bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mutant alleles differentially shape cattle complex traits and fitness

Ruidong Xiang[1,2,*], Ed J. Breen[2], Sunduimijid Bolormaa[2], Christy J. Vander Jagt[2], Amanda J. Chamberlain[2], Iona M. Macleod[2], Michael E. Goddard[1,2]

[1] *Faculty of Veterinary & Agricultural Science, The University of Melbourne, Parkville 3052, Victoria, Australia*

[2] *Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria 3083, Australia.*

[*]Corresponding Author: ruidong.xiang@unimelb.edu.au

Abstract

Classical mutations tend to be deleterious to traits and fitness. Is this the case for mutations with polygenic effects? Here, we infer ancestral and mutant alleles (MAs) for 8 million sequence variants in 113k cattle and quantify the effects of MA on 37 complex traits. Heterozygosity at sites conserved across 100 vertebrates increase fertility, stature, and milk production, positively associating these traits with fitness. MAs decrease fat and protein concentration in milk and stature but increase gestation length and somatic cell count in milk indicative of mastitis. However, the frequency of MAs that decrease fat and protein concentration and stature and increase gestation length and somatic cell count is lower than the frequency of MAs with the opposite effect. These results suggest bias in the direction of effect of mutation (e.g. towards reduced protein in milk), but selection operating to reduce the frequency of these MAs. MAs with a large-effect decrease protein and milk yield, while small-effect MAs increase the two traits. These results imply two classes of genomic sites subject to long-term selection: sites conserved across vertebrates show hybrid vigour while sites subject to less long-term selection show a bias in mutation towards alleles that are selected against.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Classical mutations, with a large effect on phenotype, tend to decrease fitness, decrease fitness-related traits and be partially recessive [1-3]. However, the majority of the genetic variance in complex traits is due to mutations of small or polygenic effect. Do these mutations show the same characteristics as classical mutations? Mutations with small effect on fitness tend to be deleterious [4]. However, how mutations affect complex traits such as body size, health and fertility is unknown.

Cattle presents a unique opportunity for studying the effects of mutation. The cattle family diverged from artiodactyls up to 30 million years ago [5]. Modern cattle are derived from at least two different subspecies of wild aurochs, i.e., *Bos primigenius primigenius* (Eurasian aurochs) and *Bos primigenius namadicus* (Indian aurochs) which diverged up to 0.5 million years ago [6-13]. Domestication of *Bos p. primigenius* led to the humpless *Bos taurus* subspecies, which has evolved some highly productive breeds for agriculture, such as the famous black-and-white Holstein breed with superior milk productivity. Besides natural selection, dairy cattle breeds experienced very recent and intensive selection for milk production traits [14,15] and stature [16]. Domestication of *Bos p. namadicus* gave rise to the humped *Bos indicus* subspecies which evolved breeds with strong resistance to hot climates, such as Brahman and Gir cattle.

A better understanding of the consequence of mutations not only updates scientific knowledge but also has practical implications. Domestic cattle support humans with food, labour, clothing material and transportation. Today, there are over 4 billion cattle across the world and over ~900 million tonnes of dairy products have been produced annually for human consumption (http://www.fao.org/3/ca8341en/CA8341EN.pdf). When practicing genomic selection, which is widely used in animal breeding [17], it would be an advantage to know a priori whether mutations are more likely to increase or decrease traits of interest. In

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

particular, if a trait is related to fitness, one might expect new mutations to be deleterious [2,18]. Traits that are related to fitness typically show inbreeding depression and heterosis caused by directional dominance. Here, we introduce a method testing for directional dominance by estimating the effect of heterozygosity on traits of cattle and use this method to identify traits that are associated with fitness. Genomic sites that are conserved across many species are likely to be functional sites [19] and mutations at conserved sites may be highly related to fitness. Therefore, we consider conserved sites and other polymorphic sites in this analysis. In the present study, we use yak, sheep and camel as outgroup species to identify cattle ancestral alleles and quantified the effect and direction of its alternative, i.e., mutant alleles, at 8M sequence variant sites on 37 traits of 4 breeds of over 113k modern cattle. We estimate the effect of heterozygosity on these traits using both conserved sites and all genomic sites, compare mutant allele effects, allele frequency and the fixation index across traits related to milk production, fertility and body size. We also expand the analysis to additional genomes of ancient and modern cattle from the 1000 Bull Genomes database [20,21], which provides a new picture of the effects of mutations on mammalian complex traits.

Results

Genomic heterosis at sites conserved across 100 vertebrates

To identify traits related to fitness, we introduced a method to estimate the effect of heterozygosity on 37 traits in over 100k animals (Described in Supplementary Table 1). We fit average heterozygosity at 317,279 sites conserved across 100 vertebrates (H'_{cons_j}) and heterozygosity from the other 15,718,164 sites ($H'_{non-cons_j}$) simultaneously (see Methods). We observed significant genomic heterosis at conserved sites for the yield of protein (Prot), fat (Fat) and milk (Milk), survival (Surv), fertility performance (Fert), stature (Stat) and angularity (related to slimness and milk yield) (Figure 1 and Supplementary Figure 1). For all

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these traits, $H'_{non-cons_j}$ was not significant when fitted together with H'_{cons_j}. This implies that milk production, fertility, survival and stature are related to fitness and that heterosis for these traits is exclusively explained by genomic sites conserved across vertebrates. To be conserved across vertebrates, mutations at these sites must be deleterious, implying extremely long-term consistent selection for the ancestral allele at these sites.

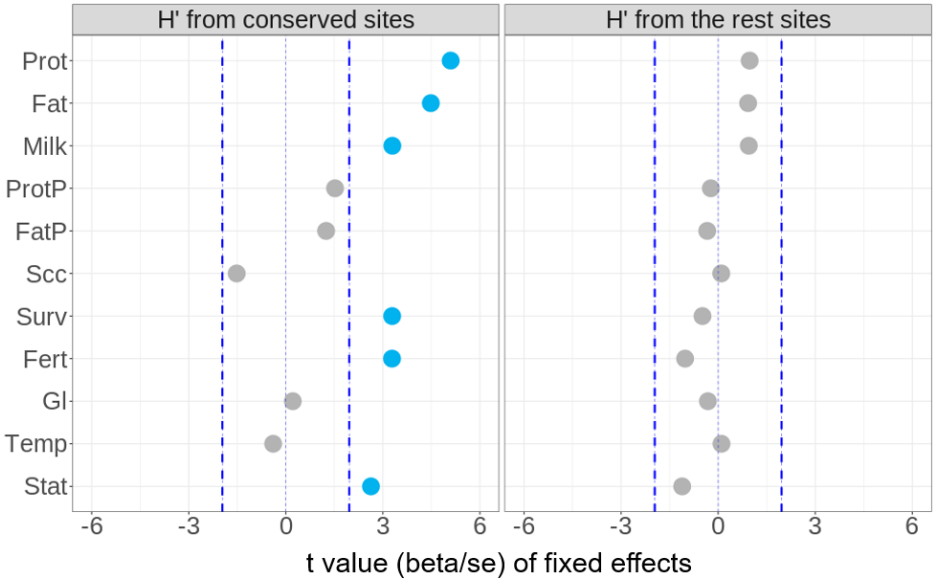

Figure 1. Genomic heterosis at conserved sites (*H'*) for traits of 104k cows. The beta values and standard errors for each trait were generated using a mixed linear model, fitting *H'* from 317,279 conserved sites (left panel) and *H'* from the rest 15,718,164 sites (right panel) together with other fixed effects (e.g., breed). Blue dashed lines indicate t value of -1.96 and 1.96 commonly used to indicate the significance.

