



## Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds

The Bovine HapMap Consortium, *et al.*

*Science* **324**, 528 (2009);

DOI: 10.1126/science.1167936

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of June 26, 2009):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/324/5926/528>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/324/5926/528/DC1>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/324/5926/528#related-content>

This article **cites 21 articles**, 11 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/324/5926/528#otherarticles>

This article has been **cited by 2 articles** hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/324/5926/528#otherarticles>

This article appears in the following **subject collections**:

Evolution

<http://www.sciencemag.org/cgi/collection/evolution>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

<sup>106</sup>Nutrition and Food Sciences, Utah State University, Logan, UT 84322, USA. <sup>107</sup>Animal Disease Research Unit, USDA–Agricultural Research Service, Pullman, WA 99164, USA. <sup>108</sup>Department of Pharmacology, 2-344 BSB, University of Iowa, 51 Newton Road, Iowa City, IA 52242, USA. <sup>109</sup>Department of Animal Science, 211 Terrill, Uni-

versity of Vermont, 570 Main Street, Burlington, VT 05405, USA.

**Supporting Online Material**

www.sciencemag.org/cgi/content/full/324/5926/522/DC1  
Materials and Methods

Figs. S1 to S23  
Tables S1 to S14  
References

10 December 2008; accepted 16 March 2009  
10.1126/science.1169588

# Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds

The Bovine HapMap Consortium\*

The imprints of domestication and breed development on the genomes of livestock likely differ from those of companion animals. A deep draft sequence assembly of shotgun reads from a single Hereford female and comparative sequences sampled from six additional breeds were used to develop probes to interrogate 37,470 single-nucleotide polymorphisms (SNPs) in 497 cattle from 19 geographically and biologically diverse breeds. These data show that cattle have undergone a rapid recent decrease in effective population size from a very large ancestral population, possibly due to bottlenecks associated with domestication, selection, and breed formation. Domestication and artificial selection appear to have left detectable signatures of selection within the cattle genome, yet the current levels of diversity within breeds are at least as great as exists within humans.

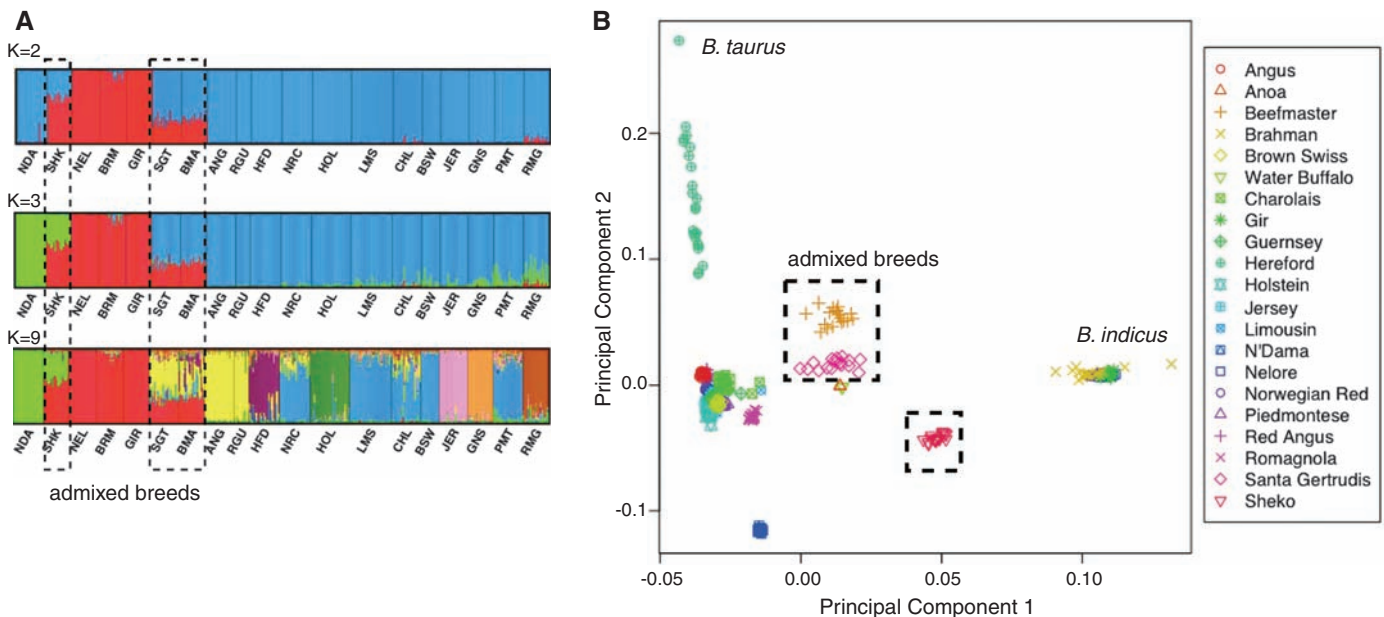
The emergence of modern civilization was accompanied by adaptation, assimilation, and interbreeding of captive animals. In cattle (*Bos taurus*), this resulted in the develop-

ment of individual breeds differing in, for example, milk yield, meat quality, draft ability, and tolerance or resistance to disease and pests. However, despite mapping and diversity studies (1–5) and the identification of mutations affecting some quantitative phenotypes (6–8), the detailed genetic structure and history of cattle are not known.

Cattle occur as two major geographic types, the taurine (humpless—European, African, and Asian) and indicine (humped—South Asian, and East African), which diverged >250 thousand years ago (Kya) (3). We sampled individuals representing 14 taurine ( $n = 376$ ), three indicine ( $n = 73$ ) (table S1), and two hybrid breeds ( $n = 48$ ), as well as two individuals each of *Bubalus quarlesi* and *Bubalus bubalis*, which diverged from *Bos taurus* ~1.25 to 2.0 Mya (9, 10). All breeds except Red Angus ( $n = 12$ ) were represented by at least 24 individuals. We preferred individuals that were unrelated for  $\geq 4$  generations; however, each breed had one or two sire, dam, and progeny trios to allow assessment of genotype quality.

