Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits

Ruidong Xiang[1,2,*], Irene Van Den Berg[1,2], Iona M. MacLeod[2], Benjamin J. Hayes[2,3], Claire P. Prowse-Wilkins[1,2], Min Wang[2,4], Sunduimijid Bolormaa[2], Zhiqian Liu[2], Simone J. Rochfort[2,4], Coralie M. Reich[2], Brett A. Mason[2], Christy J. Vander Jagt[2], Hans D. Daetwyler[2,4], Mogens S. Lund[5], Amanda J. Chamberlain[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.*

[3] *Centre for Animal Science, The University of Queensland, St Lucia 4067, Queensland, Australia.*

[4]*School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia*

[5]*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, PO Box 50, DK-8830 Tjele, Denmark*

[*]Corresponding author: Dr. Ruidong Xiang; ruidong.xiang@unimelb.edu.au;

Key words: Gene regulation, evolution, quantitative traits, animal breeding, cattle

Abstract

Many genome variants shaping mammalian phenotype are hypothesized to regulate gene transcription and/or to be under selection. However, most of the evidence to support this hypothesis comes from human studies. Systematic evidence for regulatory and evolutionary signals contributing to complex traits in a different mammalian model is needed. Sequence variants associated with gene expression (eQTLs) and concentration of metabolites (mQTLs), and under histone modification marks in several tissues were discovered from multi-omics data of over 400 cattle. Variants showing signs of being selected were identified from the 1000-bull genomes database (N=2,330). These analyses defined 30 sets of variants and for each set we estimated the genetic variance the set explained across 34 complex traits in 11,923 bulls and 32,347 cows with 17,669,372 imputed variants. The per-variant trait heritability of these sets across traits was highly consistent ($r > 0.98$) between bulls and cows. Based on the per-variant heritability, the sets of mQTL, eQTL and variants associated with non-coding RNAs ranked the highest, followed by the young variants, those under histone modification marks and selection signatures. From these results, we defined a Functional-And-Evolutionary Trait Heritability (FAETH) score indicating the functionality and predicted heritability of each variant. In 7,551 Danish cattle, the high FAETH-ranking variants had significantly increased genetic variances and genomic prediction accuracies in 3 production traits compared to the low FAETH-ranking variants. The publicly available FAETH variant score, based on regulatory and evolutionary data, provides a set of biological priors for the functional effects of variant on bovine complex traits.

Introduction

Understanding how mutations lead to phenotypic variation is a fundamental goal of genomics. With a few exceptions, complex mammalian traits with significance in evolution, medicine and agriculture are determined by many mutations and by environmental effects. Genome-wide association studies (GWAS) have been successful in finding associations between single nucleotide polymorphisms (SNPs) and complex traits (1). In most cases there are many variants, each of small effect which contribute to variation in a complex trait. Consequently, very large sample size is necessary to find significant associations which explain most of the observed genetic variation. In humans sample size has reached over 1 million (2).

To test the generality of the findings in humans it is desirable to have another species with very large sample size and cattle is a possible example. There are over 1.46 billion cattle worldwide (3) and millions are being genotyped or whole genome sequenced and phenotyped (4, 5). The two sub-species of domestic cattle, humpless taurine (*Bos taurus*) and humped zebu (*Bos indicus*), diverged approximately 0.5 million years ago from extinct wild aurochs (*Bos primigenius*) (6). These features make cattle the only GWAS model of an outbred genome with a comparable sample size to humans. In addition, cattle have a very different demographic history than humans. Whereas humans went through an evolutionary bottleneck about 10,000 to 20,000 years ago and then expanded to a population of billions, cattle have declined in effective population size due to domestication and breed formation leading to a different pattern of linkage disequilibrium (LD) to humans. Therefore, insights into the genome-phenome relationships from cattle provide a valuable addition to the knowledge from humans and other mammalian species. The knowledge of cattle genomics is also of direct practical value because rearing cattle is a major agricultural industry around the world.

Despite the huge sample sizes used in human GWAS, identification of the causal variants for a complex trait is still difficult. This is due to the small effect size of most causal variants and the LD between variants. Consequently, there are usually many variants in high LD, any one of which could be the cause of the variation in phenotype. Prioritisation of these variants can be aided by information on the function of the genomic site and its evolutionary history. For instance, mutations that change an amino acid are more likely to affect phenotype than mutations which are synonymous.

Many mutations affecting complex traits regulate gene transcription related activities. This has been demonstrated in a series of studies of human functional genomics, including but not limited to the analysis of intermediate trait quantitative trait loci (QTLs), such as metabolic QTLs (mQTLs) (7) and expression QTLs (eQTLs) (8) and analysis of regulatory elements, such as promoters (9) and enhancers (10) which can be identified with chromatin immunoprecipitation sequencing (ChIP-seq). In animals, the Functional Annotation of Animal Genomes (FAANG) project has started (11) and functional data from animal species has been accumulating (12-14). However, it is unclear which types of functional information improve the identification of causal mutations.

Mutations affecting complex traits may be subject to natural or artificial selection which leaves a 'signature' in the genome (15, 16). Given the unique evolutionary path of cattle which has been significantly shaped by human domestication (17), it is attractive to test whether variants showing signatures of selection contribute to variation in complex traits.

The aim of this study is to determine which of several possible indicators of function are most useful for predicting which sequence variants are most likely to affect 34 traits in *Bos taurus* dairy cattle. The indicators that we consider fall into 3 groups: (1) functional annotations of the bovine genome based, for instance, on ChIP-seq experiments; (2) evolutionary data such as a site being under selection; (3) GWAS data from traits that are relatively close to the

primary action of the mutation, such as gene expression. Using these indicators of function, we define 30 sets of variants and estimate the variance explained by each set across 34 traits in 44,270 cattle. We then combine the estimates of heritability per variant across traits and across functional and evolutionary categories to define a Functional-And-Evolutionary Trait Heritability (FAETH) score that ranks variants on variance explained in complex traits of dairy cattle. We then validate the FAETH score in 7,551 Danish cattle. The FAETH score of over 17 million variants is publicly available at: https://melbourne.figshare.com/s/2c5200a8333b6e759ddc.

Results

Analysis overview

Our approach was to estimate the trait variance explained by a set of variants defined by some external data, such as the mapping of the gene expression QTLs (geQTLs), RNA splicing QTLs (sQTLs), or genome annotation, for 34 traits measured in dairy cattle. Sequence variants available to this study included over 17 million SNPs and indels. Any large set of variants can explain almost all the genetic variance due to the LD between surrounding and causal variants. Therefore, we fitted each externally defined set of variants in a model together with a standard set of 600K SNPs from the bovine high-density (HD) SNP array. We combined the results from all 34 traits and all the sets of variants to derive a score for each variant based on its expected contribution to the genetic variance in these 34 traits and tested the validity of this score in an independent cattle dataset.

Our analysis had four major steps (Figure 1):

(1) The 17M sequence variants (1000 bull genome Run6 (18)) were classified according to external information from the discovery analysis of the function and evolution of each genomic site. The basis for this classification was either publicly available data or our own

data as described in the methods. The genome was partitioned 15 different ways as listed in Table 1. For example, the category of geQTL partitioned the genome variants into a set of targeted variants with geQTL p value < 0.0001 and a set of all other variants (i.e. the remaining or the 'rest' of the variants). Another partition, e.g., variant annotation, based on publicly available annotation of the bovine genome, divided variants into several non-overlapping sets, such as 'intergenic', 'intron' and 'splice sites'.

(2) For each set of variants in each partition of the genome, separate genomic relationship matrices (GRMs) were calculated among the 11,923 bulls or 32,347 cows. Where a partition included only 2 sets (e.g. geQTL and the rest) a GRM was calculated only for the targeted set (e.g. geQTL).

(3) For each of the 34 traits, the variance explained by random effects described by each GRM was estimated using restricted maximum likelihood (this analysis is referred to as a genomic REML or GREML). Each GREML analysis fitted a random effect described by the targeted GRM and a random effect described by the GRM calculated from the HD SNP chip (630,002 SNPs). Each GREML analysis estimated the proportion of genetic variance, h^2, explained by the targeted GRM in each of the 34 decorrelated traits (Cholesky orthogonalisation (19), see methods) in each sex. The h^2 explained by each targeted set of variants was divided by the number of variants in the set to calculate the h^2 per variant, i.e. per-variant h^2, and this was averaged for each variant across the 34 decorrelated traits.

(4) The FAETH score of all variants was calculated by averaging the per-variant h^2 across traits and informative partitions (12 out of 15). Partitions (3 out of 15) determined as not informative were not included in the FAETH score computation. Variance explained and the accuracy of genomic predictions (based on an independent dataset of 7,551 Danish cattle with three milk production traits) were compared between variants of high and low FAETH score.