Assignment of bovine ancestral and mutant alleles

To assign the mutant alleles in cattle, we first determined the alternative, ancestral alleles using artiodactyls, including cattle as the focal species (98 global cattle breeds from the 1000 Bull Genomes Project [20,21], Supplementary Table 2) and yak, sheep and camel as outgroup ancestor species (Ensembl 46-mammal sequence data). A probabilistic method [22] was used to assign an ancestral allele for each site mappable between 4 artiodactyl species (see Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Out of 42,573,455 equivalent sites between the 4 species, 39,998,084 sites had the ancestral allele assigned with high confidence (probability > 0.8). We compared our results with a previous study using different methods [23]. Of 1,925,328 sites that were assigned ancestral alleles with high confidence in both studies, 1,904,598 (98.7%) sites agreed. However, we have assigned ancestral alleles with high confidence to ~10 times more sites than the previous study due to the use of large sample size and whole-genome sequence data. The full results are publicly available at https://figshare.com/s/dd5985b76a413b56106b.

Biases in trait effects between ancestral and mutant alleles

We first conducted GWAS of 37 traits using over 16 million imputed sequence variants in bulls (N ~ 9k) and cows (N ~ 104k) separately (see Methods). For 7,910,190 variants where the ancestral allele was assigned, we compared the direction (increase or decrease) of the effect of the mutant alleles (MAs) on the trait (Supplementary Figure 2-3). The same comparison was also performed for variants at 317,279 conserved sites. Note that for a variant, the effect of a MA is identical to -1 × the effect of the ancestral allele. We focus the description of effects on MAs, but a MA increasing the trait is identical to an ancestral allele decreasing the trait. Within all analysed variants and conserved variants, for each trait we considered the following three variant categories for systematic comparison: 1) large-effect variants, i.e., p_{gwas} < 5e-8 and the effect direction agreed in both sexes; 2) medium-effect variants, i.e., 5e-8 <= p_{gwas} < 5e-5 and the effect direction agreed in both sexes, and 3) small-effect variants, i.e., 5e-5 <= p-value of GWAS (p_{gwas}) < 0.05 and the effect direction agreed in both sexes. Here the effect size refers to the amount of variance explained by variants which is inversely related to the p-value. The use of different effect size is because mutations of small and large effects may be different in their direction of effect. Selecting variants that have the same effect direction between independent GWAS populations [24], such

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as bulls and cows, helps to eliminate variants with spurious trait associations from the comparison. Based on a previous method [24], the True Discovery Rate by Effect Direction (TDRed) of GWAS between two sexes across 37 analysed traits for the small-, medium- and large-effect variants was 0.8, 0.98 and 0.99, respectively.

Based on GWAS results of each trait, we calculated the ratio of the number of variants where the MA increased the trait (positive effect) to the number of variants where the MA decreased the trait (negative effect). Across 37 traits and three effect-size groups, MAs showed diverse trait effect patterns (Supplementary Figure 3). Results observed from GWAS were confirmed by BayesR analysis [25], which jointly fits on average 4.3 million variants per trait (See methods and Supplementary Figure 3). Based on jointly estimated effects for a given set of variants, the significance of the effect direction bias using Kolmogorov-Smirnov was tested to estimate the p-value (p_{ks}) of the difference in the effect distribution between ancestral and mutant alleles (see Methods). We also tested the significance of bias using LD-clumped ($r^2 <$ 0.3) [26] variants to calculate the standard error (Supplementary Figure 4). In the following text, we focus on 1) MAs within of large- and small-effect categories for milk production traits as these two sets of MAs showed distinct effect direction patterns (Figure 2), and 2) MAs associated with other traits, including those with medium or small effects on somatic cell count (Scc, indicative of mastitis, medium-effect), survival (Surv, small-effect), fertility (Fert, frequency of pregnancy, small-effect), gestation length (Gl, medium effect), temperament (Temp, docility, small-effect) and stature (Stat, medium effect) (Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

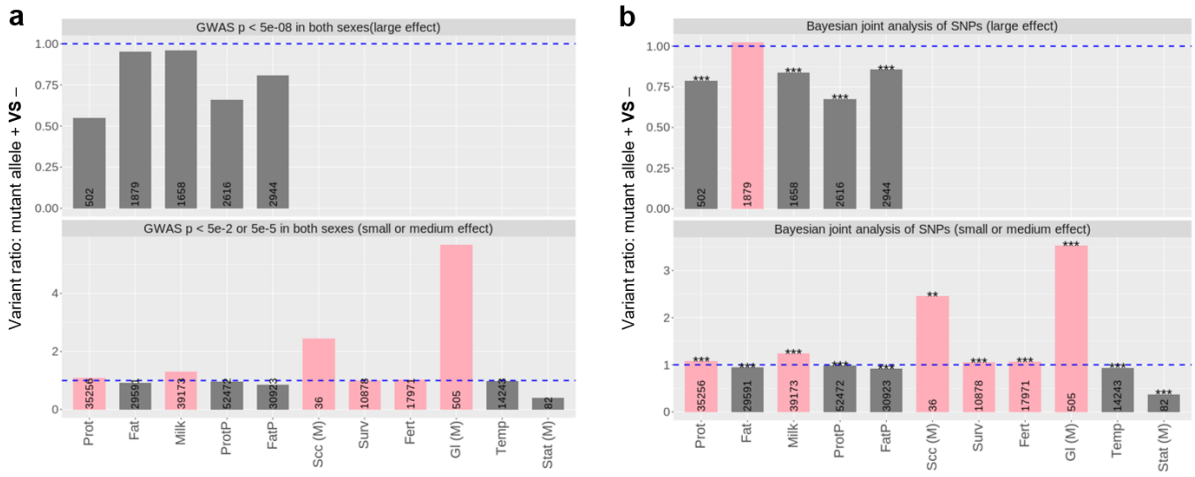

Figure 2. The ratio (y-axis) between the number of variants with mutant alleles increasing the trait (+) and the number of variants with mutant alleles decreasing the trait (−). GWAS effects of mutant alleles are shown for all variants (**a**). BayesR joint effects of mutant alleles from the same variants in (**a**) are shown for all variants (**b**). Pink colour: the majority of variants with mutant alleles tend to increase the trait (taller than the blue-dashed line). Dark grey: the majority of variants with mutant alleles tend to decrease the trait (shorter than the blue-dashed line). Numbers in bars: total number of variants significant at the given threshold. Stars: p-value for the significance of the difference in the distribution of BayesR effects between ancestral and mutant alleles, '*': $p < 0.05$, '**': $p < 0.01$, '***' $p < 0.001$. For somatic cell count (Scc), gestation length (Gl) and stature (Stat), the results are from medium-effect (M) variants and the full results are shown in Supplementary Figure 3.

