nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0056-5

Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals

Aniek C. Bouwman[1], Hans D. Daetwyler[2,3], Amanda J. Chamberlain[2], Carla Hurtado Ponce[2,4], Mehdi Sargolzaei[5,6], Flavio S. Schenkel[5], Goutam Sahana[7], Armelle Govignon-Gion[8], Simon Boitard[9], Marlies Dolezal[10], Hubert Pausch[2,11,12], Rasmus F. Brøndum[7], Phil J. Bowman[2], Bo Thomsen[9], Bernt Guldbrandtsen[7], Mogens S. Lund[7], Bertrand Servin[13], Dorian J. Garrick[14], James Reecy[14], Johanna Vilkki[15], Alessandro Bagnato[16], Min Wang[2,3], Jesse L. Hoff[17], Robert D. Schnabel[17], Jeremy F. Taylor[17], Anna A. E. Vinkhuyzen[18,19], Frank Panitz[9], Christian Bendixen[9], Lars-Erik Holm[9], Birgit Gredler[20], Chris Hozé[8,21], Mekki Boussaha[8], Marie-Pierre Sanchez[8], Dominique Rocha[8], Aurelien Capitan[8,21], Thierry Tribout[8], Anne Barbat[8], Pascal Croiseau[8], Cord Drögemüller[22], Vidhya Jagannathan[22], Christy Vander Jagt[2], John J. Crowley[23], Anna Bieber[24], Deirdre C. Purfield[25], Donagh P. Berry[25], Reiner Emmerling[26], Kay-Uwe Götz[26], Mirjam Frischknecht[20], Ingolf Russ[27], Johann Sölkner[28], Curtis P. Van Tassell[29], Ruedi Fries[11], Paul Stothard[30], Roel F. Veerkamp[1], Didier Boichard[8], Mike E. Goddard[2,4] and Ben J. Hayes[2,31]*

Stature is affected by many polymorphisms of small effect in humans[1]. In contrast, variation in dogs, even within breeds, has been suggested to be largely due to variants in a small number of genes[2,3]. Here we use data from cattle to compare the genetic architecture of stature to those in humans and dogs. We conducted a meta-analysis for stature using 58,265 cattle from 17 populations with 25.4 million imputed whole-genome sequence variants. Results showed that the genetic architecture of stature in cattle is similar to that in humans, as the lead variants in 163 significantly associated genomic regions ($P < 5 \times 10^{-8}$) explained at most 13.8% of the phenotypic variance. Most of these variants were noncoding, including variants that were also expression quantitative trait loci (eQTLs) and in ChIP-seq peaks. There was significant overlap in loci for stature with humans and dogs, suggesting that a set of common genes regulates body size in mammals.

Within each cattle population (Supplementary Table 1), the 1000 Bull Genomes Run4 reference population of 1,147 whole-genome-sequenced individuals was used to impute 630,000 SNP genotypes to 25.4 million whole-genome sequence variants[4]. A genome-wide association study (GWAS) for stature was performed in each population separately (Supplementary Table 1)[5,6]. Meta-analysis across

[1]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands. [2]AgriBio, Centre for AgriBioscience, Department of Economic Development, Jobs, Transport and Resources, Bundoora, Victoria, Australia. [3]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia. [4]Faculty of Land and Food Resources, University of Melbourne, Parkville, Victoria, Australia. [5]Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada. [6]The Semex Alliance, Guelph, Ontario, Canada. [7]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. [8]GABI, INRA, AgroParisTech, Université Paris Saclay, Jouy-en-Josas, France. [9]Section for Molecular Genetics and Systems Biology. Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark. [10]Platform of Bioinformatics and Statistics, University of Veterinary Medicine, Vienna, Austria. [11]Chair of Animal Breeding, Technische Universität München, Freising-Weihenstephan, Germany. [12]Animal Genomics, ETH Zurich, Zurich, Switzerland. [13]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet-Tolosan, France. [14]Department of Animal Science, Iowa State University, Ames, IA, USA. [15]Green Technology, Natural Resources Institute Finland (Luke), Jokioinen, Finland. [16]Department of Veterinary Medicine, University of Milan, Milan, Italy. [17]Division of Animal Sciences, University of Missouri, Columbia, MO, USA. [18]University of Queensland, Institute for Molecular Bioscience, St Lucia, Queensland, Australia. [19]University of Queensland, Queensland Brain Institute, St Lucia, Queensland, Australia. [20]Qualitas AG, Zug, Switzerland. [21]Allice, Paris, France. [22]Institute of Genetics, University of Bern, Bern, Switzerland. [23]Canadian Beef Breeds Council, Calgary, Alberta, Canada. [24]Research Institute of Organic Agriculture (FiBL), Frick, Switzerland. [25]Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Ireland. [26]Institute of Animal Breeding, Bavarian State Research Centre for Agriculture, Poing, Germany. [27]Tierzuchtforschung, Poing, Germany. [28]University of Natural Resources and Life Sciences, Vienna, Austria. [29]Animal Genomics and Improvement Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, MD, USA. [30]Department of Agricultural, Food and Nutritional Science/Livestock Gentec, University of Alberta, Edmonton, Alberta, Canada. [31]Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, University of Queensland, St Lucia, Queensland, Australia. *e-mail: b.hayes@uq.edu.au

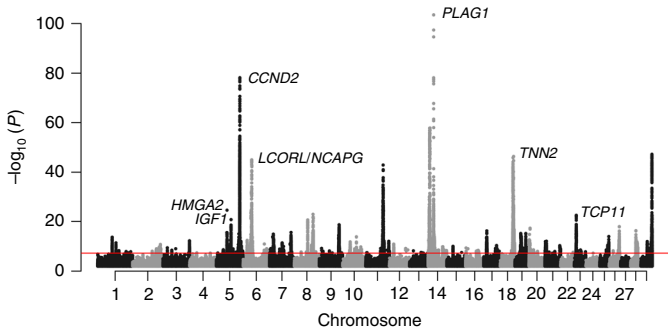

Fig. 1 | Manhattan plot for the meta-analysis of bovine stature with *n* = 58,265 animals. For each SNP, the meta-analysis calculated a *z* statistic (and *P* value) from the weighted sum of *z* statistics from GWAS in each of the 17 contributing cattle populations, with weights proportional to the square root of the number of cattle in each GWAS (Supplementary Table 1)[37]. The red line is the genome-wide significance threshold at $P = 5 \times 10^{-8}$. The most likely candidate genes in the most significantly associated regions are annotated where an obvious candidate could be identified. SNPs on odd-numbered chromosomes are in black and those on even-numbered chromosomes are in gray.

the populations found genome-wide significant ($P < 5 \times 10^{-8}$) sequence variants in 163 1-Mb regions (Fig. 1). The lead variants (most significant variants in each region) included 160 SNPs and 3 indels (Supplementary Table 2).

Three approaches were used to validate the lead variants. Association of the variants with stature was tested in 30,175 additional cattle with stature phenotypes from ten populations (Supplementary Table 3). In meta-analysis of these validation populations, 20 of 101 SNPs (the lead variants polymorphic in all populations) were validated at $P < 0.05$, giving a false discovery rate of 25% (Supplementary Table 4). SNPs were also validated within each population, as some variants were polymorphic in one or only a small number of breeds. The majority of variants (53%; 86) were validated in at least one population, and 28 were validated in more than one population (11 expected by chance) (Supplementary Table 4). The lead variants explained between 2.1% (Limousin) and 13.8% (Brown Swiss) of the phenotypic variation in stature (Table 1), significantly more than that explained by a random subset of the

Table 1 | Proportion of phenotypic variation explained by 163 lead variants in validation populations

Breed	Country	No. of animals	No. of lead SNPs polymorphic	Proportion of phenotypic variation explained by lead SNP
Simmental	Ireland	1,913	146	0.052
Limousin	Ireland	10,371	150	0.021
Hereford	Ireland	595	137	0.027
Charolais	Ireland	7,822	145	0.024
Angus	Ireland	732	139	0.039
Angus	Australia	676	125	0.054
Brown Swiss	Switzerland	5,550	160	0.138
Holstein	Australia	1,565	141	0.093

For Angus (Australia), Holstein (Australia) and Brown Swiss (Switzerland), we compared the proportion of variance explained by the lead SNP to the average variance explained by random subsets of 163 variants; this was 0.016 ± 0.003, 0.036 ± 0.004 and 0.119 ± 0.009, respectively.