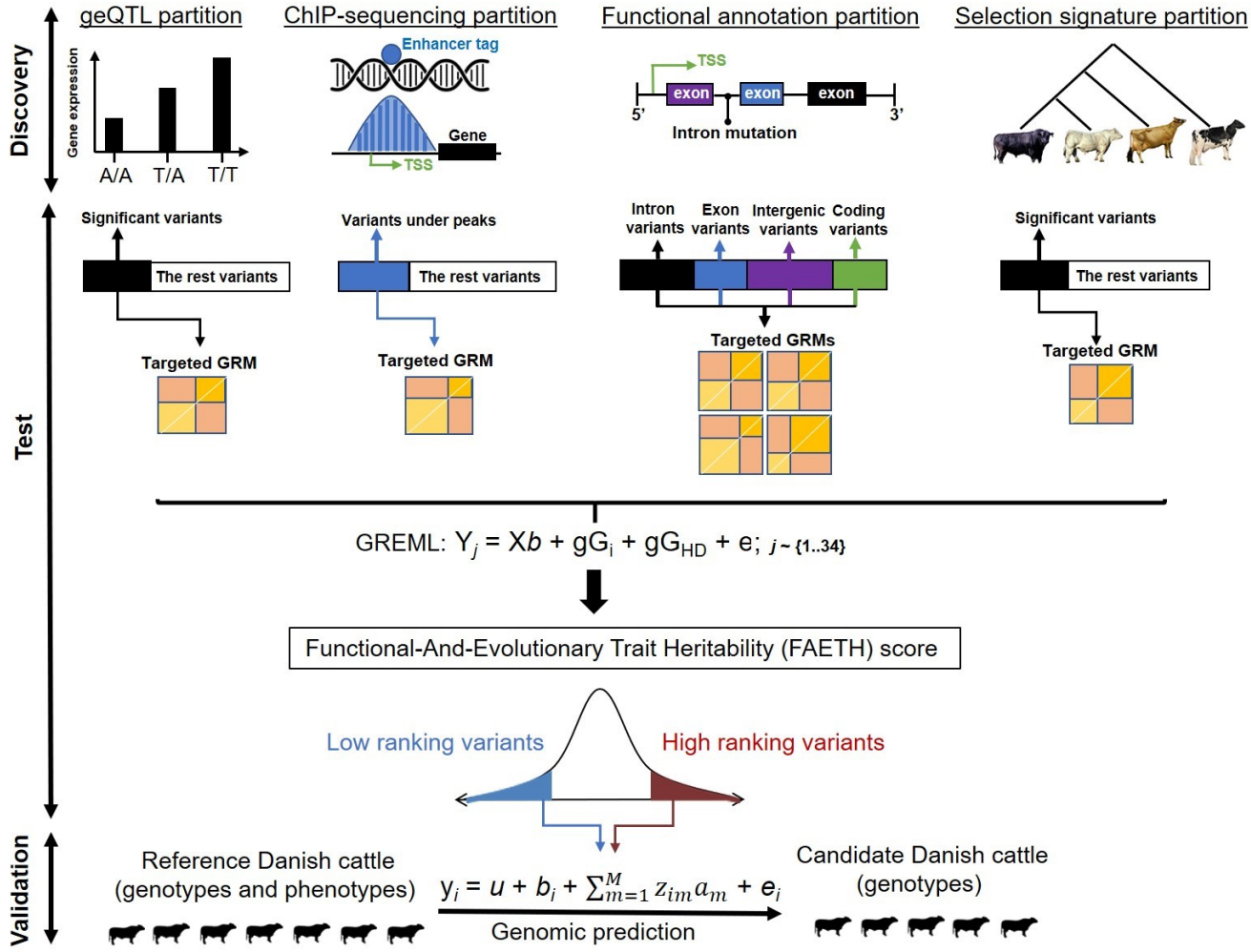

Figure 1. **Overview of the analysis.** The discovery analysis involved selection of variants from functional and evolutionary datasets, this figure shows examples of some of the datasets used. In the test analysis, each of the variant sets were used to make genomic relation matrices (GRM)s. Then, each one was analysed in genome-wide restricted maximum likelihood (GREML, gG_i) together with the high-density SNP chip GRM (gG_{HD}) for each one of the 34 traits (Y_j, j= {1..34}). Once the heritability, h^2_{set}, of each gG_i was calculated, it was averaged across traits and adjusted for the number of variants used to build the gG_i to calculate the per-variant $\overline{h^2_{set}}$. The FAETH scoring of each variant was derived based on their memberships to differentially partitioned sets and the per-variant $\overline{h^2_{set}}$. In the validation analysis, variants with high and low FAETH ranking were tested in a Danish cattle data set for GREML and genomic prediction of three production traits. The Danish reference set contained 4,911 Holstein, 957 Jersey and 745 Danish Red bulls, and the Danish validation population 500 Holstein, 517 Jersey and 192 Danish Red bulls.

Characteristics of variant sets with regulatory and evolutionary significance

Based on the 15 partitions of the genome in Table 1, we defined 30 sets of variants. The details of the discovery analysis defining these sets can be found in Methods. Briefly, regulatory variant sets including geQTLs, sQTLs and allele specific expression QTLs (aseQTLs) were discovered from multiple tissues including white blood and milk cells, liver and muscle. The polar lipid metabolites mQTLs were discovered using the multi-trait meta-analysis (20) of 19 metabolite profiles, such as phosphatidylcholine, phosphatidylethanolamine and phosphatidylserine (21), from the bovine milk fat. The ChIP-seq data used in our analysis contained previously published H3K27Ac and H3K4me3 marks in liver and muscle tissues (22, 23) and newly generated H3K4Me3 marks from the mammary gland.

Table 1. Variant sets selected from functional and evolutionary partitions.

Partitions	Targeted variant sets (the number of variants)	Animal no.
Gene expression QTLs	geQTLs with meta-analysis p < 1e-4 from blood and milk cells, liver and muscle (110,200)	209
Exon expression QTLs	eeQTLs with meta-analysis p < 1e-4 from blood and milk cells, liver and muscle (945,832)	209
Splicing QTLs	sQTLs with meta-analysis p < 1e-4 from blood and milk cells, liver and muscle (1,112,324)	209
Allele specific expression QTLs	aseQTLs with meta-analysis p < 1e-4 from blood and milk cells (1,100,446)	112
Polar lipid metabolite QTLs	mQTLs with meta-analysis p < 1e-4 from 19 types of milk metabolites (5,365)	338
ChIP-seq peaks	Under H3K4Me3 and H3K27Ac peaks from liver, muscle and mammary gland (1,166,795)	14
Variant annotation	Annotated as UTR (42,350), intergenic (11,869,145), geneend (1,007,214), intron (4,629,025), splice.sites (11,080), coding.related (105,969), noncoding.related (4,589)	na
Predicted CTCF sites	variants tagged by mapped CTCF binding motifs from humans, mice, dogs and macaques as published by (24) (252,234)	na
HPRS	Genome sites within the top 1% gkmSVM score from the Human Projection of Regulatory Regions as published by (25) (169,773)	na
Conserved sites	Genome sites with PhastCon score (26) > 0.9 calculated using genome sequences of bovine, dog, mouse and humans (100,279)	na
Selected signature	GWAS p < 1e-4 between 7 beef and 8 dairy breeds, 1000 bull genome (6,218)	1,370

Young variants	Ranked within the bottom 1% of the proportion of positive correlations (PPRR) with rare variants, 1000 bull genome (893,986)	2,330
LD score quartiles	1st quartile (4,417,033/4,416,205), 2nd quartile (4,418,731/4,419,930), 3rd quartile (4,415,633/4,415,481), 4th quartile (4,417,975/4,417,756)	44,270
Variant density quartiles	1st quartile (4,429,833), 2nd quartile (4,414,996), 3rd quartile (4,427,220), 4th quartile (4,397,323)	
MAF quartiles	1st quartile (4,414,292/4,417,036), 2nd quartile (4,421,093/4,417,428), 3rd quartile (4,416,834/4,418,157), 4th quartile (4,417,153/4,418,157)	

For the three categories of quartiles the numbers of variants on the left and right side of slash were for the bulls and cows, respectively. LD score: sum of linkage disequilibrium correlation between a variant and all variants in the surrounding 50kb region, GCTA-LDS (27). MAF: minor allele frequency. The details of the variant annotations can be found in the Table S1. The animal number are the sample size in each discovery analysis.

Figure 2 illustrates some of the properties of these variant sets. Many sQTLs with strong effects on the intron excision ratio (28) were discovered in a meta-analysis of sQTLs mapped in white blood and milk cells, liver and muscle (13) (Figure 2A). Also, a large number of significant aseQTLs were discovered using a gene-wise meta-analysis of the effects of the driver variant (dVariant) on the transcript variant (tVariant) at the exonic heterozygous sites (29) from the white blood and milk cells (Figure 2B). As shown in Figure 2C, variants tagged by the H3K4Me3 marks, a marker for promoters, were closer to the transcription start site than other variants.

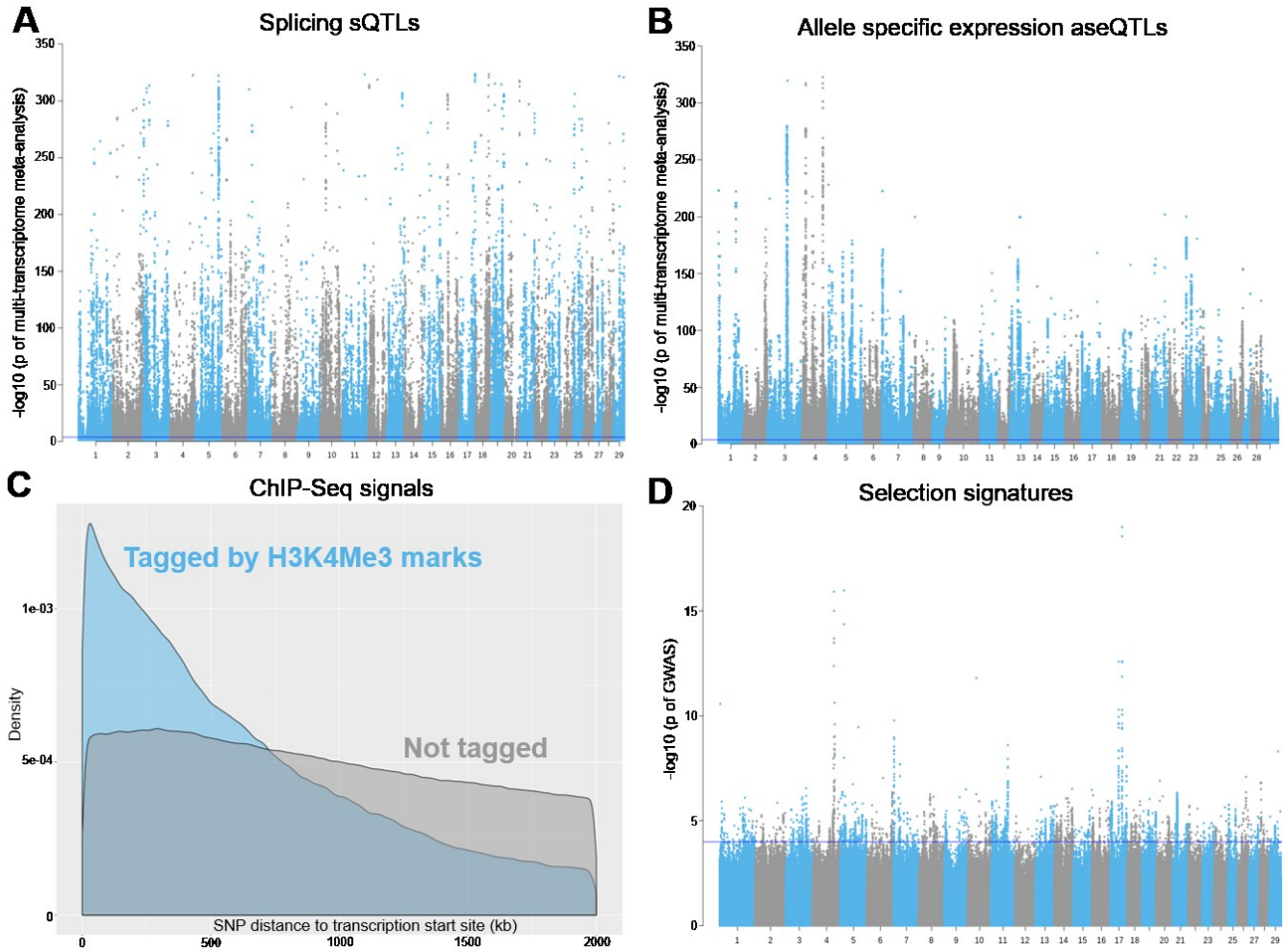

Figure 2. Examples of regulatory and evolutionary signals from the discovery analysis. A: A Manhattan plot of the meta-analysis of sQTLs from white blood and milk cells, liver and muscle tissues. B: A Manhattan plot of the meta-analysis of aseQTLs in the white blood cells. C: A distribution density plot of H3K4Me3 ChIP-seq mark tagged variants from mammary gland within 2Mb of gene transcription start site. D: artificial selection signatures between 8 dairy and 7 beef cattle breeds with the linear mixed model approach. The blue line indicates $-\log_{10}$(p value) = 4.

