



ARTICLE

<https://doi.org/10.1038/s42003-019-0454-y>

OPEN

Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls

Jicai Jiang¹, John B. Cole ², Ellen Freebern¹, Yang Da³, Paul M. VanRaden² & Li Ma ¹

A hundred years of data collection in dairy cattle can facilitate powerful studies of complex traits. Cattle GWAS have identified many associated genomic regions. With increasing numbers of cattle sequenced, fine-mapping of causal variants is becoming possible. Here we imputed selected sequence variants to 27,214 Holstein bulls that have highly reliable phenotypes for 35 production, reproduction, and body conformation traits. We performed single-marker scans for the 35 traits and multi-trait tests of the three trait groups, revealing 282 candidate QTL for fine-mapping. We developed a Bayesian Fine-MAPPING approach (BFMAP) to integrate fine-mapping with functional enrichment analysis. Our fine-mapping identified 69 promising candidate genes, including *ABCC9*, *VPS13B*, *MGST1*, *SCD*, *MKL1*, *CSN1S1* for production, *CHEK2*, *GC*, *KALRN* for reproduction, and *TMTC2*, *ARRDC3*, *ZNF613*, *CCND2*, *FGF6* for conformation traits. Collectively, these results demonstrated the utility of BFMAP, identified candidate genes, and enhanced our understanding of the genetic basis of cattle complex traits.

¹ Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA. ² Animal Genomics and Improvement Laboratory, USDA, Building 5, Beltsville, MD 20705, USA. ³ Department of Animal Science, University of Minnesota, St Paul, MN 55108, USA. Correspondence and requests for materials should be addressed to L.M. (email: lima@umd.edu)

Phenotypic records have been routinely collected in dairy cattle to facilitate selective breeding for more than one hundred years. The phenotype of a bull can be highly accurately calculated from thousands of phenotypic records of his daughters and other relatives¹. A comprehensive spectrum of phenotypes has been recorded in dairy cattle, including production, reproduction, health, and body type traits². GWAS on these traits simultaneously in the same population can provide a better understanding of the effects of underlying QTLs. Because of the intensive use of artificial insemination and strong selection in dairy bulls, there are a much smaller number of males than females in the cattle population³, and chromosome segments can be quickly traced back to an ancestral bull. The high relatedness in the cattle population can facilitate accurate imputation⁴, especially with the availability of many important ancestor bulls sequenced by the 1000 Bull Genomes project^{5–8}. These unique features of the cattle population make a large-scale GWAS with imputed sequence variants possible and valuable.

Fine-mapping of complex traits to single-variant resolution has started in human studies, e.g., ref. 9,10. Because of the high levels of linkage disequilibrium (LD) in the livestock population¹¹, fine-mapping of GWAS signals is still difficult in cattle. Additionally, existing fine-mapping methods are not easily applicable to large-scale cattle GWAS and fine-mapping studies. Some methods, e.g., CAVIARBF¹² and PAINTOR¹³, generally use a logistic model with a binary response and categorical functional annotations as covariates. Such a logistic model is then incorporated into a model search scheme that often limits the maximum number of causal variants (e.g., 3) and is computationally impractical for a locus containing thousands of sequence variants. When multiple functional data sets are to be tested, model-searching needs to be conducted separately for each set of functional annotation data, further increasing the computational burden. In cattle, Bayes and BayesRC methods have been applied to incorporate sequence data into genomic selection models, but the large amount of computation from MCMC prohibits their direct application to large-scale fine-mapping studies^{14,15}. Although GCTA-COJO is capable of fast conditional analysis for fine-mapping in cattle¹⁶, the use of summary statistics and LD data from a reference population can be suboptimal when direct genotype and phenotype data are available. To address these problems, we develop a fast Bayesian Fine-MAPping method (BFMAP) that can efficiently integrate functional annotations with fine-mapping. Specifically, BFMAP can re-use initial model search results for various functional annotations and can be employed for both fine-mapping and functional enrichment analyses. More importantly, the functional enrichment estimated from BFMAP is, by definition, the enrichment of causal effects, in contrast to the enrichment of heritability by the well-known stratified LD score regression¹⁷.

In our study, the large number of bulls with highly reliable phenotype and imputed sequence variants can facilitate powerful GWAS and fine-mapping of major GWAS signals. Although the high LD in the cattle genome makes fine-mapping and functional enrichment studies difficult, the large sample size and improved methods can help identify candidate genes of complex traits as well as biologically informative enrichment of candidate variants in functional annotation data. Specifically, we seek to use BFMAP to identify and incorporate functional annotation into the fine-mapping of 35 production, reproduction, and conformation traits in dairy cattle. The fine-mapped genes and variants can provide candidates readily testable in functional studies. The functional data enriched with variants associated with complex dairy traits will be useful for future cattle GWAS and genomic prediction studies. Additionally, the initial model search results can be reused for estimating enrichment of causal effects of dairy traits for additional

functional annotations that are being generated by the FAANG and related projects in cattle¹⁸.

