Meta-analysis of genome wide association studies for the stature of cattle reveals numerous common genes that regulate 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[12], Marlies Dolezal[9], Hubert Pausch[2,10], Rasmus F Brøndum[7], Phil J Bowman[2], Bo Thomsen[11], Bernt Guldbrandtsen[7], Mogens S Lund[7], Bertrand Servin[12], Dorian J Garrick[13], James Reecy[13], Johanna Vilkki[14], Alessandro Bagnato[15], Min Wang[2,3], Jesse L Hoff[16], Robert D Schnabel[16], Jeremy F Taylor[16], Anna AE Vinkhuyzen[17,18], Frank Panitz[11], Christian Bendixen[11], Lars-Erik Holm[11], Birgit Gredler[19], Chris Hozé[8,20], Mekki Boussaha[8], Marie-Pierre Sanchez[8], Dominique Rocha[8], Aurelien Capitan[8,20], Thierry Tribout[8], Anne Barbat[8], Pascal Croiseau[8], Cord Drögemüller[21], Vidhya Jagannathan[21], Christy Vander Jagt[2], John J Crowley[22], Intergenomics Consortium[23], Anna Bieber[24], Deirdre C Purfield[25], Donagh P Berry[25], Reiner Emmerling[26], Kay-Uwe Götz[26], Mirjam Frischknecht[19], Ingolf Russ[27], Johann Sölkner[28], Curtis P Van Tassell[29], Ruedi Fries[10], Paul Stothard[30], Roel F Veerkamp[1], Didier Boichard[8], Mike E Goddard[2,4], Ben J Hayes[2,31]

Correspondence: Ben J Hayes

E-mail: b.hayes@uq.edu.au

[1]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, PO Box 338, 6700 AH Wageningen, The Netherlands
[2]AgriBio, Centre for AgriBioscience, Department of Economic Development, Jobs, Transport and Resources, Bundoora, Victoria, Australia
[3]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, ON, Canada
[6]The Semex Alliance, Guelph, ON, Canada
[7]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark
[8]GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France
[9]Platform Bioinformatics and Statistics, University of Veterinary Medicine, 1210 Wien, Veterinärplatz 1 Geb: HA, Vienna, Austria
[10]Chair of Animal Breeding, Technische Universitaet Muenchen, 85354 Freising-Weihenstephan Germany
[11]Section for Molecular Genetics and Systems Biology. Department of Molecular Biology and Genetics. Aarhus University, Tjele, Denmark

[12]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, 31326 Castanet-Tolosan, France
[13]Department of Animal Science, Iowa State University, Ames IA, USA
[14]Green Technology, Natural Resources Institute Finland (Luke), Jokioinen, Finland
[15]Department of Veterinary Medicine, University of Milan, Italy
[16]Division of Animal Sciences, University of Missouri, Columbia MO USA
[17]University of Queensland, Institute for Molecular Bioscience, St Lucia, QLD, Australia
[18]University of Queensland, Queensland Brain Institute, St Lucia, QLD, Australia
[19]Qualitas AG, Zug, Switzerland
[20]Allice, 75595 Paris
[21]Institute of Genetics, University of Bern, Switzerland
[22]Canadian Beef Breeds Council, 6815 8 St., Calgary, Alberta, Canada
[23]Interbull center, SLU, Box 7023, Uppsala, 75007, Sweden
[24]Research Institute of Organic Agriculture (FiBL), Ackerstrasse 113, 5070 Frick, Switzerland
[25]Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Co. Cork, Ireland
[26]Institute of Animal Breeding, Bavarian State Research Centre for Agriculture, 85586 Poing, Germany.
[27]Tierzuchtforschung e.V., Senator-Gerauer-Str. 23, 85586 Poing, Germany
[28]University of Natural Resources and Life Sciences, Gregor-Mendel-Str 33, 1180 Wien, Austria
[29]Animal Genomics and Improvement Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, MD 20705
[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, Queensland, Australia

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 genetic architecture of stature to that 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 revealed that; genetic architecture of stature in cattle is similar to that in humans, as lead variants in 163 significant genomic regions ($P<5\times10^{-8}$) explained at most 13.8% of the phenotypic variance; difference in stature between miniature cattle and standard cattle of the same breed can be predicted by the lead variants; 23% of the lead variants were heterozygous in an Auroch genome; most lead variants were non-coding, including variants that were also eQTL and in ChIP-seq peaks; significant overlap existed in loci for stature with humans and other mammals; and allele frequencies of many of these variants are influenced by selection.

Stature of cattle was analyzed in 17 populations that represent 8 *Bos taurus* breeds with a total of 58,265 animals (Table S1) genotyped with either 630K SNP or 50K SNP imputed to 630K SNP. A GWAS was performed in each population separately using imputed whole-genome sequence variants with correction for population structure[4,5]. The 1,000 Bull Genomes Run4 reference population of 1,147 whole genome sequenced individuals was used to impute the 630K SNP genotypes to 25.4 million whole-genome sequence variants (SNPs and INDELS)[6]. A meta-analysis across the populations found genome-wide significant ($P<5\times10^{-8}$) sequence variants in 163 one megabase regions (Figure 1). The lead variants (most significant variants in each regions) include 160 SNPs and 3 INDELS (Table S2).

Three approaches were used to validate the 163 lead variants. Note that validation here means that the variants are associated with variation in stature, not necessarily that they are actually causative mutations.

Firstly, the association of the 163 lead variants with stature was tested in 30,175 additional cattle with stature phenotypes from 10 populations comprising 8 breeds. In a meta-analysis of these validation populations, 20 of 101 SNP (101 of the 163 variants were polymorphic in all populations) were validated at $P<0.05$, giving a false discovery rate of 25%, Table S3. We also validated the SNP within each breed, as a some variants were polymorphic in one or only small number of breeds. The majority of variants (53%, 86) were validated in at least one population, and many (17%, 28, with 11 expected by chance) were validated in more than one population, Table S3. The 163 lead variants explained between 2.1% (Limousin) to 13.8% (Brown Swiss) of the phenotypic variation in stature (Table 1), and this was significantly greater than that explained by a random subset of the same number of variants where tested. This is less than, but of a similar magnitude as, the proportion of phenotypic variance explained by significant variants in humans (~16%)[1]. The results are substantially different to those reported in dogs, where 6 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 these across dog breed analyses, within a population of village dogs[3], correcting for population structure and sex, and found the 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 there may be some loci of larger effect in dogs.

For the second approach for validation of our lead variants, we exploited the fact that for a number of cattle breeds there are miniature cattle that are several standard deviations smaller in stature than standard cattle, the result of recent strong selection for reduced stature. These animals are miniatures rather than dwarfs, as they do not display chondrodysplasia. The difference in stature of miniature cattle and standard cattle was predicted by an equation comprised of the meta-analysis effects of the 163 lead variants (effects in Table S2). In all three breeds where we had genome sequence from standard cattle and miniature cattle of the same breed, the prediction equation correctly predicted that the miniature animals had substantially shorter stature (Figure 2A, B). In the third validation approach, the same prediction equation accurately predicted differences in stature between seven breeds with sequence data but not included in the meta-analysis (r^2=0.80), Figure 2C. This is in spite of the fact that our meta-analysis was strictly within breed, as mean stature in all populations was set to zero prior to the meta-analysis.