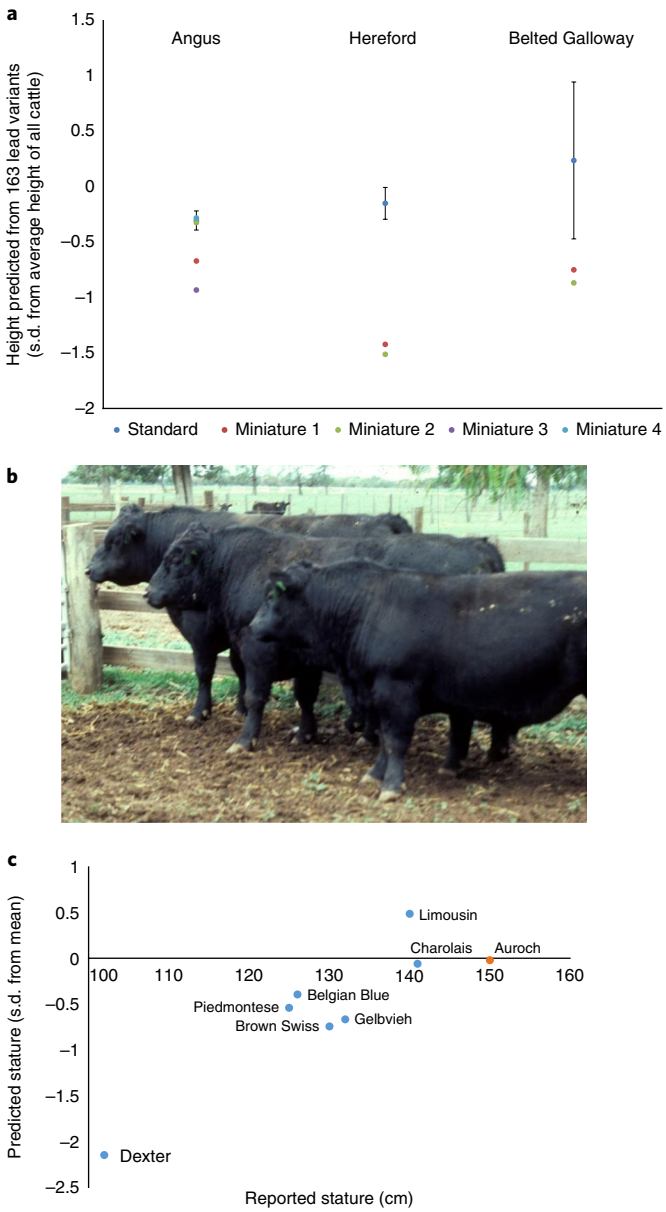

Fig. 2 | Validation of lead variants. a, The 163 lead variants predict differences within breeds between miniature and standard cattle. There were *n* = 4 miniature Angus animals, *n* = 2 miniature Hereford animals and *n* = 2 miniature Belted Galloway animals sequenced together with *n* = 48 standard Angus animals, *n* = 30 standard Hereford animals and *n* = 2 standard Belted Galloway animals. For miniature cattle, individual predicted height is plotted. For standard cattle, the values plotted are average predicted height, and standard errors are the s.d. of predicted height divided by the square root of the number of standard cattle for each breed. The average heights of standard and miniature cattle are approximately 116 cm and 108 cm; 120 cm and 105 cm; and 120 cm and 110 cm for Angus, Belted Galloway and Hereford animals, respectively[38–41]. **b**, Standard and miniature Angus cattle. The three animals (back to front) correspond to animals from a selected high-growth line, a control line and a low-growth line. The low-growth line is the origin of miniature Angus cattle. Photo courtesy of R. Herd and P. Arthur (NSW Department of Primary Industries, Australia). **c**, Predicted average stature of seven breeds (not included in the original meta-analysis), where stature was predicted from the 163 lead SNPs and their effects, as compared to average reported stature for these breeds. The average reported stature was from three-breed comparison studies[39–41]. Standard errors of breed average reported statures were approximately 6 cm.

same number of variants where tested. This amount of variation is of a similar magnitude to the proportion of phenotypic variance explained by variants significantly associated with height in humans (~16%)[1]. The results are substantially different from those in dogs, where six loci have been reported to explain the majority of variance in body size[2]. However, the analyses in dogs have largely been across breeds, rather than within breeds (with one exception[3]). We estimated the proportion of variance accounted for by 17 loci previously identified in across-dog-breed analyses within a population of village dogs[3], correcting for population structure and sex, and found that these 17 loci explained 13.5% of the variation in body size. This is of similar magnitude to the proportion of variance explained within cattle breeds by all 163 lead variants in cattle, suggesting that there may be some loci of larger effect in dogs.

For the second validation approach, we exploited the fact that, for a number of cattle breeds, there are miniature cattle that are several s.d. smaller in stature than standard cattle, from recent strong selection. These animals are miniatures rather than dwarfs, as they do not display chondrodysplasia. In all three breeds where we had genome sequence data from standard cattle and miniatures, a prediction equation comprising the effects of the 163 lead variants correctly predicted that the miniature animals had substantially shorter stature, for all but one of the miniature animals (Fig. 2a,b). In the third validation approach, the same equation accurately predicted differences in stature between seven breeds that had sequence data but were not included in the meta-analysis ($r^2 = 0.80$; Fig. 2c).

The most significant variant in the meta-analysis was a SNP in intron 3 of *PLAG1* (AC_000171.1:g.25015640G>T, rs109815800, association with stature $P < 1 \times 10^{-104}$) on BTA14, one of eight putative causative mutations previously identified in or close to this gene[7]. PLAG1 initiates transcription of *IGF2*, a mitogenic hormone important for fetal growth and development, and has been implicated in genetic variation of stature in humans as well as cattle[1,7–9]. In the population used by Karim et al.[7], the eight candidate variants were in perfect linkage disequilibrium (LD). In our study with additional breeds and more animals, these SNPs were not in complete LD (Supplementary Table 5) and SNP rs109815800 was more strongly associated with stature ($P < 1 \times 10^{-104}$) than the others proposed[7]. These results demonstrate the power of the meta-analysis conducted here to directly identify a small number of variants as putative causative mutations. While GWAS analysis with imputed sequence data has identified causal mutations in cattle, imperfect imputation (Supplementary Figs. 1–5) may result in the causal mutation not being identified as the most highly associated variant, especially if the variant is rare. Here the accuracy of imputation was >0.9 for variants with a minor allele frequency (MAF) >0.10 and for most of the variants in the *PLAG1* region (Supplementary Table 5; note that the rs109815800 variant was among those genotyped on the 630K array in some populations)[10].

