.1088470

Lee, Y. S. et al. (2016). Genetic diversity and phylogeography of Siberian roe deer, *Caproulus pygargus*, in central and peripheral populations. *Ecology and Evolution*, *6*, 7286–7297.

Lepais, O., & Weir, J. T. (2014). SimRAD: An R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, *14*, 1314–1321. https://doi.org/10.1111/1755-0998.12273

Li, H. A. (2011). Statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*, 493–496. https://doi.org/10.1038/nature10231

Li, Z., Lin, Z., Ba, H., Chen, L., Yang, Y., Wang, K., … Li, G. (2017). Draft genome of the reindeer (*Rangifer tarandus*). *Gigascience*, *6*, 1–5. https://doi.org/10.1093/gigascience/gix102

Lischer, H. E., & Excoffier, L. (2012). PGDSPIDER: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*, 298–299. https://doi.org/10.1093/bioinformatics/btr642

Lister, A. M., Grubb, P., & Sumner, S. R. M. (1998). In P. Duncan, J. D. C. Linell, & R. Andersen (Eds.), *The European Roe Deer: The biology of success* (pp. 23–46). Scandinavian University Press.

Liu, X. M., & Fu, Y. X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*, 1099–1099.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Research*, *17*, 142–152.

Luthi, D., Le Floch, M., Bereiter, B., Blunier, T., Barnola, J. M., Siegenthaler, U., … Stocker, T. F. (2008). High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature*, *453*, 379–382.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*, 345–352.

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TIGRSCAN and GLIMMERHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*, 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Matosiuk, M., Borkowska, A., Świsłocka, M., Mirski, P., Borowski, Z., Krysiuk, K., … Ratkiewicz, M. (2014). Unexpected population genetic structure of European roe deer in Poland: An invasion of the mtDNA genome from Siberian roe deer. *Molecular Ecology*, *23*, 2559–2572. https://doi.org/10.1111/mec.12745

Mennecart, B., DeMiguel, D., Bibi, F., Rössner, G. E., Métais, G., Neenan, J. M., … Costeur, L. (2017). Bony labyrinth morphology clarifies the origin and evolution of deer. *Sci. Rep-uk*, *7*, 13176. https://doi.org/10.1038/s41598-017-12848-9

Mugal, C. F., Wolf, J. B. W., & Kaj, I. (2012). Why time matters: Codon evolution and the temporal dynamics of dN/dS. *Molecular Biology and Evolution*, *31*, 212–231. https://doi.org/10.1093/molbev/mst192

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, *32*, 1749–1751. https://doi.org/10.1093/bioinformatics/btw044

Nilsen, E. B., Gaillard, J.-M., Andersen, R., Odden, J., Delorme, D., van Laere, G., & Linnell, J. D. C. (2009). A slow life in hell or a fast life in heaven: Demographic analyses of contrasting roe deer populations. *Journal of Animal Ecology*, *78*(3), 585–594. https://doi.org/10.1111/j.1365-2656.2009.01523.x

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Patton, A. H., Margres, M. J., Stahlke, A. R., Hendricks, S., Lewallen, K., Hamede, R. K., … Storfer, A. (2019). Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian devils. *Molecular Biology and Evolution*, *36*(12), 2906–2921. https://doi.org/10.1093/molbev/msz191

Pedersen, C.-E., Albrechtsen, A., Etter, P. D., Johnson, E. A., Orlando, L., Chikhi, L., … Heller, R. (2018). A southern African origin and cryptic structure in the highly mobile plains zebra. *Nature Ecology & Evolution*, *2*, 491–498. https://doi.org/10.1038/s41559-017-0453-7

Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). STAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*, *13*, 946–952.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*, e37135. https://doi.org/10.1371/journal.pone.0037135

Plakhina, D. A., Zvychanaya, E. Y., Kholodova, M. V., & Danilkin, A. A. (2014). Identification of European (*Capreolus capreolus* L.) and Siberian (*C. pygargus* Pall.) roe deer hybrids by microsatellite marker analysis. *Genetika*, *50*, 862–867. https://doi.org/10.1134/S1022795414070151

Ptak, G. E., Tacconi, E., Czernik, M., Toschi, P., Modlinski, J. A., & Loi, P. (2012). Embryonic diapause is conserved across mammals. *PLoS ONE*, *7*, e33027. https://doi.org/10.1371/journal.pone.0033027

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Qiu, J., Zhou, Y., Mao, L., Ye, C., Wang, W., Zhang, J., … Lu, Y. (2017). Genome variation associated with local adaptation of weedy rice during de-domestication. *Nature Communications*, *8*, 15323.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Randi, E., Pierpaoli, M., & Danilkin, A. (1998). Mitochondrial DNA polymorphism in populations of Siberian and European roe deer (*Capreolus pygargus* and *C. capreolus*). *Heredity*, *80*, 429–437. https://doi.org/10.1046/j.1365-2540.1998.00318.x

