

Supplementary Methods

Sample

We analysed data from 19,900 individuals belonging to 70 local cattle breeds (*Bos taurus*) from Turkey to Portugal and across Northern Europe (Fig. 1A). Great care was taken to include only local native (autochthonous) breeds and to avoid recently improved breeds (e.g., Holstein-Friesian). The cattle breeds were: Aboundance, Alentejana, Alistana, Anatolian Black, Arouquesa, Asturiana, Aubrac, Avileña, Barrosã, Black Pied Aosta, Black Pied Danish, Black Pied Jutland, Blonde d'Aquitaine, Braunvieh, Breton Pie-Noir, Brown Swiss, Bu_a, Cachena, Charolais, Chianina, Danish Red, Døle, East Finncattle, Flamande, Fleckvieh, Gelbvieh, Grigia Alpina, Hinterwälder, Hungarian Grey Steppe, Limpurger, Limousin, Mertolenga, Maremmana, Marinhova, Maronesa, Minhota, Mirandesa, Montbéliarde, Mostrenga, Normande, North Finncattle, Oropa, Partenaise, Podolica, Piemontese, Pinzgauer, Polish Red Cattle, Red Pied Aosta, Red Polled Østland, Red Pied Swedish, Reggiana, Rotbunte, Rotvieh, Rodopsk, Rubia Galega, Salers, Sayaguesa, Schwarzbunte, Swedish Red Pooled, Tarantaise, Telemark, Trondheim, Turkish Grey Steppe, Ukrainian Grey Steppe, Vestland, Vestland Fjord, West Finncattle, Varzese, Vorderwälder, Vosgienne. Sample size per breed averaged 284 individuals and ranged from 31 to 1512. The NCE group consists of 19 breeds from France (Flamande, Vosgienne, Montbéliarde, Tarantaise), Germany (Braunvieh, Fleckvieh, Gelbvieh, Hinterwälder, Rotbunte, Rotvieh, Schwarzbunte, Vorderwälder), Denmark (Black Pied Danish, Danish Red, Black Pied Jutland), Sweden (Red Pied Swedish, Swedish Red Pooled), Norway (Døle), Poland (Polish Red Cattle).

Genotyping procedures and statistical analysis

Sixty percent of the breeds (forty-two) were genotyped for this study (see Table 1, supplemental material) and the other data were assembled from published studies¹⁻³. Genotyping was conducted using the following methods: sequencing, PCR-SSCP, PCR-RFLP, and IEF. Allelic frequencies and sample sizes per breed are summarised in supplementary Table 1. The six genes code for the following proteins: the caseins - α_{s1} -casein (*CSN1S1*), β -casein (*CSN2*), α_{s2} -casein (*CSN1S2*) and *k*-casein (*CSN3*) - and the whey proteins - β -lactoglobulin (*LGB*) and α -lactalbumin (*LAA*). The regions screened are located on three chromosomes, covering 10 exons, comprising 17 single nucleotide polymorphisms (SNP's) and one 16 base-pairs deletion, across these six genes (for details see Supplementary Table 2). The NCE breeds contain six coding-gene variants (amino-acid replacement substitutions across five gene loci) that were not found in other breeds from Europe or Turkey. They have twice the number of alleles per locus (4.1, on average) compared to breeds from the Mediterranean and Near East (Turkish) regions (2.3; $P < 0.01$, Wilcoxon signed-ranks test). In contrast, only a single unique allele was observed in 12 of 51 non-NCE breeds (at the α -lactalbumin locus in breeds from Mediterranean Europe and Turkey). Much of this high genetic distinctiveness of NCE breeds results from allele frequency inversion i.e. alleles with high frequency in the NCE show a significantly lower frequency outside and *vice-versa* ($P < 0.01$ at each of four – *CSN1S1*, *CSN2*, *CSN3*, *LGB*- of the six loci; χ^2 permutation test).

Geostatistical analysis

To determine and visualise the geographical distribution of genetically distinct groups of cattle breeds, we constructed a European synthetic map based on a Principal Components Analysis (PCA). This graphical approach has been recently applied to domestic animals⁴. For the first and second principal components (that represent 36% of the total variance), interpolation maps were drawn using the “Spatial Analyst Extension” of ArcView version 3.2 (Fig. 1B, Fig. 2), available at www.esri.com/software/arcview/. The “Inverse Distance Weighted” (IDW) option with a power of two was used for the interpolation of the surface. IDW assumes that each input point has a local influence that diminishes with distance. The geographic location used is the centre of the distribution area (geo-reference), from where the individual samples of each breed were collected. The factor scores from the PCA were grouped in seven equal groups as depicted by the seven colours.