Results

Data description. We imputed over 3 million selected sequence variants to 27,214 Holstein bulls after quality control edits, using the 1000 Bull Genomes data as reference. These bulls were selected to have highly reliable breeding values (predicted transmitting abilities; PTA) for 35 production, reproduction, and body conformation traits, with an average reliability of 0.71 across traits (Table 1). The number of bulls available for individual traits ranged from 11,713 to 27,161, with >20,000 animals having data for 32 traits (Table 1). The 27,214 bulls had over 31.6 million daughters with records for milk production, and the counts were lower for other traits. This large, high-quality bull data set enables our following GWAS and fine-mapping studies with great power and precision.

Single-trait GWAS. We used a mixed-model approach implemented in the software MMAP¹⁹ that can incorporate reliability variation across individual bulls. The mixed-model used in our GWAS was robust against population structure and familial relatedness. As shown in Supplemental Data 1, 27 of the 35 traits had a genomic control factor between 0.95 and 1.05.

Using a genome-wide significance level of $P < 5E-8$, we found many clear association signals for the 35 dairy traits (Supplementary Fig. 1). In total, there were 286 unique QTL regions associated with the 35 traits, and the number of associations for individual traits ranged from <3 for leg and foot traits to 23 for protein percentage (Supplemental Data 1 and 2). As compared to the Cattle QTLdb release 35²⁰, we found that 123 associations (43%) had been previously reported while 163 associations (57%) were newly discovered in this study. We identified 15 new association signals (out of 68) even for the five production traits that had been extensively studied previously, and 92 new associations (out of 125) for type traits that drew less attention in previous studies (Fig. 1 and Supplemental Data 2). While a proportion of these newly discovered QTLs were identified to be associated with new traits, these results demonstrated the superior power of our GWAS in dairy cattle.

Multi-trait association analysis. Consistent with trait definition, hierarchical clustering of the 35 traits based on the absolute correlation coefficients identified three trait clusters: production, reproduction, and body type (Fig. 2). Interestingly, rump angle, teat length, and dairy form were clustered into reproduction traits, although they are type traits by definition, indicating a close genetic correlation between these three traits and cattle reproduction. Even after removing the potential distortion from net merit, rump angle, and teat length were still clustered in the reproduction group while dairy form was clustered in production traits (Supplementary Fig. 2).

From multi-trait association analyses of the three trait clusters, we identified 33, 21, and 39 associations for production, reproduction, and type traits using $P < 5E-8$, respectively (Fig. 3 and Supplemental Data 3). While multi-trait analysis is generally more powerful than single-trait GWAS for pleiotropic QTLs^{21,22}, we found fewer associations from the multi-trait analyses than in single-trait results (76 vs 286 unique QTLs). This is likely due to the proportion of QTLs with pleiotropic effects on related traits is less than expected, and/or the limited benefit of including additional traits in cattle studies where individual traits are already highly accurate (Table 1). Although the majority of the multi-trait associations were already identified from single-trait GWAS, we found ten associations that were missed by single-trait

Table 1 Number of Holstein bulls, mean and standard deviation (SD) of PTAs, and reliabilities for 35 dairy traits