Seabury, C. M., Bhattarai, E. K., Taylor, J. F., Viswanathan, G. G., Cooper, S. M., Davis, D. S., … Seabury, P. M. (2011). Genome-wide polymorphism and comparative analyses in the white-tailed deer (*Odocoileus virginianus*): A model for conservation genomics. *PLoS ONE*, *6*, e15811. https://doi.org/10.1371/journal.pone.0015811

Shiu, S. H., Byrnes, J. K., Pan, R., Zhang, P., & Li, W. H. (2006). Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proceedings of the National Academy of Sciences USA*, *103*, 2232–2236. https://doi.org/10.1073/pnas.0510388103

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A. F. A., & Hubley, R. (2018). Retrieved from http://www.repeatmasker.org

Sommer, R. S., Fahlke, J. M., Schmolcke, U., Benecke, N., & Zachos, F. E. (2009). Quaternary history of the European roe deer *Capreolus capreolus*. *Mammal Reviews*, *39*, 1–16.

Stamatakis, A. (2014). RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*, W435–439. https://doi.org/10.1093/nar/gkl200

Sun, Y. B., An, Z. S., Clemens, S. C., Bloemendal, J., & Vandenberghe, J. (2010). Seven million years of wind and precipitation variability on the Chinese Loess Plateau. *Earth and Planetary Science Letters*, *297*, 525–535.

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols* Bioinformatics Chapter 4, Unit, 4, 10.

Templeton, A. R. (2008). The reality and importance of founder speciation in evolution. *BioEssays*, *30*, 470–479. https://doi.org/10.1002/bies.20745

van der Molen, J., & van Dijck, B. (2000). The evolution of the Dutch and Belgian coasts and the role of sand supply from the North Sea. *Global and Planetary Change*, *27*, 223–244. https://doi.org/10.1016/S0921-8181(01)00068-6

Vernesi, C., Pecchioli, E., Caramelli, D., Tiedemann, R., Randi, E., & Bertorelle, G. (2002). The genetic structure of natural and reintroduced roe deer (*Capreolus capreolus*) populations in the Alps and central Italy, with reference to the mitochondrial DNA phylogeography of Europe. *Molecular Ecology*, *11*, 1285–1297.

Vorobieva, N. V., Sherbakov, D. Y., Drushkova, A. S., Stanyon, R., Tsybankov, A. A., Vasilev, S. K., … Graphodatsky, A. S. (2011). Genotyping of *Capreolus pygargus* fossil DNA from Denisova cave reveals phylogenetic relationships between ancient and modern samples. *PLoS ONE*, *6*, e24045.

Wang, Y. U., Zhang, C., Wang, N., Li, Z., Heller, R., Liu, R., … Qiu, Q. (2019). Genetic basis of ruminant headgear and rapid antler regeneration. *Science*, *364*, eaav6202. https://doi.org/10.1126/science.aav6335

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*, 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for analysis of population structure. *Evolution*, *38*(6), 1358–1370.

Wilson, G. A., & Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, *163*, 1177–1191.

Woolley, S., Johnson, J., Smith, M. J., Crandall, K. A., & McClellan, D. A. (2003). TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics*, *19*(5), 671–672. https://doi.org/10.1093/bioinformatics/btg043

Xu, Z., & Wang, H. (2007). LTR _ FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*, W265–W268. https://doi.org/10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088

Zeileis, A., & Grothendieck, G. (2005). ZOO: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1–27. http://www.jstatsoft.org/v14/i06/

Zhang, C., Chen, L., Zhou, Y., Wang, K., Chemnick, L. G., Ryder, O. A., … Qiu, Q. (2018). Draft genome of the milu (*Elaphurus davidianus*). *Gigascience*, *7*. https://doi.org/10.1093/gigascience/gix130

Zhang, X., Deng, C., Ding, J., Ren, Y., Zhao, X., Qin, S., … Zhu, L. (2016). Comparative genomics and metagenomics analyses of endangered Père David's deer (*Elaphurus davidianus*) provide insights into population recovery. *bioRxiv*. https://doi.org/10.1101/073528

Zhu, L., Deng, C., Zhao, X., Ding, J., Huang, H., Zhu, S., … Yang, Z. (2018). Endangered Pere David's deer genome provides insights into population recovering. *Evolutionary Applications*, *11*, 2040–2053.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: de Jong MJ, Li Z, Qin Y, et al. Demography and adaptation promoting evolutionary transitions in a mammalian genus that diversified during the Pleistocene. *Mol Ecol*. 2020;00:1–16. https://doi.org/10.1111/mec.15450