The most significant variant in the meta-analysis was a SNP in intron 3 of *PLAG1* (*rs109815800*, $P<10^{-104}$) 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 the genetic variation of stature in humans as well as cattle[1,7,8,9]. 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 SNP were not in complete LD (in the sequenced animals, Table S4), and SNP *rs109815800* was more strongly associated with stature ($P<10^{-104}$) than the others proposed[7]. The results demonstrate the power of the meta-analysis conducted here to directly identify a small number of SNPs as putative causative mutations, capitalizing on different allele phase relationships in different cattle populations. Imperfect imputation (Figure S1, Figure S2) may result in the causal mutation not being identified as the most highly associated variant, especially if the variant is rare (accuracy of imputation was >0.9 for variants with MAF>0.10, and for most of the variants in the *PLAG1* region, Table S4) (the *rs109815800* variant among those genotyped by the 630K array in some populations). However it has been demonstrated using the 1000 bull genomes data set that imputation of sequence variants followed by genome wide association was able to detect known causal mutations in our data set, for other phenotypes, namely protein and fat percentages in milk (Figures S3, S4 and S5)[10]. The polygenic architecture of stature in cattle is exemplified by the fact that SNP *rs109815800* explained only 0.14% and 0.2% of the phenotypic variance in the Angus and Hereford validation populations, where MAF was 0.07 and 0.16 respectively. This is in part because the MAF of this SNP is low in most breeds, however none of the other variants explained more than 1% of the variation.

To investigate what type of variants affect stature in cattle, genome annotation, eQTL and ChIP-Seq data was used. Note that these analyses do depend on an at least an enrichment of our lead variants for causative mutations, and bootstrap re-sampling suggested a considerable

proportion of our variants were unlikely to be merely linked variants with the smallest P-value in the meta-analysis due to sampling effects, Table S5 (25 of our variants were the lead variant in greater than 50% of bootstrap samples). Of the 163 lead variants identified in our cattle meta-analysis, 5 were missense variants, a 7 fold enrichment of missense variants in the lead variants compared with what would be expected by chance (Table 2). The missense variants included one in *HMGA2*, a well documented human stature gene. *HMGA2* directly regulates the RNA binding protein IGF2BP2 (IGF2 binding protein 2), which in turn enhances *IGF2* translation[11]. Another missense variant was found in *LCOR* (Ligand-dependent corepressor), which is broadly expressed in fetal and adult tissues to regulate development and homeostasis[12,13,14]. In many species, including humans, mice, and rats (and in bovine in this study), a small genomic region that includes *LCORL* (ligand dependent nuclear receptor corepressor like) and *NCAPG* 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 high levels of LD among SNP in these regions (also observed in this study). The identification of a missense variant in *LCOR* in our study, a gene with very high homology and potentially similar function to *LCORL*, to be associated with stature, provides 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. The hypothesis that many of these SNP are in regulatory regions in humans is supported by the recent observation that GWAS associations are enriched in regions of open chromatin[16]. Interestingly eight of the 83 intergenic variants found were located in bovine ChIP-Seq peaks, which is more than expected by chance ($P<0.05$). These were identified from H3K27 acetylation and H3K4 trimethylation histone modification assays of bovine liver, which indicates that these variants are in enhancers, repressors, or promoters and may therefore alter the expression of nearby genes[17].

To further investigate the hypothesis that many of our lead variants are regulatory, we performed an expression QTL (eQTL) study using RNASeq data from white blood cells from 93 Holstein cows. While gene expression in fetal tissue would presumably have been more informative than blood in mature cows for this study, recent evidence suggests a reasonable overlap of eQTL across tissues[18]. Ten of the 163 lead stature variants from the meta-analysis were also eQTL in white blood cells, an 18 fold enrichment over the number expected by chance (Table 2, Table S2). It is possible that the 163 stature variant regions may be enriched in eQTL even if functionally unrelated, due to non-random clustering of genes for example. We assessed evidence for a functional relationship (either pleiotropy or causality) with the HEDI (heterogeneity in dependent instruments) test[19]. Seven out of the ten eQTL/stature variants showed no heterogeneity of effects with linkage disequilibrium, suggesting these mutations could be either causal for both the gene expression levels and stature, or pleiotropic for these traits (this is still a very significant enrichment). One such variant, BTA4 32075456 bp, associated ($P<1x10^{-5}$) with the expression of *IGF2BP3* (Insulin-like growth factor 2 binding protein 3), is an interesting candidate, as the IGF2BP3 protein suppresses translation

of *IGF2* during late fetal development[20,21,22,23,24]. The direction of effects was consistent with this mechanism – the allele associated with increased expression of *IGF2BP3* was associated with decreased bovine stature. Additional evidence that this SNP is an eQTL is provided by the observation of allele specific expression for *IGF2BP3* in multiple tissues in a cow heterozygous for the SNP [25].

We next investigated if there was a greater overlap of loci affecting stature in bovine and in humans than would be expected by chance. Of the 92 genes overlapping with or within (±5 kb) of the 163 lead variants, eleven were identified by Wood et al.[1] as affecting stature in humans (Table S2), a significant enrichment compared to the overlap expected by chance alone ($P<10^{-12}$, chi-square test). This test is stringent, as it requires the lead variant to be within or very close to the causal gene. An approximate confidence interval for each QTL region was defined (see methods). Resulting confidence intervals averaged 527 kb (Table S2). These QTL confidence regions overlapped with 26 of the genes that have been identified as associated with stature or body size in humans and/or dogs (Table S2). 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 defined confidence intervals in our study[3,26].

Considering that many of the stature variants were only segregating in one or two breeds (Figure S6), an interesting question arises as to whether the stature variants are recent mutations (for example arising after breed formation), or ancient standing variation that have been 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 genome sequence of a 6,750 year old Auroch genome[27]. Of the 163 lead variants, 134 had six or more reads covering the variant so could be the genotype 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 pre-domestication and certainly pre-breed formation (though it must be noted that only a proportion of our lead variants might be actual causal mutations). Interestingly the stature of the Auroch (from the effects of our lead variants) was predicted to be larger than all but one of the modern breeds, Figure 2C, consistent with the large skeletal size of Aurochs from the fossil record[28]. 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 (Table S2), where the ancestral allele was determined from sequence comparisons between Bos taurus, American Bison (*Bison bison*), Yak (*Bos grunniens*) and Water Buffalo (*Bubalis bubalis*). The observation that some polymorphisms with an ancestral allele that increases stature still segregates in multiple breeds may also be because the direction of selection for stature has not been consistent among cattle breeds (effectively balancing the effects of selection). As cattle were domesticated, there was selection for reduced stature compared to that of wild Aurochs populations (either directly, or as a correlated response to selection for early sexual maturity, or both) as evidenced by the bone lengths of ancient domestic versus contemporaneous wild

cattle[29,30]. Selection for reduced stature continued until at least the 15th century when Northern European cattle measured less than one meter in stature at the withers[27,29]. More recently, there has been very strong selection for increased stature in some breeds, with for example Holstein, Brown Swiss and Fleckvieh all increasing in stature by approximately 2mm per year in the last decade[32,33,34].

Additional evidence that sequence variants affecting stature have been subject to selection since domestication and breed formation was that nearly 50% of the 163 variants were in selection signatures identified in the 1000 bull genome Run4 1,147 bull whole genome sequences[35,36], a 30 fold enrichment compared to other (non stature-associated) SNPs, Figure S7. Selection for stature is exemplified by the detection of selective sweeps for the same haplotype in 5 breeds for *NCAPG-LCORL* and in ten breeds for *PLAG1* (Figure 3). Interestingly the *PLAG1* allele that increases stature is almost fixed in tall breeds (e.g. Limousin, Charolais, Holstein), while in breeds of short and moderate stature the degree of fixation was variable (Jersey, Brown Swiss, Angus, Montbeliarde, Fleckvieh). Note that our analysis does not rule out selection signatures arising from selection on a trait with pleiotropic affects with the 163 lead variants for stature.