The variant annotation partition had 7 merged sets (Table 1, Table S1) based on the Variant Effect Prediction of Ensembl (30) and NGS-variant (31). Additional information of variant function annotation was obtained from the Human Projection of Regulatory Regions (HPRS) as published in (25) and predicted CTCF sites as published in (24).

The evolutionary variant sets were discovered from cross- and within- species genome analyses. Variants within cross-species conserved sites were selected based on the PhastCon score (26) calculated using the genomes of bovine, dog, mouse and humans (Table 1). The within-species analysis used the whole genome sequence variants from Run6 of the 1000 bull genomes database (32). Those variants with higher frequency in dairy than in beef breeds ('selection signature', Table 1, Figure 2D and Figure S1) were detected from a GWAS where the breed-type was modelled as a binary phenotype in the linear mixed model (33) of 15 beef and dairy breeds.

With the 1000 bull genome data, we used a novel statistic to identify variants possibly subject to artificial and/or natural selection, PPRR, the *Proportion of Positive correlations* (*r*) *with Rare variants*. Figure S2A illustrates a coalescence where a mutation has been positively selected, i.e. is relatively young, and increased in frequency rapidly. In this coalescence the selected mutation was seldomly on the same branch as rare mutations and so the LD *r* between the selected mutation and rare alleles was typically negative. This was similar to the logic employed by (34). In this partition of the genome, the 1% of variants with the lowest PPRR, after correcting for the variants' own allele frequency (see Figure S2 and methods) were defined as young variants.

The quartile categories partitioned the genome variants into four sets of variants of similar size based on either their LD score (sum of LD r^2 between a variant and all the variants in the surrounding 50kb region, GCTA-LDS (27)), or the number of variants within a 50kb window or their minor allele frequency (MAF) (35) (Table 1).

The proportion of genetic variance for 34 traits explained by each set of variants

In the test datasets of 11,923 bulls and 32,347 cows,, common variants (MAF>=0.001) of the sets described above were used to make GRMs (33). Each of these GRMs were then fitted together with the high-density variant chip GRM (variant number = 632,002) in the GREML

analysis to estimate the proportion of additive genetic variance explained by each functional and evolutionary set of variants, h^2_{set}, in each of the 34 decorrelated traits separately in bulls and cows (Table 2). Overall, the ranking of the averaged h^2_{set} across 34 traits, $\overline{h^2_{set}}$, was highly consistent between bulls and cows (Spearman ranking correlation rho = 0.985). All the $\overline{h^2_{set}}$ estimates, except that of the intergenic variants, were higher for bull traits than cow traits, consistent with the higher heritability of phenotypic records in bulls than in cows (36) because bull phenotypes are actually the average of many daughters of the bull. When the HD variants were fitted alone they explained on average 17.8% (±2.7%) of the variance in bulls and 4.7% (±1.4%) in cows (Table S2). In general, the $\overline{h^2_{set}}$ estimates increase with the number of variants in the set. For example, expression QTLs, including exon expression eeQTLs, splicing sQTLs and allele specific expression aseQTLs, which included around 5% of the total variants explained 11~15% of trait variance in bulls and 2.5%~4% trait variance in cows. The young variants inferred by the novel statistic PPRR, which accounted for 0.54% of the total variants, explained 0.78% trait variance in bulls and 0.12% trait variance in cows.

Table 2. The relative proportion of selected variant in sets compared to the total number of variants analysed (Genome fraction) and their averaged heritability ($\overline{h^2_{set}}$) in bulls and cows, across 34 traits with the standard error in the parenthesis.

Category	Genome fraction	$\overline{h^2}$ in bulls	$\overline{h^2}$ in cows
eeQTLs	4.77%	14.52% (2.2%)	3.96% (1.2%)
sQTLs	5.57%	15.08% (2.5%)	3.88% (1.2%)
aseQTLs	5.21%	11.0% (2.0%)	2.47% (0.7%)
mQTLs	0.03%	0.71% (0.2%)	0.12% (0.04%)
geQTLs	0.53%	1.54% (0.4%)	0.19% (0.06%)
ChIPseq	6.60%	4.21% (0.8%)	0.90% (0.3%)
noncoding.related	0.03%	0.06% (0.02%)	0.013% (0.004%)
Splice.sites	0.06%	0.08% (0.02%)	0.02% (0.005%)
UTR	0.24%	0.18% (0.03%)	0.03% (0.01%)
Coding.related	0.60%	0.26% (0.06%)	0.04% (0.012%)
Geneend	5.70%	3.76% (0.8%)	0.80% (0.2%)
Intron	26.2%	5.56% (0.7%)	1.53% (0.3%)
Intergenic	67.2%	10.3% (1.3%)	17.3% (2.2%)
Predicted CTCF sites	1.43%	0.36% (0.08%)	0.046% (0.02%)
HPRS	0.96%	0.31% (0.08%)	0.045% (0.02%)
Conserved sites	0.57%	0.23% (0.05%)	0.030% (0.01%)
Selection signatures	0.02%	0.011% (0.004%)	0.002% (0.0008%)
Young variants	0.54%	0.78% (0.2%)	0.12% (0.05%)
LD score q1	25%	4.57% (0.6%)	1.18% (0.3%)
LD score q2	25%	5.56% (0.7%)	1.45% (0.3%)
LD score q3	25%	6.38% (0.8%)	1.75% (0.4%)
LD score q4	25%	6.94% (0.9%)	2.01% (0.5%)
Variant density q1	25%	5.59% (0.7%)	1.49% (0.3%)
Variant density q2	25%	5.42% (0.7%)	1.45% (0.3%)
Variant density q3	25%	5.72% (0.7%)	1.55% (0.3%)
Variant density q4	25%	5.99% (0.7%)	1.65% (0.4%)
MAF q1	25%	1.36% (0.2%)	0.35% (0.08%)
MAF q2	25%	11.5% (1.3%)	3.51% (0.7%)
MAF q3	25%	29.2% (2.4%)	10.3% (1.8%)
MAF q4	25%	40.5% (2.8%)	15.6% (2.4%)

q1~q4 were the genome partitions based on the 1st, 2nd, 3rd and 4th quartiles of minor allele frequency (MAF), LD score and the number of variants (variant density) per 50kb windows.

The h^2_{set} increased greatly from MAF quantile 1 to 4. However, the dramatically low $\overline{h^2_{set}}$ estimates for the 1st MAF quartile may be associated with the reduced imputation accuracy for low MAF variants. By contrast $\overline{h^2_{set}}$ increased only slightly with LD score and even less with variant density.

Estimates of $\overline{h^2_{set}}$ were divided by the number of variants in the set to calculate the per-variant $\overline{h^2_{set}}$ allowing comparison of the genetic importance of variant sets made with varied

number of variants. Since the per-variant $\overline{h^2_{set}}$ was estimated independently in bulls and cows yet showed high consistency between sexes (Figure S3), the average per-variant $\overline{h^2_{set}}$ across sexes was used to rank each variant set (Figure 3). The set of mQTLs made the top of the rankings (Figure 3), due to its concentrated $\overline{h^2_{set}}$ (0.71% in bulls and 0.12% in cows, Table 2) in a relatively small genome fraction (0.03%, Table 2). The high ranking of the mQTLs set was followed by several expression QTLs sets, including eeQTLs, sQTLs, geQTLs and aseQTLs (Figure 3). Similar rankings were achieved by the 'non.coding related' set (0.03% of genome variants, included variants annotated as 'non_coding_transcript_exon_variant' and 'mature_miRNA_variant' (Table S1), the 'splice.site' set (0.06% of genome variants, including all the variants annotated as associated with splicing functions) and the set of young variants (0.54% of genome variants). The 'UTR' set, which included variants annotated as within 3' and 5' untranslated regions of genes, and the 'geneend' set, which included variants annotated as within the downstream and upstream of genes, both had modest rankings along with the ChIP-seq set and selection signatures. The 'coding.related' set, which included variants annotated as synonymous and missense, ranked higher than conserved sites, top 1% HPRS, intergenic variants and predicted CTCF sites. Intron and the 1st quartile MAF set had the lowest per variant h^2.