Single-nucleotide polymorphisms (SNPs) that were polymorphic in many populations were primarily derived by comparing whole-genome sequence reads representing five taurine and one indicine breed to the reference genome assembly obtained from a Hereford cow (10) (table S2). This led to the ascertainment of SNPs with high minor allele frequencies (MAFs) within the discovery breeds (table S5). Thus, as expected, with trio progeny removed, SNPs discovered within the taurine breeds had higher average MAFs

\*The full list of authors with their contributions and affiliations is included at the end of the manuscript.



**Fig. 1. (A)** Population structure assessed by InStruct. Bar plot, generated by DISTRICT, depicts classifications with the highest probability under the model that assumes independent allele frequencies and inbreeding coefficients among assumed clusters. Each individual is represented by a vertical bar, often partitioned into colored segments with the length of each segment representing the proportion of the individual's genome from  $K = 2, 3$ , or  $9$  ancestral populations. Breeds are separated by black

lines. NDA, N'Dama; SHK, Sheko; NEL, Nelore; BRM, Brahman; GIR, Gir; SGT, Santa Gertrudis; BMA, Beefmaster; ANG, Angus; RGU, Red Angus; HFD, Hereford; NRC, Norwegian Red; HOL, Holstein; LMS, Limousin; CHL, Charolais; BSW, Brown Swiss; JER, Jersey; GNS, Guernsey; PMT, Piedmontese; RMG, Romagnola. **(B)** Principal components PC1 and PC2 from all SNPs. Taurine breeds remain separated from indicine breeds, and admixed breeds are intermediate.

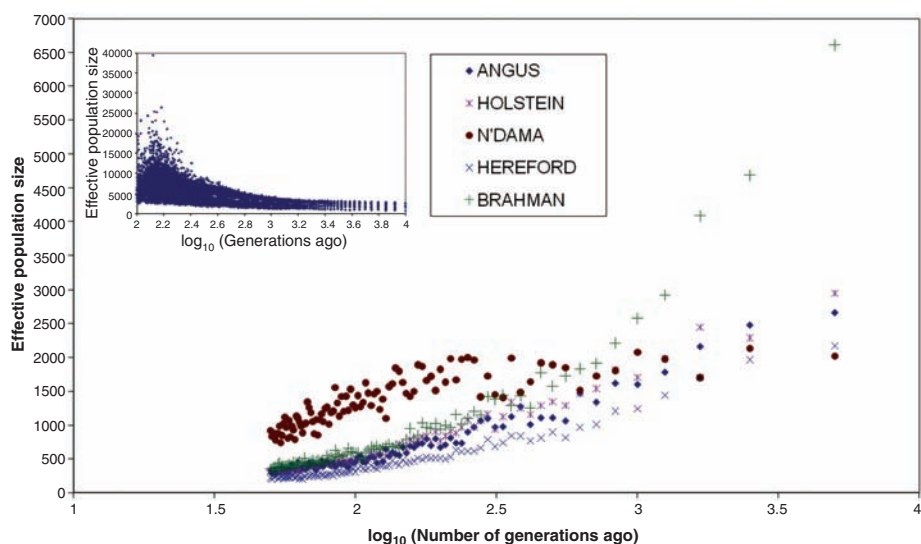
within the taurine than the indicine breeds, and vice versa (table S5); about 30% of SNPs had MAFs >0.3 within the taurine breeds, whereas only about 19% had MAFs >0.3 within the indicine breeds (table S4). The proportions of SNPs in intergenic, intronic, and exonic regions were 63.74, 34.9, and 1.35%, respectively, similar to their representation within the genome. We found that as few as 50 SNPs were necessary for percentage assignment and proof of identity (table S9). Additionally, when we compared ancestries

based on pedigree and allele-sharing between individuals, we were able to predict accurately the extent of ancestry when the pedigree was not known (fig. S24), which could be a useful tool for the management of endangered bovine populations.

To examine relatedness among breeds, we analyzed SNP genotype frequencies with InSTRUCT (11) and performed principal component analysis (PCA) using Eigenstrat (12) (Fig. 1 and fig. S27). Varying the number of presumed ancestral pop-

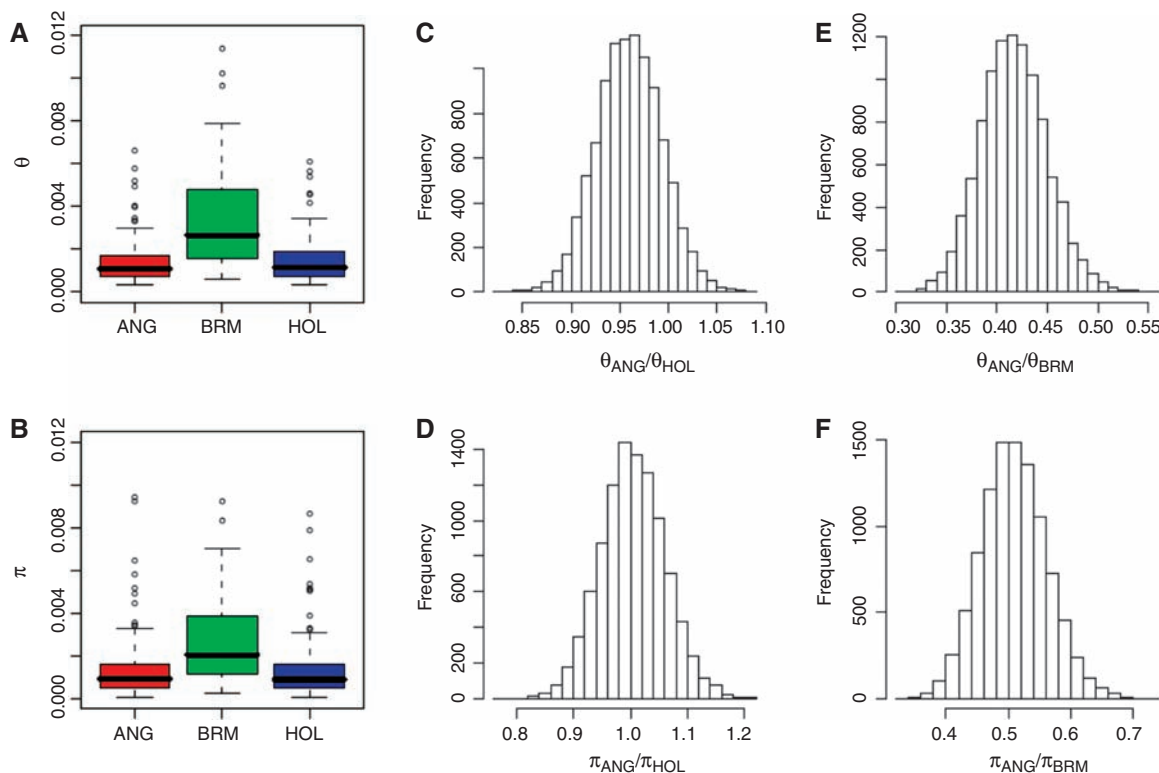
ulations ( $K$ ) within InSTRUCT revealed clusters consistent with the known history of cattle breeds (Fig. 1A). The first level of clustering ( $K = 2$ ) reflects the primary, predomestication division of taurine from indicine cattle. Consequently, breeds derived from indicine and taurine crosses (Beefmaster, Santa Gertrudis, and Sheko) show signatures of admixture with both approaches. At  $K = 3$ , the African breeds N'Dama and Sheko separate from the European breeds—a division that reflects an early, possibly predomestication, divergence. PCA recapitulated these findings (Fig. 1B). At higher levels of  $K$ , we observed clusters that identify single breeds as closed endogamous breeding units. For example, at  $K = 9$ , Jersey, Hereford, Romagnola, and Guernsey each form unique clusters.

