

Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics

Jared E. Decker^a, J. Chris Pires^b, Gavin C. Conant^a, Stephanie D. McKay^a, Michael P. Heaton^c, Kefei Chen^d, Alan Cooper^d, Johanna Vilkkii^e, Christopher M. Seabury^f, Alexandre R. Caetano^g, Gary S. Johnson^h, Rick A. Brennemanⁱ, Olivier Hanotte^{j,k}, Lori S. Eggert^b, Pamela Wiener^l, Jong-Joo Kim^m, Kwan Suk Kimⁿ, Tad S. Sonstegard^o, Curt P. Van Tassell^o, Holly L. Neibergs^p, John C. McEwan^q, Rudiger Brauning^q, Luiz L. Coutinho^r, Masroor E. Babar^s, Gregory A. Wilson^t, Matthew C. McClure^a, Megan M. Rolfe^a, JaeWoo Kim^a, Robert D. Schnabel^a, and Jeremy F. Taylor^{a,1}

Divisions of ^aAnimal Sciences, ^bBiological Sciences, and ^cDepartment of Veterinary Pathobiology, University of Missouri, Columbia MO 65211; ^dUnited States Department of Agriculture, Agricultural Research Service, Meat Animal Research Center, Clay Center NE 68933; ^eAustralian Centre for Ancient DNA, School of Earth and Environmental Sciences, University of Adelaide, Adelaide SA 5005, Australia; ^fAgrifood Research Finland MTT, FIN-31600, Jokioinen, Finland; ^gVeterinary Pathobiology, Texas A&M University, College Station TX 77843; ^hEmbrapa Recursos Genéticos e Biotecnologia, Brasília-DF, C.P. 02372, 70770-900, Brasil; ⁱHenry Doorly Zoo, Omaha NE 68107; ^jInternational Livestock Research Institute, P.O. Box 30709, Nairobi 00100, Kenya; ^kSchool of Biology, University of Nottingham, Nottingham NG7 2RD, United Kingdom; ^lThe Roslin Institute, and R(D)SVS, University of Edinburgh, Roslin, Midlothian EH25 9PS, United Kingdom; ^mSchool of Biotechnology, Yeungnam University, Gyeongsan, Republic of Korea; ⁿDepartment of Animal Science, Chungbuk National University, Cheongju, Republic of Korea; ^oUnited States Department of Agriculture, Agricultural Research Service, Bovine Functional Genomics Laboratory, Beltsville MD 20705; ^pDepartment of Animal Sciences, Washington State University, Pullman WA 99164; ^qAnimal Genomics, AgResearch, Invermay, PB 50034, Mosgiel, New Zealand; ^rDepartamento de Zootecnia, ESALQ-USP, Av. Pádua Dias, 11, Piracicaba, SP 13418-900, Brasil; ^sDepartment of Livestock Production, University of Veterinary and Animal Sciences, Lahore 54000, Pakistan; and ^tCanadian Wildlife Service, 200 4999 98th Avenue Northwest, Edmonton AB, Canada T6B 2X3

Edited by James E. Womack, Texas A&M University, College Station, TX, and approved September 14, 2009 (received for review April 29, 2009)

The Pecorans (higher ruminants) are believed to have rapidly speciated in the Mid-Eocene, resulting in five distinct extant families: Antilocapridae, Giraffidae, Moschidae, Cervidae, and Bovidae. Due to the rapid radiation, the Pecoran phylogeny has proven difficult to resolve, and 11 of the 15 possible rooted phylogenies describing ancestral relationships among the Antilocapridae, Giraffidae, Cervidae, and Bovidae have each been argued as representations of the true phylogeny. Here we demonstrate that a genome-wide single nucleotide polymorphism (SNP) genotyping platform designed for one species can be used to genotype ancient DNA from an extinct species and DNA from species diverged up to 29 million years ago and that the produced genotypes can be used to resolve the phylogeny for this rapidly radiated infraorder. We used a high-throughput assay with 54,693 SNP loci developed for *Bos taurus taurus* to rapidly genotype 678 individuals representing 61 Pecoran species. We produced a highly resolved phylogeny for this diverse group based upon 40,843 genome-wide SNP, which is five times as many informative characters as have previously been analyzed. We also establish a method to amplify and screen genomic information from extinct species, and place *Bison priscus* within the Bovidae. The quality of genotype calls and the placement of samples within a well-supported phylogeny may provide an important test for validating the fidelity and integrity of ancient samples. Finally, we constructed a phylogenomic network to accurately describe the relationships between 48 cattle breeds and facilitate inferences concerning the history of domestication and breed formation.