The classical model [1-3] predicts that the majority of MAs, or mutations, are deleterious or slightly deleterious. In our study, MAs consistently showed biases towards decreasing protein and fat concentration (Figure 2 and Supplementary Figure 3,4), docility and stature, and towards increasing somatic cell count (an indicator of mastitis) and gestation length. For milk yield and protein yield the bias in the direction of MA depends on the size of MA effect. Large-effect MAs tended to decrease milk and protein yield whereas small-effect MAs tended to increase them. A possible explanation is that mutation seldom has a large positive effect on milk protein yield or fertility but small positive effect mutations occur and are increased in frequency by artificial selection.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Also, there was a slight majority of small-effect MAs which tended to increase fertility and survival. The effects of these sets of MAs is partially due to pleiotropy, i.e., the effect of these MAs on multiple traits (Supplementary Table S3). For instance, while small-effect MAs increasing milk yield decreased fat yield, protein % and fat %, they also increased protein yield. Also, while small-effect MAs increasing fertility increased gestation length, they also increased stature.

The simplest explanation for the bias in the direction of MA effects is that it is due to a bias in the direction of mutation. For instance, that mutation more often leads to a decrease in fat% rather than an increase. However, it is also possible that mutations that decrease fat% are selected and therefore more likely to be discovered than mutations that increase fat%. Below we exclude this possibility by comparing the allele frequency at variants where the MA increases or decreases the trait.

Allele frequency of mutant alleles in modern and ancient cattle

Across all variable sites, the allele frequency of MAs was lower than the allele frequency of ancestral alleles (Supplementary Figure 5). Also, the frequency of MAs at conserved sites (0.27) was lower than the frequency of MAs across all sites (0.32). This is consistent with the selection for the ancestral allele which is necessary to maintain conservation of the same allele across vertebrates.

We grouped variants based on their mutant allele reducing (MAs−) or increasing the trait (MAs+) and compared their allele frequency in over 110k Holstein, Jersey, crossbreds and Australian Red bulls and cows (Figure 3a,b). To account for LD, we estimated the error of MA frequency based on LD-clumped ($r^2 < 0.3$) [26] variants. As an external validation, we also

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

considered this analysis in a selection of 7 subspecies/breeds of 1,720 ancient and modern cattle from the 1000 Bull Genomes Project [20,21] (Figure 3c,d).

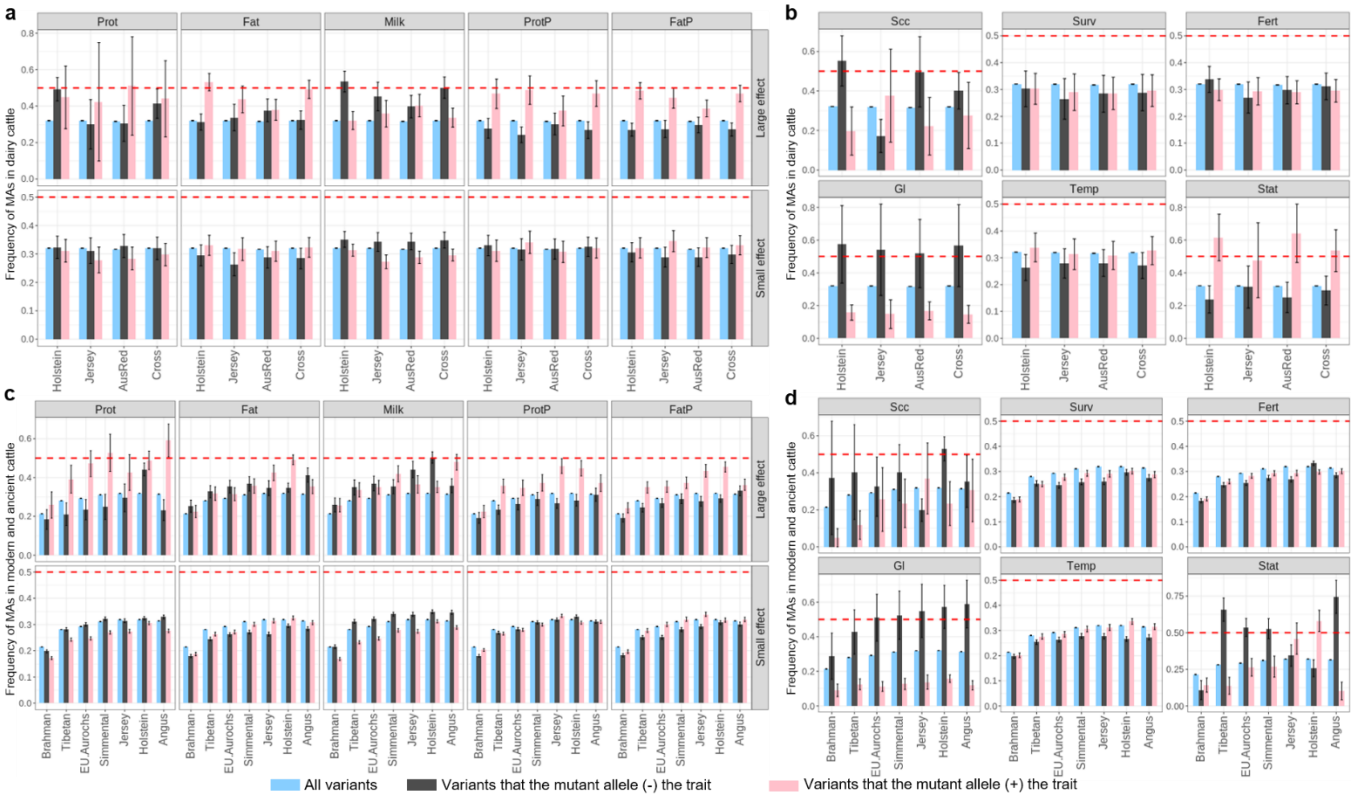

Figure 3. The allele frequency of mutant alleles (MAs) in cattle. The average frequency of variants associated with different traits is shown with standard error bars based on LD clumped variants. All variants include 7.9M variants assigned with mutant alleles. Red dashed line represents the frequency of 0.5. In the dairy cattle selection (**a** and **b**), 90,627 Holstein, 13,465 Jersey, 3,358 Australian Red (AusRed) and 4,649 crossbreds were used. In the ancient and modern cattle (**c** and **d**), 210 Brahman, 25 Tibetan, 10 Eurasian Aurochs, 242 Simmental, 95 Jersey, 840 Holstein and 287 Angus were used. For panels **b** and **d**, results for survival (Surv), fertility (Fert) and temperament (Temp) were from small-effect MAs while results for somatic cell count (Scc), gestation length (Gl) and stature (Stat) were from medium-effect MAs.