If modern breeds arose from bottlenecks from a large ancestral population, we should detect bottleneck signatures within patterns of linkage disequilibrium (LD) and effective population size. We found that the decline of  $r^2$  with genetic distance varied among breeds, although the decline was generally rapid (fig. S10). The extent of LD in cattle is greater than human (13) but less than dog (14). The Jersey and Hereford breeds had higher  $r^2$  than other breeds across the range of distances separating loci. N'Dama had the highest  $r^2$  values at short distances and the lowest  $r^2$  at long distances, which suggested that they were derived from a relatively small ancestral population not subjected to very narrow bottlenecks. The indicine breeds had lower  $r^2$  values at short distances and intermediate  $r^2$  values at longer distances, which indicated that their ancestral popula-



**Fig. 2.** Effective population size in the past estimated from linkage disequilibrium data. Inset graph shows effective population size for the European humans over the same period; from (13). Breeds as in Fig. 1.

**Fig. 3.** Nucleotide diversity across five ENCODE regions resequenced in 47 animals from ANG, Angus; BRM, Brahman; and HOL, Holstein. (A) Watterson's estimate ( $\theta$ ) of the population mutation rate per base pair (pooled across regions). (B) Average pairwise nucleotide distance ( $\pi$ ) within breeds. (C and E) Nonparametric bootstrap estimates of diversity ratios among the three populations on the basis of  $\theta$ . (D and F) Nonparametric bootstrap estimates of diversity ratios among the three populations on the basis of  $\pi$ .



tion was much larger than that from which taurine cattle were domesticated (Fig. 2). As the MAFs for utilized SNPs were generally high and the estimates of LD did not require phased chromosomes, these results should be robust.

When breeds were combined, the decline in LD was more rapid, which reflected a lack of conserved phase relations across breeds. We characterized the extent of haplotype-sharing among breeds between pairs of adjacent SNPs using the *r* statistic. A high correlation between *r* values between two breeds indicates that the same haplotypes tend to persist within both breeds. Correlations between *r* values for SNPs separated by 10 kb were high among the taurine and indicine breeds but were low between these groups (fig. S11). Once SNPs are separated by 100 to 250 kb, we found little haplotype sharing between breeds. Clearly, phase relations dissipated as populations diverged despite the relatively young origin of all breeds. Breeds known to have a recent shared ancestry, notably, Angus and Red Angus; Holstein and Norwegian Red; and Beefmaster and Santa Gertrudis, showed a high correlation among *r* values for SNPs separated by 100 to 250 kb.

Breeds were expected to differ for effective population sizes (*N<sub>e</sub>*) on the basis of differences in the decline of *r<sup>2</sup>* with genetic distance (13). We estimated *N<sub>e</sub>* at various times in each breed's history by setting average *r<sup>2</sup>* values equal to their expectation (15) (Fig. 2 and table S1). *N<sub>e</sub>* has recently declined for all breeds, which reflects bottlenecks associated with domestication, breed formation, and, in some breeds, recent intense selection for milk or beef production. In contrast, human *N<sub>e</sub>* has expanded exponentially over the same period (inset to Fig. 2).

A smaller *N<sub>e</sub>* suggests lower genetic diversity, which is of concern for species viability. To assess genetic diversity free from SNP ascertainment bias, we used the polymerase chain reaction to amplify and sequence 119 closely spaced fragments

from five genomic regions on two chromosomes. Two of these regions were known to harbor quantitative trait loci (QTL). Following the amplification of these regions from 18 Angus, 16 Holstein, and 5 Brahman, the individual segments were Sanger-sequenced to detect SNPs. Of the 1201 discovered SNP, only 258 were common to taurine and indicine breeds, consistent with their age of divergence. Remarkably, 569 SNP (47.4%) were unique to Brahman, and 365 SNP (30.4%) were found only in Angus or Holstein, with 169 SNP (46.3%) common to both breeds. This suggests that breeds represent partly overlapping subsamples within the taurine diversity. However, seven times as many taurine animals had to be sequenced to uncover 75.3% as many SNPs as were discovered in indicine animals. Estimates of the unascertained genomic distributions of SNPs by MAFs within taurine and indicine breeds are in fig. S19.

Diversities as measured by the population mutation rate ( $\theta$ ) and pairwise nucleotide heterozygosity ( $\pi$ ) were also estimated for the 119 fragments and compared between the three breeds (Fig. 3). Angus and Holstein have similar levels of nucleotide diversity measured by both statistics ( $\sim 1.4 \times 10^{-3}$ ) and have  $\sim 40\%$  more nucleotide variation than is found in human populations ( $\sim 1.0 \times 10^{-3}$ ). Brahman variation was even higher, with average estimates of  $\theta$  and  $\pi$  of  $3.35 \times 10^{-3}$  and  $2.74 \times 10^{-3}$ , respectively. These correspond to densities of 1 SNP every 714 bp for pairs of Angus or Holstein chromosomes and 1 SNP every 285 bp for pairs of Brahman chromosomes. These results demonstrate that genetic diversity in cattle is not low despite the decline in *N<sub>e</sub>*.