ancient DNA | Pecorans | domestication

The Pecorans are one of the most diverse groups of mammals, ranging in size from the diminutive duiker (adult weight 9–24 kg, shoulder height 0.45–0.51 m) to the giant giraffe (adult weight 500–1,250 kg, shoulder height 4.5–5.8 m). They are indigenous to all continents except South America and Australia (1) and live in a wide variety of environments. The ruminants are believed to have rapidly radiated in the Mid-Eocene (1), and due to this rapid radiation, the Pecoran phylogeny has proven difficult to resolve, with 11 of the 15 possible rooted phylogenies describing relationships among the Antilocapridae, Giraffidae, Cervidae, and Bovidae having been argued as representations of the true phylogeny (2, 3). A supermatrix analysis of nucleotide sequence data from 16 genes has resolved some of the nodes within the Pecoran “Tree of Life (3)” and has provided the most

strongly supported available phylogeny to which we compare the results of our analyses. However, many of the nodes within this phylogeny either have little support or are completely unresolved (e.g., the genus *Caprinae*), and extinct taxa have yet to be phylogenetically placed with confidence (e.g., aurochs). These weakly supported phylogenies have hampered evolutionary studies and conservation efforts for this intriguingly diverse group.

The number and location of prehistoric domestication events for the extinct aurochs (*Bos primigenius*) has also been controversial (4–8), and the ancestry of many of the derived modern breeds of cattle is unknown. Genome-wide single nucleotide polymorphism (SNP) data captured using high-throughput assays provide a method to perform rapid genomic surveys and have recently been used to resolve the history of human populations (9, 10). However, these studies were restricted to a single species, and the remarkable power of these analyses (with >500,000 informative sites) was not fully captured because population relationships depicted using neighbor-joining trees fail to identify multiple ancestral relationships for historically admixed populations. We report an inter-generic, large-scale phylogenomic analysis which applied a genome-wide SNP assay developed for one species to many distantly related species. We also report the application of a genome-wide SNP assay to capture data for ancient DNA samples.

Results

Genotype Fidelity. We have genotyped 16,353 animals representing 61 cattle breeds and 70 species, as divergent from *Bos taurus* as the Savannah elephant (Table S1), with the Illumina BovineSNP50 BeadChip (11, 12) according to Illumina protocols (13). To examine the quality of genotype calls in these outgroup species, we first sequenced the SNP site and flanking regions for

Author contributions: J.E.D. designed research; J.E.D., S.D.M., K.C., A.C., M.C.M., M.M.R., J.W.K., R.D.S., and J.F.T. performed research; M.P.H., K.C., A.C., J.V., C.M.S., A.R.C., G.S.J., R.A.B., O.H., L.S.E., P.W., J.-J.K., K.S.K., T.S.S., C.P.V.T., H.L.N., L.L.C., M.E.B., and G.A.W. contributed new reagents/analytic tools; J.E.D., J.C.P., G.C.C., J.C.M., R.B., R.D.S., and J.F.T. analyzed data; and J.E.D. and J.F.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

¹To whom correspondence should be addressed. E-mail: taylorjerr@missouri.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904691106/DCSupplemental.