Our results reveal that the genetic architecture of stature within domestic cattle breeds is highly polygenic, similar to the genetic architecture of stature observed in humans (and other complex traits in cattle[38]). Results of the new analysis within village dogs indicate a larger number of loci will be required to explain variation in body size than previously reported. In dogs a small number of loci explain some of the across breed differences in body size, while in cattle 163 variants were required to explain stature differences between standard and miniature cattle. The difference between genetic architecture in cattle and dogs reflects both population history and selection history. The effective population size of most dog breeds is much smaller than most cattle breeds, as demonstrated by the substantially greater extent of linkage disequilibrium in dog breeds[39] than in cattle breeds[40], no doubt exacerbated by the typically larger litter size for dogs and more rapid turnover of generations. In addition, there has been very strong selection in dogs for loci with extreme effect on stature, such as dwarfing mutations. In cattle (and humans), these mutations are selected against because of undesirable pleiotropic effects, while in dogs they become a breed defining feature, for example chondrodysplasia in Dachshunds which results from a duplication of the *FGF4* gene[41]. Finally, our results support the hypothesis that there are numerous common genes that affect size in mammals.

Data

Sequence for miniature cattle can be found at Bioproject PRJNA238491 (1000 bull genomes project),

Biosample accession numbers are: 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 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.

References

1. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46,** 1173-86 (2014).
2. Rimbault, M., Beale, H.C., Schoenebeck, J.J., et al. Derived variants at six genes explain nearly half of size reduction in dog breeds. *Genome Research* **23**, 1985-1995 (2013).
3. Hayward, J.J., Castelhano, M.G., Oliveira, K.C., Corey, E., Balkman, C., Baxter, T.L., Casal, M.L., Center, S.A., Fang, M., Garrison, S.J., Kalla, S.E., Korniliev, P., Kotlikoff, M.I., Moise, N.S., Shannon, L.M., Simpson, K.W., Sutter, N.B., Todhunter, R.J. & Boyko, A.R. Complex disease and phenotype mapping in the domestic dog. *Nat Commun*. **7**:10460 (2016).
4. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. & Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-54 (2010).
5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E. &

Visscher, P.M. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565-569 (2010).

6. Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., VanTassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R. & Hayes, B.J. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics* **46**, 858-865 (2014).
7. Karim, L, Takeda, H., Lin, L., Druet, T., Arias, J.A., Baurain, D., Cambisano, N., Davis, S.R., Farnir, F., Grisart, B., Harris, B.L., Keehan, M.D., Littlejohn, M.D., Spelman, R.J., Georges, M. & Coppieters, W. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nature Genetics* **43**, 405-13 (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., Kemper, K., Sasazaki, S., Reverter, A., Pryce, J.E., Barendse, W., Bunch, R., McCulloch, R., Harrison, B., Bolormaa, S., Zhang, Y.D., Hawken, R.J., Goddard, M.E. & Lehnert, S.A. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. *Animal Genetics* **44**, 636-647 (2013).
10. Pausch, H., Emmerling, R., Gredler-Grandl, B., Fries, R., Daetwyler, H.D. & Goddard, M.E. Meta-Analysis Of Sequence-Based Association Studies Across Three Cattle Breeds Reveals 25 QTL For Fat And Protein Percentages In Milk At Nucleotide Resolution. bioRxiv 143404. doi:10.1101/143404 (2017).
11. Li, Z., Gilbert, J. A., Zhang, Y., Zhang, M., Qiu, Q., Ramanujan, K. et al. An HMGA2-IGF2BP2 Axis Regulates Myoblast Proliferation and Myogenesis. *Developmental Cell*, **23**, 1176–1188 (2012).
12. Fernandes, I., Bastien, Y., Wai, T., Nygard, K., Lin, R., Cormier, O., Lee, H.S., Eng, F., Bertos, N.R., Pelletier, N., Mader, S., Han, V.K., Yang, X.J. & White, J.H. Ligand-dependent nuclear receptor corepressor LCoR functions by histone deacetylase-dependent and -independent mechanisms. *Mol Cell* **11**, 139-150 (2003).
13. Calderon, M.R., Verway, M., An, B.S., DiFeo, A., Bismar, T.A., Ann, D.K., Martignetti, J.A., Shalom-Barak, T. & White, J.H. 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., Verway, M., Benslama, R.O., Birlea, M., Bouttier, M., Dimitrov, V., Mader, S. & White, J.H. Ligand-dependent corepressor contributes to transcriptional repression by C2H2 zinc-finger transcription factor ZBRK1 through association with KRAB-associated protein-1. Nucleic Acids Research 42, 7012-7027 (2014).
15. Kemper, K.E., Visscher, P.M. & Goddard ME. Genetic architecture of body size in mammals. *Genome Biology* **13**, 244 (2012).
16. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F.R; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI

Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J.R., Okada, Y., Raychaudhuri, S., Daly, M.J., Patterson, N., Neale, B.M. & Price AL. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228-1235 (2015).

17. Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., et al. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566 (2014).
18. GTEx Consortium.. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **84**,648-660 (2015).
19. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., & Yang, J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. **48**:481-487 (2016).
20. Akhtar, M., Holmgren, C., Gondor, A. et al. Cell type and context-specific function of PLAG1 for IGF2 P3 promoter activity. *International Journal of Oncology* **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 Research* **60**, 106-13 (2000).
23. Nielsen, J., Christiansen, J., Lykke-Andersen, J., Johnsen, A.H., Wewer, U.M., & Nielsen, F.C. A Family of Insulin-Like Growth Factor II mRNA-Binding Proteins Represses Translation in Late Development. *Molecular and Cellular Biology* **19**, 1262-1270 (1999).
24. Reik, W., Constancia, M., Dean, W., Davies, K., Bowden, L., Murrell, A., Feil, R., Walter, J. & Kelsey G. Igf2 imprinting in development and disease. *International Journal Developmental Biology* **44**, 145-50 (2000)
25. Chamberlain, A.J., Vander Jagt, C.J., Hayes, B.J., Khansefid, M., Marett, L.C., Millen, C.A., Nguyen, T.T. & Goddard, M.E. Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**, 993 (2015).
26. Sutter, N.B., Bustamante, C.D., Chase, K., Gray, M.M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P.G., Quignon, P., Johnson, G.S., Parker, H.G., Fretwell, N., Mosher, D.S., Lawler, D.F., Satyaraj, E., Nordborg, M., Lark, K.G., Wayne, R.K. & Ostrander, E.A. A single IGF1 allele is a major determinant of small size in dogs. *Science* **316**, 112-115 (2007).
27. Park, S.D., Magee, D.A., McGettigan, P.A., Teasdale, M.D., Edwards, C.J., Lohan, A.J., Murphy, A., Braud, M., Donoghue, M.T., Liu, Y., Chamberlain, A.T., Rue-Albrecht, K., Schroeder, S., Spillane, C., Tai, S., Bradley, D.G., Sonstegard, T.S., Loftus, B.J. & MacHugh, D.E. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome Biology* **16**, 234 (2015).
28. Clutton-Brock, J. 1987, A Natural History of Domesticated Mammals. Cambridge University Press. The Edinburgh Building, Cambridge CB2 2RU. ISBN 0 521 6324