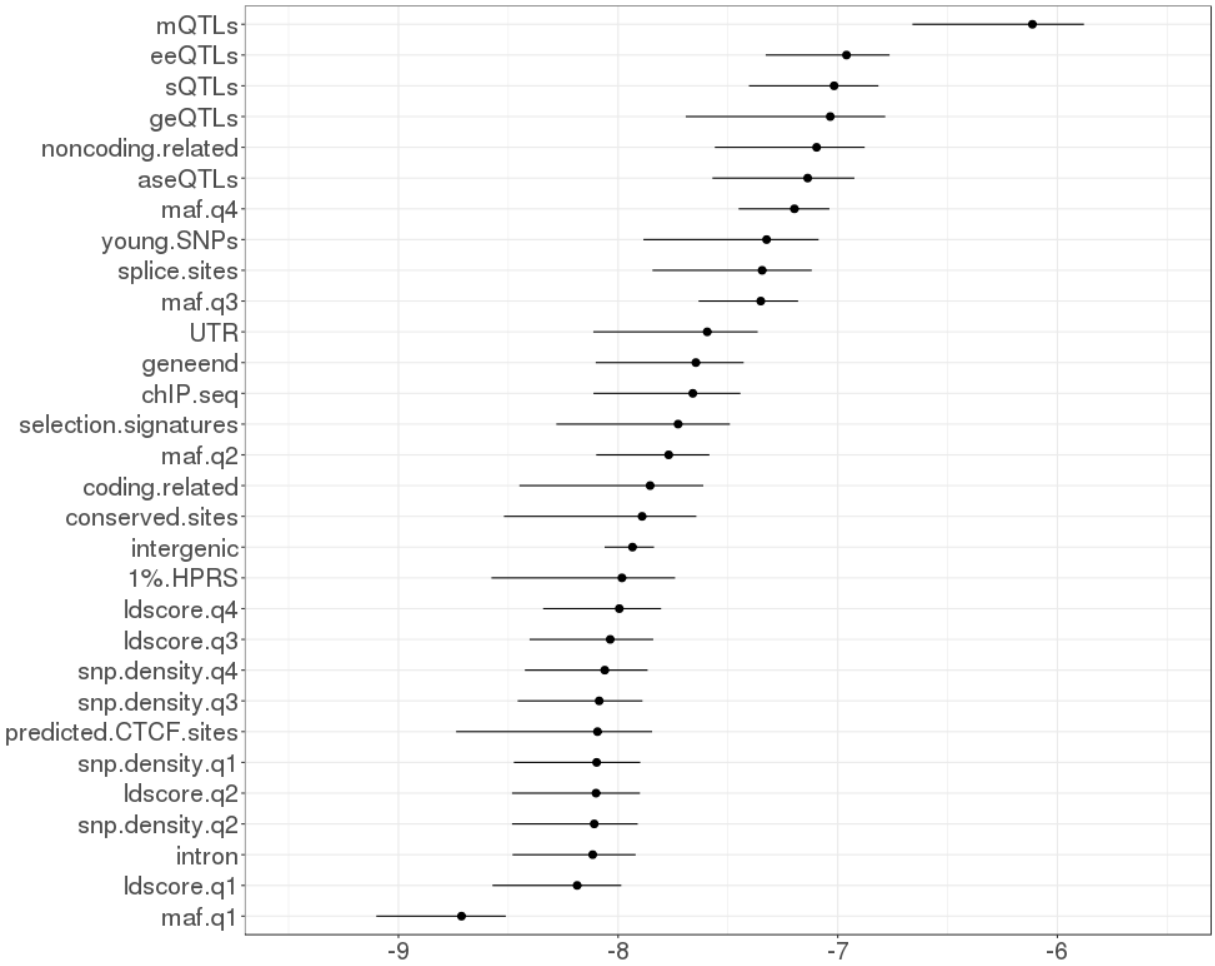

Figure 3. Proportion of genetic variances explained by sets of variants selected from functional and evolutionary categories. The ranking of variant sets based on the log10 scale of per-variant $\overline{h^2_{set}}$, averaged across bulls (left error bar) and cows (right error bar).

Variants from sets of high-ranking per-variant $\overline{h^2_{set}}$ were highlighted in important QTL regions with the multi-trait GWAS results (Figure 4). In the expanded region of beta-casein (*CSN2*), a major but complex QTL for milk protein due to the existence of multiple QTL with strong LD, different high-ranking variant sets tended to tag variants with strong effects from multiple locations (Figure 4A). Many variants with the strongest effects and close to *CSN2* were tagged by sQTLs. Several clusters of variants from up and downstream of *CSN2* with slightly weaker effects were tagged by sets of ChIP-Seq marks, young variant and mQTLs. Conversely, for the expanded region of microsomal glutathione S-transferase 1 (*MGST1*), a

major QTL for milk fat, variants from high-ranking sets were more enriched in two major locations (Figure 4A). The top variant within the *MGST1* gene was again an sQTL, confirming the previous results (13). The same region was also enriched with aseQTLs and ChIP-Seq mark tagged variants (Figure 4B). The ChIP-Seq and young variant sets appear to have tagged a different variant cluster around 0.7Mb downstream from *MGST1* (Figure 4B).

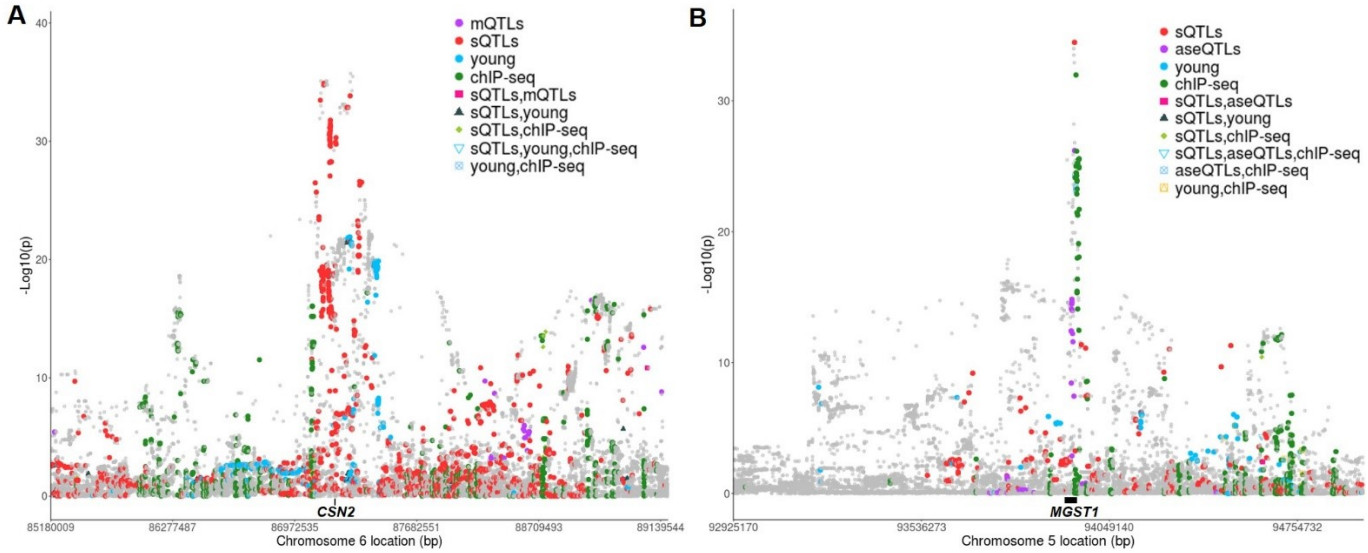

Figure 4. Example of top ranked variant sets in important bovine trait QTL. A: Manhattan plot of the meta-analysis of GWAS of 34 traits in the ±2Mb region surrounding the beta casein (CSN2) gene, a major QTL for milk protein yield. B: Manhattan plot of the meta-analysis of GWAS of 34 traits in the ±1Mb region of the microsomal glutathione S-transferase 1 (MGST1) gene, a major QTL for milk fat yield. The dots are coloured based on their set memberships. The black bar between the grey dots and the X-axis indicates the gene locations.

The FAETH score of sequence variants

To quantify the importance of variants using a combination of functionality, selection signatures as well as their trait heritability, a novel framework was introduced to score variants based on their memberships of the sets of variants. Each time the genome variants were partitioned into non-overlapping sets, each variant was a member of only one set and

was assigned the per-variant $\overline{h^2_{set}}$ of that variant. Therefore, all variants were assigned the same number (12 partitions) of per-variant $\overline{h^2_{set}}$ and the average of these 12 partitions was calculated for each variant and called the FAETH score. A criterion of per-variant $\overline{h^2_{set}}$ > per-variant $\overline{h^2_{rest}}$ was also imposed to determine whether the variant set was informative. This criterion determined that three variant sets (Conserved sites, HPRS and predicted CTCF sites) were not informative and they were not included in the FAETH scoring (see methods). The FAETH score of 17,669,372 sequence variants for their genetic contribution to complex traits has been made publicly available at https://melbourne.figshare.com/s/2c5200a8333b6e759ddc.

Variants with high FAETH score have consistent effects across breeds

In the analyses reported above the effect of a variant was estimated across all breeds. However, it is possible to fit a nested model in which both the main effect of the variant and an effect of the variant nested within a breed is included in the model. If a variant is causal or in high LD with a causal variant we might expect the effect to be similar in all breeds. Whereas if the variant is merely in LD with the causal variant, the effect might vary between breeds. Based on the FAETH score, the top 1/3 and bottom 1/3 ranked sequence variants in the Australian data were selected as 'high' and 'low' ranking variants, respectively. Figure 5A showed the estimates of across breed and within breed variant variances for both high- and low-ranking variants. In both cases the within breed variance is small, but the high-ranking variants have a larger across breed variant variance and a smaller within-breed variant variance than the low-ranking variants. This implied the consistency of the FAETH ranking of variants across breeds.

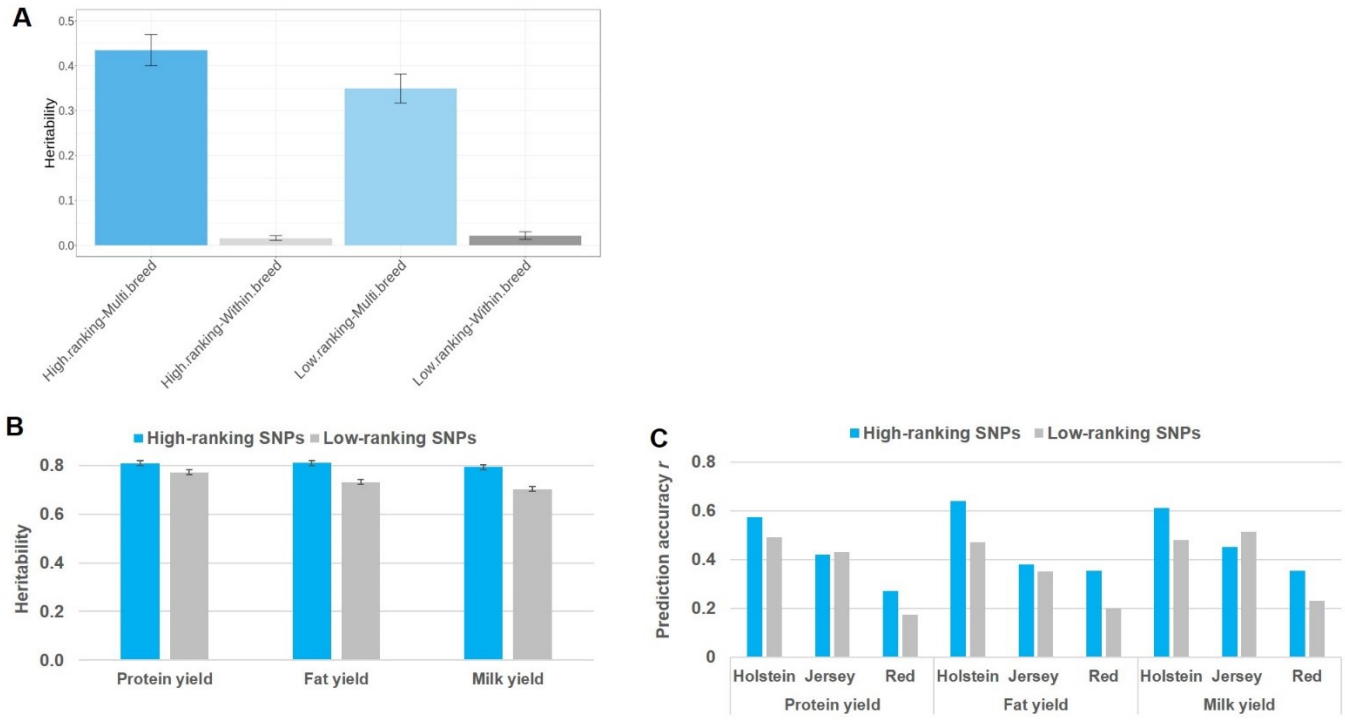

Figure 5. Further tests of the variant FAETH score. A: The heritability of high and low FAETH ranking variants for the multi-breed GRM and the within-breed GRM (2 GRMs fitted together) estimated across 34 traits in the Australian data. The error bars are the standard error of heritability calculated across 34 traits. B: The heritability of high and low FAETH variants for 3 production traits in Danish data. The error bars are the standard error of the heritability of each GREML analysis. C: prediction accuracy of gBLUP of 3 production traits in Danish data using high and low FAETH variants (averaged between bulls and cows).