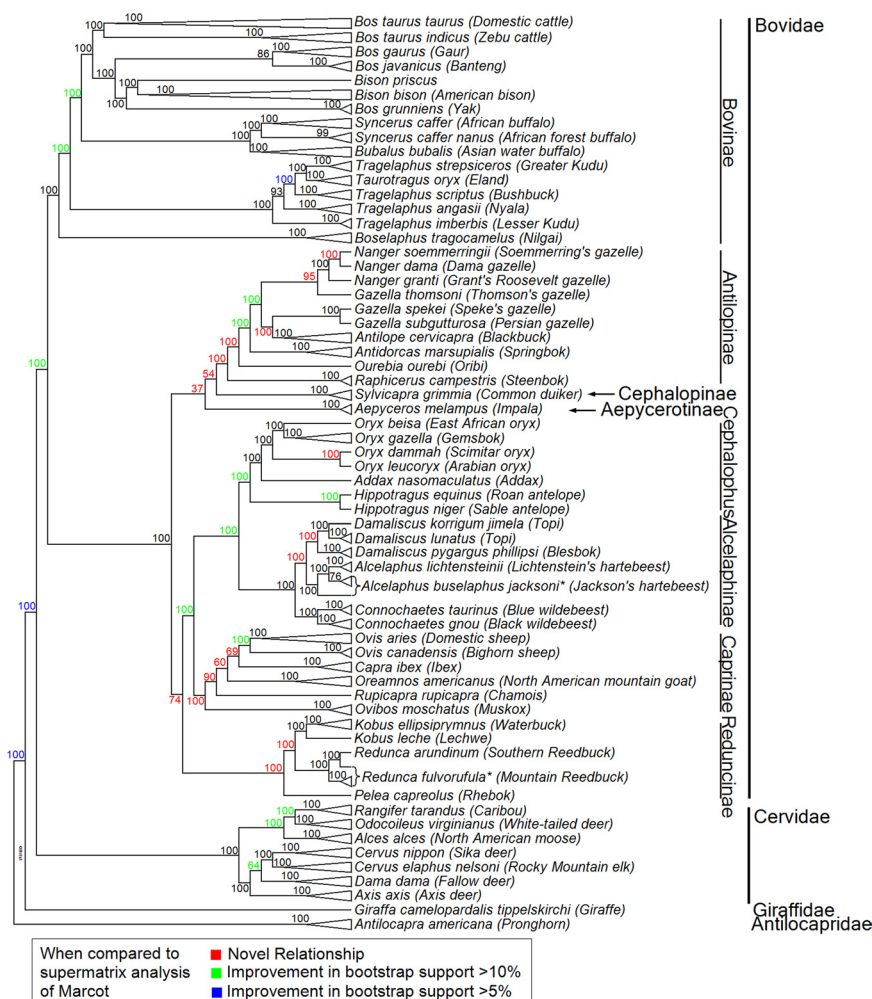


Fig. 1. Strict consensus cladogram (no branch lengths) of 17 most parsimonious trees based on 40,843 SNP genotypes. *, Denotes paraphyletic group.

rs17871403 in 14 species, with pronghorn the most divergent of the sequenced species (Table S2). This SNP was chosen because it has been well characterized in cattle and is a member of a SNP panel that is widely used for parentage analysis (14). Of the genotypes produced by the BovineSNP50 assay (Illumina) for this SNP in these species, 99.13% were concordant with the sequence when we allowed for genotype ambiguity (i.e., *WW* and *SS*) (see *Methods*). One of the six genotyped North American mountain goats and one of the eight genotyped caribou had discordant BovineSNP50 and sequence-based genotype calls (Table S2). This analysis of a single SNP across multiple species suggests a genotyping error rate for BovineSNP50 loci of only 0.87%.

We next aligned all 40,843 SNP probe sequences, which are 50 bases in length, to the international sheep genomics consortium (www.sheepmap.org) genome assembly (available at <https://isgdata.agresearch.co.nz/> and in an annotated form at <http://www.livestockgenomics.csiro.au/sheep/oar1.0.php>) and found that only 26,098 (63.9%) could be uniquely aligned, primarily due to the incomplete status of the assembly. Of these SNP, 829 had an unknown base (*N*) identified at the position of the SNP, and for the remaining 25,269 SNPs, there were 308,518 genotypes called in 17 sheep. Genotype calls were in agreement with the genotype predicted from the respective sequence base for 298,311 genotypes (96.7%). There were 1,834 heterozygous genotypes and 8,373 genotypes that were homozygous for an allele not predicted by the sequence assembly. This suggests a

BovineSNP50 genotyping error rate of between 2.7 and 3.3% in the outgroup species.

Finally, when minor allele frequencies (MAF) averaged over 40,843 SNPs were plotted against average genotype call rates, samples from outgroup species with the lowest call rates had higher than expected MAF (Fig. S1). This appears to be indicative of DNA quality issues since, for example, DNA for the *Capra ibex* samples was extracted from irradiated blood samples that had been stored under refrigeration for several years. On removing these samples, there was almost no correlation between MAF and call rate (Fig. S1). This indicates that as genetic distance from cattle increases and call rate decreases, spurious heterozygote and alternate homozygote genotype calls rarely arise, indicating support for the quality of these data.