29. Vretemark, M. From bones to livestock, Stockholm University (1997).
30. 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).
31. Svensson, E.M., Anderung, C., Baubliene, J., Persson, P., Malmström, H., Smith, C., Vretemark, M., Daugnora, L. & Götherström A. Tracing genetic change over time using nuclear SNPs in ancient and modern cattle. *Animal Genetics* **38**, 378-83 (2007).
32. Krogmeier, D. Zusammenhänge zwischen Nutzungsdauer und Körpergröße unter besonderer Berücksichtigung des Stallsystems bei Braunvieh und Fleckvieh. *Züchtungskunde*, **81**, 328–340 (2009).
33. Beavers, L. & Van Doormaal, B. A closer look at stature. CDN Report. February 2016.
34. Laumay, A., & le Mezec P. Bilan de l'indexation des races bovines laitieres. Resultats de la Campagne 2014. INRA Report 0015202017 (2015).
35. Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S. & San Cristobal M. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* **186**, 241-262 (2010).
36. Fariello, M. I., Boitard, S., Naja, H., San Cristobal, M. & Servin, B. Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics* **193**, 929-941 (2012).
37. Boitard, S., Boussaha, M., Capitan, A., Rocha, D. & Servin, B. Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds. *Genetics* **203**:433-50 (2016).
38. 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).
39. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B. et. al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**:803-19 (2005).
40. Bovine HapMap Consortium., Gibbs, R.A., Taylor, J.F., Van Tassell, C.P., Barendse, W., Eversole, K.A. et. al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**:528-32 (2009).
41. Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG, Maslen CL, Acland GM, Sutter NB, Kuroki K, Bustamante CD, Wayne RK & Ostrander EA An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**, 995-998.
42. 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. *Livestock Production Science* **49**, 305-312 (1997).
43. 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**:3133-41 (2002).
44. 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).
45. 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).
46. Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A. & Goddard, M.E. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J Dairy Sci*. **95**:4114-41129 (2012).
47. Browning, B.L. & Browning, S.R. Genotype imputation with millions of reference samples. American Journal Human *Genetics* **98:**116-126 (2016).
48. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* **34:**816–834 (2010).
49. Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15:**478 (2014).
50. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genome wide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
51. Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R., & Butler D. ASReml user guide release 3.0. Hemel Hempstead: VSN International Ltd; 2009.
52. Harris, R.S. Improved pairwise alignment of genomic DNA. Ph.D. thesis, Pennsylvania State University (2007).
53. 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**:483-6 (2014).
54. Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Tad S Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., James A Yorke, J.A. & Salzberg, S.L. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology* **10**, R42 (2009).
55. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14:**R36 (2013).
56. Anders, S., Pyl, P.T. & Huber, W. HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31:**166-169 (2015).
57. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11:**R106 (2010).
58. John D. Storey with contributions from Andrew J. Bass, Alan Dabney and David Robinson (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.6.0.
59. 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)

60. Blott, S., Kim, J-J., Moisio, S., Schmidt-Küntzel, A., Cornet, A., Berzi, P., 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**:253–66 (2003).
61. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, 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).
62. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee J-H, 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).
63. Garrick, D.J., Taylor, J.F. & Fernando, R.L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol* **41:**55 (2009).

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 e.V., the Arbeitsgemeinschaft österreichischer Fleckviehzüchter and ZuchtData EDV Dienstleistungen GmbH for providing genotype data. A.B. acknowledges EU Collaborative Project LowInputBreeds (Grant agreement No 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-ID: PA 2789/1-1). D.B. and D. P. acknowledge funding from 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, U.S. Department of Agriculture, under awards number 2013-68004-20364 and 2015-67015-23183.

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 the manuscript. M.S., F.S.S., G.S., D.C.P., H.P., J.V., B.G., 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 DYDs and YDs of Fleckveh animals and I.C. contributed genotypes. A.G.G., C.H., M.P.S., A.C., T.T., A.B., P.C. and A.B. prepared phenotypes and genotypes for French cattle, and ran GWAS. M.F., I.R. and J.S. prepared phenotypes and genotypes for Swiss and Austrain cattle, and ran GWAS A.A.E.V. contributed to across species identification of stature genes. M.B., M.W., P.S., D.R., V.J., R.D.S. performed variant annotation. B.J.H., D.J.G., J.F.T., C.B., J.R., A.B., F.P., B.T., L.E.H., C.D., R.F., C.P.C.T., R.V., D.B., P.S., M.E.G., B.G. and M.L. conceived the experimental design, analysed stature data for contributed breeds, and wrote the manuscript.

Competing financial interests

None to declare

FIGURES AND TABLES

Figure 1. Manhattan plot for the meta-analysis of bovine stature. The red line is the genome-wide significance threshold at P-value=5×10^{-8}. The most likely candidate gene in the most significant regions is given, where an obvious candidate could be identified.

Figure 2. (A) The 163 lead variants predict differences within breeds between miniature and standard cattle. Stature was predicted as $2\sum_{i=1}^{163}\bar{p}_i\,\hat{\beta}_i$,where for variant *i* $\bar{p}_i$ is the average allele frequency of miniature or standard animals for the *i*th SNP, and $\hat{\beta}_i$ is the effect of the variant from the meta-analysis. There were four miniature Angus, two miniature Herefords, and two miniature Belted Galloway cattle sequenced, and 48 standard Angus, 30 standard Herefords, and two standard Belted Galloway animals sequenced. Average height of standard and miniature cattle is approximately 116 cm, 108cm, 120cm, 105cm, 120cm, and 110cm for Angus, Belted Galloway and Hereford respectively[42-45]. (B) Standard and miniature Angus cattle, photo courtesy of Dr Paul Arthur, NSW Department of Primary Industries, Australia[42]. (C) Predicted average stature of seven breeds (not included in the original meta-analysis), where stature was predicted from the 163 lead SNPs as $2\sum_{i=1}^{163}\bar{p}_i\,\hat{\beta}_i$,where for variant *i* $\bar{p}_i$ is the average allele frequency of animals in the breed for the *i*th variant, and $\hat{\beta}_i$ is its effect estimated in the meta-analysis, compared to average reported stature for these breeds. The average reported stature was from three breed comparison studies[43-45]. Standard errors of breed average reported stature were approximately 6 cm. (D) Size of effect against allele frequency in the meta-population (including all breeds).

Figure 3. Haplotype diversity for 15 cattle breeds in two genomic regions (*NCAPG-LCORL*, *PLAG1*) where selection signatures match segregation of stature QTLs. For each panel, each color represents a local haplotype cluster. The *PLAG1* gene is located on chromosome 14 25,007,291-25,009,296 bp, *NCAPG* on chromosome 6: 38,765,969-38,812,051 bp, and *LCORL* on chromosome 6: 38,840,894-38992,112. 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, and the different colors represent different haplotypes[36]. For example, Angus (ANG) is nearly fixed for the yellow haplotype at *PLAG1*, while Gelbvieh (GEL) segregate for a number of different haplotypes. Breeds were 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.

589

590

Table 1. Proportion of phenotypic variation explained by 163 lead variants in validation populations. For Angus (Australia), Holstein (Australia) and Brown Swiss (Switzerland) we compared this to the average of average of random subsets of 163 variants, this was 0.016±0.003, 0.036±0.004 and 0.119±0.009.

Breed	Country	Number animals	Number lead SNP polymorphic	Proportion of phenotypic variation explained by lead SNP
Simmental	Ireland	1913	146	0.052
Limousin	Ireland	10371	150	0.021
Hereford	Ireland	595	137	0.027
Charolais	Ireland	7822	145	0.024
Angus	Ireland	732	139	0.039
Angus	Australia	676	125	0.054
Brown Swiss	Switzerland	5550	160	0.138
Holstein	Australia	1565	141	0.093

Table 2. Annotation of the most significant sequence variants in 163 genomic regions affecting stature in cattle, proportion of all variants in 1000 bull genomes Run4 with this annotation, level of enrichment/depletion of lead variants in each class, and significance of enrichment/depletion.

Annotation class	Number of lead variants	Proportion of lead variants	Proportion of all variants in genome with this annotation ***	Fold Enrichment /Depletion	P-value ****
intergenic_variant	83	0.459	0.663	0.69	0.63
upstream_gene_variant	11	0.061	0.035	1.74	0.33
5_prime_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*	8	0.044	0.024	1.85	0.049
White blood cell eQTL**	10	0.055	0.003	18.33	0.00001

*ChIP-Seq peaks identified from H3K27 acetylation and H3K4 trimethylation histone modification assays of bovine liver[17]

**See supplementary materials for details

***From Run4 of 1000 bull genomes

****Based on a Chi-Squared test comparing observed and expected number of variants in each class, with one degree of freedom.