Validation of the FAETH score in Danish cattle

An independent dataset of 7,551 Danish cattle of multiple breeds were used to test the FAETH score. The Australian high- and low- ranking variant sets were mapped in the Danish data. In the GREML analysis of Danish data, the high-ranking variants had significantly higher heritability than the low-ranking variants across three production traits (Figure 5B). The genomic best linear unbiased prediction (gBLUP) of Danish traits were also evaluated where the models were trained in the multiple-breed reference data to predict three production traits in each of three breeds (3 × 3 = 9 scenarios, Figure 5C). Out of these 9

scenarios, high-ranking variants had higher accuracies than the low-ranking variants in 7 scenarios.

Discussion

GWAS have been very successful in finding variants associated with complex traits but they have been less successful in identifying the causal variants because often there are a large group of variants, in high LD with each other (particularly in livestock), that are all associated with the trait. To distinguish among these variants, it would be useful to have information, external to the traits being analysed, that point to variants which are likely to have an effect on phenotype. In this paper we have evaluated 30 sources of external information based on genome annotation, evolutionary data and intermediate traits such as gene expression and milk metabolites. Then, we assessed the variance that each set of variants explained when they were included in a statistical model that also included a constant set of 600k SNPs from the bovine HD SNP array. The purpose of this method is to find sets of variants which add to the variance explained by the HD SNPs presumably because they are in higher LD with the causal variants than the HD SNPs are. Since, the causal variants themselves are likely to be among the sequence variants analysed, this method is a filter for classes of variants that are enriched for causal variants or variants in high LD with them.

Our analysis highlights the importance of intermediate trait QTL, including QTLs for metabolic traits and gene expression (mQTLs, geQTLs, eeQTLs, sQTLs and aseQTLs). This is not a surprising result as the significant contribution of different intermediate trait QTLs to complex trait variations have been reported in humans (7, 28, 37-39) and cattle (13, 40-42). An advantage of these intermediate traits over conventional phenotypes is that individual QTL explain a larger proportion of the variance. For instance, cis eQTL tend to have a large

effect on gene expression. This reduces the noise-to-signal ratio and so increases power to distinguish causal variants from variants in partial LD with them. However, an intermediate QTL mapping study requires a large amount of resources, especially when considering different metabolic profiles and tissues with a large sample size. In the current analysis we utilised several methods to combine results from individual studies of intermediate QTL mapping (19, 20, 29) (equation 1, 2, 3, 5 in Methods). This could reduce the noise from individual analysis and this is likely to increase the chance of finding causal mutations.

To our knowledge, no study has systematically compared the genetic importance of mQTLs with eQTLs. The high ranking of mQTLs over eQTLs in our study might be related to the fact that the mQTLs were discovered from the milk fat and the analysed phenotype in the test data contained several milk production traits. However, out of the 5,365 chosen mQTL variants, 961 variants were from the ±2Mb region of DGAT1 gene while no mQTLs were from chromosome 5 which harbors the MGST1 gene (Table S3, Figure 4B), both of which are known major milk fat QTL. This suggests that many variants from the mQTL set, not only influence milk fat production, but may have other functions including contributing to variations in general fat synthesis which is active in many mammalian tissues. Several large-scale human studies have highlighted the importance of mQTLs in various complex traits (7, 43).

Consistent with previous studies in cattle and humans (13, 28, 39), splicing sQTLs and its related eeQTLs ranked slightly higher than other expression QTL sets (Figure 3). A previous study in cattle found that aseQTLs and geQTLs had a similar magnitude of enrichment with trait QTL (29) consistent with the current observation.

Genomic sites that are conserved over evolutionary time probably affect some phenotype and hence fitness. However, we did not find conserved sites strongly associated with cattle complex traits. In humans, conserved sites across 29 mammals (44) showed strong

enrichment of variants associated with complex traits and many diseases (45). Conserved sites between humans and mice (46) and between humans and primates (47) showed enrichments in disease trait associations. A mutation that increases risk of disease is likely to be selected against (purifying selection) whereas a mutation that increases milk yield, for instance, may be positively or negatively selected depending on the environment and background genotype (subject to stabilising selection). Therefore, conserved sites might be more enriched for effects on disease than on traits subject to stabilising selection. Interestingly, conserved sites across 45 genomes of placental mammals and primates were not enriched for long noncoding RNAs associated with human complex traits (48). Perhaps the number and the types of species selected for the genome conservation estimation is another factor to consider when searching for trait associated conserved sites. Other methods exist for finding sites that have been subject to conservation and these sites are sometimes enriched for associations with human diseases traits (44, 47).

We proposed a novel method to identify variants that are young but at moderate frequency and found this set was enriched for effects on quantitative traits (Figure 3, Figure 4). However, Kemper et al (49) showed that variants identified by selection signatures using traditional methods, such as Fst (50) and iHS (51) had little contribution to complex traits in cattle. In the current study, the selection signatures between beef and dairy cattle ('Selection signature' set as shown in Table 1) explained some genetic variation in complex traits, although its quantity is relatively small (Table 2, Figure 3). It is possible that the inclusion of many non-production traits in the current study increased the chance of finding the trait-related sequence variants that are under artificial selection. The use of sequence variants in the current study may also have increased power compared with the study conducted by Kemper et al that used HD chip variants (49).

The set of variants with low PPRR ('young variants') had a higher ranking of genetic importance to the complex traits than the other artificial selection signatures (Figure 3). The identification of relatively young variants is based on the theory that very recent selection will increase the frequency of the favoured alleles (34). Thus, the young variant set could contain variants that were either under artificial selections and/or recently appeared and this may be the reason that it explained more trait variation than the artificial selection signatures. As shown in Figure 4, many young variants can be found in major production trait QTL. Genome regulatory elements such as enhancers and promoters are important regulators of gene expression and they can be identified by ChIP-seq assays. In humans, ChIP-seq tagged binding QTLs (bQTL) showed significant enrichments in complex and disease traits (52). We did not have enough individuals with ChIP-seq data to identify bQTLs in the current study. However, with only a limited amount of ChIP-seq data included, variants tagged by H3K4me3 ChIP-seq showed closer distance to the transcription start sites (Figure 2C) and H3K4me3 and H3K27ac together tagged variants had some contribution to complex trait variation (Figure 3). Also, the FAETH ranking of the ChIP-seq tagged variant set was similar to the ranking of variant annotation sets of gene end (variants within down- and up- stream of genes) and UTR (variants within 3' and 5' UTR). It is logical that variants with the potential to affect promoters and/or enhancers are annotated as close to genes or located in gene regulatory regions.

The variant annotation sets of non-coding related and splice sites ranked relatively high for their contribution to trait variation (Figure 3). Previously, variants annotated as splice sites also had a high ranking of genetic importance to cattle complex traits (53). The majority of the variants from the non-coding related set are 'non_coding_transcript_exon_variant' (Table S1) which is 'a sequence variant that changes non-coding exon sequence in a non-coding transcript' according to VEP (30). This group of variants can be associated with long non-

coding RNAs and their important contributions to complex traits in humans (48) and cattle (54) have been reported. Variants that were annotated as coding related, of which the majority of variants are missense and synonymous (Table S1), had relatively low ranking of genetic importance to complex traits (Figure 3). It seems a surprising result, but Koufariotis et al also reported similar observations in cattle (53). Perhaps coding variants that have an effect on phenotype are subject to purifying selection and hence have low heterozygosity and hence low contribution to variance.

The contribution of the variants with different LD properties to complex traits is an ongoing debate in humans (55-57). In our analysis of cattle, a domesticated species which tends to have strong LD between variants, negligible influences of variant LD differences to complex traits were observed (Table 2). Also, variants within regions that have more variants (variant density) did not explain more trait variation. Common variants, as expected (58), had substantial amount of contribution to complex traits (Table 2, Figure 3).

Based on the variant membership to differentially partitioned genome sets and the value of the per-variant $\overline{h^2_{set}}$, the FAETH score of sequence variants combined the information of evolutionary and functional significance and heritability estimates across multiple complex traits for each variant. This novel analytical framework provides simple but effective and comprehensive ranking for each variant that entered the analysis. Additional information of functional and/or evolutionary datasets can be easily integrated and linked to the variant contributions to multiple complex traits. A single score for each variant also makes the potential use of FAETH score easy and straightforward. For example, variants can be categorised as high and low FAETH ranking to create biological priors to inform Bayesian modelling for genomic selection (59).

The utility of the FAETH score estimated in the Australian data was tested independently in a Danish dataset. Overall, variants with high FAETH ranking explained significantly more

genetic variance in protein, fat and milk yield in the Danish data, compared to the variants with low FAETH ranking (Figure 5B). This validates the enriched genetic information in the variants ranked high by FAETH. When evaluated in the genomic prediction trained in multiple breeds and predicted into single breeds, high-ranking variants had increased prediction accuracy compared to low-ranking variants for all 3 traits in Danish Holstein and Red breeds and for fat yield in the Danish Jersey breed (Figure 5C). By building the within-breed GRM and comparing it with the multi-breed GRM (Figure 5A), our analysis suggested that the variants with the high FAETH ranking contained variants with consistent effects across different breeds. Future systematic analysis with increased breed diversity will provide better evaluation of the performance of the FAETH ranked variants in cross-breed genomic models.

In humans, Finucane et al (45) combined many sources of data to calculate a prior probability that a variant affects a phenotype. Our approach is different to theirs in some respects. They used GWAS summary data and stratified LD score regression, whereas we used raw data and GREML. They fitted all sources of information simultaneously whereas we fitted one at a time in competition with the HD variants. We were unable to fit all sources at once with GREML for computational reasons but also because the extensive LD in cattle makes it harder to separate the effects of multiple variant sets. On the other hand, GREML is more powerful than LD score regression (60). Our results are similar to theirs in some cases but not all. For instance, we both found expression QTL to be enriched for associations with complex traits, but we did not find enrichment in conserved sites or coding sites.