To assess the statistical significance of breed groupings, a Tukey-multiple comparison test (a form of a general linear model) was used to compare allele frequency distributions on breeds located in the different coloured isopleths. We also tested for allele frequency differences between pairs of randomly chosen breeds from different isopleths (coloured regions) using a χ^2 - test corrected for small counts in some allele classes. In all cases, statistical significance was obtained between breeds from the two extreme groups (light yellow versus red; $P < 0.05$). Comparisons between breeds from all the intermediate groups gave nearly-significant or non-significant test results. The frequencies for the human lactase persistence allele were gathered from the literature⁵⁻⁷ and were used to construct the synthetic map using the same software and same interpolation option for the surface as for the cattle milk protein genes.

Correlation tests

Correlations between PC1 scores, lactase persistence frequencies and Neolithic Pastoralist distribution⁸ (Funnel Beaker culture) were estimated using the *Spearman* coefficient, which is the most appropriate (non parametric) method to test for correlation between non-normally distributed variables (i.e. presence/absence of the Funnel Beaker culture). Standard formulas were used for calculating the *F*-value associated with the multiple *r*, and for the *t*-values associated with the regression coefficients.

$$t = r \sqrt{\frac{n'-2}{1-r^2}}$$

Spatial autocorrelation correction

Autocorrelation does not affect the value of a correlation coefficient⁹, but increases the probability of Type I errors (because points are not independent, so degrees of freedom are overestimated). Taking this into account, we correct for spatial autocorrelation using the Dutilleul method¹⁰ as implemented by Legendre (2000)¹¹, that computes a reduced sample size (and degrees of freedom) that can be applied directly to a *t*-test used for testing the significance of a correlation coefficient. This method is recommended¹² to correct for autocorrelation, even when some structure is present in the geographic distributions of the values (i.e. gradients). This correction depends on the distance classes used in the autocorrelation analysis, and the range of degrees of freedom refers to different analysis using from 6 to 12 classes, with either equal number of comparisons or equal intervals.

Selection tests

We performed tests for selection at each locus in each breed using the Ewens-Watterson, Tajima's D and Fu's F_s tests in ARLEQUIN (S. Scheneider, D. Roessli, L. Excoffier, Arlequin Ver. 2000: A software for population genetics data analysis, Genetics and Biometrics Laboratory, University of Geneva, Switzerland).

Acknowledgements

We thanks to F. Goyache and all the breeder associations for help in sampling collection.

References

1. Lien, S. *et al.* Comparison of milk protein allele frequencies in Nordic cattle breeds. *Anim. Genet* **30**, 85-90 (1999).
2. Grosclaude, F. Le polymorphisme génétique des principales lactoprotéines bovines. *INRA Prod Anim* **1**, 5-17 (1988).
3. DiStasio, L. in *Austrian Society of Animal Production. A. S. Branch* 54-66. (1993).
4. Hanotte, O. *et al.* African pastoralism: genetic imprints of origins and migrations. *Science* **296**, 336-339 (2002).
5. Schrimshaw, N. S. & Murray, E. B. The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. *Am J Clin Nutr* **48**, 1059-1179 (1988).
6. Kozlov, A. I. & Lisitsyn, D. M. in *Archaeogenetics: DNA and the Population History of Europe* (eds. Renfrew, C. & Boyle, K.) 309-20. (MacDonald Institute Cambridge, Cambridge, 2000).
7. Hollox, E. J. *et al.* Lactase haplotype diversity in the old world. *Am J Hum Genet* **68**, 160-172 (2001).
8. Midgley, M. S. *TRB culture: the first farmers of the North European Plain* (Edinburgh University Press, Edinburgh, 1992).
9. Cliff, A. & Ord, J. *Spatial processes: models and applications* (Pion, London, 1981).
10. Dutilleul, P. Modifying the T-test for assessing the correlation between 2 spatial processes. *Biometrics* **49**, 305-314 (1993).
11. Legendre, P. *Modtest* (University of Montreal, Montreal, 2000)
12. Perry, J. N. *et al.* Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography* **25**, 578-600 (2002).