606

607

MATERIALS AND METHODS

Meta-analysis was performed on GWAS results from 17 populations that represented 8 *Bos taurus* breeds. Within each population, animals were genotyped with either the Illumina Bovine SNP50v2.0 (50K SNP) or BovineHD (777k) SNP (with the majority ongenotyped with 50K). Genotype calls with GenTrain score (GenCall) <0.6 were excluded, 55 SNP with duplicate map positions. Approximately 630K SNP remained depending on population for the HD SNP and 43K BovineSNP50v2 SNP. Some SNP were re-ordered based on their LD mapped position, as described by Erbe[46]. Imputation of animals genotyped for 43K SNP to 630 K SNP was performed with Beagle[47], Minimac[48] or Fimpute[49], and was very accurate (>0.95, assessed by cross validation)[46].

All sequenced animals were used as a reference when imputing whole genome sequence genotypes in each population. 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 630k 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 Table S1.

Variant effect and standard error of the effect from the GWAS were standardized per population by dividing them by the phenotypic standard deviation. The individual population GWAS results for variants with a MAF<0.005 and an/or an effect size of more than 5 standard deviations 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 per variant. Meta-analysis was performed using the inverse variance fixed-effects method in METAL with genomic control (for λ_{GC} see Table S1)[50].

Definition of significant loci and confidence intervals. A quantitative trait locus 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 based on 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 SNP had a difference in –$\log_{10}$(P-value) from the lead SNP of greater than 4 (if the drop in –log10(P-value) was greater than 4, then decreased again, the procedure continued until all further SNP had a difference in –log10(Pvalue) from the lead SNP greater than 4). The maximum distance considered was 0.5Mb either side of the lead variant.

Validation

The 163 lead SNPs were validated in ten populations, Table S3. 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-sibs of these animals.

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

$$\boldsymbol{y} = \mathbf{1}_{\boldsymbol{n}}\mu + \boldsymbol{X}b + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ was a vector of phenotypes, $\mathbf{1}_{\mathrm{n}}$ was a vector of ones, μ was the mean, X was a vector of genotypes for the tested lead variant, b was the effect of the variant, $\boldsymbol{Z}$ was a design matrix allocating phenotypes to animals, $\boldsymbol{u}$ was a vector of breeding values, and $\boldsymbol{e}$ was 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 $\boldsymbol{G}$ was the genomic relationship matrix (used to control for population substructure including familial relationships) and σ_g^2 was the additive genetic variance. The model was fitted in EMMAX[4].

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[5]. The proportion of variance explained by this matrix was determined by fitting the model

$$\boldsymbol{y} = \mathbf{1}_{\boldsymbol{n}}\mu + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ was a vector of phenotypes, $\mathbf{1}_{\mathrm{n}}$ was a vector of ones, μ was the mean, $\boldsymbol{Z}$ was a design matrix allocating phenotypes to animals, $\boldsymbol{u}$ was a vector of breeding values, and $\boldsymbol{e}$ was 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 $\boldsymbol{G}^*$ was the genomic relationship matrix created from genotypes at the 163 lead SNPs and σ_g^{2*} was the additive genetic variance explained by the 163 lead SNPs. Variance components were estimated with ASREML[51]. To determine the proportion of variance expected to be explained 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 comprised of 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 $2\sum_{i=1}^{163}\overline{p}_i\,\hat{\beta}_i$, where $\overline{p}_i$ is the average allele frequency of miniature or standard animals for the i^{th} SNP, and $\hat{\beta}_i$ is 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 ten fold 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, 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[42].

For Miniature Belted Galloways, 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 125cm 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 120cm at hip". (http://www.galloway.asn.au/miniaturegalloways.html). This compares to standard female Belted Galloways which have an average of 126cm hip height, with a standard deviation of 2cm (http://www.beltie.org/breed-surveys-data.php). For Miniature Herefords, the desired height for the breed is 100cM, though bulls up to 110cm have been registered by the breed association (http://www.miniatureherefords.org.au/). This compares to a standard Hereford with average height of 120cm[43].

In the third validation approach, average height of seven breeds was predicted from their whole genome sequences, and compared to height reported in three experiments measuring height of these breeds[43-45]. There were two Dexter sequences, 33 Charolais sequences, 10 Belgian Blue sequences, and 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 $2\sum_{i=1}^{163}\overline{p}_i\,\hat{\beta}_i$ with terms defined above.

Proportion of variation accounted for by 17 previously identified loci within village dogs.

We re-analysed the village dog dataset from Hayward et al[3]. The data set we analysed included 330 village dogs measured for body weight. Using 160,727 variants, the first 10 principal components of the genomic relationship matrix were derived and fitted in a multiple regression model to account for population structure within the 330 dogs (5 principal components were significant). Sex was also fitted as a fixed effect. The multiple regression model included the 17 SNP (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 $\sum_{i=1}^{17}2p_i(1-p_i)\alpha_i^2 \quad /\sigma_p^2$ 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 i^{th} SNP, and α_i^2 is the allele substitution effect of the i^{th} SNP.

Bootstrap analysis

Boostrap 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. Boostrap 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 re-run for the 25.4 million variants using METAL[50] as described above. There were 100 bootstrap samples.

Ancestral allele determination

To determine the ancestral allele, the following genome assemblies were used

1) Cattle UMD3.1 reference genome sequence (Btau6 version): http://hgdownload.soe.ucsc.edu/goldenPath/bosTau6/bigZips.

2) Bison (Bison_UMD1.0/bisBis1) genome assembly (bisBis1, U. Maryland): http://hgdownload-test.cse.ucsc.edu/goldenPath/bisBis1/bigZips.

3) Sheep (Ovis aries) genome assembly (Oar_v3.1 version) : https://www.ncbi.nlm.nih.gov/assembly/GCF_000298735.1/

4) Yak (Bos grunniens) genome assembly (Yak genome 1.1 version) : http://me.lzu.edu.cn/yak/#main_tabs=3

5) Water buffalo (Bubalus bubalis) genome assembly (UMD_CASPUR_WB_2.0): https://www.ncbi.nlm.nih.gov/assembly/GCA_000471725.1/#/st

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[52]. LASTZ documentation can be found at the following link : http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html.

The following parameters to run LASTZ:

--nogapped: Skip gapped extension when doing alignment

--notransition: Don't 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[53].

White Blood Cell eQTL

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 3 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 white blood cell (WBC) stabilisation procedure in the RiboPure™ blood kit (Ambion by Life Technologies) protocol. Whole blood cell samples were then transferred to the main laboratory on ice, then stored at -20^{o}C.

RNA was extracted from WBC using the RiboPure Blood Kit (Ambion) according to 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 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. One hundred fifty base paired-end reads were called with bcltofastq and output in fastq format. Sequence quality was assessed using FastQC. QualityTrim (https://bitbucket.org/arobinson/qualitytrim) was used to trim and filter 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, or greater than 3 no calls (Ns), or with greater than three consecutive bases having quality score less than 15, or 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[54,55]. Custom computer scripts were used to assess sequencing performance, library quality and alignment quality. Alignment files (.bam) for WBC libraries with >12.5 million read pairs (after quality control filtering) and also having >80% mapping rate were retained for gene count matrix generation. Gene counts for the aforementioned alignment files were created using the python package HTSeq[56]. Counts were combined to form a gene by sample count matrix. This count matrix was then normalised to take into account library size using the R software package, DESeq[57].

Whole genome sequence data were imputed into 630K genotypes for the cows using the bull whole genome sequences in Run4 of the 1,000 bull genomes project. After removing variants that had a minor allele frequency less than 0.05 for the cows in the experiment, 10.4 million variants remained. Only genes that were expressed in the WBC for more than 25% of the cows were analysed, 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 were tested (ignoring trans effects on other chromosomes). That is, 11,089 genome wide association analyses were run, with up to 690,000 variants (eg. for chromosome 1, there were this many sequence variants tested for each gene). Association testing was performed with EMMAX[4] fitting the genomic relationship matrix among cows to control for population structure, and 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 (due to very low counts 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 SNP. If a threshold of $P<10^{-5}$ was used, the false discovery rate is 1.3%, Table S7, which seemed reasonable.