Our study demonstrates that the increasing amount of genomic and phenotypic data is making the cattle model a robust and critical resource of testing genetic hypotheses for large mammals. A recent large-scale study for cattle stature also supports the general utility of the cattle model in GWAS (5). In the current study, we highlight the contribution of the variants

associated with intermediate QTLs and non-coding RNAs to complex traits and this is consistent with many observations in human studies (8, 9, 28). However, we also provide contrasting evidence to results found from humans. We found that the conserved sites had little contribution to cattle complex traits, which is in contrast to its reported significant contribution in human complex and disease traits (45). Also, we found LD property of variants had negligible influences on trait heritability, contrasting the recent evidence for the strong influences of LD property on human complex traits (55). In addition, variants under artificial selection, which are absent from humans where natural selection clearly operates on complex traits (61), had limited contributions to bovine complex traits. While the reasons for these contrasting results are yet to be studied, our findings from cattle add valuable insights into the ongoing discussions of genetics of mammalian complex traits.

Our study is not without limitations. While some discovery analyses of the intermediate QTLs used relatively large sample size, the number of tissues and/or types of 'omics data included for discovering expression QTLs and mQTLs is yet to be increased. Also, in the discovery analysis, the selection criteria for informative variants to be included as targeted sets for building GRMs were relatively simple. In the test analysis, the heritability estimation for different GRMs used the GREML approach which has been under some debate because of its potential bias (56, 62). Analysis of functional categories by the genomic feature models with BLUP has been previously tested (63), although this methods can be computationally intensive. However, we aimed to treat each discovery dataset as equally as possible and all GRMs were analysed in the test dataset the same systematic way. The positive results from the validation analysis also suggest that informative variants have been well captured in the discovery and test analyses.

Conclusions

We provide the first extensive evaluation of the contribution of sequence variants with functional and evolutionary significance to multiple bovine complex traits. While developed using genomic and phenotypic data in the cattle model, the novel analytical approaches for the functional and evolutionary datasets and the FAETH framework of variant ranking can be well applied in other species. With their utility demonstrated, the publicly available variant FAETH scores will provide effective and simple-to-implement prior data for advanced genome-wide mapping and prediction.

Materials and Methods

Discovery analysis

A total of 360 cows from a three-year experiment at the Ellinbank research facility of Agriculture Victoria in Victoria, Australia, were used to generate functional datasets including RNA-seq, ChIP-seq and milk fat metabolites. All the animal experiments were approved by the Agriculture Victoria Animal Ethics Committee (2013–23).

The data of geQTLs, eeQTLs and sQTLs in each tissue of white blood and milk cells in a total of 131 Holstein and Jersey cows as previously published (NCBI Bioproject accessionPRJNA305942 (13)) were used. In addition, the data of geQTLs, eeQTLs and sQTLs from liver and semitendinosus muscle samples were also used (13) (NCBI Bioproject accession PRJNA392196). Previously, the geQTLs, eeQTLs and sQTLs were identified using the expression level of genes, exons and excision ratio of introns calculated using leafcutter (28), respectively, with imputed whole genome sequence (accuracy $r > 0.92$) analysed by Matrix eQTL (13, 64). From this analysis in each tissue, each variant had an estimate of the effect *b* and standard error (*se*) allowing for the multi-transcriptome meta-analysis in the current study. Such meta-analysis combining information from all four tissues followed the formula: $\chi^2_{(1)} = [\sum_{n=1}^{N} \frac{t_n}{\sqrt{N}}]^2$ (equation 1, published in (13)). N = the number of tissues (N = 4

in this case) where the single-transcriptome variant t values (b/se) were estimated. Variants with the p value < 0.0001 for the meta-analysis of 4 tissues in the analysis of geQTLs, eeQTLs and sQTLs were chosen for the geQTL, eeQTL and sQTL sets, respectively.

The aseQTLs were discovered using the RNA-seq data of white blood and milk cells in a total of 112 Holstein cows (5). The allele specific expression status of variants in heterozygous sites was tested based on the framework of transcript tVariant and driver dVariants proposed by Khansefid et al (29). Briefly, the model: $y_{acr} = X_1 b_1 + e$ (equation 2, adopted from (29)) was used, where y_{acr} was an N × 1 vector of $\log_{10}$ allele count ratio between parental genomes for the heterozygous exonic tVariant; N was the number of heterozygous animals at the tVariant; X_1 was an N × 1 vector coding the genotype of each animal at a dVariant which may drive the differential allele expression of the tVariant; b_1 as the regression coefficient, i.e., effects of the dVariant, for X_1 and e was the residual. dVariants were defined as all the variants within ±1MB distance to the tVariant and thus for a given phenotype as the allele count ratio at a tVariant, local (±1MB) linear models were performed for all dVariants. Then, tested dVariants had estimates of b_1 and the p values allowing for weighted meta-analysis for each gene using the formula: $\bar{z} = \frac{\sum_{i=1}^{N} z_i}{\sqrt{N}}$ (equation 3, published in (29)); where N was the number of times that b_1 of the dVariant was calculated by equation 2; $z_i = \Phi^{-1}(p_i)$ where Φ was the cumulative standard normal distribution and p_i was the p value of b_1 for each tested dVariant from equation 2. Variants with the p value < 0.0001 for the meta-analysis in both blood and milk cells were chosen for the aseQTL set.

The discovery of milk fat polar lipids metabolites mQTLs was based on the mass-spectrometry quantified concentration of 19 polar lipids from 338 Holstein cows. The bovine milk were collected as described above and polar lipids were extracted from bovine milk following the previously developed protocols (21). The chromatographic separation of polar lipids used a Luna HILIC column (250×4.6 mm, 5 µm, Phenomenex) maintained at 30 °C.

The lipids were detected by the LTQ-Orbitrap mass spectrometer (Thermo Scientific) operated in electrospray ionization positive (for most polar lipid classes) or negative (for analysis of PI) Fourier transform mode. The identification of lipid species present in milk was performed as previously reported (21). Quantification of selected polar lipid species was based on peak area of parent ions after normalization by the internal standard. GWAS of the concentration of each polar lipid was conducted using the model: $y_{lipids} = X\beta + Zu + wa + e$ (equation 4), where $\boldsymbol{y_{lipids}}$ was the vector of concentration of polar lipids of analysed individuals; $\boldsymbol{\beta}$ was the vector of fixed effects (analytical batches); $\mathbf{X}$ was a design matrix relating phenotypes to their fixed effects; u was the vector of animal effects where $\mathbf{u} \sim N(0, \mathbf{G}\sigma_g^2)$, $\mathbf{G}$ was the genomic relationship matrix between individuals $\boldsymbol{Z}$ was the incidence matrix; *w* was the vector of imputed sequence genotypes (over 10.1 million sequence variants) coded as 0, 1 or 2 (representing the genotypes aa, Aa or AA) and a was the effect of the variant; e was the vector of residual effects. For each GWAS, each variant had an estimate of the effect *b* and *se* allowing for multi-trait meta-analysis of variant effects across 19 traits with the formula: $\chi^2_{(N)} = t_i' V^{-1} t_i$ (equation 5, published in (20)). N = the number of single-trait GWAS conducted; t_i was a N×1 vector of the signed t-values (b/se) of variant$_i$ for the N traits; t_i' was a transpose of vector t_i (1×N); V^{-1} was an inverse of the N × N correlation matrix where the correlation between two traits was the correlation over all analysed variant t values of the two traits. Variants with the p value < 0.0001 for the meta-analysis of 19 polar lipids were chosen for the mQTL set.

ChIP-seq marks indicative of enhancers and promoters from a combination of experimental and published datasets were used. Trimethylation at lysine 4 of histone 3 (H3K4me3) ChIP-seq peak data of 9 bovine muscle samples (23) (NCBI GEO accession: GSE61936) and of H3K4me3 and acetylation at lysine 27 of histone 3 (H3K27ac) ChIP-seq peak data from 4

bovine liver samples (22) (EMBL Array Express accession: E-MTAB-2633) were downloaded.

H3K4me3 ChIP-seq peaks from mammary tissue of a lactating Holstein Dairy cow was generated as follows. 50mg of ground frozen tissue was fixed for 10 minutes and chromatin prepared using the MAGnify Chromatin Immunoprecipitation kit (Thermofisher) as per the manufacturer's instructions. Chromatin immunoprecipitation was performed with the same kit. 0.25 and 0.5µg of H3K4Me3 antibody (Abcam) was used for each immunoprecipitation with chromatin from 200,000 cells per reaction in triplicate. Libraries were made from ChIP product from all 3 reactions combined and input DNA (non-immunoprecipitated chromatin) using the NEBNext library prep kit (New England BioLabs). Libraries were sequenced on the HiSeq 3000 (Illumina) in a 150 cycle paired end run. More than 100 million reads were produced for the ChIP and input samples. Raw sequence reads were trimmed of adapter and poor-quality bases using Trimmomatic (65) using options ILLUMINACLIP: ADAPTER.fa:2:30:3:1:true LEADING:20 TRAILING:20 SLIDINGWINDOW:3:15 MINLEN:50. Reads were then aligned to the genome using BWA mem algorithim (66). Duplicates were marked with Picard (v2.6.0) MarkDuplicates (http://broadinstitute.github.io/picard) and reads with low mapping quality filtered using Samtools (v1.8) view (67) with -q 15 option. Narrow peak-calling was performed using MACS2 (v2.1.1, https://github.com/taoliu/MACS) based on default settings. DeepTools (v2.5.4) plotFingerprint (68) was used to plot cumulative sums of reads to assess ChIP quality. Phantompeakqualtools (v1.1) (69) was used to calculate cross-strand correlation metrics as another measure of ChIP quality. Bovine sequence variants within these peaks were defined as the ChIP-seq tagged variants and tagged variants from all samples were merged to one list of ChIP-seq tagged set.

The discovery of variant sets with evolutionary significance was based on the whole genome sequences of Run 6 of the 1000 bull genomes project (32). The selection signature analysis used a subset of 1,370 cattle of 15 dairy and beef breeds (Figure S1) with a linear mixed model approach. 18,446,470 sequence variants were used after filtering for Hardy–Weinberg equilibrium $p < 0.0001$ and MAF <0.005. For each animal, a binary phenotype (1/0) was created based on the assignment of the animal as a 'dairy' or 'beef' breed. This breed phenotype was analysed in the GWAS model: $y_{breed} = X\beta + Zu + wa + e$ (equation 6), where **y** was the vector of binary phenotype of breed (1/0) of analysed individuals; **β** was the vector of fixed effects (types of sequence assays); **X** was a design matrix relating phenotypes to their fixed effects; u was the vector of animal effects where $\mathbf{u} \sim N(0, \mathbf{G}\sigma_g^2)$, **G** was the genomic relationship matrix between the 1,370 individuals; $\boldsymbol{Z}$ was the incidence matrix; w was the vector of whole genome sequence genotypes coded as 0, 1 or 2 (representing the genotypes aa, Aa or AA) and a was the effect of the variant; e was the vector of residual effects. To improve the power of the GWAS, the leave-one-chromosome-out approach implemented in GCTA (33) was used and variants with the p value < 0.0001 for the GWAS were chosen for the selection signatures set.