Resolution of the Pecoran Phylogeny. Using genotypes for 40,843 SNPs scored with the BovineSNP50 BeadChip (see *Methods*), we produced a completely bifurcating tree with highly supported nodes for 61 Pecoran species, that contains species that diverged up to 29 million years ago (Fig. 1) (15). There were 39,695 parsimony-informative characters using all 678 animals and, remarkably, 21,019 with cattle excluded. Within the Bovidae, only nine nodes had support <100%. We propose 17 relationships and increase the support for 16 previously proposed nodes within the infraorder, when compared to the supermatrix phylogeny of Marcot (3). A striking observation from the phylogeny is that taxonomic classifications of families and subfamilies

mirror the topology of the cladogram, since higher taxa form monophyletic groups. This is an improvement over earlier phylogenies, as previously questionable groupings are now shown to be monophyletic.

Ancient DNA Samples. Currently, PCR-based and non-PCR-based multiple strand displacement amplification (MDA) approaches are used to perform whole genome amplification (16, 17). MDA requires high-quality DNA over 2 kb in length and was found to be inefficient for the ancient bison DNA. Consequently, we used a universal linker-based PCR amplification performed with the GenomePlex Whole Genome Amplification kit (Sigma-Aldrich) to amplify the minute amounts of damaged DNA preserved in bone samples from two ancient Russian *Bison priscus* specimens and test whether the Illumina iSelect platform could be used to analyze samples derived from extinct species. The first, sample BS662, was collected from permafrost deposits at Alyoshkina Zaimka, Siberia, and is approximately 20,000 years old (18). The second, ACAD012, was collected from Sur'ya 5 cave in the Ural Mountains and has been accelerator mass spectrometry radiocarbon dated to $34,460 \pm 290$ years BP. Due to the low amounts of DNA from the ancient specimens and the short DNA fragment lengths produced in the whole genome amplification of degraded ancient samples, the genotype call rates for these samples were much lower than for modern bison (Table S1). However, when these ancient samples were included in the Bovini phylogeny (Fig. 1), BS662 was basal to the modern *Bison bison* clade as expected, but ACAD012 fell within the modern Hereford cattle clade. When we sequenced several overlapping fragments that had been individually amplified from the hyper-variable mitochondrial control region of sample ACAD012, we identified variability within the overlapping regions. This is consistent with the sample having been contaminated with modern DNA or being extremely degraded, as also suggested by our genotype data and consequently the sample was removed from the study. A replicate whole genome amplification (library identification KCMU02) was produced from the *B. priscus* sample used to generate BS662, and when this sample was included in the data set, it was sister to BS662, and both remained sister to modern bison within the phylogeny. However, in the preparation of this library, we avoided the initial DNA fragmentation step within the amplification protocol that appeared to greatly improve the quality and quantity of produced genotypes, as KCMU02 produced a higher genotype call rate (54.9 vs. 45.8%) and far lower heterozygosity (11.5 vs. 39.6%) than did BS662 (Table S3). While only 76.1% of the 12,279 genotypes that were called in both samples were identical, 99.7% of the homozygous genotypes, the only genotype class that has the potential to be phylogenetically informative (see *Methods*), were identical between the replicates.

Relationships Among Cattle Breeds. Phylogenetic relationships were also inferred for 48 cattle breeds ($n = 372$ animals) (Table S1) using parsimony, with most nodes being highly supported (bootstrap values $>70\%$). To accommodate heterozygotes, data were first coded with heterozygotes as polymorphic (noninformative) and then as an independent character state (see *Methods*). When coded as polymorphic, the topology of the cladogram corresponded to the known geographic origins of breeds (Fig. 2A). Interestingly, however, when heterozygotes were coded as distinct characters, the topology changed and no longer clearly reflected the biogeography of breed origins (Fig. 2B).

To further resolve the issue of breed origins, we constructed phylogenetic networks which can reveal conflicting signals in the data (Fig. 3 and Fig. S2). In Fig. S2, *Bos taurus indicus* and *Bos taurus taurus* are distinct groups with long edges between the subspecies. Within *B. t. taurus*, using the Reynolds et al. (19) distance metric and parsimony cladograms (Fig. 2), African

taurine cattle were inferred to be more divergent from European cattle than are the Asian *B. t. taurus* breeds, with 100% bootstrap support in cladograms (Fig. 2 and Figs. S2 and S3). Because SNP were almost exclusively discovered from European *B. t. taurus* samples (12), there is a strong ascertainment bias toward SNP common within European *B. t. taurus* on the BovineSNP50 BeadChip, leading to severe biases in estimates of genetic distance that have prevented us from accurately dating the nodes separating European, African, and Asian cattle (Figs. S3 and S4). Furthermore, the data were recalcitrant to correction for ascertainment (see *Methods*). The network with individuals at node tips (Fig. 3) appears to accurately depict the admixed nature of many populations, for example, the relationship of Belgian Blue to Holsteins and Shorthorns, and Jersey to Iberian and British breeds. The network also reveals pedigree relationships, with sire HO020740 being an interior node to son HO020879.