Although 73,840 significant variants were detected at the $P<10^{-6}$ threshold, they were associated with only 659 genes. This indicates that multiple variants, in strong linkage disequilibrium, 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 (Figure S8).

Selection signature analysis

Genome scans for selection were performed using FLK[35] and hapFLK[36], two tests that identify regions of high differentiation between populations. Fifteen populations were considered, listed in Table S8, and unrelated animals were selected within each population. The selection was done by excluding animals found outliers from their reported breed, based on their PCA coordinates. Then, within each breed, unrelated animals were selected based on the genomic relationship matrix kinship coefficients, computed using GCTA[5].

FLK and hapFLK were calculated with the hapflk software (https://forge-dga.jouy.inra.fr/projects/hapflk) , using the 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[57] and SNPs corresponding to an FDR of 5% were called significant.

Enrichment analysis

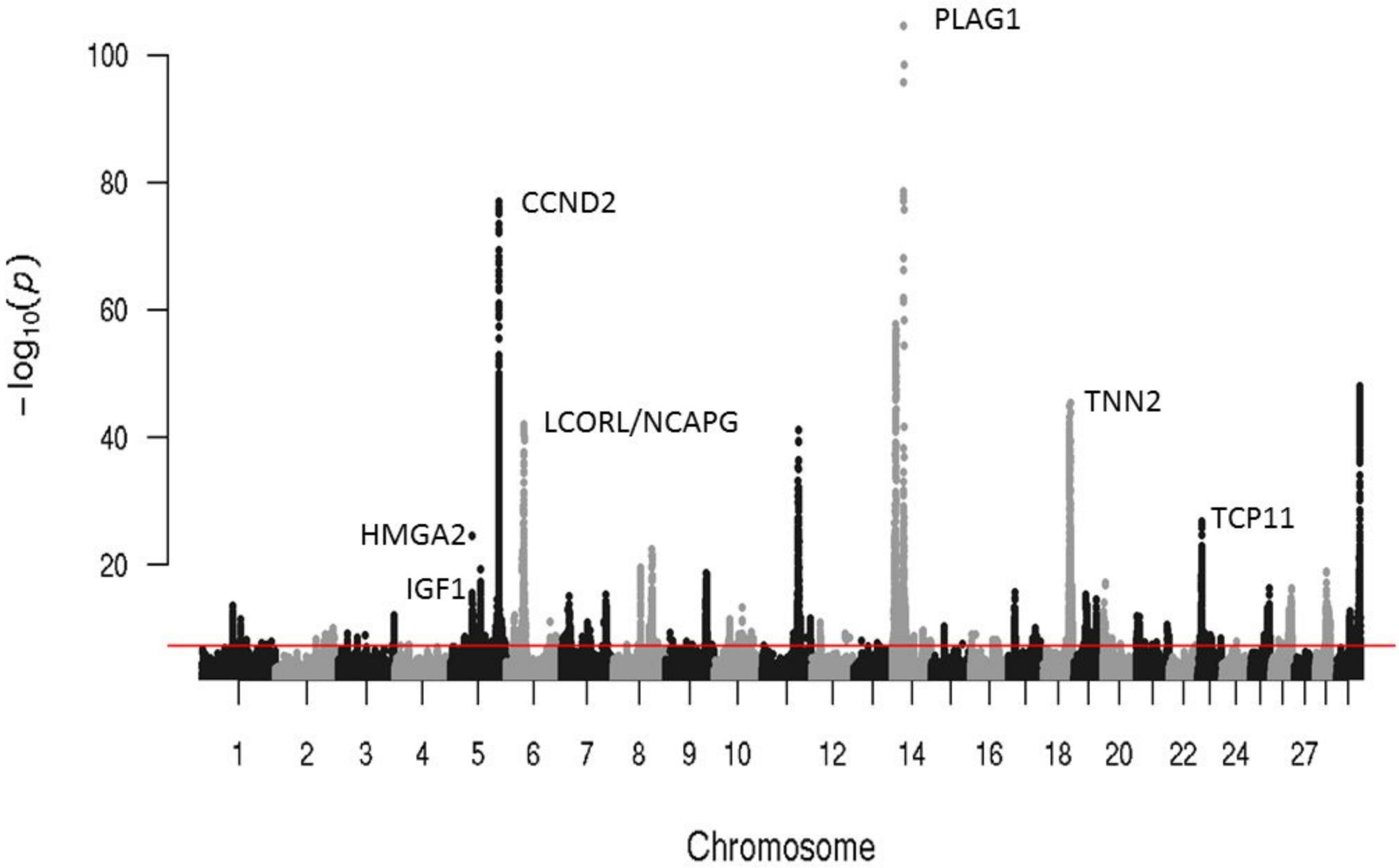
An enrichment analysis among GWAS hits was performed based on a stratified FDR approach[58]. FLK p-values of all SNPs were divided into two sets: a set of GWAS hits, and the set of non-GWAS hits. Within each set, the proportion of true positives ($1-\pi_0$) was estimated with the qvalue R package. The enrichment in the GWAS set 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 using the 163 SNPs in place of all GWAS hits.

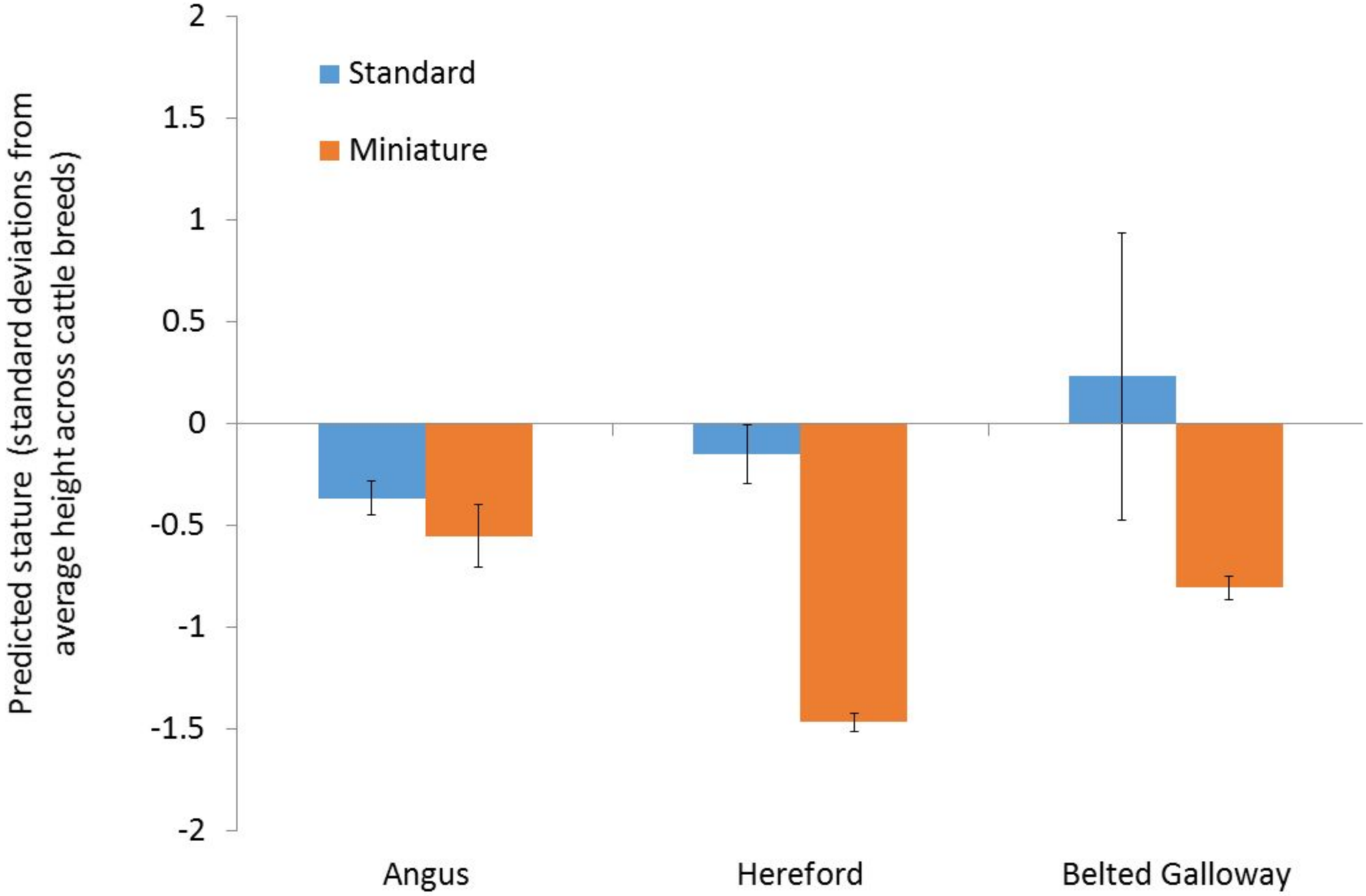
To assess the significance of the enrichment of selection signatures in cattle GWAS hits, the same procedure was applied to human GWAS regions. We extracted human GWAS hits from the human GWAS catalog (https://www.ebi.ac.uk/gwas/)[59]. We considered only the 35 traits that had more than 150 hits in the 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, to define 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 Table S9. Only human traits with enrichment > 1 are shown.