To investigate what type of variants affect stature in cattle, genome annotation, eQTL and ChIP–seq data were used. These analyses depend on at least an enrichment of our lead variants for causative mutations. Bootstrap resampling suggested that a reasonable proportion of our variants could be causal and not merely linked with the causal variant (Supplementary Table 6). Of the 163 lead variants, 5 were missense, representing a sevenfold enrichment of missense variants among the lead variants as compared with what would be expected by chance (Table 2). The missense variants included one in *HMGA2*, a well-documented stature-associated gene in humans. The protein encoded by *HMGA2* regulates the RNA-binding protein IGF2BP2 (IGF2-binding protein 2), which in turn enhances translation of the *IGF2* gene[11]. Another missense variant was found in *LCOR* (ligand-dependent co-repressor), which is broadly expressed in fetal and adult tissues to regulate development and homeostasis[12–14]. In many species, including humans, mice and rats (and cattle, this study), a small genomic region that includes *LCORL* (ligand-dependent nuclear receptor co-repressor like) and *NCAPG* (non-SMC condensin I complex subunit G) is associated with variation in height and body size[1,15]. Determining which of these two genes is responsible for variability in height has not been possible because of the close proximity of these genes and the high levels of LD among SNPs in these regions (also observed in this study). The identification in our study of a missense variant in *LCOR*, a gene with very high homology and potentially similar function to *LCORL*, as being associated with stature provides some evidence supporting *LCORL* as the causative gene in other species.

The majority of lead variants from the 163 stature-associated regions were not coding variants (Table 2), consistent with observations from GWAS for height in humans[16]. Eight of the 83 intergenic variants were located within bovine ChIP–seq peaks, more than expected by chance. ChIP–seq peaks were identified from H3K27 acetylation and H3K4 trimethylation histone modification assays of bovine liver, which indicates that these variants may be in enhancers, repressors or promoters[17].

To further investigate the hypothesis that many of our lead variants are regulatory, we performed an eQTL study using RNA-seq data from white blood cells (WBCs) in 93 Holstein cows. Although gene expression in fetal tissue would be more informative than that in blood from mature cows for this study, recent evidence suggests a reasonable overlap of eQTLs across tissues[18].

Table 2 | Annotation of the most significant sequence variants in 163 genomic regions affecting stature in cattle

Annotation class	No. of lead variants	Proportion of lead variants	Proportion of all variants in genome with this annotation[c]	Fold enrichment/ depletion	*P* value[d]
Intergenic variant	83	0.459	0.663	0.69	0.63
Upstream gene variant	11	0.061	0.035	1.74	0.33
5′ UTR variant	1	0.006	0.0004	15.00	0.0002
Intron variant	55	0.304	0.261	1.16	0.59
Missense variant	5	0.028	0.004	7.00	0.01
Downstream gene variant	8	0.044	0.030	1.47	0.43
ChIP-seq peaks[a]	8	0.044	0.024	1.85	0.049
WBC eQTL[b]	10	0.055	0.003	18.33	0.00001

Annotations of the 163 lead SNP, the proportion of all sequence variants in 1000 Bull Genomes Run4 with the corresponding annotation, the level of enrichment or depletion for this annotation class in the 163 lead variants, and the significance of enrichment/depletion from a chi-squared test. [a]ChIP-seq peaks identified from H3K27 acetylation and H3K4 trimethylation histone modification assays of bovine liver[17]. [b]See "White blood cell eQTLs" in the Methods for details. [c]From Run4 of 1000 Bull Genomes. [d]Based on a chi-squared test comparing the observed and expected number of variants in each class, with 1 degree of freedom.

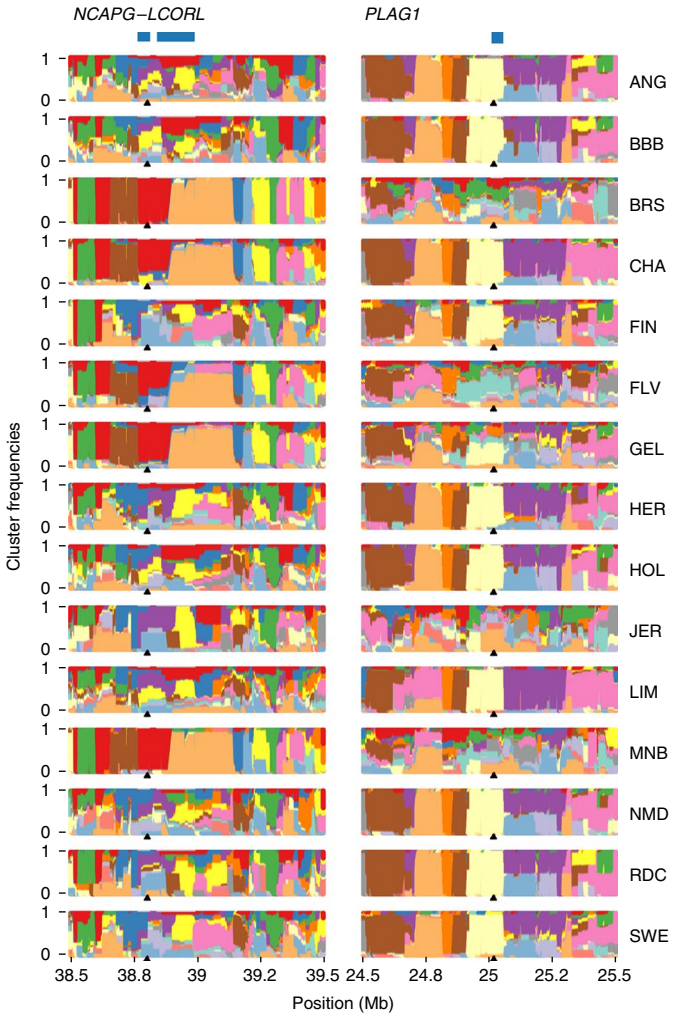

Fig. 3 | Haplotype diversity for 15 cattle breeds in two genomic regions (*NCAPG–LCORL* and *PLAG1*) where selection signatures match segregation of QTLs for stature. In each panel, each color represents a local haplotype cluster. The *PLAG1* gene is located on chromosome 14 at 25,007,291–25,009,296 bp, *NCAPG* is located on chromosome 6 at 38,765,969–38,812,051 bp and *LCORL* is located on chromosome 6 at 38,840,894–38,992,112 bp. The blue bars indicate the positions of these genes. At each position in the panels, the height of the color band represents the frequency of the corresponding haplotype in the population; the different colors represent different haplotypes[35]. For example, Angus is nearly fixed for the yellow haplotype at *PLAG1*, while Gelbvieh segregates for a number of different haplotypes. ANG, Angus; BBB, Belgian Blue; BRS, Brown Swiss; CHA, Charolais; FIN, Finnish Ayrshire; FLV, Fleckvieh; GEL, Gelbvieh; HER, Hereford; HOL, Holstein; JER, Jersey; LIM, Limousin; MNB, Montbeliard; NMD, Normande; RDC, Danish Red; SWE, Swedish Red.

Ten lead stature-associated variants were also eQTLs in WBCs, representing an 18-fold enrichment over the number expected by chance (Table 2 and Supplementary Table 2). It is possible that the genome regions containing the lead variants may be enriched in eQTLs even if these eQTLs are functionally unrelated to stature, owing to non-random clustering of genes. We assessed evidence for a functional relationship (either pleiotropy or causality) with the HEDI (heterogeneity in dependent instruments) test[19]. Seven of the ten eQTL/stature-associated variants showed no heterogeneity of effect with LD, suggesting that these variants could be causal for both gene expression levels and stature or pleiotropic for these traits. One such variant, AC_000161.1:.g.32075456C>T associated ($P < 1 \times 10^{-5}$) with expression of *IGF2BP3* (insulin-like growth factor 2–binding protein 3), is an interesting candidate, as the IGF2BP3 protein suppresses translation of the *IGF2* gene during late fetal development[20–24]. The direction of effect for the variant was consistent with this mechanism—the allele associated with increased expression of *IGF2BP3* was associated with decreased bovine stature.