The lower genetic diversity within modern taurine cattle could reflect a lower diversity within the predomestication ancestral population, and/or postdomestication effects of stronger bottlenecks at breed formation and stronger selection for docility and productivity. Selection is unlikely to be the primary cause, because the

diversity distributions for  $\theta$  and  $\pi$  were similar for all five sequenced regions, and only one region revealed a signature of selection. On the other hand, Fig. 2 suggests that the predomestication *N<sub>e</sub>* of indicine cattle, which originated in southern Asia, a center of species diversity, was much larger than that of taurine cattle. Finally, the process of breed formation in European taurine cattle involved sequential limited migrations from the center of domestication in west Asia (5). Diversity declines with distance from primary sites of domestication (4) and ancient DNA from domesticated cattle and aurochs in Europe show that there was essentially no gene flow from the aurochs into domesticated cattle (5). Therefore, the evidence suggests that the current difference in diversity is mainly due to progenitor population diversity and bottleneck effects at, and before, breed formation rather than differences in the intensity of natural or artificial selection postdomestication.

Cattle have been marked by selection during domestication, breed formation, and ongoing selection to enhance performance and productivity. We utilized three methods to detect genomic selection in cattle: (i) the iHS statistic, which identifies regions of increased local LD (16) suggestive of directional selection; (ii) the *F<sub>ST</sub>* statistic, a measure of the degree of differentiation between subpopulations (17); and (iii) the composite likelihood ratio test (CLR) (18), which assumes a selective sweep model (10). The iHS method was limited by low SNP density and our inability to completely specify ancestral SNP allele states (10). However, despite these limitations, we found evidence for selective sweeps on chromosomes 2, 6, and 14 (table S8 and fig. S20). We identified selection near *MSTN*, in which mutations can cause double muscling (6). Similarly, high iHS values were found in the region near *ABCG2* in which mutations cause differences in milk yield and composition (8). A peak in iHS values was also identified within a gene poor region of chromo-

**Table 1.** Genomic regions associated with extreme *F<sub>ST</sub>* values with gene content consistent with domestication. *F<sub>ST</sub>* values averaged over eight adjacent SNP. Gene functions from OMIM and NCBI Gene database, except for *R3HDM1* described in (2).

Genes	Index SNP	<i>F<sub>ST</sub></i>	BTA	Location	Effect or important phenotypes
<b>High values</b>					
<i>ZRANB3, R3HDM1</i>	rs29021800	0.31	2	64740286...64931017	Feed efficiency
<i>WIF1</i>	BTA-27454	0.29	5	52696749...53098507	Mammalian mesoderm segmentation
<i>SPOCK1</i>	BTA-142690	0.30	7	47501122...47899778	Proteoglycan—synaptic fields of the developing CNS
<i>NBEA</i>	BTA-153392	0.34	12	25884192...26189285	Human idiopathic autism
<i>NMT1, DCAKD, C1QL1</i>	BTA-45533	0.31	19	46088946...46157261	Activator of serum complement system
<i>DACH2, CHM, POU3F4, BRWD3</i>	BTA-161991	0.39	X	41471338...44478564	Human mental retardation
<i>NLGN3 to DGAT2L6</i>	BTA-164256	0.36	X	49279035...50192452	Severe combined immunodeficiency
<b>Low values</b>					
<i>PPARGC1A, DHX15, SOD3</i>	BTC-039516	0.04	6	45354707...45415844	Antioxidative extracellular protection
No known gene	BTC-049723	0.05	14	4569804...5204473	
<i>DNAH9</i>	rs29018632	0.05	19	30943404...31220868	Multisubunit molecular motor
<i>POU5F1, MHC</i>	BTA-55856	0.05	23	27895932...28145846	Major histocompatibility complex
<i>ZNF187</i>	rs29024230	0.04	23	30241236...30502690	Expressed in olfactory tissues
<i>AUTS2</i>	BTC-074065	0.04	25	31773107...32498861	Human autism susceptibility candidate
<i>RYR2</i>	rs29011563	0.05	28	8736599...8772178	Stress- and exercise-induced sudden cardiac death

some 14 adjacent to a region containing genes from *KHDRB3* to *TG*, associated with intramuscular fat content in beef (19).

Calculation of  $F_{ST}$  across all populations for each SNP detected both balancing and divergent selection (fig. S20). Some of the highest and lowest average  $F_{ST}$  values were found in genes associated with behavior, the immune system, and feed efficiency (Table 1). Domestication most likely required the selection of smaller and more docile animals that could resist pathogens and adapt to a human-controlled environment (20). One region under selection contains *R3HDMI* and is associated with efficient food conversion and intramuscular fat content in some breeds (2). In addition to the *R3HDMI* gene (21), this region is also under selection in Europeans, most likely because it contains *LCT*, mutations of which allow the digestion of lactose in adults (22). These results suggest that mutations in this region may affect energy homeostasis. Furthermore, we detected selection between beef and dairy breeds with both CLR and iHS, represented by a broad, high  $F_{ST}$  peak across the region, centered on *SPOCK1* (Table 1). As several QTL have been mapped to this region, multiple loci could be under divergent selection (1), although this peak does not encompass *CAST*, which affects meat quality (23).

Our high resolution examination of cattle shows that unlike the dog—which has restricted diversity and high levels of inbreeding—domesticated cattle had a large ancestral population size and that more aurochs must have been domesticated than wolves; reducing the severity of the domestication bottleneck. SNP diversity within taurine breeds was similar to that of humans, but was significantly less than diversity within indicine breeds, which suggested that the Indian subcontinent was a major site of cattle domestication and predomestication diversity. Selection first for domestication and then for agricultural specialization have apparently reduced breed effective population sizes to relatively small numbers. The recent decline in diversity is sufficiently rapid that loss of diversity should be of concern to animal breeders. Despite this, population levels of LD are unexpectedly low considering the relatively small  $N_e$ , which indicates that effective population sizes were much larger in the very recent past.