Discussion

The genotype validation results suggest that BovineSNP50 genotype errors are uncommon, are randomly distributed, and are independent of call rate in the outgroup species. While *Ovis aries* and *B. taurus* are not the most distantly related species surveyed in this study (Fig. 1), their most recent common ancestor was at the base of the Bovidae clade. The use of *O. aries* as a representative for the other species is supported by its 67.2% genotype call rate (Table S1), which was similar to ($\pm 7\%$), or lower than, that for all species and breeds, with the exceptions of Axis deer, Ibex, and Pronghorns, which had call rates $<60\%$.

Despite large amounts of missing data within outgroup species or for the ancient DNA samples, by constructing a larger initial data matrix, which includes more taxa and data than used in previous analyses (20–23), we have produced a highly-resolved phylogeny for a rapidly radiated infraorder, which includes extant and extinct species and in which relationships between and within families have been unresolved. Common ancestry can confound studies of speciation and the evolutionary origins and importance of particular traits; the highly resolved phylogeny presented here can control for this issue by allowing the use of phylogenetically independent contrasts (24). Further, it facilitates informed conservation efforts, as both ancestral relationships and diversity are clearly defined (25), allowing the identification of species and populations within species to target for preservation. With small data sets, the estimated bootstrap support values can be biased due to the presence of a strong correlation between the samples. Large data sets, such as reported here, accurately estimate the support for internal nodes, since nearly independent pseudosamples can be generated for the construction of bootstrap trees.

We demonstrate that reliable genotypes can be produced from ancient DNA samples, but that more work is needed to optimize amplification and genotyping protocols. We suspect that the much higher than expected heterozygosities for these samples are due either to template damage or the nonspecific binding of small, possibly exogenous, DNA fragments to the SNP probes. Despite challenges in library optimization, we placed replicate *B. priscus* samples as sister to modern bison with strong support and have therefore established the feasibility of high-throughput genotyping of ancient samples. Our results also suggest that the fidelity of the produced genotypes may be assessed by their incorporation into a well-resolved phylogeny and that samples producing unreliable genotypes may be identified and removed from further analysis by this process.

Incongruence between the two breed phylogenies occurred as a result of persistent signatures of admixture, which has been well documented in the histories of several breeds. Thus, the conflicting breed phylogenies oversimplify the complex relationships that exist among populations due to geographic isolation, introgression, migration, and admixture. Networks were effective