We next investigated whether there was greater overlap of loci affecting stature in cattle and humans than would be expected by chance. Of the 92 genes overlapping or within ±5 kb of the 163 lead variants, 11 were identified by Wood et al.[1] as affecting stature in humans (Supplementary Table 2), more than expected by chance ($P < 1 \times 10^{-12}$, chi-squared test). This test is stringent, as it requires the lead variant to be within or very close to the causal gene. QTL confidence regions (Supplementary Table 2) overlapped with 26 of the genes identified as associated with stature or body size in humans and/or dogs (Supplementary Table 2). For example, variants in *GHR*, *HMGA2*, *SMAD2*, *STC2*, *IGF1* and *IGF1R* are strongly associated with differences in size between dog breeds; of these genes, only *GHR* and *SMAD2* were not found within the QTL confidence intervals in our study[3,25].

Considering many of the lead stature-associated variants were only segregating in one or two breeds (Supplementary Fig. 6), an interesting question arises as to whether the stature-associated variants are recent mutations (for example, arising after breed formation) or ancient standing variation recently fixed by selection or drift in some breeds. Aurochs were the wild ancestor of modern cattle. We investigated both the heterozygosity of our lead variants and stature prediction using the sequence of a 6,750-year-old Auroch genome[26]. Of the 163 lead variants, 134 had six or more reads covering the variant position and so could be called. Of these, 31 were heterozygous. This result (close to the expectation for one animal if all lead variants were segregating in the population) indicates that many of the lead variants arose before domestication and certainly before breed formation (although it must be noted that only a proportion of our lead variants might be actual causal mutations). Interestingly, the predicted stature of the Auroch from our lead variants was greater than that for all but one of the modern breeds (Fig. 2c), consistent with the large skeletal size of Aurochs from the fossil record[27]. The hypothesis that most of the genomic variation affecting stature is ancient standing variation rather than recent mutations is supported by the fact that, even for some of the variants with the largest effects, it is the ancestral allele rather than the derived allele that has the effect of increasing stature (Supplementary Table 2). The observation that some variants with an ancestral allele that increases stature still segregate in multiple breeds may also be due to the direction of selection for stature not being consistent across cattle breeds (effectively balancing the effects of selection). As cattle were domesticated, there was selection for reduced stature in comparison to wild Auroch populations (directly, as a correlated consequence to selection for early sexual maturity, or both), as evidenced by bone lengths of ancient domestic versus contemporaneous wild cattle[28,29]. Selection for reduced stature continued until at least the fifteenth century, when northern European cattle measured less than 1 m in stature[28,30]. More recently, there has been very strong selection for increased stature in some breeds, with Holstein, Brown Swiss and Fleckvieh all increasing in stature by approximately 2 mm per year over the last decade[31–33].

Additional evidence that sequence variants affecting stature have been subject to selection since domestication and breed formation comes from the finding that nearly 50% of the 163 variants are in selection signatures identified in 1000 Bull Genomes sequences[34,35], representing a 30-fold enrichment as compared to random SNPs (Supplementary Fig. 7). Selection for stature is exemplified by selective sweeps for the same haplotype in five breeds for *NCAPG–LCORL* and in ten breeds for *PLAG1* (Fig. 3). The *PLAG1* allele

that increases stature is almost fixed in tall breeds (for example, Limousin, Charolais and Holstein), while in breeds of shorter stature the degree of fixation is variable (Jersey, Brown Swiss, Angus, Montbeliarde and Fleckvieh).

Our results show that the genetic architecture of stature within cattle breeds is highly polygenic, similar to the genetic architecture of stature observed in humans (and other complex traits in cattle[36]). Results of a new analysis of village dogs indicate that, within dog breeds, a larger number of loci are likely to be required to explain variation in body size than previously reported. Many of the loci associated with stature are shared across the three species, supporting the hypothesis that there are numerous common genes that regulate body size in mammals. These common genes include a striking number of regulators of expression of the *IGF2* gene, which encodes a key hormone for fetal growth and development.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0056-5.

Received: 24 October 2016; Accepted: 3 January 2018;
Published online: 19 February 2018

References

1. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
2. Rimbault, M. et al. Derived variants at six genes explain nearly half of size reduction in dog breeds. *Genome Res.* **23**, 1985–1995 (2013).
3. Hayward, J. J. et al. Complex disease and phenotype mapping in the domestic dog. *Nat. Commun.* **7**, 10460 (2016).
4. Daetwyler, H. D. et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* **46**, 858–865 (2014).
5. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
6. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
7. Karim, L. et al. Variants modulating the expression of a chromosome domain encompassing *PLAG1* influence bovine stature. *Nat. Genet.* **43**, 405–413 (2011).
8. Pryce, J. E., Hayes, B. J., Bolormaa, S. & Goddard, M. E. Polymorphic regions affecting human height also control stature in cattle. *Genetics* **187**, 981–984 (2011).
9. Fortes, M. R. et al. Evidence for pleiotropism and recent selection in the *PLAG1* region in Australian Beef cattle. *Anim. Genet.* **44**, 636–647 (2013).
10. Pausch, H. et al. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. *BMC Genomics* **18**, 853 (2017).
11. Li, Z. et al. An HMGA2–IGF2BP2 axis regulates myoblast proliferation and myogenesis. *Dev. Cell* **23**, 1176–1188 (2012).
12. Fernandes, I. et al. Ligand-dependent nuclear receptor corepressor LCoR functions by histone deacetylase–dependent and –independent mechanisms. *Mol. Cell* **11**, 139–150 (2003).
13. Calderon, M. R. et al. Ligand-dependent corepressor (LCoR) recruitment by Kruppel-like factor 6 (KLF6) regulates expression of the cyclin-dependent kinase inhibitor *CDKN1A* gene. *J. Biol. Chem.* **287**, 8662–8674 (2012).
14. Calderon, M. R. et al. Ligand-dependent corepressor contributes to transcriptional repression by C2H2 zinc-finger transcription factor ZBRK1 through association with KRAB-associated protein-1. *Nucleic Acids Res.* **42**, 7012–7027 (2014).
15. Kemper, K. E., Visscher, P. M. & Goddard, M. E. Genetic architecture of body size in mammals. *Genome Biol.* **13**, 244 (2012).
16. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
17. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
18. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
19. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
20. Akhtar, M. et al. Cell type and context-specific function of PLAG1 for *IGF2* P3 promoter activity. *Int. J. Oncol.* **41**, 1959–1966 (2012).
21. DeChiara, T. M., Efstratiadis, A. & Robertson, E. J. A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. *Nature* **345**, 78–80 (1990).
22. Voz, M. L., Agten, N. S., Van de Ven, W. J. & Kas, K. *PLAG1*, the main translocation target in pleomorphic adenoma of the salivary glands, is a positive regulator of IGF-II. *Cancer Res.* **60**, 106–113 (2000).
23. Nielsen, J. et al. A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Mol. Cell. Biol.* **19**, 1262–1270 (1999).
24. Reik, W. et al. *Igf2* imprinting in development and disease. *Int. J. Dev. Biol.* **44**, 145–150 (2000).
25. Sutter, N. B. et al. A single *IGF1* allele is a major determinant of small size in dogs. *Science* **316**, 112–115 (2007).
26. Park, S. D. et al. Genome sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography and evolution of cattle. *Genome Biol.* **16**, 234 (2015).
27. Clutton-Brock, J. *A Natural History of Domesticated Mammals* (Cambridge University Press, Cambridge, UK, 1987).
28. Vretemark, M. *From Bones to Livestock* (Stockholm University, City, 1997).
29. Manning, K., Timpson, A., Shennan, S. & Crema, E. Size reduction in early European domestic cattle relates to intensification of Neolithic herding strategies. *PLoS One* **10**, e0141873 (2015).
30. Svensson, E. M. et al. Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. *Anim. Genet.* **38**, 378–383 (2007).
31. Krogmeier, D. Zusammenhänge zwischen Nutzungsdauer und Körpergröße unter besonderer Berücksichtigung des Stallsystems bei Braunvieh und Fleckvieh. *Zuchtungskunde* **81**, 328–340 (2009).
32. Beavers, L. & Van Doormaal, B. A closer look at stature (CDN Report) (Publisher, City, 2016).
33. Laumay, A. & le Mezec, P. Bilan de l'indexation des races bovines laitieres. Resultats de la campagne 2014 (INRA Report 0015202017) (Publisher, City, 2015).
34. Bonhomme, M. et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* **186**, 241–262 (2010).
35. Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* **193**, 929–941 (2013).
36. Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J. & Hayes, B. J. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. Biol. Sci.* **283**, 1835 (2016).
37. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
38. Arthur, P. F., Parnell, P. F. & Richardson, E. C. Correlated responses in calf body weight and size to divergent selection for yearling growth rate in Angus cattle. *Livest. Prod. Sci.* **49**, 305–312 (1997).
39. Arango, J. A., Cundiff, L. V. & Van Vleck, L. D. Comparisons of Angus-, Braunvieh-, Chianina-, Hereford-, Gelbvieh-, Maine Anjou–, and Red Poll–sired cows for weight, weight adjusted for body condition score, height, and body condition score. *J. Anim. Sci.* **80**, 3133–3141 (2002).
40. Arango, J. A., Cundiff, L. V. & Van Vleck, L. D. Breed comparisons of Angus, Charolais, Hereford, Jersey, Limousin, Simmental, and South Devon for weight, weight adjusted for body condition score, height, and body condition score of cows. *J. Anim. Sci.* **80**, 3123–3132 (2002).
41. Arango, J. A., Cundiff, L. V. & Van Vleck, L. D. Comparisons of Angus, Charolais, Galloway, Hereford, Longhorn, Nellore, Piedmontese, Salers, and Shorthorn breeds for weight, weight adjusted for condition score, height, and condition score of cows. *J. Anim. Sci.* **82**, 74–84 (2004).