For fat%, protein%, docility and stature MAs that increase the trait had higher allele frequency than MAs that decrease the trait. For somatic cell count and gestation length, the reverse is true. That is, MAs increasing somatic cell count and gestation length have lower

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

allele frequency than MAs that decrease the trait (Figure 3). Thus, although MAs more commonly decrease fat% than increase it, the allele frequency is higher at sites where the MAs increase fat%. This implies that selection acts against MA that decreases fat% or favours MAs that increases fat%. Consequently, the higher incidence of MAs that decrease fat% cannot be due to selection favouring them but must be due to mutation more often resulting in an allele that decreases fat% than increases it. Comparing results in Figure 2 and 3 shows that this is the usual pattern – the more common direction of effects of mutation generates alleles that are selected against and hence have a reduced allele frequency.

For other traits, the results are less clear-cut. For milk yield, the majority MAs of large effect tended to decrease the trait (Figure 2). Interestingly, these large-effect milk-decreasing MAs, which were deleterious, had a higher frequency than those MAs increasing milk yield (Figure 3). On the other hand, the majority of MAs of small effect tended to increase the milk yield (Figure 2). Yet, these small-effect MAs that increase milk were at a lower frequency than MAs that decrease milk yield (Figure 3). Interpretation of these results is helped by remembering that milk yield is negatively correlated with fat% and protein% (Supplementary Table 3).

Selection of trait-associated mutant alleles in modern and ancient cattle

The above results for MA frequency at trait-associated variants imply selection. The selection could be consistent across breeds which would limit the divergence of allele frequency between breeds or it could be different between breeds leading to divergence in allele frequency. We compared the average of Wright's fixation index ($\overline{F_{ST}}$), for MA+ variants and MA− variants calculated using dairy cattle (Figure 4a,b) and ancient and modern cattle (Figure 4c,d). To account for LD, we estimated the error of $\overline{F_{ST}}$ based on LD-clumped ($r^2 <$ 0.3) [26] variants.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

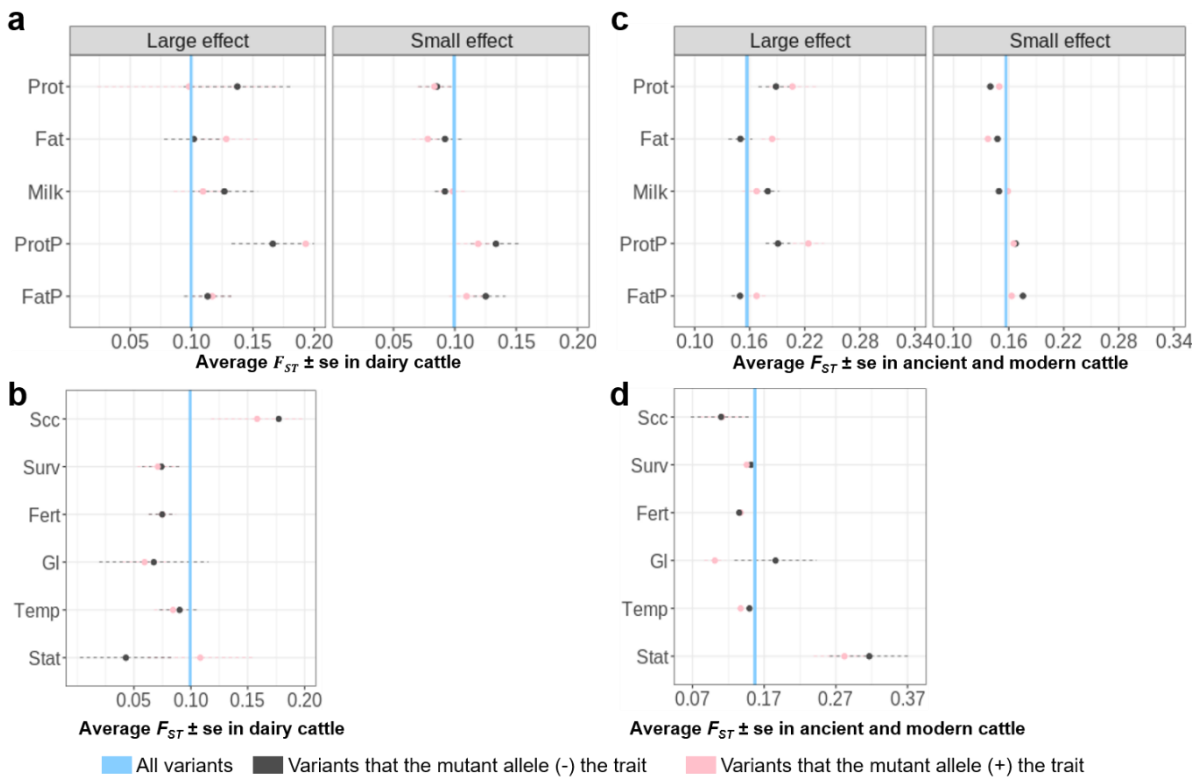

Figure 4. Selection (average Wright's fixation index $\overline{F_{ST}}$) of variants with mutant alleles that increase or decrease the trait in dairy cattle (**a,b**) and ancient and modern cattle (**c**,**d**). The $\overline{F_{ST}}$ is shown as dots with its standard error bars estimated using LD clumped variants. The blue line represents the $\overline{F_{ST}}$ for 7.9M variants analysed (0.1±4.3e-05) in dairy cattle in **a** and **b**; and $\overline{F_{ST}}$ = 0.157±5e-05 in ancient and modern cattle in **c** and **d**). For panels **b** and **d**, results for survival (Surv), fertility (Fert) and temperament (Temp) were from small-effect MAs while results for somatic cell count (Scc), gestation length (Gl) and stature (Stat) were from medium-effect variants.

In general, variants associated with milk production traits (including somatic cell count, Figure 4a) showed higher than average F_{ST} among dairy breeds implying divergent selection, while variants associated with other traits, including survival and fertility, tended to have below-average F_{ST} indicating convergent selection (Figure 4b). $\overline{F_{ST}}$ for gestation length was below average especially for MA+, probably due to selection against mutations that increase gestation length in all breeds (Figure 4d).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Among ancient and modern cattle, $\overline{F_{ST}}$ is high for both MA+ and MA− variants for stature indicating divergent selection for height (Figure 4b). The allele frequency of MAs decreasing height was the least frequent in Holstein cattle and was the most frequent in Tibetan cattle living at high altitude and Angus cattle selected for beef production (Figure 3d). This suggests that the direction of selection could vary across cattle breeds under different environmental conditions and/or artificial selection.

Discussion

Classical mutations, with large phenotypic effects, are typically deleterious, decrease traits related to fitness and are partially recessive. Our results on polygenic mutations affecting complex traits are only partially in agreement with this pattern.

In agreement with the expectation, we observed directional dominance for milk, fat and protein yield, stature, survival and fertility which might be considered fitness traits. However, this directional dominance was only significant at sites that were conserved across vertebrates. Thus it only occurred at sites subject to extremely long-term selection for the same ancestral allele. This long term conservation of the ancestral allele implies that mutation generates less-fit alleles and that the fittest allele is the same regardless of species. This long term selection cannot be directly on traits involving lactation since the same allele is conserved in vertebrates other than mammals.