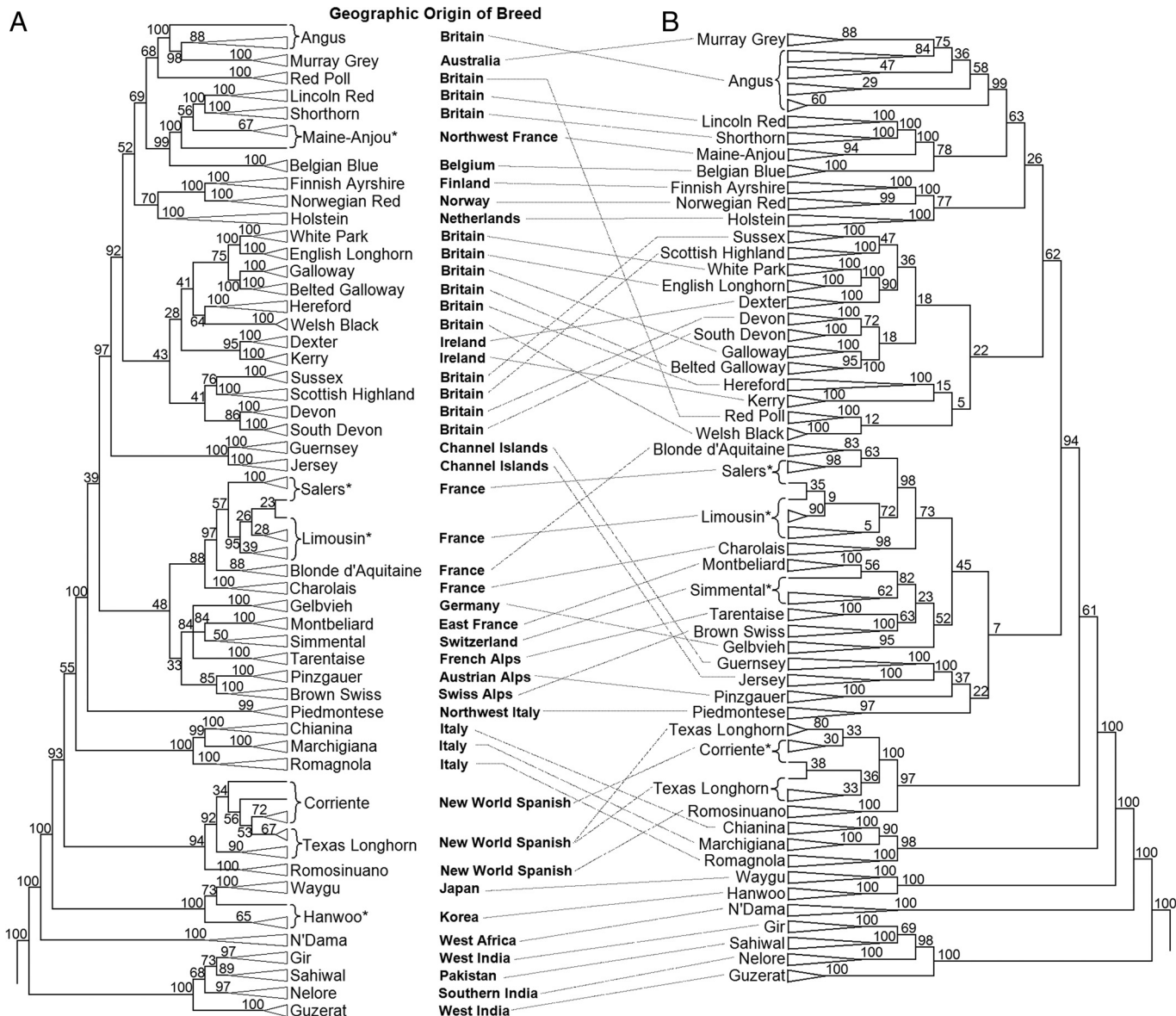


Fig. 2. Consensus of most parsimonious cladograms of 48 cattle breeds. (A) Most parsimonious cladogram of 48 cattle breeds with heterozygotes coded as polymorphic. Geographic origins were retrieved from the literature (21). (B) Most parsimonious cladogram of 48 cattle breeds with heterozygotes coded as a third and separate character state. Values at nodes are percent bootstrap support from 1,000 pseudoreplicates. Dotted lines connect clades of a breed between the two cladograms. *B. t. indicus* is represented by the Gir, Sahiwal, Nelore, and Guzerat breeds, with all other breeds being *B. t. taurus* (Table S1). *, Denotes paraphyletic group.

tive in revealing both geographic isolation and admixture. There were long branches between *B. t. taurus* and *B. t. indicus*, indicating divergence long before domestication. The networks are also consistent with the biogeography of breeds, with European, East Asian, and African taurine cattle forming separate clusters reflecting a predomestication or early postdomestication divergence for these lineages. The West African *B. t. taurus* N'Dama breed diverges from edges shared with *B. t. indicus* in Fig. 3, and admixture proportions from 0.2–8.6% with African *B. t. indicus* have previously been estimated for N'Dama populations (26). Fig. 3 also reveals the biogeographical history of European cattle, which is based upon migrations out of the Fertile Crescent, with domesticated cattle moved sequentially through Turkey, the Balkans, and Italy (27), then radiating through Central Europe and France, and finally into the British Isles (Figs. 2 and 3 and Figs. S2 and S3). These data also support

a second route to the Iberian peninsula by sea from Africa or the Fertile Crescent leading to subsequent admixture with European cattle (4), as the Spanish breeds found in the New World are basal to German and French breeds (Figs. 2 and 3). This pattern of geographic dispersal is interrupted only in a few cases in which breed histories document admixture, such as the Belgian Blue, which was formed between 1840 and 1890 by the crossing of local cattle with Friesian and Shorthorn imported from the Netherlands and England, respectively (28) (Fig. 3). Fig. 3 reveals numerous breed relationships, such as the relationship of the Jersey to both Iberian and British breeds (28), indicating that many exportations and crossbreeding experiments were performed by early pastoralists. Importantly, this figure reveals that the history of breed formation in cattle has been complicated and has involved bottlenecks, evolution in isolation, coancestry, migration, and admixture.