Acknowledgements

H.D.D., A.J.C., P.J.B. and B.J.H. would like to acknowledge the Dairy Futures Cooperative Research Centre for funding. H.P. and R.F. acknowledge funding from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr 'Synbreed—Synergistic Plant and Animal Breeding' (grant 0315527B). H.P., R.F., R.E. and K.-U.G. acknowledge the Arbeitsgemeinschaft Süddeutscher Rinderzüchter, the Arbeitsgemeinschaft Österreichischer Fleckviehzüchter and ZuchtData EDV Dienstleistungen for providing genotype data. A. Bagnato acknowledges the European Union (EU) Collaborative Project LowInputBreeds (grant agreement 222623) for providing Brown Swiss genotypes. Braunvieh Schweiz is acknowledged for providing Brown Swiss phenotypes. H.P. and R.F. acknowledge the German Holstein Association (DHV) and the Confederación de Asociaciones de Frisona Española (CONCAFE) for sharing genotype data. H.P. was financially supported by a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft (DFG) (grant PA 2789/1-1). D.B. and D.C.P. acknowledge funding from the Research Stimulus Fund (11/S/112) and Science Foundation Ireland (14/IA/2576). M.S. and F.S.S. acknowledge the Canadian Dairy Network (CDN) for providing the Holstein genotypes. P.S. acknowledges funding from the Genome Canada project entitled 'Whole Genome Selection through Genome Wide Imputation in Beef

Cattle' and acknowledges WestGrid and Compute/Calcul Canada for providing computing resources. J.F.T. was supported by the National Institute of Food and Agriculture, US Department of Agriculture, under awards 2013-68004-20364 and 2015-67015-23183. A. Bagnato, F.P., M.D. and J.W. acknowledge EU Collaborative Project Quantomics (grant 516 agreement 222664) for providing Brown Swiss and Finnish Ayrshire sequences and genotypes. A.C.B. and R.F.V. acknowledge funding from the public–private partnership 'Breed4Food' (code BO-22.04-011-001-ASG-LR) and EU FP7 IRSES SEQSEL (grant 317697). A.C.B. and R.F.V. acknowledge CRV (Arnhem, the Netherlands) for providing data on Dutch and New Zealand Holstein and Jersey bulls.

Author contributions

A.C.B. conducted the meta-analysis and contributed to writing the manuscript. H.D.D., A.J.C. and C.V.J. ran the 1000 Bull Genomes pipeline and extracted sequence variants, and A.J.C. and C.V.J. performed the eQTL analysis. C.H.P. sourced samples for miniature cattle and generated whole-genome sequence alignments for these. M.S., D.P.B., P.J.B. and F.S.S. contributed to genotype imputation and writing of the manuscript. M.S., F.S.S., G.S., D.C.P., H.P., J.V., B. Gredler, J.J.C., J.L.H. and R.F.B. performed GWAS analysis. S.B., B.S. and M.D. performed selection signature analysis. R.E. and K.-U.G. prepared daughter yield deviations and yield deviaitons of Fleckvieh animals, and the Intergenomics Consortium contributed genotypes. A.G.-G., C.H., M.-P.S., A.C., T.T., A. Bieber, P.C. and A. Barbat prepared phenotypes and genotypes for French cattle and ran GWAS. M.F., I.R. and J.S. prepared phenotypes and genotypes for Swiss and Austrian cattle and ran GWAS. A.A.E.V. contributed to across-species identification of stature-related genes. M.B., M.W., P.S., D.R., V.J. and R.D.S. performed variant annotation. B.J.H., D.J.G., J.F.T., C.B., J.R., A. Bagnato, F.P., B.T., L.-E.H., C.D., R.F., C.P.V.T., R.F.V., D.B., P.S., M.E.G., B. Guldbrandtsen and M.S.L. conceived the experimental design, analyzed stature data for contributed breeds and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0056-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.J.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Meta-analysis. Meta-analysis was performed on GWAS results from 17 populations that represented eight *Bos taurus* breeds. Within each population, animals were genotyped with either the Illumina Bovine SNP50v2.0 (50K SNP) or BovineHD (777k) SNP array (with the majority genotyped with the 50K array). Genotype calls with a GenTrain score (GenCall) <0.6 were excluded, including 55 SNPs with duplicate map positions. Approximately 630,000 SNPs remained depending on population for the HD SNP and 43K BovineSNP50v2 SNP arrays. Some SNPs were reordered on the basis of their LD-mapped position, as described by Erbe[42]. Imputation of animals genotyped for 43,000 SNPs to 630,000 SNPs was performed with Beagle[43], Minimac[44] or Fimpute[45] and was very accurate (>0.95, as assessed by cross-validation)[42].

All sequenced animals were used as a reference when imputing whole-genome sequence genotypes in each population.

Ancestral allele determination. To determine the ancestral allele, the following genome assemblies were used: (i) the cattle UMD3.1 reference genome sequence (Btau6 version) (see URLs); (ii) the bison (Bison_UMD1.0/bisBis1) genome assembly (bisBis1, University of Maryland) (see URLs); (iii) the sheep (*Ovis aries*) genome assembly (Oar_v3.1 version) (see URLs); (iv) the yak (*Bos grunniens*) genome assembly (Yak genome 1.1 version) (see URLs); and (v) the water buffalo (*Bubalus bubalis*) genome assembly (UMD_CASPUR_WB_2.0) (see URLs).

Pairwise alignments of the bovine genome sequence to the yak, water buffalo, bison and sheep genome sequences were carried out using the LASTZ sequence alignment program[46] (see URLs for LASTZ documentation).

LASTZ was run with the parameters --nogapped (skip gapped extension when doing alignment); --notransition (do not allow any match positions in seeds to be satisfied by transitions); --step = 20 (offset between the starting positions of successive target words considered for potential seeds); --format = maf (specifies the output format (the maf format in our study)).

A custom Python script was subsequently used to predict the yak, water buffalo, bison and sheep putative ancestral allelic state of the 164 SNPs[47]. The script is available from the authors upon request.