We observed a different kind of mutational bias affecting fat%, protein%, somatic cell count and gestation length. Mutation more often decreased fat% and protein% and increased somatic cell count and gestation length and selection acted to suppress the allele frequency of these mutations. However, this pattern occurred at non-conserved and conserved sites and these traits did not show significant directional dominance. Selection for the ancestral allele which increases fat% and protein% is unlikely to be due to artificial selection since selection

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for milk volume leads to decreased fat% and protein%. On the other hand, natural selection might act against mutations that decrease fat% and protein% because this would decrease the nutrient value of the milk to the calf. This interpretation is supported by the consistency of the bias in mutation and allele frequency across dairy cattle and other subspecies/breeds including the *Bos indicus* breed Brahman. There are mutations at *DGAT1* and *GHR* loci that increase milk yield but decrease fat% and protein% (Supplementary Figure 6). These are only at appreciable frequency in domesticated cattle especially breeds artificially selected for milk volume.

For milk, fat and protein yield the results differ between mutations of large and small effect. Mutations with a large effect on milk protein yield more often decrease protein yield than increase it perhaps because the physiology supporting milk protein synthesis has been optimised in part at least. Mutations with a small effect on protein yield are almost equally likely to increase or decrease yield perhaps because natural selection favours an intermediate level of milk protein yield because too high a yield drains the cow of nutrients needed for maintenance and reproduction.

The selection which we have observed affecting the frequency of mutations of positive and negative effect could be both natural selection acting over a long period before and since the domestication of cattle, and artificial selection acting over the last 10,000 years and, more intensely, over the last ~100 years in dairy cattle. Artificial selection may differ between breeds and generate high Fst between breeds. For fat%, protein% and stature at least one class of mutation is more common than random mutations and the overall F_{ST} between breeds tended to be high. Our analysis also highlighted some specific breeds. For example, the selection of variants associated with somatic cell count led to high F_{ST} among dairy cattle but low F_{ST} in our other breeds. Holstein cattle have been selected to be tall [16] and this is reflected in the low frequency of MAs decreasing stature in Holstein. On the other hand, the

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

high frequency of MAs decreasing stature in Tibetan cattle (Figure 4d) may be due to its adaptation to high altitude [27].

Pleiotropy may also explain some surprising results. For instance, MAs decreasing fertility tended to be most frequent in the Holstein breed (Figure 3d), perhaps because these alleles tended to increase milk yield and stature.

Although mutation is biased in its effect on some traits, the bias is small for most traits. That is, mutations decreasing protein yield are only slightly more common than mutations that increase protein yield. Also, although conserved sites explain directional dominance and are enriched for polymorphisms affecting complex traits [19], they do not explain the majority of the genetic variance. That is, there are many sites affecting traits, such as milk yield and stature, at which the allele carried varies between species implying that the fittest allele varies depending on the environment and the background genotype of the species.

The sequence variants associated with a complex trait are not necessarily causal but likely to be in high LD with the causal variants. This tends to dilute the signal that might be discovered if causal variants were used. However, variants in high LD may share a similar evolutionary history and therefore show some of the same characteristics. We used BayesR which jointly fits variants and LD-clumping to account for LD. However, we acknowledge that we cannot completely remove the effects of LD on our results. Therefore, future studies with even larger sample sizes, e.g., ~1 million, may update our results.

In conclusion, we provide a new picture of the effects of mutation and selection on mammalian complex traits. This picture shows that biases in the effect direction of mutations on complex traits exist but, despite this, mutations with both positive and negative effects exist for all traits examined. Overall, our results imply that selection that is consistent over the long term maintains the best allele at each site and so mutations at these sites are negative for fitness and often for traits related to fitness. At sites conserved across vertebrates, the

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutations are also partially recessive for fitness-related traits. There is also bias at sites that are not conserved across vertebrates but perhaps have experienced consistent selection in cattle so that mutations are more often negative than positive for instance for fat% in milk. These sites do not show directional dominance. However, there are many sites where the direction of selection varies between species and between breeds within species and where the effect direction of mutations can be either positive or negative.

Methods

Data preparation for calling bovine ancestral alleles. The assignment of bovine ancestral alleles was based on a model comparison of alleles from cattle with alleles from outgroups of yak (*Bos grunniens*), sheep (*Ovis aries*) and camel (*Camelus dromedarius*). For the cattle species, we used whole-genome sequence data of 98 individuals from the Run 7 version of the 1000 Bull Genomes Project [20,21]. Each individual represents a breed collected by the consortium. Only those whole-genome sequence samples with coverage > 10x were selected and if multiple individuals were found for a breed, the whole-genome sequence sample with the highest coverage was chosen. Both *Bos taurus* and *Bos indicus* subspecies were included (Supplementary Table 2). Whole-genome sequence data in VCF format for these 98 cattle, as a subset from the 1000 Bull Genomes database, was generated for further analysis.

For the outgroup species, we used whole-genome sequence data of 46 mammals stored in the Multiple Alignment File generated by Ensembl EPO pipeline (http://asia.ensembl.org/info/genome/compara/multiple_genome_alignments.html). The 46-mammal EPO Multiple Alignment File was downloaded. Then, the software WGAbed (https://henryjuho.github.io/WGAbed/) from python v2.7 was used to retrieve sequence data for cattle, yak, sheep and camel in bed file format. Only sites with sequence data available in at least one outgroup species were kept. Using the cattle coordinates in the 4-species

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

WGAbed files, the sequence data of the outgroup species were matched with the 98 cattle. As a result, 42,573,455 sites found in the 98 cattle and in at least one outgroup species were found. Sequence data on these 42,573,455 sites across 4 species were used to determine the bovine ancestral alleles.

Probabilistic determination of bovine ancestral alleles. We used the method proposed by Keightley et al [22] with the model choice of The Kimura two-parameter (K2) which accounts for allele frequency of the focal species to determine the probability of an allele being ancestral at each available site. The method was implemented in estsfs [22] and the K2 model was chosen due to its equivalent accuracy to other models but better computation efficiency. As described above, the sequence data of three outgroup species were used. The order of phylogenetic tree topology was cattle → yak → sheep → camel. As requested by the software, allele counts of A, C, G and T were determined for the focal species (cattle) and for out species at each available site. For cattle, the total allele count for each site was 196 (98 ×2). For each outgroup species, the total allele count for each site was up to 1. Missing sequence data in the outgroup species were treated as 0 counts. For each site, estsfs produced a probability (P_{ancs}) of the major allele in the focal species being ancestral. We then determined alleles which were major at a site with $P_{ancs} > 0.8$ or those alleles which were minor at a site with $P_{ancs} < 0.2$ to be ancestral. For those sites where the major or minor alleles could not be determined but the $P_{ancs} > 0.8$ or < 0.2, the cattle allele with the highest frequency in the 3 out species was assigned ancestral. The rest of the sites were determined as ambiguous where no clear ancestral alleles could be determined. The detailed results of ancestral alleles for those 42,573,455 sites across 4 species and the probability of the alleles being ancestral or ambiguous is publicly available at: https://figshare.com/s/dd5985b76a413b56106b.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequence variants under conserved sites across 100 vertebrate species. The variant selection followed previous procedures [19]. Briefly, conservation was determined by the criteria of PhastCon score [28] > 0.9 based on the sequence data of those 100 species. The conserved sites were primarily determined using the human genome coordinates (hg38) and were lifted over to the bovine genome ARS-UCD1.2 using the LiftOver software (https://genome.ucsc.edu/cgi-bin/hgLiftOver) with a lift-over rate > 92%. In total, 317,279 variants in the current study were assigned as the conserved variants.