To fully utilise the data of 1000 bull genomes the metric PPRR, *proportion of positive correlations* (*r*) *with rare variants* (MAF<0.01), was developed to infer the variant age. Our idea was based on the coalescent theory where the history of haplotypes of a sample can be represented by branching structures with the root being their common ancestor (70). The distribution of sequence variants is related to the branch lengths for the coalescence, and as demonstrated in humans (34) very recent selection decreased the branch lengths and increased the frequency of the favoured allele, compared to a neutral expectation. Therefore, haplotypes with favoured alleles had reduced number of 'singleton mutations' (34), i.e., the rarest type of variants which has only been seen once. This highlighted the negative

relationship between allele rarity and favourability (Figure S2A) and thus inspired our proposal: variants that appeared and/or are selected recently, i.e., relatively young, in a population could be enriched in regions with a reduced number of positive relationships with rare variants. PPRR was then calculated as $\pi_{+r} = \frac{N_k[+r(w_c,w_{rare})]}{N_k[r(w_c,w_{rare})]}$ (equation 7), where π_{+r} was the PPRR; $N_k[+r(w_c, w_{rare})]$ was the count (N) of all the positive correlations (r) between the genotypes of common variants (w_c) and the genotypes of rare variants (w_{rare}) in a given window with a size of k (k = 50kb for this study for computational efficiency). $N_k[r(w_c, w_{rare})]$ was the count of all correlations regardless of the sign. The calculation of π_{+r} can be easily and effectively performed using plink1.9 (www.cog-genomics.org/plink/1.9/). To reduce noises only correlations with $|r| > 0.0002$ were considered. An example of the distribution of the PPRR across the allele frequency for bovine chromosome 25 was given in the Figure S2B. In the end, variants within the top 1% of the reversed ranking of PPRR in each 10% allele frequency bin, e.g., AF $\in$ $\{(0, 0.1], (0.1, 0.2], \ldots, (0.9, 1)\}$, were selected to represent the young variant set.

Conserved genome sites, as another measure of the evolutionary significance, were also determined using the PhastCon program (26). The analysis used the reference genome sequences of cattle (UMD3.1), dog (CanFam3.1), mouse (GCRm38.p6) and human (GRCh38.p12) from Ensembl (https://www.ensembl.org/). Variants within the bovine genome sites with PhastCon score > 0.9 were chosen for the conserved site variant set. However, the conserved site variant set was not informative in our analysis (detailed in the *test analysis*) and was not included in the final ranking of variants.

Several datasets with annotated variant functions were also used to partition genome variants. The variant annotation category shown in Table 1 was based on predictions from Ensembl variant Effect Predictor (30) in conjunction with NGS-variant (31). Several variant annotations were merged from the original annotations to achieve reasonable sizes for

GREML. Merged annotations included geneend, splice.sites, coding related and non.coding related and the details of the size of original variant annotation sets can be found in the Table S1. Two additional variant functional annotations were also considered. One was the gkm SVM score of bovine genome sites from the HPRS (25) where each bovine genome site had a score of predicted regulatory potential. Variants in our study that overlapped HPRS and the were within the top 1% of the SVM score ranking (169,773 variants) were selected as the HPRS variant set. Another annotation dataset was the predicted CTCF sites published by Wang et al (24). Variants that overlapped with predicted bovine CTCF sites from (24) were chosen to be the CTCF variant set (252,234 variants). The HPRS and CTCF variant sets were not informative in our analysis (detailed in the *test analysis*) and were not included in the final ranking of variants.

Variant sets based on their distribution of LD score, density and MAF were created using GCTA-LDMS method (35) based on imputed genome sequences of the test dataset of 11,923 bulls and in 32,347 cows (detailed below). Over 17.6 million genome variants were partitioned into four quartiles of LD score per region (region size = 50kb), number of variants per window (window size = 50kb) and MAF sets of variants which were used to make GRMs. The quartile partitioning of sequence variants followed the default setting of the GCTA-LDMS. As a by-product of GCTA LD score calculation, the number of variants per 50kb window was computed and the quartiles of the value of variant number per region for each variant was used to generate the variant density sets.

Test analysis

An Australian dataset of 11,923 bulls and 32,347 cows from Holstein (9,739 ♂ / 22,899 ♀), Jersey (2,059 ♂ / 6,174 ♀), mixed breed (0 ♂ / 2,850 ♀) and Red dairy breeds (125 ♂ / 424 ♀) obtained from DataGene (http://www.datagene.com.au/) with 34 phenotypic traits (trait deviations for cows and daughter trait deviations for bulls (19)), including 5 production, 2

reproduction, 3 management and 24 type traits, were used for the test analysis (Table S2). All the traits were ordered by their number of non-missing records and transformed by Cholesky factorisation (19), so that they had minimal correlations with each other. Briefly, the formula of $C_n = L^{-1}g_n$ (equation 8, published in (19)) was used where $C_n$was a k (number of traits)×1 vector of Cholesky scores for the animal n; L was the k×k matrix of the Cholesky factor which satisfied $LL^t = COV$, the k×k covariance matrix of raw scores after standardisation as z-scores, g_n was an k×1 vector of traits for animal n. As a result, the kth Cholesky transformed trait can be interpreted as the k^{th} original trait corrected for the preceding k-1 traits and each Cholesky transformed trait had a variance of close to 1 (Table S2).

A total of 17,669,372 imputed sequence variants with Minimac3 imputation accuracy (REF) R^2 > 0.4 in above described bulls and cows using 1000 bull genome data (5, 32) as the reference set were used in the test analysis. Lists of variant sets selected from the discovery analysis with MAF > 0.001 in 11,923 bulls and in 32,347 cows were used to make targeted GRMs using GCTA (33). A GRM of the high-density (HD) variant chip (630,002 variants) was also made. Each targeted GRM was analysed in the 2-GRM REML model as : $y_{tr_i} = X\beta + Z_{set}u_{set} + Z_{HD}u_{HD} + e$ (equation 9); where $\boldsymbol{y}_{tr_i}$ was the vector of trait $\underline{i}^{th}$ phenotypic trait of analysed individuals; $\boldsymbol{\beta}$ was the vector of fixed effects (breeds); $\mathbf{X}$ was a design matrix relating phenotypes to their fixed effects; u_{set} was the vector of animal effects for the targeted GRM where $\mathbf{u}_{set}$~N$(0, \mathbf{G}_{set}\sigma_g^2)$, $\mathbf{G}_{set}$ was the GRM between the analysed individuals made of the targeted variant set; $\mathbf{Z}_{set}$ was the incidence matrix made of the targeted variant set; u_{HD} was the vector of animal effects for the GRM made of the HD variants where $\mathbf{u}_{HD}$~N$(0, \mathbf{G}_{HD}\sigma_g^2)$, $\mathbf{G}_{HD}$ was the GRM between the analysed individuals made of the HD variants (630,002); e was the vector of residual. GREML was analysed using MTG2 (71) for each trait separately in different sexes to calculate the heritability, h_{set}^2, of the targeted GRM. For each GRM within each sex, the $\overline{h_{set}^2}$ was calculated as the average across

34 traits. The per-variant $\overline{h^2_{set}}$ was calculated as the $\overline{h^2_{set}}$ divided by the number of variants in the targeted GRM.

To calculate the FAETH variant ranking, for genome partitions where one set of variants was chosen, i.e., sets of eeQTLs, geQTLs, sQTLs, mQTLs, ChIP-seq, selection signatures, young variants, conserved site variant, HPRS and CTCF, the heritability of the set of rest variants, h^2_{rest}, was calculated as $h^2_{all\ SNPs} - h^2_{set}$. This allowed that for each genome partition, each variant had a membership to a set. For each trait, $h^2_{all\ SNPs}$ was calculated using the same model as equation 9, except that $\mathrm{Z}_{set}u_{set}$ was replaced by $\mathrm{Z}_{all\ SNPs}u_{all\ SNPs}$. $u_{all\ SNPs}$ was for the GRM where $\mathbf{u}_{all\ SNPs}\sim\mathrm{N}(0, \mathbf{G}_{all\ SNPs}\sigma^2_g)$, $\mathbf{G}_{all\ SNPs}$ was the GRM between the analysed individuals made of all variants considered with MAF > 0.001 (over 16.1 million variants); $\mathbf{Z}_{set}$ was the incidence matrix made of the all variant set. Then, this allowed for the calculation of $h^2_{rest} = h^2_{all\ SNPs} - h^2_{set}$ for each trait and the $\overline{h^2_{rest}}$ as the average across 34 traits. The per-variant $\overline{h^2_{rest}}$ was then calculated as the $\overline{h^2_{rest}}$ divided by the number of variants in the remaining ('rest'') set as the difference between the total number of variants and the number of variants in the targeted set. A criterion of per-variant $\overline{h^2_{set}}$ > per-variant $\overline{h^2_{rest}}$ was used to determine whether the variant set was informative. Based on this criterion, the sets of conserved site variant, HPRS and CTCF were determined not informative and their per-variant $\overline{h^2}$ estimates were not included in the FAETH ranking.

The FAETH ranking of variant sets used the estimates per-variant $\overline{h^2_{set}}$ and the ranking of each variant was derived based on the variant membership to the non-overlapping sets within each partition. If a variant belonged to a targeted set or a rest set in the partition, the estimate of per-variant $\overline{h^2_{set}}$ or the per-variant $\overline{h^2_{rest}}$ was assigned to the variant accordingly. In the end, the variant FAETH ranking was based on the average of the 12 genome partitions retained (as shown in Table 1).

Validation analysis

The validation used variants within the top 1/3 (high) and bottom 1/3 (low) ranking from the Australian analysis to make GRMs in a total of 7,551 Danish bulls of Holstein (5,411), Jersey (1,203) and Danish Red (937) with a total of 8,949,635 imputed sequence variants in common between the Danish and Australian datasets, with a MAF ≥ 0.002 and imputation accuracy measured by the info score provide by IMPUTE2 ≥ 0.9 in the Danish data (72). Deregressed proofs (DRP) were available for all animals in the Danish dataset for milk, fat and protein yield. The Danish dataset was divided into a reference and validation set, where the reference set include 4,911 Holstein, 957 Jersey and 745 Danish Red bulls and the validation set included 500 Holstein, 517 Jersey and 192 Danish Red bulls. Over 1.25 million high-ranking variants and over 1.25 million low-ranking variants were used to make the high- and low- ranking GRMs. For the individuals in the reference set, each trait of protein, milk and fat yield was analysed with the GREML model $y_{Dan} = \mathrm{X}\beta + \mathrm{Z}_{Dan}u_{Dan} + e$ (equation 10) using GCTA (33), where $\boldsymbol{y_{Dan}}$ was the vector of DRP of analysed Danish individuals; $\boldsymbol{\beta}$ was the vector of fixed effects (breeds); $\mathbf{X}$ was a design matrix relating phenotypes to their fixed effects; u was the vector of animal effects where $\mathbf{u}_{Dan} \sim \mathrm{N}(0, \mathbf{G}_{Dan}\sigma_g^2)$, $\mathbf{G}_{Dan}$ was the genomic relationship matrix between Danish individuals, $\mathbf{Z}_{Dan}$ was the incidence matrix; e was the vector of residual. This allowed the estimate of h^2 of high- and low- ranking variants in the Danish data.