## References and Notes

1. S. M. Kappes et al., *J. Anim. Sci.* **78**, 3053 (2000).
2. W. Barendse et al., *Genetics* **176**, 1893 (2007).
3. D. G. Bradley et al., *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5131 (1996).
4. R. T. Loftus et al., *Mol. Ecol.* **8**, 2015 (1999).
5. C. J. Edwards et al., *Proc. R. Soc. London B. Biol. Sci.* **274**, 1377 (2007).
6. L. Grobet et al., *Nat. Genet.* **17**, 71 (1997).
7. B. Grisart et al., *Genome Res.* **12**, 222 (2002).
8. M. Cohen-Zinder et al., *Genome Res.* **15**, 936 (2005).
9. A. Schreiber et al., *J. Hered.* **90**, 165 (1999).
10. Materials and methods are available as supporting materials on Science Online.
11. H. Gao, S. Williamson, C. D. Bustamante, *Genetics* **176**, 1635 (2007).
12. N. Patterson, A. L. Price, D. Reich, *PLoS Genet.* **2**, e190 (2006).
13. A. Tenesa et al., *Genome Res.* **17**, 520 (2007).
14. K. Lindblad-Toh et al., *Nature* **438**, 803 (2005).
15. J. A. Sved, *Theor. Popul. Biol.* **2**, 125 (1971).
16. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, *PLoS Biol.* **4**, e72 (2006).
17. J. M. Akey, G. Zhang, K. Zhang, L. Jin, M. D. Shriver, *Genome Res.* **12**, 1805 (2002).
18. R. Nielsen et al., *Genome Res.* **15**, 1566 (2005).
19. W. Barendse et al., *Aust. J. Exp. Agric.* **44**, 669 (2004).
20. J. Clutton-Brock, *A Natural History of Domesticated Mammals* (Cambridge Univ. Press, Cambridge, 2nd ed., 1999), 238 pp.
21. P. C. Sabeti et al., *Nature* **449**, 913 (2007).
22. S. A. Tishkoff et al., *Nat. Genet.* **39**, 31 (2007).
23. W. Barendse et al., *Genetics* **176**, 2601 (2007).
24. The Bovine Hapmap Consortium funded by the National Human Genome Research Institute of the National Institutes of Health, U.S. Department of Agriculture's Agricultural Research Service (USDA-ARS) and Cooperative State Research, Education and Extension Service (USDA CSREES), the Research Council of Norway, as well as: American Angus Association, American Hereford Association, American Jersey Cattle Association, AgResearch (New Zealand), Beef CRC and Meat and Livestock Australia for the Australian Brahman Breeders Association, Beefmaster Breeders United, The Brazilian Agricultural Research Corporation (Embrapa), Brown Swiss Association, Commonwealth Scientific and Industrial Research Organization, Dairy InSight, GENO Breeding and Artificial Insemination Association—Norway, Herd Book/France Limousin Selection, Holstein Association USA, International Atomic Energy Agency (IAEA)—United Nations Food and Agriculture Organization (FAO)/IAEA Vienna, International Livestock Research Institute—Kenya, Italian Piedmontese Breeders—Parco Tecnologico Padano, Italian Romagnola Society—Università Cattolica del Sacro Cuore, Livestock Improvement Corporation, Meat and Wool New Zealand, North American Limousin Foundation, Red Angus Association of America, Roslin Institute for UK Guernsey, and Sygen (now Genus). See SOM for additional acknowledgements. The genome sequence (AAF03000000) and SNPs (SOM) are available from NCBI.

## The Bovine HapMap Consortium

**Overall project leadership:** Richard A. Gibbs,<sup>1,2\*</sup> Jeremy F. Taylor,<sup>3\*</sup> Curtis P. Van Tassel<sup>4\*</sup>

**HapMap project group leaders:** William Barendse,<sup>5,6</sup> Kellye A. Eversole,<sup>7</sup> Richard A. Gibbs,<sup>1,2</sup> Clare A. Gill,<sup>8</sup> Ronnie D. Green,<sup>9</sup> Debora L. Hamernik,<sup>10</sup> Steven M. Kappes,<sup>9</sup> Sigbjørn Lien,<sup>11</sup> Lakshmi K. Matukumalli,<sup>12,4</sup> John C. McEwan,<sup>13</sup> Lynne V. Nazareth,<sup>1,2</sup> Robert D. Schnabel,<sup>3</sup> Jeremy F. Taylor,<sup>3</sup> Curtis P. Van Tassel,<sup>4</sup> George M. Weinstock,<sup>1,2</sup> David A. Wheeler<sup>1,2</sup>

**Breed champions:** Paolo Ajmone-Marsan,<sup>14</sup> William Barendse,<sup>5,6</sup> Paul J. Boettcher,<sup>15</sup> Alexandre R. Caetano,<sup>16</sup> Jose Fernando Garcia,<sup>15,17</sup> Clare A. Gill,<sup>8</sup> Ronnie D. Green<sup>9</sup> (leader), Olivier Hanotte,<sup>18</sup> Sigbjørn Lien,<sup>11</sup> Paola Mariani,<sup>19</sup> John C. McEwan,<sup>13</sup> Loren C. Skow,<sup>20</sup> Tad S. Sonstegard,<sup>4</sup> Curtis P. Van Tassel,<sup>4</sup> John L. Williams<sup>19,21</sup>

**Pedigree analysis and breed sampling:** Alexandre R. Caetano,<sup>16</sup> Boubacar Diallo,<sup>22</sup> Ronnie D. Green,<sup>9</sup> Lemecha Hailemariam,<sup>23</sup> Olivier Hanotte,<sup>18</sup> Mario L. Martinez,<sup>24</sup> Chris A. Morris,<sup>25</sup> Luiz O. C. Silva,<sup>26</sup> Richard J. Spelman,<sup>27</sup> Jeremy F. Taylor<sup>3</sup> (leader), Curtis P. Van Tassel<sup>4,28</sup> (leader), Woudyalew Mulatu,<sup>28</sup> Keyan Zhao<sup>29</sup>