Animals and phenotypes for variant-trait association analysis. Data was collected by farmers and processed by DataGene Australia (http://www.datagene.com.au/) for the official May 2020 release of National breeding values. No live animal experimentation was required. Phenotype data was based on trait deviations for cows and daughter trait deviations for bulls. Daughter trait deviations were the average trait deviations of a bull's daughters and all phenotypes were pre-corrected for known fixed effects, with processing done by DataGene. Phenotype data used included a total of 8,949 bulls and 103,350 cows from DataGene. . Holstein (6,886♂ / 87,003♀), Jersey (1562♂ / 13,353♀), cross-breed (36♂ / 5,037♀) and Australian Red dairy (265♂ / 3,379♀)breeds were included. In total, 37 traits related to milk production, mastitis, fertility, temperament and body conformation were studied (Supplementary Table 1). Larger trait values of fertility (Fert), ease of birth (Ease), temperament (Temp), milking speed (MSpeed), likeability (Like) meant poor performances, so to assist the interpretability of the study, we have corrected the trait direction so that larger values of Fert, Ease, Temp, MSpeed and Like meant increased fertility performance (calving frequency), labour ease, docility, milking speed and the overall preference as a dairy cow (Supplementary Table 1).

Genotype data for association analysis. The genotypes used in the current study included a total of 16,035,443 imputed bi-allelic sequence variants with Minimac3 [29,30] imputation

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

accuracy $R^2 > 0.4$ and the minor allele frequency (MAF) > 0.005 in both sexes. Most bulls were genotyped with a medium-density SNP array (50K: BovineSNP50 Beadchip, Illumina Inc) or a high-density SNP array (HD: BovineHD BeadChip, Illumina Inc) and most cows were genotyped with a low-density panel of approximately 6.9k SNPs overlapping with the standard-50K panel The low-density genotypes were first imputed to the Standard-50K panel and then all 50K genotypes were imputed to the HD panel using Fimpute v3 [19,31]. Then, all HD genotypes were imputed to sequence using Minimac3 with Eagle (v2) to pre-phase genotypes ([30,32]). The sequence reference set for imputation used Run7 of the 1000 Bull Genomes Project [20] aligned to the ARS-UCD1.2 reference bovine genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1/) [21,33].

Genome-wide association studies. Above mentioned traits were analysed one trait at a time independently in each sex with linear mixed models using GCTA [34]:

$$\mathbf{y} = \mathbf{mean} + \mathbf{breed} + \mathbf{bx} + \mathbf{a} + \mathbf{error} \quad (equation\ 1)$$

where **y** = vector of phenotypes for bulls or cows, **breed** = three breeds for bulls, Holstein, Jersey and Australian Red and four breeds for cows (Holstein, Jersey, Australian Red and MIX); **bx** = regression coefficient *b* on variant genotypes **x**; **a** = polygenic random effects ~N(0, $\mathbf{G}\sigma_g^2$) where **G** = genomic relatedness matrix based on all variants.

Bayesian mixture model analysis. In the above-described GWAS, sequence variants, many of which are in high LD, were analysed one at a time. In order to assess variant effects and account for LD, we fitted selected variants jointly in BayesR [25]. For each trait, variants that showed the same sign between bulls and cows (regardless of p-value) and can be assigned with an ancestral allele were analysed with BayesR. Across 37 traits, the number of variants analysed ranged from 3,961,180 to 4,737,492. To reduce the computational burden of

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

BayesR, we estimated the joint effects of these variants for each trait in bulls. BayesR models the variant effects as mixture distribution of four normal distributions including a null distribution, $N(0, 0.0\sigma^2{}_g)$, and three others: $N(0, 0.0001\sigma^2{}_g)$, $N(0, 0.001\sigma^2{}_g)$, $N(0, 0.01\sigma^2{}_g)$, where $\sigma^2{}_g$ was the additive genetic variance for the trait. The starting value of $\sigma^2{}_g$ for each trait was estimated using GREML implemented in the MTG2 [35] with a single genomic relationship matrix made of all 16M sequence variants. The statistical model used in the single-trait BayesR was:

$$\mathbf{y} = \mathbf{Wv} + \mathbf{Xb} + \mathbf{e} \text{ (equation 2)}$$

where **y** was a vector of phenotypic records; **W** was the design matrix of marker genotypes; centred and standardised to have a unit variance; **v** was the vector of variant effects, distributed as a mixture of the four distributions as described above; **X** was the design matrix allocating phenotypes to fixed effects; **b** was the vector of fixed effects, including breeds; **e** = vector of residual errors. As a result, the effect *b* for each variant jointly estimated with other variants were obtained for further analysis.

The difference in effect distribution between ancestral and mutant alleles. For an analysed variant, one allele is ancestral and then the other is mutant. If there is a bias in effect direction in ancestral alleles or mutant alleles in a given set of variants, the effect distribution of the ancestral and mutant alleles would be different. We tested if the distribution of the effect of ancestral alleles estimated from BayesR was significantly different from that of mutant alleles using the two-sample Kolmogorov-Smirnov test implemented ks.test() in R v3.6.1. The coding was ks.test(a,m) where a was the vector of variant effects based on the ancestral alleles and m was a vector of variant effects based on the mutant alleles. To be more conservative, we also tested the significance of biases using LD-clumped ($r^2 < 0.3$ within 1Mb windows) variants with small, medium and large effects using default settings in plink1.9 [26].

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Heterozygosity of individuals at conserved sites. It is widely accepted that higher genomic heterozygosity is linked to gene diversity, therefore, fitness. However, it is not clear at which set of genes or variants heterozygosity is more related to fitness. Also, the simple estimation of heterozygosity, i.e., assigning allele counts of 0 or 2 as homozygous and 1 as heterozygous, leads to biases as the estimation is not independent of additive effects (illustrated later). Our previous work showed conserved sites across 100 vertebrate species significantly contribute to trait variation [19,36] and it is also logical to assume mutations at conserved sites tend to have strong effects on fitness. Therefore, we firstly partitioned the genome into 317,279 conserved and 15,718,164 non-conserved variants. Then, we re-parameterised the genotype allele count for each variant commonly used to model the dominance deviation, so that the estimation of dominance deviation is independent of the additive effects. We focused on cows because their traits were largely measured on themselves, contrasting to bull traits which were based on their daughters' traits. We estimated the variant-wise sum of the re-parameterised allele count value for dominance deviation which was later termed as z'_{D_i} for each variant *i* in cows. The sum was averaged by the number of variants and this average value based on re-parameterised dominance allele count for the individual *j* was termed as H'_j to represent the individual heterozygosity. We estimated the individual heterozygosity from conserved sites (H'_{cons_j}) and non-conserved sites ($H'_{non-cons_j}$) and these computations are specified in the following text.