To further test the utility of the variant ranking, genomic prediction with gBLUP was also performed by dividing the Danish individuals into reference and validation datasets. The –blup-variant option in GCTA (33) was used to obtain variant effects from the GREML analyses, that were used to predict GEBV in the validation population. Prediction accuracies were computed for each of the breeds in the validation population, as the correlation between GEBV and DRP.

The high- and low- ranking variants were also evaluated for their utility in across-breed and within breed analysis in the Australian dataset. The high- and low- ranking variants were used to make GRMs in Australian bulls. The within-breed GRM was built following the intuition from (73) by setting the across-breed elements, i.e., the relationship between individual pairs from different breeds, of the original GRM to the mean of the breed block. For each of the 34 traits, the original multi-breed GRM was fitted together with the within-breed GRM in the 2-GRM REML model similar to equation 9. Then, the $\overline{h^2}$ of multi-breed and within breed GRMs of high- and low- ranking variants across 34 traits were calculated.

Footnotes

M.E.G. conceived the project. R.X. and I.M.M. implemented the design of the analysis. C.P.W. and A.J.C. performed sample collections and ChIP sequencing experiments. C.M.R. and B.A.M. contributed to the genotyping work. Z.L. and S.J.R. contributed to the data generation of milk fat metabolites. S.B., I.M.M. and H.D.D. provided data and assisted with study design. R.X., I.M.M., B.J.H., C.P.P., M.W. H.D.D., C.J.V. and M.E.G. analysed data. I.B. and M.S.L. conducted validation analysis. R.X. and M.E.G. wrote the paper. R.X., M.E.G., B.J.H., C.P.W. A.J.C., I.B., I.M.M., H.D.D and M.S.L. revised the paper. All authors read and approved the final manuscript.

Acknowledgements

Australian Research Council's Discovery Projects (DP160101056) supported R.X. and M.E.G. Dairy Futures CRC supported the generation of the Holstein and Jersey transcriptome data. DairyBio (a joint venture project between Agriculture Victoria and Dairy Australia) funded the generation of the mammary ChIPseq data. I.B. was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark

(grant 0603-00519B). No funding bodies participated in the design of the study nor collection, analysis, or interpretation of data nor in writing the manuscript. We thank DataGene for access to data used in this study and Gert Nieuwhof, Kon Konstantinov and Timothy P. Hancock for preparation and provision of data. We thank partners from the 1000-bull genome project for the data access. We thank Dr. Majid Khansefid for the discussion of aseQTL analysis.

References:

1. Visscher PM, *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* 101(1):5-22.
2. Nielsen JB, *et al.* (2018) Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature genetics* 50(9):1234.
3. FAO (2018) http://www.fao.org/faostat.
4. Taylor JF, Taylor KH, & Decker JE (2016) Holsteins are the genomic selection poster cows. *Proceedings of the National Academy of Sciences*:201608144.
5. Bouwman AC, *et al.* (2018) Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics* 50(3):362.
6. MacHugh DE, Shriver MD, Loftus RT, Cunningham P, & Bradley DG (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus). *Genetics* 146(3):1071-1086.
7. Yousri NA, *et al.* (2018) Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nature communications* 9(1):333.
8. Consortium G (2017) Genetic effects on gene expression across human tissues. *Nature* 550(7675):204.
9. Lizio M, *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* 16(1):22.
10. Andersson R, *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455-461.
11. Andersson L, *et al.* (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology* 16(1):57.
12. Clark EL, *et al.* (2017) A high resolution atlas of gene expression in the domestic sheep (Ovis aries). *PLoS Genetics* 13(9):e1006997.
13. Xiang R, *et al.* (2018) Genome variants associated with RNA splicing variations in bovine are extensively shared between tissues. *BMC Genomics* 19(1):521.
14. Giuffra E, Tuggle CK, & FAANG Consortium T (2018) Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annual review of animal biosciences*.

15. Zeng J, *et al.* (2018) Signatures of negative selection in the genetic architecture of human complex traits. *Nature genetics* 50(5):746.
16. Yang J, *et al.* (2017) Genetic signatures of high-altitude adaptation in Tibetans. *Proceedings of the National Academy of Sciences*:201617042.
17. Xu L, *et al.* (2014) Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Molecular biology and evolution* 32(3):711-725.
18. Hayes BJ & Daetwyler HD (2018) 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual review of animal biosciences*.
19. Xiang R, MacLeod IM, Bolormaa S, & Goddard ME (2017) Genome-wide comparative analyses of correlated and uncorrelated phenotypes identify major pleiotropic variants in dairy cattle. *Scientific Reports* 7(1):9248.
20. Bolormaa S, *et al.* (2014) A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. *PLOS Genetics* 10(3):e1004198.
21. Liu Z, Moate P, Cocks B, & Rochfort S (2015) Comprehensive polar lipid identification and quantification in milk by liquid chromatography–mass spectrometry. *Journal of Chromatography B* 978:95-102.
22. Villar D, *et al.* (2015) Enhancer evolution across 20 mammalian species. *Cell* 160(3):554-566.
23. Zhao C, *et al.* (2015) Genome-Wide H3K4me3 Analysis in Angus Cattle with Divergent Tenderness. *PLOS ONE* 10(6):e0115358.
24. Wang M, *et al.* (2018) Putative bovine topological association domains and CTCF binding motifs can reduce the search space for causative regulatory variants of complex traits. *BMC genomics* 19(1):395.
25. Nguyen QH, *et al.* (2018) Mammalian genomic regulatory regions predicted by utilizing human genomics, transcriptomics, and epigenetics data. *GigaScience* 7(3):gix136.
26. Siepel A, *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* 15(8):1034-1050.
27. Yang J, *et al.* (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics* 47(10):1114.
28. Li YI, *et al.* (2018) Annotation-free quantification of RNA splicing using LeafCutter. *Nature genetics* 50(1):151.
29. Khansefid M, *et al.* (2018) Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC genomics* 19(1):793.
30. McLaren W, *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biology* 17(1):122.
31. Grant JR, Arantes AS, Liao X, & Stothard P (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* 27(16):2300-2301.
32. Daetwyler HD, *et al.* (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics* 46(8):858.
33. Yang J, Lee SH, Goddard ME, & Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 88(1):76-82.
34. Field Y, *et al.* (2016) Detection of human adaptation during the past 2000 years. *Science* 354(6313):760-764.

35. Yang J, *et al.* (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47(10):1114-1120.
36. Kemper KE, *et al.* (2015) Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution* 47(1):29.
37. Ongen H, *et al.* (2017) Estimating the causal tissues for complex traits and diseases. *Nature genetics* 49(12):1676.
38. Consortium G (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348(6235):648-660.
39. Zhernakova DV, *et al.* (2017) Identification of context-dependent expression quantitative trait loci in whole blood. *Nature genetics* 49(1):139.
40. Kemper K, *et al.* (2016) Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. *BMC genomics* 17(1):858.
41. Sanchez M-P, *et al.* (2017) Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. *Genetics Selection Evolution* 49(1):68.
42. Littlejohn MD, *et al.* (2016) Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. *Scientific Reports* 6:25376.
43. Shin S-Y, *et al.* (2014) An atlas of genetic influences on human blood metabolites. *Nature genetics* 46(6):543.
44. Lindblad-Toh K, *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476.
45. Finucane HK, *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* 47(11):1228.
46. Gjoneska E, *et al.* (2015) Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 518(7539):365.
47. Srinivasan S, *et al.* (2016) Genetic markers of human evolution are enriched in schizophrenia. *Biological psychiatry* 80(4):284-292.
48. Tan JY, *et al.* (2017) Cis-acting complex-trait-associated lincRNA expression correlates with modulation of chromosomal architecture. *Cell reports* 18(9):2280-2288.
49. Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, & Goddard ME (2014) Selection for complex traits leaves little or no classic signatures of selection. *BMC genomics* 15(1):246.
50. Depaulis F & Veuille M (1998) Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular biology and evolution* 15(12):1788-1790.
51. Akey JM, *et al.* (2010) Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences* 107(3):1160-1165.
52. Tehranchi AK, *et al.* (2016) Pooled ChIP-seq links variation in transcription factor binding to complex disease risk. *Cell* 165(3):730-741.
53. Koufariotis LT, Chen Y-PP, Stothard P, & Hayes BJ (2018) Variance explained by whole genome sequence variants in coding and regulatory genome annotations for six dairy traits. *BMC genomics* 19(1):237.
54. Cai W, *et al.* (2018) Genome Wide Identification of Novel Long Non-coding RNAs and Their Potential Associations With Milk Proteins in Chinese Holstein Cows. *Frontiers in genetics* 9.

55. Speed D, *et al.* (2017) Reevaluation of SNP heritability in complex human traits. *Nature genetics* 49(7):986.
56. Yang J, Zeng J, Goddard ME, Wray NR, & Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. *Nature genetics* 49(9):1304.
57. Evans LM, *et al.* (2018) Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics* 50(5):737.
58. Yang J, *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565-569.
59. MacLeod I, *et al.* (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC genomics* 17(1):144.
60. Ni G, *et al.* (2018) Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. *The American Journal of Human Genetics*.
61. Guo J, *et al.* (2018) Global genetic differentiation of complex traits shaped by natural selection in humans. *Nature communications* 9(1):1865.
62. Kumar SK, Feldman MW, Rehkopf DH, & Tuljapurkar S (2016) Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences* 113(1):E61-E70.
63. Fang L, *et al.* (2017) Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC genomics* 18(1):604.
64. Shabalin AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10):1353-1358.
65. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
66. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26(5):589-595.
67. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
68. Ramirez F, *et al.* (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research* 44(W1):W160-165.
69. Kharchenko PV, Tolstorukov MY, & Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology* 26(12):1351.
70. Wakeley J (2009) *Coalescent theory: an introduction*.
71. Lee SH & Van der Werf JH (2016) MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* 32(9):1420-1422.
72. van den Berg I, Boichard D, & Lund MS (2016) Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genetics Selection Evolution* 48(1):83.
73. Khansefid M, *et al.* (2014) Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. *Journal of animal science* 92(8):3270-3283.