**Sample acquisition and DNA preparation:** Colette A. Abbey,<sup>8</sup> Morris Agaba,<sup>18</sup> Flábio R. Araujo,<sup>26</sup> Rowan J. Bunch,<sup>5,6</sup> James Burton,<sup>30</sup> Clare A. Gill<sup>8</sup> (leader), Chiara Gorni,<sup>19</sup> Ronnie D. Green,<sup>9</sup> Hanotte Olivier,<sup>18</sup> Blair E. Harrison,<sup>5,6</sup> Sigbjørn Lien,<sup>11</sup> Bill Luff,<sup>31</sup> Marco A. Machado,<sup>24</sup> Paola Mariani,<sup>19</sup> John C. McEwan,<sup>13</sup> Chris A. Morris,<sup>25</sup> Joel Mwakaya,<sup>18</sup> Graham Plastow,<sup>32</sup> Warren Sim,<sup>5,6</sup> Loren C. Skow,<sup>20</sup> Timothy Smith,<sup>33</sup> Tad S. Sonstegard,<sup>4</sup> Richard J. Spelman,<sup>27</sup> Jeremy F. Taylor,<sup>3</sup> Merle B. Thomas,<sup>5,6</sup> Alessio Valentini,<sup>34</sup> Curtis P. Van Tassel,<sup>4</sup> Paul Williams,<sup>5</sup> James Womack,<sup>35</sup> John A. Woolliams<sup>21</sup>

**Genome assembly:** Yue Liu,<sup>1,2</sup> Xiang Qin,<sup>1,2</sup> Kim C. Worley<sup>1,2</sup> (leader)

**SNP discovery:** Chuan Gao,<sup>8</sup> Clare A. Gill,<sup>8</sup> Huaiyang Jiang,<sup>1,2</sup> Yue Liu,<sup>1,2</sup> Stephen S. Moore,<sup>32</sup> Lynne V. Nazareth,<sup>1,2</sup> Yanru Ren,<sup>1,2</sup> Xing-Zhi Song,<sup>1,2</sup> David A. Wheeler<sup>1,2</sup> (leader), Kim C. Worley<sup>1,2</sup>

**ENCODE resequencing:** Carlos D. Bustamante,<sup>29</sup> Ryan D. Hernandez,<sup>29</sup> Donna M. Muzny,<sup>1,2</sup> Lynne V. Nazareth,<sup>1,2</sup> Shobha

Patil,<sup>1,2</sup> Yanru Ren,<sup>1,2</sup> Anthony San Lucas,<sup>1,2</sup> David A. Wheeler<sup>1,2</sup> (leader)

**Genotyping:** Qing Fu,<sup>1,2</sup> Matthew P. Kent,<sup>11</sup> Sigbjørn Lien,<sup>11</sup> Stephen S. Moore,<sup>32</sup> Lynne V. Nazareth<sup>1,2</sup> (leader), Richard Vega,<sup>1,2</sup> David A. Wheeler<sup>1,2</sup> (leader)

**Database & Web site development:** Colette A. Abbey,<sup>8</sup> Chuan Gao,<sup>8</sup> Clare A. Gill,<sup>8</sup> Ronnie D. Green,<sup>9</sup> Lakshmi K. Matukumalli<sup>12,4</sup> (leader), Aruna Matukumalli,<sup>4</sup> Sean McWilliam,<sup>5,6</sup> Curtis P. Van Tassel<sup>4</sup>

**QA/QC:** Colette A. Abbey,<sup>8</sup> Clare A. Gill,<sup>8</sup> Matthew P. Kent,<sup>11</sup> Sigbjørn Lien,<sup>11</sup> Lakshmi K. Matukumalli<sup>12,4</sup> (leader), Robert D. Schnabel<sup>3</sup> (leader), Gert Sclpe,<sup>19</sup> Jeremy F. Taylor<sup>3</sup>

**Allele frequency analysis:** Paolo Ajmone-Marsan,<sup>14</sup> Katarzyna Bryc,<sup>29</sup> Carlos D. Bustamante<sup>29</sup> (leader), Jungwoo Choi,<sup>8</sup> Hong Gao,<sup>29</sup> John J. Grefenstette,<sup>12</sup> Lakshmi K. Matukumalli<sup>12,4</sup> (leader), Brenda Murdoch,<sup>32</sup> Stephen S. Moore,<sup>32</sup> Lynne V. Nazareth<sup>1,2</sup> (leader), Alessandra Stella,<sup>19</sup> Curtis P. Van Tassel,<sup>4</sup> Rafael Villa-Angulo,<sup>12</sup> David A. Wheeler<sup>1,2</sup> (leader), Mark Wright<sup>29</sup>

**Map data provision and analysis:** Jan Aerts,<sup>21,36</sup> Oliver Jann,<sup>21</sup> Lakshmi K. Matukumalli,<sup>12,4</sup> Riccardo Negrini,<sup>14</sup> Tad S. Sonstegard,<sup>4</sup> John L. Williams<sup>19,21</sup>

**Haplotype estimation:** Paolo Ajmone-Marsan,<sup>14</sup> John J. Grefenstette<sup>12</sup> (leader), Lakshmi K. Matukumalli,<sup>12,4</sup> Riccardo Negrini,<sup>14</sup> Robert D. Schnabel,<sup>3</sup> Jeremy F. Taylor,<sup>3</sup> Rafael Villa-Angulo<sup>12</sup>

**Long-range LD analysis:** John J. Grefenstette,<sup>12</sup> Lakshmi K. Matukumalli,<sup>12,4</sup> Curtis P. Van Tassel<sup>4</sup> (leader), Rafael Villa-Angulo<sup>12</sup>

**LD persistence across breeds:** Mike E. Goddard,<sup>37,38</sup> Ben J. Hayes<sup>37</sup> (leader)

**Selective sweeps:** William Barendse<sup>5,6</sup> (leader), Daniel G. Bradley,<sup>39</sup> Paul J. Boettcher,<sup>15</sup> Carlos D. Bustamante,<sup>29</sup> Jungwoo Choi,<sup>8</sup> Marcos Barbosa da Silva,<sup>4,24</sup> Clare A. Gill,<sup>8</sup> Lilian P. L. Lau,<sup>39</sup> George E. Liu,<sup>4</sup> David J. Lynn,<sup>39,40</sup> Francesca Panzitta,<sup>19</sup> Gert Sclpe,<sup>19</sup> Mark Wright<sup>29</sup>

**Applications:** Carlos D. Bustamante<sup>29</sup> (leader), Ken G. Dodds,<sup>13</sup> John C. McEwan<sup>13</sup> (leader), Jeremy F. Taylor<sup>3</sup> (leader), Curtis P. Van Tassel<sup>4</sup>

<sup>1</sup>Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

<sup>2</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. <sup>3</sup>Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211–5300, USA.