According to quantitative genetics theory [37-39], the genetic value (G') of an individual can be partitioned into the mean (μ), additive genetic value (A) arising from additive effect (*a*) and dominance genetic value (D) arising from dominance deviation (*d*). At a single locus, let the allele frequency of the three genotype classes of AA, AB and BB be p^2, $2pq$ and q^2, respectively. In a simple genetic model, the genetic value can be decomposed as:

$$G' = \mu + A + D + e = \mu + x_{A_i}a + z_{D_i}d + e \text{ (equation 3)}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Where x_{A_i} was the allele count for genotype AA, AB and BB for locus or variant *i* which were usually coded as 0, 1, 2, respectively, to represent the additive component, and z_{D_i} was usually coded as 0, 1, 0, for genotype AA, AB and BB for variant *i*, respectively, which differentiates the homozygous and heterozygous to represent the dominance component. Therefore, in the simplest form, the genome-wide heterozygosity of the individual *j* can be calculated as:

$$H_j = {\sum_i^N z_{D_i}}\Big/{N} \text{ (equation 4)}$$

where H_j is the simple genome-wide heterozygosity of individual j, N is the total number of variants. Note that such calculation of H_j can also be used to derive inbreeding depression, where $I_j = (\sum_i^N 2p_i q_i) \times H_j$. I_j was the inbreeding coefficient for the j_{th} individual.

In equation 3, however, due to the non-zero correlation between x_A and z_D under Hardy-Weinberg equilibrium (HWE), the estimation of *a* and *d* is not independent. This then resulted in the estimation of H_j was not orthologous to the additive components. Therefore, we proposed to re-parameterise this model to estimate *a* and *d* independently.

According to Falconer [38] at this locus, the additive effects can be derived using the regression of genetic value on the number of A alleles, where $A'_{AA} = 2q \times \alpha$, $A'_{AB} = (p - q) \times \alpha$ and $A'_{BB} = -2p \times \alpha$. A' is the re-parameterised additive genetic value and α is the allele substitution effect: $\alpha = a + (p - q)d$. Because the dominance deviation is the difference between the genetic value and the mean plus the additive value, the dominance effects can be derived as $D'_{AA} = -2p^2 \times d$, $D'_{AB} = 2pq \times d$ and $D'_{BB} = -2q^2 \times d$. D' is the re-parameterised dominance genetic value. Therefore, equation 3 can be re-parameterised as:

$$G' = \mu + A' + D' + e = \mu + x'_{A_i}\alpha + z'_{D_i}d + e \text{ (equation 5)}$$

Where x'_A was coded as $2q$, $p - q$ and $-2p$ for genotype of AA, AB and BB of variant *i*, respectively, to represent the additive component and z'_D was coded as $-2p^2$, $2pq$, $-2q^2$ for

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genotype of AA, AB and BB of variant *i*, respectively, to represent the dominance component. Such re-parametrisation has the following features: 1) The covariance between the additive and dominance effects is zero; 2) the variance of the additive effects gives the additive variance; and 3) The variance of the dominance deviations gives the dominance variance. Equation 5 then leads to:

$$H'_j = \sum_i^N z'_{D_i} \Big/ N \text{ (equation 6)}$$

Where H'_j was the re-parameterised genome-wide heterozygosity for individual *j*, z'_D was $-2p^2$, $2pq$, $-2q^2$ for genotype of AA, AB and BB of variant *i* and N was the total number of variants. We then applied equation 6 to conserved and non-conserved variants to estimate individual heterozygosity from conserved sites (H'_{cons_j}) and non-conserved sites ($H'_{non-cons_j}$). We then fitted H'_{cons_j} and $H'_{non-cons_j}$ as fixed effects together with the fixed effects of breed jointly in GREML similar to equation 1. The difference was that there is no fixed effect of variants but more fixed effects due to the fitting of H'_{cons_j} and $H'_{non-cons_j}$. The GREML analysis used the implementation with MTG2 [35].

Mutant allele frequency and F_{ST} in different breeds/subspecies. Two sets of data were used for this analysis. The first dataset was the Australian dairy cattle (8,949 bulls and 103,350 cows, Holstein, Jersey, Australian Red and crossbreds) used for GWAS as described above. The second data set used for the analysis of mutant allele frequency and F_{ST} was the curated whole-genome sequence data of 1,720 cattle from the 1000 Bull Genomes database (Run 7) [20,21], which we refer to as modern and ancient cattle. Samples that met the quality criteria of the 1000 Bull Genomes project were selected and they included 210 Brahman, 25 Tibetan, 10 Eurasian Aurochs, 242 Simmental, 95 Jersey, 840 Holstein and 287 Angus. Genome sequences from 6 Gir and 12 Nellore cattle from the 1000 Bull Genomes database were also analysed to support the results of mutant allele frequency of *Bos indicus*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequence data at 7,910,190 variants assigned with mutant alleles were retrieved for these animals to make a plink (v1.9) binary genotype file. The A1 allele of the plink genotypes was set to the mutant allele and its frequency was calculated using the '--freq' function for different selections of populations and variant sets. Average mutant allele frequency and the standard error were calculated for different selections of variants, e.g., variants with mutant alleles increasing or decreasing traits. Standard errors for frequency and F_{ST} (described below) were all estimated using LD-clumped variants in the same procedure in plink [26] as described above. For variants associated with milk production traits, i.e., the yield of milk protein, fat and milk and percentage of protein and fat, we selected variants with large (GWAS p-value < 5e-8 in both sexes) and small (GWAS p-value < 5e-2 and p-value > 5e-5 in both sexes) effects to focus on. For other trait-associated variants, the group with the largest effects available were selected for this comparison. For example, for stature, there were no variants with p-value < 5e-8 in both sexes, we then selected the medium-effect variants (GWAS p-value < 5e-5 and p-value > 5e-8 in both sexes). For fertility, there was no variants with p-value < 5e-5 in both sexes, we then selected the small-effect variants (GWAS p-value < 5e-2 and p-value > 5e-5 in both sexes) for the comparison. Average mutant allele frequency and the standard error were also calculated for all 7.9M variants analysed as the baseline. The analysis procedure for allele frequency on the Australian dairy cattle was applied to these 1000 Bull Genomes individuals.