White blood cell eQTLs. 360 Holstein cows from the 'Novel Strategies to Breed Dairy Cattle for Adaptation and Reduced Methane Emissions' Australian project were sampled during a 3-year project, 120 cows per year, in three batches of 40 cows. Whole-blood cell samples were taken from all cows at the DEDJTR Ellinbank research facility at weeks 2 and 4 of the trial period, with approval from the DEDJTR Animal Ethics Committee (2013–14), as follows. Blood was collected by venipuncture of the coccygeal vein after routine morning milking and was processed according to the blood fractionation and WBC stabilization procedure in the protocol for the RiboPure Blood kit (Ambion by Life Technologies). Whole-blood cell samples were then transferred to the main laboratory on ice and stored at –20 °C.

RNA was extracted from WBCs using the RiboPure Blood kit according to the manufacturer's instructions. 112 Holstein cows were selected whose RNA integrity number was greater than 6, balancing for sire, number of lactations, days in milk and the sampling date. RNA-seq libraries were prepared using the SureSelect Strand Specific RNA Library Prep kit (Agilent) according to the manufacturer's instructions. Each library was uniquely barcoded and randomly assigned to one of four pools and sequenced on a HiSeq 3000 (Illumina) in a 150-cycle paired-end run. 150-base paired-end reads were called with bcltofastq and output in fastq format. Sequence quality was assessed using FastQC. QualityTrim (see URLs) was used to trim and filter out poor-quality bases and sequence reads. Adaptor sequences and bases with a quality score less than 20 were trimmed from the ends of reads. Reads were discarded with mean quality scores less than 20, more than three no calls (Ns), greater than three consecutive bases having a quality score less than 15 or a final length less than 50 bases. Only paired reads were retained for alignment.

Paired RNA reads for each sample were aligned to the UMD3.1 bovine genome assembly using TopHat2, allowing for two mismatches[48,49]. Computer scripts were used to assess sequencing performance, library quality and alignment quality; these scripts are available from the authors upon request. Alignment files (.bam) for WBC libraries with >12.5 million read pairs (after quality control filtering) and also having a mapping rate >80% were retained for gene count matrix generation. Gene counts for the aforementioned alignment files were created using the Python package HTSeq[50]. Counts were combined to form a gene-by–sample count matrix. This count matrix was then normalized to take into account library size using the R software package, DESeq[51].

Statistical analysis. *Meta-analysis.* Subsequently, GWAS was performed within each population on the imputed whole-genome sequence variants (SNPs and short insertions and deletions) using mixed linear models that included each population's genomic relationship matrix (GRM), which were constructed with at least 630,000 SNPs (BovineHD chip) to account for population stratification and familial relationships. Association was tested by linear regression of phenotypic measures on the number of copies of the alternate allele, assuming additive effects. More details about the populations and individual GWAS can be found in Supplementary Table 1.

Variant effect and standard error of the effect from the GWAS were standardized for each population by dividing them by the phenotypic s.d. The GWAS results from individual populations for variants with a minor allele frequency (MAF) <0.005 and/or an effect size of more than 5 s.d. from the mean were not included in the meta-analysis. In total, 58,265 animals were included in the meta-analysis of 25,406,107 variants, but the total sample size varied by variant. Meta-analysis was performed using the inverse-variance fixed-effects method in METAL with genomic control (for λ_{GC}, see Supplementary Table 1)[37].

A QTL was defined as a chromosomal region where adjacent pairs of significant variants were less than 1 Mb from each other. Within each locus, the most significant variant was taken as the lead variant. From the lead variant within such a locus, a more conservative QTL locus was defined on the basis of a $-\log_{10}$ (P value) drop off of 4, i.e., the difference between the $-\log_{10}$ (P value) of the lead variant and variants on either side moving further until all SNPs had a difference in $-\log_{10}$ (P value) from the lead SNP of greater than 4 (if the drop in $-\log_{10}$ (P value) was greater than 4, then decreased again, the procedure continued until all further SNPs had a difference in $-\log_{10}$ (P value) from the lead SNP of greater than 4). The maximum distance considered was 0.5 Mb on either side of the lead variant.

Validation. The 163 lead SNPs were validated in ten populations (Supplementary Tables 3 and 4). Phenotypes were corrected for fixed effects, including herd, age and year of measurement. Care was taken in selection of validation animals to ensure that none of the validation animals were the same as those used in the meta-analysis, nor were they full or half-siblings of these animals.

Sequence genotypes were imputed from 630,000 genotypes on all of the validation animals to test the significance of the SNPs. The model fitted within each population was

$$\boldsymbol{y} = 1_n\mu + \boldsymbol{X}b + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\mathbf{1}_n$ is a vector of ones, μ is the mean, $\boldsymbol{X}$ is a vector of genotypes for the tested lead variant, b is the effect of the variant, $\boldsymbol{Z}$ is a design matrix allocating phenotypes to animals, $\boldsymbol{u}$ is a vector of breeding values and $\boldsymbol{e}$ is a vector of random residuals. The breeding values $\boldsymbol{u}$ were assumed to be derived from a multivariate normal distribution $\boldsymbol{u} \sim \mathrm{N}(0, \boldsymbol{G}\sigma_g^2)$, where $\boldsymbol{G}$ is the GRM (used to control for population substructure including familial relationships) and σ_g^2 is the additive genetic variance. The model was fitted in EMMAX[5].

In three validation populations (Australian Angus, Australian Holstein and Swiss Brown Swiss), an additional analysis was performed to determine the proportion of variation explained by the 163 lead SNPs. Genotypes for the 163 lead SNPs were extracted, and a genomic relationship matrix was formed using these SNPs[6]. The proportion of variance explained by this matrix was determined by fitting the model

$$\boldsymbol{y} = 1_n\mu + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\mathbf{1}_n$ is a vector of ones, μ is the mean, $\boldsymbol{Z}$ is a design matrix allocating phenotypes to animals, $\boldsymbol{u}$ is a vector of breeding values and $\boldsymbol{e}$ is a vector of random residuals. The breeding values $\boldsymbol{u}$ were assumed to be derived from a multivariate normal distribution $\boldsymbol{u} \sim \mathrm{N}(0, \boldsymbol{G}^*\sigma_g^{2*})$, where $\boldsymbol{G}^*$ is the GRM created from genotypes at the 163 lead SNPs and σ_g^{2*} is the additive genetic variance explained by the 163 lead SNPs. Variance components were estimated with ASREML[52]. To determine the proportion of variance expected to be explained by chance, another 163 SNPs with the same allele frequencies as the 163 lead variants were randomly sampled from the sequence data and the model above was fitted. This process was repeated five times, and the proportions of explained variance were averaged.

A second validation approach evaluated whether the prediction equation comprising the effects for the 163 lead SNPs from the meta-analysis could predict the differences in stature between standard and miniature cattle from the same breed. Stature was predicted as $2\sum_{i=1}^{163}\overline{p}_i\widehat{\beta}_i$, where $\overline{p}_i$ was the average allele frequency of miniature or standard animals for the ith SNP and $\widehat{\beta}_i$ was the effect of the SNP from the meta-analysis. There were four miniature Angus, two miniature Hereford and two miniature Belted Galloway cattle each sequenced to approximately tenfold coverage. SNP genotypes were called in these animals using the same pipeline that was used for the 1000 Bull Genomes project[6]. In the original experiment where the miniature Angus cattle were bred, the mature weight and height of cows were 497 ± 6 kg and 115.7 ± 0.6 cm for the standard line and 418 ± 6 kg and 108.3 ± 0.6 cm for the miniature line[38]. For miniature Belted Galloway animals, the breed specification is "bulls at 10 to 12 months of age to be no more than 110 cm at hip height; maximum height for showing, at any age, is 125 cm at hip. Females at 10 to 12 months of age to be no more than 105 cm at hip height; maximum height for showing, at any age is 120 cm at hip" (Miniature Herefords; see URLs). This compares to standard female Belted Galloway animals that have an average hip height of 126 cm, with an s.d. of 2 cm (see URLs). For miniature Hereford animals, the desired height for the breed is 100 cm, although bulls up to 110 cm in height have been registered by the breed association (see URLs). This compares to a standard Hereford with an average height of 120 cm[39].