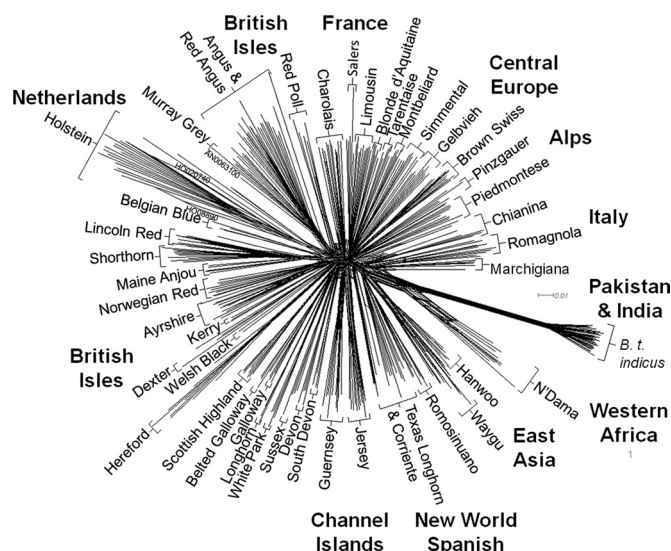


Fig. 3. Phylogenetic network depicting common ancestry for 372 animals representing 48 cattle breeds.

In all analyses, African cattle were the earliest diverged taurine cattle. Consequently, our results now confine the domestication debate to two distinct hypotheses: (i) The occurrence of major domestication events in the Fertile Crescent and Indus Valley (7) were followed by minor captures of aurochs in Africa, East Asia, and Europe (4, 6) or (ii) three separate domestication events occurred in the Fertile Crescent, Indus Valley, and Africa, with a fourth independent domestication in East Asia less likely (5, 8).

The largest previous supermatrix analysis of artiodactyls included 3,823 parsimony-informative characters and required several years of data collection (3). We produced 21,019 parsimony-informative characters at a rate of 1,152 samples in 6 days for \$100 per sample. Where high-density SNP assays are available for sister species, our approach could affordably be applied to the analysis of other orders and families. Such rapid and inexpensive data generation will transform studies of evolution and domestication through the creation of highly resolved phylogenies, including both extant and extinct species. Genome-wide SNP genotyping assays developed for one species can be used for rapid phylogenomic analysis across a broad taxonomic range and are powerful tools for population and evolutionary studies.

Methods

Whole Genome Amplification of Ancient DNA. Ancient DNA was extracted from fossil bison bone specimens using the standard phenol/chloroform/Amicon Ultra-4 method (17). DNA extractions, omniplex library preparations, and PCRs were set-up and performed in a geographically isolated, dedicated ancient DNA facility at the University of Adelaide, Australia. To generate a library of genomic fragments from limited ancient DNA extract, DNA was amplified using the PCR-based GenomePlex Whole Genome Amplification kit (WGA2; Sigma-Aldrich) according to the following protocol: 10 μ L DNA were thoroughly mixed with 2 μ L library preparation buffer and 1 μ L library stabilization solution, and denatured at 95 $^{\circ}$ C for 2 min. After denaturation, 1 μ L library preparation enzyme was added to generate omniplex libraries, followed by a series of incubations at 16 $^{\circ}$ C for 20 min, 24 $^{\circ}$ C for 20 min, 37 $^{\circ}$ C for 20 min, and 75 $^{\circ}$ C for 5 min in a thermal cycler (Corbett Life Science). The omniplex libraries were next amplified using a limited number of genomic amplification cycles. PCR amplification was conducted in a 75- μ L reaction volume containing 14 μ L omniplex library, 7.5 μ L amplification master mix, 48.5 μ L nuclease-free water, and 5 μ L WGA DNA polymerase. The PCR amplification conditions were initial denaturation at 95 $^{\circ}$ C for 3 min, followed by 15 cycles of 94 $^{\circ}$ C for 15 s and 65 $^{\circ}$ C for 5 min. GenomePlex-amplified ancient DNA products were finally purified using the GenElute PCR Clean-Up kit (Sigma-

Aldrich). Ancient DNA libraries were verified by PCR amplification and sequencing of the hypervariable mtDNA control region before analysis with the BovineSNP50 BeadChip (Illumina). A second amplification, labeled KCMU02, of the sample that produced B5662 was constructed using the same protocol as above, except the genomic fragmentation step within the WGA2 protocol was omitted.