<sup>4</sup>Bovine Functional Genomics Laboratory, U.S. Department of Agriculture (USDA) Agricultural Research Service (USDA-ARS), Beltsville Agricultural Research Center (BARC)—East, Beltsville, MD 20705, USA. <sup>5</sup>Commonwealth Scientific and Industrial Research Organization (CSIRO), Livestock Industries, 306 Carmody Road, St. Lucia, Queensland 4067, Australia.

<sup>6</sup>Cooperative Research Center (CRC) for Beef Genetic Technologies, University of New England, Armidale, NSW 2351, Australia. <sup>7</sup>Alliance for Animal Genome Research, 5207 Wyoming Road, Bethesda, MD 20816, USA. <sup>8</sup>Department of Animal Science, Texas A&M University, 2471 TAMU, College Station, TX 77843–2471, USA. <sup>9</sup>National Program Staff, USDA-ARS, 5601 Sunnyside Avenue, Beltsville, MD 20705, USA. <sup>10</sup>USDA Cooperative State Research, Education, and Extension Service, 1400 Independence Avenue, SW, Washington, DC 20250–2220, USA. <sup>11</sup>Centre for Integrative Genetics and Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Arbotreiveien 6, Ås 1432, Norway. <sup>12</sup>Department of Bioinformatics and Computational Biology, George Mason University, 10900 University Boulevard, Manassas, VA 20110, USA. <sup>13</sup>Animal Genomics, AgResearch, Invermay, Post Box 50034, Mosgiel 9053, New Zealand. <sup>14</sup>Istituto di Zootecnica, Università Cattolica del Sacro Cuore, via East Parmense, 84, Piacenza 29100, Italy. <sup>15</sup>Joint United Nations Food and Agriculture Organization (FAO)—International Atomic Energy Agency (IAEA), Division of Nuclear Techniques in Food and Agriculture, IAEA, Post Office Box 100, Wagramstrasse 5, Vienna A1400, Austria. <sup>16</sup>Embrapa Genetic Resources and Biotechnology Center, Final Avenida W5 Norte, Brasília, DF 70770-900, Brazil. <sup>17</sup>Universidade Estadual Paulista (UNESP), Department of Animal Production and Health, IAEA Collaborating Centre in Animal Genomics and Bioinformatics, Aracatuba, SP 16050-680, Brazil. <sup>18</sup>Animal Genetics Resources Characterization, International Livestock Research Institute, Post Office Box 30709, Nairobi 00100, Kenya. <sup>19</sup>Parco Tecnologico Padano, Via Einstein, Polo Universitario, Lodi 26900, Italy. <sup>20</sup>Department of Veterinary Integrative Biological Sciences, College of Veterinary Medicine and

Biomedical Sciences, Texas A&M University, College Station, TX 77843-4461, USA. <sup>21</sup>The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, Midlothian, E25 9PS, UK. <sup>22</sup>Direction Nationale de l'Élevage, Post Office Box 559, Conakry, Guinea. <sup>23</sup>Ethiopian Institute of Agricultural Research, Post Office Box 2003, Addis Ababa, Ethiopia. <sup>24</sup>Embrapa Dairy Cattle Center, Rua Eugênio do Nascimento, 610, Juiz de Fora, MG 36038-330, Brazil. <sup>25</sup>Animal Genomics, AgResearch, Ruakura, Post Box 3123, Hamilton 3240, New Zealand. <sup>26</sup>Embrapa Beef Cattle Center, Rod. BR 262, km 4, Campo Grande, MS 79002-970, Brazil. <sup>27</sup>Research and Development, LIC, Post Box 3016, Hamilton 3240, New Zealand. <sup>28</sup>Animal Genetics Resources Characterization, International Livestock Research Institute, Post Office Box 5689, Addis Ababa, Ethiopia. <sup>29</sup>Department of Biological Statistics and Computational Biology, Cornell University, 101 Biotechnology Building, Ithaca, NY 14853, USA. <sup>30</sup>Veterinary Biomedical Sciences, Royal (Dick) School of Veterinary Studies, The University of Edinburgh Summerhall, Edinburgh,

EH9 1QH Scotland. <sup>31</sup>World Guernsey Cattle Federation, The Hollyhocks, 10 Clos des Goddards, Rue des Goddards, Castel, Guernsey, GY5 7JD, Channel Islands, UK. <sup>32</sup>Agricultural Food and Nutritional Science, University of Alberta, 410 AgFor Centre, Edmonton, AB, T6G 2P5, Canada. <sup>33</sup>Molecular Genetics Research Unit, USDA-ARS, U.S. Meat Animal Research Center, Post Office Box 166, Clay Center, NE 68933, USA. <sup>34</sup>Dipartimento di Produzioni Animali, Università della Tuscia, via de Lellis, Viterbo 01100, Italy. <sup>35</sup>Department of Veterinary Pathobiology, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843-4461, USA. <sup>36</sup>Genome Dynamics and Evolution, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK. <sup>37</sup>Animal Genetics and Genomics, Department of Primary Industries, 475 Mickelham Road Attwood, VIC 3031, Australia. <sup>38</sup>Faculty of Land and Food Resources, University of Melbourne, Royal Parade, Parkville, VIC 3010, Australia. <sup>39</sup>Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland. <sup>40</sup>Department of Molecular Biology and Biochemistry,

Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada.

\*To whom correspondence and requests for materials should be addressed. E-mail: curt.vantassell@ars.usda.gov (C.P.V.T.), taylorjerr@missouri.edu (J.F.T.), and agibbs@bcm.tmc.edu (R.A.G.)

†In memoriam.

### Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5926/528/DC1

Materials and Methods

Figs. S1 to S27

Tables S1 to S9

References

31 October 2008; accepted 16 March 2009

10.1126/science.1167936

# Revealing the History of Sheep Domestication Using Retrovirus Integrations

Bernardo Chessa,<sup>1,2</sup> Filipe Pereira,<sup>3</sup> Frederick Arnaud,<sup>1</sup> Antonio Amorim,<sup>3</sup> Félix Goyache,<sup>4</sup> Ingrid Mainland,<sup>5</sup> Rowland R. Kao,<sup>1</sup> Josephine M. Pemberton,<sup>6</sup> Dario Beraldi,<sup>6</sup> Michael J. Stear,<sup>1</sup> Alberto Alberti,<sup>2</sup> Marco Pittau,<sup>2</sup> Leopoldo Iannuzzi,<sup>7</sup> Mohammad H. Banabazi,<sup>8</sup> Rudovick R. Kazwala,<sup>9</sup> Ya-ping Zhang,<sup>10</sup> Juan J. Arranz,<sup>11</sup> Bahy A. Ali,<sup>12</sup> Zhiliang Wang,<sup>13</sup> Metehan Uzun,<sup>14</sup> Michel M. Dione,<sup>15</sup> Ingrid Olsaker,<sup>16</sup> Lars-Erik Holm,<sup>17</sup> Urmas Saarma,<sup>18</sup> Sohail Ahmad,<sup>19</sup> Nurbiy Marzanov,<sup>20</sup> Emma Eythorsdottir,<sup>21</sup> Martin J. Holland,<sup>22,23</sup> Paolo Ajmone-Marsan,<sup>24</sup> Michael W. Bruford,<sup>25</sup> Juha Kantanen,<sup>26</sup> Thomas E. Spencer,<sup>27</sup> Massimo Palmarini<sup>1\*</sup>

The domestication of livestock represented a crucial step in human history. By using endogenous retroviruses as genetic markers, we found that sheep differentiated on the basis of their “retrotype” and morphological traits dispersed across Eurasia and Africa via separate migratory episodes. Relicts of the first migrations include the Mouflon, as well as breeds previously recognized as “primitive” on the basis of their morphology, such as the Orkney, Soay, and the Nordic short-tailed sheep now confined to the periphery of northwest Europe. A later migratory episode, involving sheep with improved production traits, shaped the great majority of present-day breeds. The ability to differentiate genetically primitive sheep from more modern breeds provides valuable insights into the history of sheep domestication.

The first agricultural systems, based on the cultivation of cereals, legumes, and the rearing of domesticated livestock, developed within Southwest Asia ~11,000 years before present (yr B.P.) (1, 2). By 6000 yr B.P., agro-pastoralism introduced by the Neolithic agricultural revolution became the main system of food production throughout prehistoric Europe, from the Mediterranean north to Britain, Ireland, and Scandinavia (3); south into North Africa (4); and east into West and Central Asia (5).

Sheep and goats were the first livestock species to be domesticated (6). Multiple domestication events, as inferred by multiple mitochondrial lineages, gave rise to domestic sheep and similarly other domestic species (7–10). Initially, sheep were reared mainly for meat but, during the fifth millennium B.P. in Southwest Asia and the fourth millennium B.P. in Europe, specialization

for “secondary” products such as wool became apparent. Sheep selected for secondary products appear to have replaced more primitive domestic populations. Whether specialization for secondary products occurred first in Southwest Asia or occurred throughout Europe is not known with certainty, owing to the lack of definitive archaeological evidence for the beginning of wool production (6, 11, 12).

For this study, we used a family of endogenous retroviruses (ERVs) as genetic markers to examine the history of the domestic sheep. ERVs result from the stable integration of the retrovirus genome (“provirus”) into the germline of the host (13) and are transmitted vertically from generation to generation in a Mendelian fashion. The sheep genome contains at least 27 copies of ERVs related to the exogenous and pathogenic Jaagsiekte sheep retrovirus (enJSRVs) (14–16). Most enJSRVs loci are fixed in domestic sheep,

but some are differentially distributed between breeds and individuals (i.e., they are insertionally polymorphic) (14). enJSRVs can be used as highly informative genetic markers because the presence of each endogenous retrovirus in the host

<sup>1</sup>Institute of Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Glasgow G61 1QH, UK.

<sup>2</sup>Dipartimento di Patologia e Clinica Veterinaria, Università degli Studi di Sassari, 07100 Sassari, Italy.

<sup>3</sup>Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Faculdade de Ciências da Universidade do Porto, 4200-465 Porto, Portugal.

<sup>4</sup>Área de Genética y Reproducción Animal, SERIDA-Somío, E-33203 Gijón, Spain.

<sup>5</sup>Division of Archaeological, Geographical and Environmental Sciences, University of Bradford, Bradford BD7 1DP, UK.

<sup>6</sup>Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK.

<sup>7</sup>National Research Council (CNR), ISPAAM, 80147 Naples, Italy.

<sup>8</sup>Department of Biotechnology, Animal Science Research Institute of Iran (ASRI), 3146618361 Karaj, Iran.

<sup>9</sup>Department of Veterinary Medicine and Public Health, Sokoine University of Agriculture, Morogor, Tanzania.

<sup>10</sup>State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.

<sup>11</sup>Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, 24071 León, Spain.

<sup>12</sup>Genetic Engineering and Biotechnology Research Institute, Mubarak City for Scientific Research and Technology Applications, New Borg El-Arab City, 21934, Alexandria, Egypt.

<sup>13</sup>National Diagnostic Center for Exotic Animal Diseases, China Animal Health and Epidemiology Centers, Qingdao 266032, China.

<sup>14</sup>School of Health Science, Canakkale Onsekiz Mart University, 17100 Canakkale, Turkey.

<sup>15</sup>International Trypanotolerance Centre, Banjul, Gambia.

<sup>16</sup>Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, 0033 Oslo, Norway.

<sup>17</sup>Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, University of Aarhus, 8830 Tjele, Denmark.

<sup>18</sup>Department of Zoology, Institute of Ecology and Earth Sciences, University of Tartu, 51014 Tartu, Estonia.

<sup>19</sup>NWFP Agricultural University, Peshwar, Pakistan.

<sup>20</sup>All-Russian Research Institute of Animal Husbandry, Russian Academy of Agricultural Sciences, Dubrovitsy 142132, Russia.

<sup>21</sup>Agricultural University of Iceland, Hvanneyri, IS-311 Borgarnes, Iceland.

<sup>22</sup>Medical Research Council Laboratories, Fajara, Banjul, Gambia.

<sup>23</sup>London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.

<sup>24</sup>Istituto di Zootecnica, Università Cattolica del Sacro Cuore, 29100 Piacenza, Italy.

<sup>25</sup>School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK.

<sup>26</sup>Biotechnology and Food Research, MIT Agrifood Research Finland, 31600 Jokioinen, Finland.

<sup>27</sup>Center for Animal Biotechnology and Genomics, Texas A&M University, College Station, TX 77843, USA.

\*To whom correspondence should be addressed. E-mail: m.palmarini@vet.gla.ac.uk