In the third validation approach, the average height of seven breeds was predicted from their whole-genome sequences and compared to height reported in three experiments measuring the height of these breeds[38–40]. There were two Dexter sequences, 33 Charolais sequences, 10 Belgian Blue sequences, 59 Brown Swiss sequences, 34 Gelbvieh sequences, 31 Limousin sequences and 5 Piedmontese sequences. Allele frequencies for each breed calculated from these sequences were used in the prediction equation $2\sum_{i=1}^{163}\overline{p}_i\widehat{\beta}_i$ with the terms defined as above.

Proportion of variation accounted for by 17 previously identified loci within village dogs. We reanalyzed the village dog dataset from Hayward et al.[3]. The dataset we analyzed included 330 village dogs measured for body weight. Using 160,727 variants, the first ten principal components of the GRM were derived and fitted in a multiple regression model to account for population structure within the 330 dogs (five principal components were significant). Sex was also fitted as a fixed effect. The multiple regression model included the 17 SNPs (fitted simultaneously) identified in Hayward et al.[3] and in other publications in other dog breeds as having a significant effect on body size. The proportion of variance explained by the markers was calculated as $\sum_{i=1}^{17}2p_i(1-p_i)\alpha_i^2/\sigma_p^2$, where σ_p^2 is the phenotypic variance of weight (with the effect of sex and the principal components removed), p_i is the allele frequency of the ith SNP and α_i^2 is the allele substitution effect of the ith SNP.

Bootstrap analysis. Bootstrap sampling was performed to contribute evidence that the lead variants could be causative mutations. We recorded the proportion of bootstrap samples in which the lead variant from the original meta-analysis remained the lead variant in the bootstrap sample. Bootstrap sampling was performed by sampling 17 populations with replacement from the 17 populations used in the meta-analysis. Once the 17 populations were sampled, the meta-analysis was rerun for the 25.4 million variants using METAL[37] as described above. There were 100 bootstrap samples.

eQTL analysis. Whole-genome sequence data were imputed into 630,000 (imputed from 43,000) genotypes for the cows using the bull whole-genome sequences in Run4 of the 1000 Bull Genomes project. After removing variants that had a MAF less than 0.05 for the cows in the experiment, 10.4 million variants remained. Only genes that were expressed in WBCs for more than 25% of the cows were analyzed, to avoid spurious associations due to very low read counts. For each of 11,089 genes that satisfied this criterion, association of expression level (sequence counts) with all of the variants on the chromosome that contained that gene was tested (ignoring trans effects on other chromosomes). That is, 11,089 genome-wide association analyses were run, with up to 690,000 variants (for example, for chromosome 1, there were this many sequence variants tested for each gene). Association testing was performed with EMMAX[5], fitting the GRM among cows to control for population structure, fixed effects of parity, days in milk, sampling day and RNA sequencing batch. Read counts were transformed as $\log(x+1)$, where x was the read count of a particular gene for a cow.

On average, 56 million reads were generated per WBC library. On average, 88.4% of reads passed quality control, of which an average of 91.73% mapped to the reference genome. Quality filtering after alignment to the reference genome resulted in 15 samples being excluded from the count matrix (owing to very low counts as compared to other samples).

We used the experiment-wise false discovery rate—the proportion of significant variants that are actually false positive results—to determine which threshold was appropriate when testing individual SNPs. If a threshold of $P < 1\times10^{-5}$ was used, the false discovery rate was 1.3% (Supplementary Table 7), which seemed reasonable.

Although 73,840 significant variants were detected at the $P < 1\times10^{-6}$ threshold, they were associated with only 659 genes. This indicates that multiple variants, in strong LD, are detecting the same eQTL.

There was a trend for the most significant variant to be closer to the gene for which the expression level was the phenotype (Supplementary Fig. 8).

Selection signature analysis. Genome scans for selection were performed using FLK[34] and hapFLK[35], two tests that identify regions of high differentiation between populations. Fifteen populations were considered, listed in Supplementary Table 8, and unrelated animals were selected within each population. Selection was done by excluding animals found as outliers from their reported breed, on the basis of their principal-component analysis coordinates. Then, within each breed, unrelated animals were selected on the basis of the GRM kinship coefficients, computed using GCTA[6].

FLK and hapFLK were calculated with hapFLK software (see URLs), using ancestral allele information to root the population tree. *P* values were estimated for each test using procedures documented with the software. *q* values were calculated using the qvalue R package, and SNPs corresponding to a false discovery rate of 5% were called significant.

Enrichment analysis. An enrichment analysis among GWAS hits was performed based on a stratified FDR approach[53]. FLK *P* values for all SNPs were divided into two sets: a set of GWAS hits and the set of non-GWAS hits. Within each set, the proportion of true positives ($1-\pi_0$) was estimated with the qvalue R package. The enrichment in the GWAS set as compared to the non-GWAS set was calculated as the ratio of the GWAS hits value to the non-GWAS hits values. The same approach was used for lead variants with the 163 SNPs in place of all GWAS hits.

To assess the significance of the enrichment of selection signatures in cattle GWAS hits, the same procedure was applied to human GWAS regions. We extracted human GWAS hits from the human GWAS catalog (see URLs)[54]. We considered only the 35 traits that had more than 150 hits in the GWAS catalog, to match our 163 lead variants. For each trait, we used the reported closest genes to all GWAS hits to map the human association to the cattle genome, using Ensembl and RefSeq annotations of UMD 3.1. This allows, for each human trait, the definition of a set of homologous cattle genes within which we retrieved FLK *P* values. In the set of SNPs included in these genes, we estimated the enrichment in selection signatures as explained above. Results of the analyses are given in Supplementary Table 9. Only human traits with enrichment >1 are shown.

Tests for detection of known causal mutations affecting fat and protein percentage in the milk of dairy cattle. We performed association tests between the imputed sequence variant genotypes and protein percentage and fat percentage in milk in Holstein, Fleckvieh and Brown Swiss cattle. The known mutations included a mutation in the growth hormone receptor gene (*GHR*; encoding p.Tyr279Phe, chromosome 20[54]), a mutation encoding p.Ala232Lys in the *DGAT1* gene[55] on chromosome 14 and a mutation encoding p.Tyr851Ser in the *ABCG2* gene[56] on chromosome 6. The *GHR* mutation segregates in Holstein, Fleckvieh and Brown Swiss, the *DGAT1* mutation segregates in Holstein and Fleckvieh, and the *ABCG2* mutation segregates at very low frequency in Holsteins only.

The analysis is presented in Pausch et al.[10]. However, figures demonstrating that imputed sequence data could discover known causative mutations were not presented in that the manuscript and are presented here. 214 Brown Swiss and 345 Holstein animals were genotyped using the Illumina BovineHD BeadChip that comprises 777,962 SNPs. All other animals were genotyped using the Illumina BovineSNP50 BeadChip that comprises 54,609 SNPs. The Brown Swiss, Holstein and Fleckvieh animals were imputed to higher density using FImpute[45] (Brown Swiss) and Minimac[44] (Holstein and Fleckvieh). The final dataset included 573,650 and 564,374 autosomal SNPs. Sequence variant genotypes were imputed in 6,777 Fleckvieh, 5,204 Holstein and 1,646 Brown Swiss animals using the 1000 Bull Genomes Run4 multi-breed reference population with Minimac[44]. Association tests were performed between imputed sequence variant genotypes on chromosomes 6 and 20 and daughter-derived values for protein percentage, and on chromosome 14 and daughter-derived values for fat percentage. Association testing was carried out with EMMAX[5] using the '-Z' flag to consider predicted allele dosages for the imputed sequence variants.