Sample Selection. Table S1 shows the numbers of animals genotyped from each species or cattle breed. In taxa or breeds where <10 animals were genotyped, all animals were sampled. If >10 animals were genotyped, animals with the highest genotype call rates and earliest birth dates were selected. When pedigree information was available, closely related animals were avoided, except in Angus and Holstein where 10 old animals (born in the 1950s, 1960s, and 1970s) and 10 recently born animals (born in the late 1990s and 2000s) were selected. When more than 50 animals within a breed had call rates of at least 98% and no pedigree information was available, 10 animals were sampled at random. Samples belonging to recently formed crossbred breeds were removed from the analysis, as these samples distort parsimony phylogenies. Genotypes for the two ancient Bison samples were included despite their much lower genotype call rates, which were expected due to DNA degradation and fragmentation, and the use of whole genome amplification, which affect the fidelity of the Infinium assay. The provenance of all samples included in the analyses is provided in Table S4.

SNP Selection. The BovineSNP50 BeadChip (Illumina) consists of SNP primarily discovered by the sequencing of reduced representation libraries (11), the alignment of random shotgun reads from six cattle breeds to the Hereford assembly, or from the draft assembly of the bovine genome (12). To improve genotype quality for *B. t. indicus* and the outgroup species, we manually adjusted genotype call clusters in Illumina BeadStudio to improve genotype calls. Where pedigree information was available, such as in *O. aries* and *B. bison*, the rate of misinheritances was minimized. A set of 40,843 SNP was selected from the 54,693 loci queried by the assay. Loci selected for analysis were all located on autosomes, had a call rate of at least 80% in 36 (75%) *B. t. taurus* breeds, and were not monomorphic in all breeds. This strategy was effective in selecting informative SNP with few genotype errors (Table S5). Data are available at <http://animalsciences.missouri.edu/animalgenomics/publications/php>.

Genotype Calls in Outgroup Species. Almost 96% of the beads on the BovineSNP50 BeadChip query Infinium II SNP, in which adenine and thymine share a fluorescent probe and guanine and cytosine share a different fluorescent probe. For samples in which all four bases are present at a single locus, AA, AT, and TT genotypes produce indistinguishable fluorescence intensities, as do GG, GC, and CC. Thus, A/T or C/G SNP discovered in *B. t. taurus* were limited in the assay design (1.8 and 2.2%, respectively, and use Infinium I chemistry). However, in species diverged from *B. t. taurus* where all four bases could be present, genotypes are WW (W is the IUPAC code for A or T bases) for one homozygote class, SS (S is the IUPAC code for G or C bases) for the alternate homozygote, and NN (ambiguous) for the heterozygote class. This ambiguity is evident when sequences and genotypes for outgroup species were compared (Table S2). The WW and SS genotypes were identified in BeadStudio as AA and BB genotype calls.

Phylogenetic Analysis. Most parsimonious trees were inferred from the genotypes using TNT version 1.1 (29). In the analyses involving the outgroup species, phylogenetic signal was obtained only from the homozygous genotypes, and AA homozygotes were coded as "0," BB homozygotes were coded as "1," heterozygotes were coded as a polymorphic character state (i.e., "[0,1]"), and missing genotypes were coded as "?." However, in the analyses of the cattle breeds, an additional data set was created in which heterozygotes were identified by a unique character state (i.e., AA = 0, AB = 1, BB = 2). A heuristic search was conducting using the search technology in TNT, and the search level was initially set to 20. Specifically, we used the SPR-TBR algorithm followed by random sectorial searches, constrained sectorial searches, exclusive sectorial searches, and 10 rounds of tree-drifting. The complete search was replicated 20 times, with 10 rounds of tree fusing at the conclusion of these 20 replicates. A subset of the samples from the tribe Bovini was independently analyzed along with the ancient bison samples to validate the quality of the data generated from these ancient samples. A data set with 714 samples from all taxon groups was first used to construct the most parsimonious trees. After excluding samples with low quality DNA, low bootstrap support, and/or nonsensical placement in the cladogram (i.e., elephant and horse as sister to *B. taurus*), a final data set with 678 samples was used to construct most parsimonious trees. The cladogram was rooted with *Antilocapra*