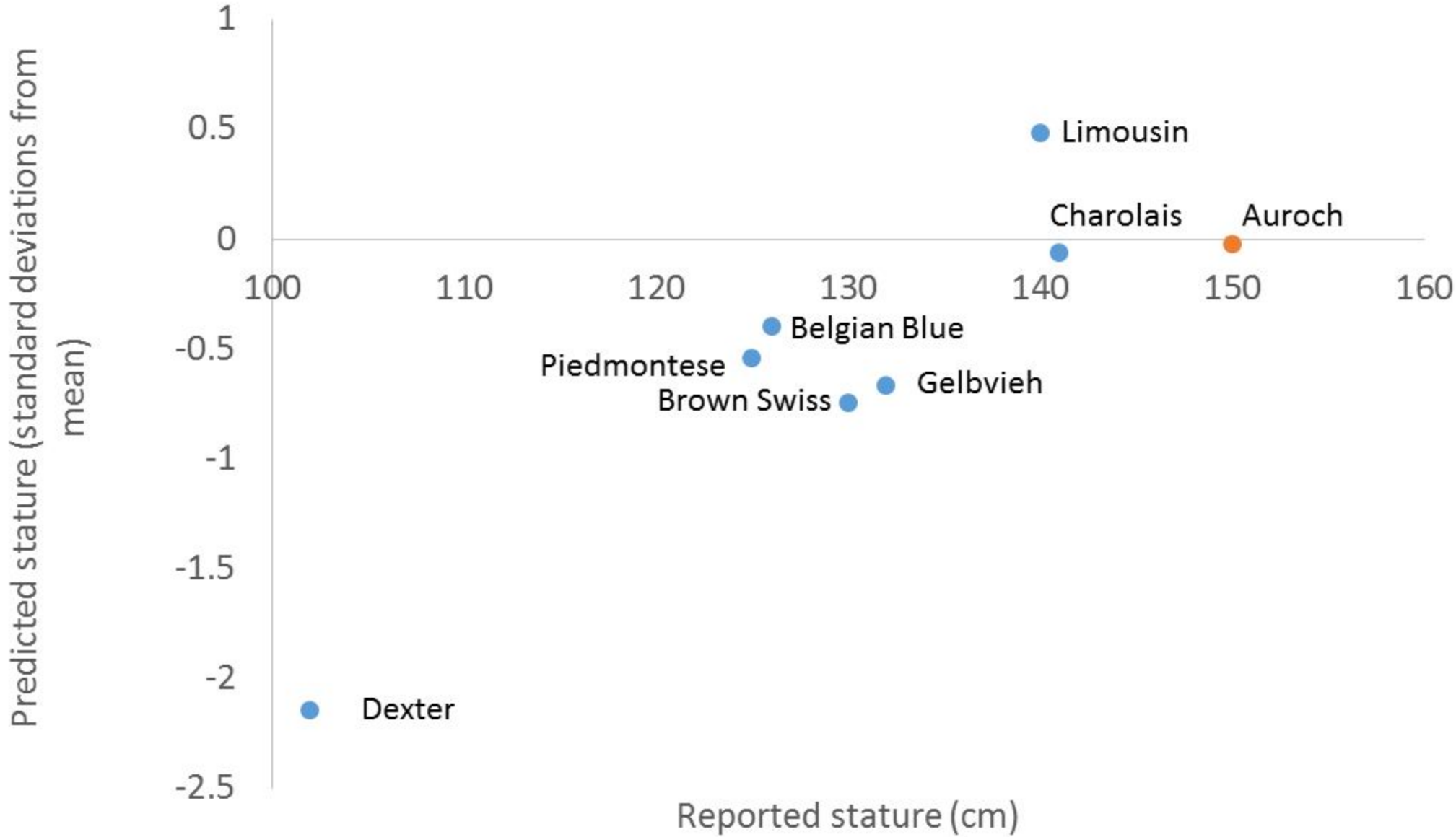
Tests for detection of known casual mutations affecting fat and protein percentage in milk of dairy cattle

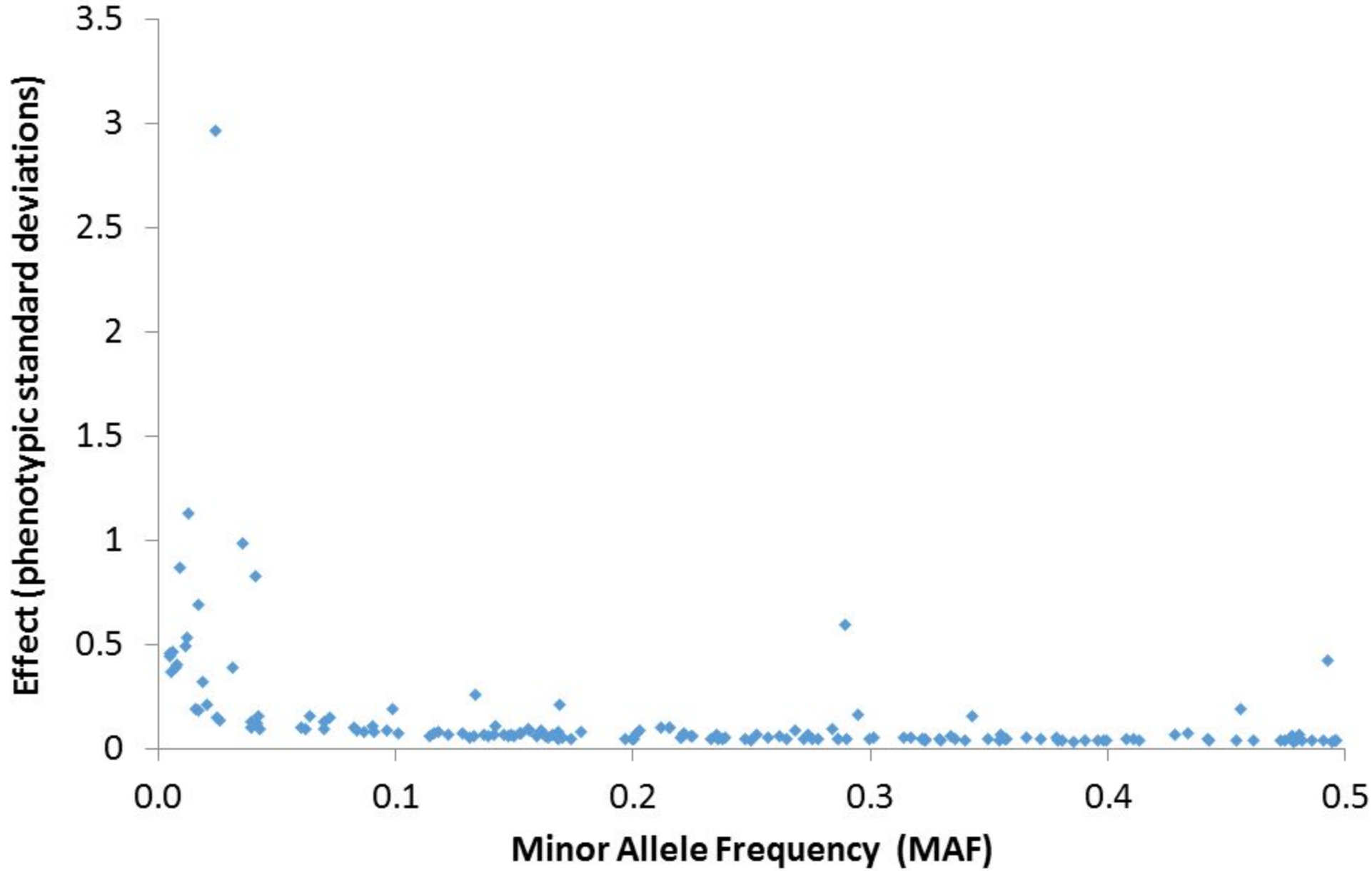
We performed association tests between the imputed sequence variant genotypes and protein percentage and fat percentage in milk in Holstein, Fleckvieh and Brown Swiss cattle. The known mutations were in the Growth hormone receptor gene (GHR GHR:p.Y279F-mutation, chromosome 20[60]), p.A232K in the DGAT1 gene[61] on chromosome 14, and p.Y851S mutation in the ABCG2 gene[62] 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.