URLs. Australian Galloway Association, http://www.galloway.asn.au/miniaturegalloways.html; Belted Galloway Society, http://www.beltie.org/breed-surveys-data.php; Miniature Herefords, http://www.miniatureherefords.org.au/; Bos_taurus_UMD_3.1/bosTau6 assembly of the cow genome, http://hgdownload.soe.ucsc.edu/goldenPath/bosTau6/bigZips; Bison_UMD1.0/bisBis1 assembly of the bison genome, http://hgdownload-test.cse.ucsc.edu/goldenPath/bisBis1/bigZips; sheep reference genome Oar_v3.1, https://www.ncbi.nlm.nih.gov/assembly/GCF_000298735.1; *Bos mutus* isolate yakQH1 genome, https://www.ncbi.nlm.nih.gov/nuccore/AGSK01000000; water buffalo reference genome UMD_CASPUR_WB_2.0, https://www.ncbi.nlm.nih.gov/assembly/GCA_000471725.1/#/st; LASTZ, http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html; QualityTrim, https://bitbucket.org/arobinson/qualitytrim; hapFLK software, https://forge-dga.jouy.inra.fr/projects/hapflk; human GWAS catalog, https://www.ebi.ac.uk/gwas/.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Sequence for miniature cattle can be found at NCBI BioProject PRJNA238491 (1000 Bull Genomes project).

BioSample accession numbers are as follows: SAMN05861856, SAMN05861898, SAMN05861943, SAMN05861857, SAMN05861944, SAMN05861858, SAMN05861899, SAMN05861859, SAMN05861900, SAMN05861901, SAMN05861860, SAMN05861945, SAMN05861902, SAMN05861903, SAMN05861861, SAMN05861862, SAMN05861863, SAMN05861946, SAMN05861864, SAMN05861865, SAMN05861866, SAMN05861904, SAMN05861905, SAMN05861906, SAMN05861907, SAMN05861947, SAMN05861867, SAMN05861948, SAMN05861908, SAMN05861909, SAMN05861910, SAMN05861868, SAMN05861911, SAMN05861912, SAMN05861949, SAMN05861950, SAMN05861951, SAMN05861913, SAMN05861869, SAMN05861914, SAMN05861915, SAMN05861870, SAMN05861916, SAMN05861917, SAMN05861871, SAMN05861872, SAMN05861873, SAMN05861918, SAMN05861874, SAMN05861919, SAMN05861875, SAMN05861876, SAMN05861920, SAMN05861877, SAMN05861878, SAMN05861921, SAMN05861879, SAMN05861880, SAMN05861922, SAMN05861881, SAMN05861952,

SAMN05861882, SAMN05861953, SAMN05861923, SAMN05861924, SAMN05861925, SAMN05861883, SAMN05861926, SAMN05861927, SAMN05861928, SAMN05861954, SAMN05861955, SAMN05861956, SAMN05861957, SAMN05861958, SAMN05861884, SAMN05861885, SAMN05861929, SAMN05861886, SAMN05861887, SAMN05861959, SAMN05861888, SAMN05861960, SAMN05861930, SAMN05861961, SAMN05861931, SAMN05861932, SAMN05861889, SAMN05861933, SAMN05861934, SAMN05861935, SAMN05861890, SAMN05861891, SAMN05861892, SAMN05861893, SAMN05861894, SAMN05861936, SAMN05861937, SAMN05861962, SAMN05861938, SAMN05861939, SAMN05861963, SAMN05861940, SAMN05861941, SAMN05861895, SAMN05861896, SAMN05861942, SAMN05861964, SAMN05861897.

RNA sequence for the eQTL experiment can be found under BioProject PRJNA305942, SRP067373: SAMPLE 210004817-W2-Blood-RNA, SRS1206435; SAMPLE 210004817-W2-Milk-RNA, SRS1206437; SAMPLE Y10ST0027-W2-Blood-RNA, SRS1206444; SAMPLE Y10ST0027-W2-Milk-RNA, SRS1206446; SAMPLE Y10ST0106-W2-Blood-RNA, SRS1206447; SAMPLE Y10ST0106-W2-Milk-RNA, SRS1206629.

References

42. Erbe, M. et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **95**, 4114–4129 (2012).
43. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
44. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
45. Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478 (2014).
46. Harris, R.S. *Improved Pairwise Alignment of GenomicDNA*. PhD thesis, Pennsylvania State Univ. (2007).
47. Rocha, D., Billerey, C., Samson, F., Boichard, D. & Boussaha, M. Identification of the putative ancestral allele of bovine single-nucleotide polymorphisms. *J. Anim. Breed. Genet.* **131**, 483–486 (2014).
48. Zimin, A. V. et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* **10**, R42 (2009).
49. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
50. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
51. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
52. Gilmour, A. R., Gogel, B., Cullis, B., Thompson, R. & Butler, D. *ASReml User Guide Release 3.0*. (Hemel Hempstead: VSN International, Stockholm, 2009).
53. Sun, L., Craiu, R. V., Paterson, A. D. & Bull, S. B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet. Epidemiol.* **30**, 519–530 (2006).
54. Blott, S. et al. Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**, 253–266 (2003).
55. Grisart, B. et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Res.* **12**, 222–231 (2002).
56. Cohen-Zinder, M. et al. Identification of a missense mutation in the bovine *ABCG2* gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res.* **15**, 936–944 (2005).

nature research

Corresponding author(s): Ben Hayes

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

Materials and Methods. Sample size was chosen based on previous genome wide association studies in humans and cattle, which suggested effects of individual SNP on stature would be small. Therefore a very large sample size was used (58,265) so that small effect sizes could be detected

2. Data exclusions

Describe any data exclusions.

Materials and methods. Animals were screened based on SNP call rates, as described in previous publications

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

Independent validation studies with very large numbers of animals (as many as in the in the original discovery experiment) were conducted to determine if the results could be replicated.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was used, field data was used

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding, field data was used

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Existing software was used, including Fimpute, Beagle3, Minimac, METAL, EMMAX, ASREML.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	Materials and Methods. Sequence for miniature cattle can be found at Bioproject PRJNA238491 (1000 bull genomes project), Biosample accession numbers are: SAMN05803879, BTAUGLWAUSF000000AG0134, SAMN05803880 BTAUGLWAUSF000000AG0135, SAMN05803881, BTAUHERAUSF000000AG0137, SAMN05803882 BTAUHERAUSF000000AG0139, SAMN05803883 BTAUDXTAUSM000000AG0144, SAMN05803884 BTAUAANAUSF000000AG0149, SAMN05803885, BTAUAANAUSF000000AG0150, SAMN05803886 BTAUAANAUSM000000AG0152, SAMN05803887 BTAUAANAUSM000000AG0154, SAMN05803888, BTAUDXTAUSF000000AG0156. RNA Sequence for the eQTL experiment can be found at Bioproject PRJNA305942, SRP067373, SAMPLE 210004817-W2-Blood-RNA, SRS1206435, SAMPLE 210004817-W2-Milk-RNA, SRS1206437, SAMPLE Y10ST0027-W2-Blood-RNA, SRS1206444, SAMPLE Y10ST0027-W2-Milk-RNA, SRS1206446, SAMPLE Y10ST0106-W2-Blood-RNA, SRS1206447, SAMPLE Y10ST0106-W2-Milk-RNA SRS1206629.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Not applicable

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Not applicable
b. Describe the method of cell line authentication used.	Not applicable
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	Not applicable

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	The eQTL experiment described in the paper was approved by the Department of Economic Development, Jobs, Transport and Resources Animal Ethics Committee (2013-14). All other data was collected for routine genetic evaluation.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Not applicable