With the same plink binary genotype file described above and the population structure for dairy cattle (4 dairy breeds) and for ancient and modern cattle (7 breeds/subspecies), GCTA [34] was used to calculate the F_{ST} value with the method described in Weir [40] with the option of '--fst' and '--sub-pop'. The average F_{ST} value with standard errors was then calculated for different selections of variants in the same fashion for selecting variant groups to compare the mutant allele frequency as described above.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data and Code availability

The information of cattle ancestral alleles for those 42,573,455 sites is publicly available at: https://figshare.com/s/dd5985b76a413b56106b. Multiple alignment data used to determine cattle ancestral alleles are publicly available via Ensembl EPO pipeline (http://asia.ensembl.org/info/genome/compara/multiple_genome_alignments.html). DataGene Australia (http://www.datagene.com.au/) are custodians of the raw phenotype and genotype data of Australian farm animals. The DNA sequence data as part of the 1000 Bull Genomes Consortium consortium [16,20,21] included NCBI BioProjects PRJNA431934, PRJNA238491, PRJDB2660, PRJEB18113, PRJEB1829, PRJEB27309, PRJEB28191, PRJEB9343, PRJNA294709, PRJNA316122, PRJNA474946, PRJNA477833, PRJNA494431, PRJDA48395, PRJNA431934, PRJNA238491. Other supporting data are shown in the supplementary materials of the current manuscript. The probability of ancestral allele assignment used the software published by [22]. The linear mixed model used GCTA [34] and MTG2 [35]. The Bayesian analysis used BayesR [41]. The R code of estimating heterozygosity across conserved sites will be made public upon publication.

References:

1 Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96-98 (1973).

2 Eyre-Walker, A., Keightley, P. D., Smith, N. G. & Gaffney, D. Quantifying the slightly deleterious mutation model of molecular evolution. *Molecular Biology and Evolution* **19**, 2142-2149 (2002).

3 Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610-618 (2007).

4 Mehlhoff, J. D. *et al.* Collateral fitness effects of mutations. *Proceedings of the National Academy of Sciences* **117**, 11597-11607 (2020).

5 Jiang, Y. *et al.* The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* **344**, 1168-1173, doi:10.1126/science.1252806 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6 Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M. & Cunningham, P. Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences* **91**, 2757-2761 (1994).

7 Bradley, D. G., MacHugh, D. E., Cunningham, P. & Loftus, R. T. Mitochondrial diversity and the origins of African and European cattle. *Proceedings of the National Academy of Sciences* **93**, 5131-5135 (1996).

8 Troy, C. S. *et al.* Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410**, 1088-1091 (2001).

9 Chen, S. *et al.* Zebu cattle are an exclusive legacy of the South Asia Neolithic. *Molecular biology and evolution* **27**, 1-6 (2010).

10 Utsunomiya, Y. *et al.* Genomic clues of the evolutionary history of Bos indicus cattle. *Animal genetics* **50**, 557-568 (2019).

11 Verdugo, M. P. *et al.* Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* **365**, 173-176 (2019).

12 Upadhyay, M. *et al.* Genetic origin, admixture and population history of aurochs (Bos primigenius) and primitive European cattle. *Heredity* **118**, 169-176 (2017).

13 Park, S. D. *et al.* Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. *Genome biology* **16**, 234 (2015).

14 García-Ruiz, A. *et al.* Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences* **113**, E3995-E4004 (2016).

15 Coffey, E., Horan, B., Evans, R. & Berry, D. Milk production and fertility performance of Holstein, Friesian, and Jersey purebred cows and their respective crosses in seasonal-calving commercial farms. *Journal of Dairy Science* **99**, 5681-5689 (2016).

16 Bouwman, A. C. *et al.* Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics* **50**, 362 (2018).

17 Meuwissen, T., Hayes, B. & Goddard, M. Genomic selection: A paradigm shift in animal breeding. *Animal frontiers* **6**, 6-14 (2016).

18 Chen, J., Glémin, S. & Lascoux, M. From drift to draft: how much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution? *Genetics* **214**, 1005-1018 (2020).

19 Xiang, R. *et al.* Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences* **116**, 19398-19408, doi:10.1073/pnas.1904159116 (2019).

20 Daetwyler, H. D. *et al.* Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics* **46**, 858 (2014).

21 Hayes, B. J. & Daetwyler, H. D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual review of animal biosciences* **7**, 89-102, doi:10.1146/annurev-animal-020518-115024 (2019).

22 Keightley, P. D. & Jackson, B. C. Inferring the probability of the derived vs. the ancestral allelic state at a polymorphic site. *Genetics* **209**, 897-906 (2018).

23 Rocha, D., Billerey, C., Samson, F., Boichard, D. & Boussaha, M. Identification of the putative ancestral allele of bovine single-nucleotide polymorphisms. *Journal of Animal Breeding and Genetics* **131**, 483-486 (2014).

24 Xiang, R., van den Berg, I., MacLeod, I. M., Daetwyler, H. D. & Goddard, M. E. Effect direction meta-analysis of GWAS identifies extreme, prevalent and shared pleiotropy in a large mammal. *Commun Biol* **3**, 88, doi:10.1038/s42003-020-0823-6 (2020).

25 Erbe, M. *et al.* Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science* **95**, 4114-4129 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

27 Chen, N. *et al.* Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nature Communications* **9**, 1-13 (2018).

28 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050 (2005).

29 Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-784 (2014).

30 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955 (2012).

31 Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478, doi:10.1186/1471-2164-15-478 (2014).

32 Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* **48**, 1443 (2016).

33 Rosen, B. D. *et al.* De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* **9**, giaa021 (2020).

34 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).

35 Lee, S. H. & Van der Werf, J. H. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* **32**, 1420-1422 (2016).

36 Xiang, R. *et al.* Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. *Nature Communications* **12**, 860, doi:10.1038/s41467-021-21001-0 (2021).

37 Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52**, 399-433 (1919).

38 Falconer, D. S. & Mackay, T. F. C. *Introduction to quantitative genetics*. (Longman, 1996).

39 Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer Associates, 1998).

40 Weir, B. S. & Ott, J. Genetic data analysis II. *Trends in genetics* **13**, 379 (1997).

41 Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS genetics* **11**, e1004969 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.19.440546; this version posted April 19, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

Australian Research Council's Discovery Projects (DP160101056 and DP200100499) supported R.X. and M.E.G. DairyBio, a joint venture project between Agriculture Victoria (Melbourne, Australia), Dairy Australia (Melbourne, Australia) and the Gardiner Foundation (Melbourne, Australia), funded computing resources used in the analysis. The authors also thank the University of Melbourne, Australia for supporting this research. No funding bodies participated in the design of the study nor analysis, or interpretation of data nor in writing the manuscript. DataGene provided access to the reference data used in this study and the 1000 Bull Genomes consortium provided access to cattle sequence data. We thank Gert Nieuwhof and Kon Konstantinov (DataGene) for the preparation and provision of data. We thank Prof. Hans D. Daetwyler for assisting the access to the 1000 Bull Genomes data. We thank Prof. Naomi Wray for a critical read of the maunsscript.

Author contributions

M.E.G. and R.X conceived the study. R.X. performed all analyses. E.J.B. contributed to the BayesR analysis. I.M.M., assisted with data curation. S.B., C.J.J., and A.J.C. contributed to the imputation of sequence variants. R.X. and M.E.G. wrote the paper. R.X., M.E.G., E.J.B. and I.M.M. revised the paper. All authors read and approved the final manuscript.

Competing Interests

The authors declare no competing interests.