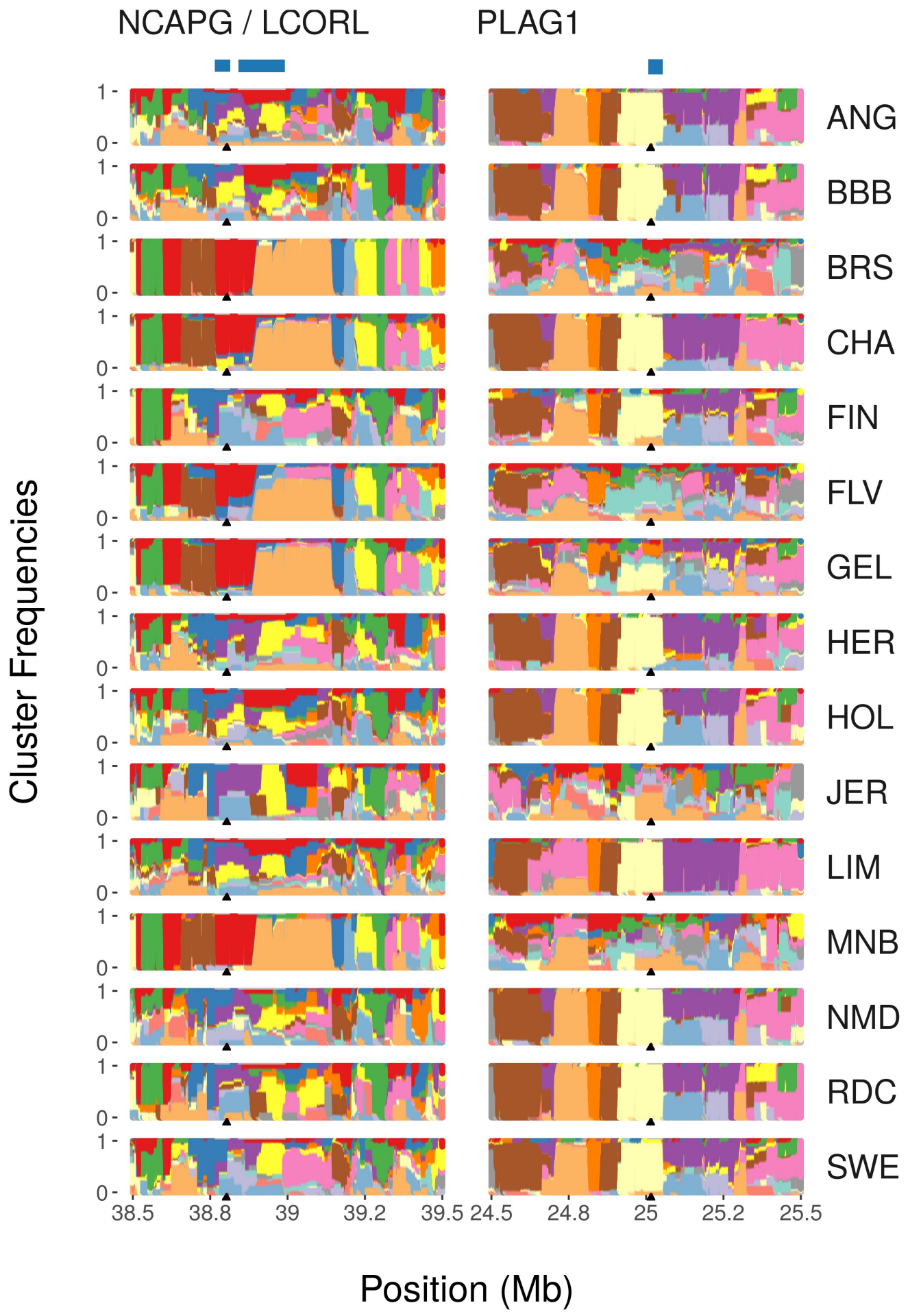
The analysis is that presented in Pausch et al[10]. However Figures demonstrating that imputed sequence data could discover known causative mutations were not presented in that manuscript and are presented here. There were 214 BSW and 345 HOL animals were genotyped using the Illumina BovineHD Beadchip that comprises 777,962 SNPs. All other animals were genotyped using the Illumina BovineSNP50 Beadchip that comprises 54,609 SNPs. The BSW, and HOL and FLECK animals were imputed to higher density using FImpute[49] and Minimac[48], respectively. The final dataset included 573,650 and 564,374 autosomal SNPs. Sequence variant genotypes were imputed in 6777 Fleckvieh, 5204 Holstein and 1646 Brown Swiss animals using the 1000 bull genomes Run4 multi-breed reference population with Minimac[48]. 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*[4] using the '-Z'-flag to consider predicted allele dosages for the imputed sequence variants.


PLAG1
CCND2
LCORL/NCAPG
TNN2
TCP11
HMGA2
IGF1
100
80
60
40
20
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
27
Chromosome


Predicted stature (standard deviations from average height across cattle breeds)
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Standard
Miniature
Angus
Hereford
Belted Galloway


Predicted stature (standard deviations from mean)
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
100
110
120
130
140
150
160
Reported stature (cm)
Limousin
Charolais
Auroch
Belgian Blue
Piedmontese
Gelbvieh
Brown Swiss
Dexter


Effect (phenotypic standard deviations)
3.5
3
2.5
2
1.5
1
0.5
0
0.0
0.1
0.2
0.3
0.4
0.5
Minor Allele Frequency (MAF)


NCAPG / LCORL
PLAG1
ANG
BBB
BRS
CHA
FIN
FLV
GEL
HER
HOL
JER
LIM
MNB
NMD
RDC
SWE
1
0
38.5
38.8
39
39.2
39.5
24.5
24.8
25
25.2
25.5
Cluster Frequencies
Position (Mb)