Trait Name	Abbreviation	N of Bulls	Deregressed PTA		Reliability	
			Mean	SD	Mean	SD
Milk yield	Milk	27,156	-245.86	850.58	0.860	0.082
Fat yield	Fat	27,156	-5.92	30.52	0.860	0.082
Protein yield	Protein	27,156	-5.31	23.84	0.863	0.083
Fat percentage	Fat_Percent	27,156	0.0136	0.107	0.860	0.082
Protein percentage	Pro_Percent	27,156	0.0086	0.0464	0.863	0.083
Net merit	Net_Merit	27,161	-106.91	278.63	0.763	0.110
Productive life	Prod_Life	26,727	-1.367	3.461	0.682	0.145
Somatic cell score	SCS	27,143	3.027	0.235	0.786	0.110
Age at first calving	AFC	16,314	-0.446	11.855	0.439	0.258
Days to first breeding ^a	DFB	11,713	0.534	2.825	NA	NA
Daughter pregnancy rate	Dtr_Preg_Rate	25,699	-0.593	3.025	0.618	0.185
Heifer conception rate	Heifer_Conc_Rate	19,334	-0.660	9.610	0.377	0.210
Cow conception rate	Cow_Conc_Rate	20,380	-1.053	6.879	0.597	0.202
Sire calving ease	Sire_Calv_Ease	26,345	7.959	2.461	0.671	0.224
Daughter calving ease	Dtr_Calv_Ease	23,263	9.141	3.182	0.594	0.176
Sire stillbirth	Sire_Still_Birth	21,543	8.190	1.831	0.495	0.249
Daughter stillbirth	Dtr_Still_Birth	20,424	8.085	2.958	0.508	0.222
Final score	Final_score	25,638	-0.817	1.484	0.702	0.140
Stature	Stature	25,641	-0.482	1.532	0.844	0.079
Strength	Strength	25,633	-0.278	1.513	0.743	0.147
Dairy form	Dairy_form	25,615	-0.492	1.745	0.752	0.132
Foot angle	Foot_angle	25,626	-0.742	2.263	0.664	0.198
Rear legs (side view)	Rear_legs(side)	25,641	-0.009	1.734	0.754	0.137
Body depth	Body_depth	25,636	-0.413	1.622	0.720	0.180
Rump angle	Rump_angle	25,641	0.038	1.482	0.828	0.089
Rump width	Rump_width	25,641	-0.504	1.543	0.766	0.114
Fore udder attachment	Fore_udder_att	25,640	-0.908	1.852	0.781	0.112
Rear udder height	Rear_ud_height	25,640	-0.885	2.095	0.737	0.136
Udder depth	Udder_depth	25,631	-0.653	1.665	0.836	0.082
Udder cleft	Udder_cleft	25,641	-0.720	1.980	0.718	0.156
Front teat placement	Front_teat_pla	25,641	-0.562	1.663	0.781	0.106
Teat length	Teat_length	25,631	0.104	1.482	0.815	0.087
Rear legs (rear view)	Rear_legs(rear)	24,763	-0.759	2.709	0.605	0.178
Feet and legs composite	Feet_and_legs	25,608	-0.928	2.501	0.600	0.208
Rear teat placement	Rear_teat_pla	25,492	-0.436	1.900	0.762	0.103

^aFor DFB, we used PTA as reliability was unavailable

analyses (Supplemental Data 4). Interestingly, we noticed that the top variant in multi-trait analysis could be >1 Mb away from the top variants in single-trait GWAS (Supplementary Fig. 3), so the multi-trait results were combined with single-trait analyses to refine candidate QTL regions for fine-mapping.

Fine-mapping. To facilitate fast fine-mapping analyses, we developed a fast Bayesian Fine-MAPping method (BFMAP) that calculates a posterior probability of causality (PPC) for variants in candidate regions. We picked QTL regions for fine-mapping from both single- and multi-trait GWAS results. Initially, we fine-mapped 434 association signals for 282 QTLs using a significance threshold of $5E-7$ (Supplemental Data 5). The observed number of fine-mapped signals in a QTL is approximately exponentially distributed, consistent with our expectation of more causal mutations with a lower probability in a QTL region (Fig. 4). After further quality control edits, we finally fine-mapped 308 association signals for 32 traits (Supplemental Data 6). Specifically, there were more than 20 independent association signals identified on chromosomes 5, 6, 14, 18, and 29, while very few were identified on chromosomes 12, 22, and 27.

We investigated the impacts of incorporation of SnpEff-inferred effect impact (commonly used functional annotation) on fine-mapping performance. First, incorporating variant impacts resulted in a substantial change of PPC for variants in the 308 fine-mapped association signals. Variants with moderate impact

had a considerable increase in PPC when functional information was included in the calculation, while modifier variants generally had a decreased PPC (Fig. 5a). Second, fine-mapping by incorporating variant impacts generated significantly smaller 95% credible variant sets than that using an equal prior for all variants ($P = 0.01$, Wilcoxon signed-rank test; Fig. 5b). These two features make the incorporation of functional annotation favored in our fine-mapping analyses.

Enrichment analysis. To verify the quality of our fine-mapped variants and characterize their distribution on the cattle genome, we investigated the enrichment of fine-mapped variants with different functional annotation data available to cattle, including location in protein-coding gene, effect predicted by SnpEff²³, and evolutionary constrain predicted by GERP²⁴. Our enrichment analysis estimated the probability of a causal variant being in a functional category and the probability of a non-causal variant being in the category. The ratio of the two probabilities was used to measure the enrichment of causal variants for this functional category²⁵, with a value larger than one indicating higher enrichment than the genome background. This enrichment analysis has also been implemented in BFMAP.

We first categorized variants into five groups based on their locations regarding protein-coding genes, i.e., CDS, 5' UTR+2 kb upstream, intron, 3' UTR+2 kb downstream, and other (intergenic or non-protein-coding genic regions). Despite the strong

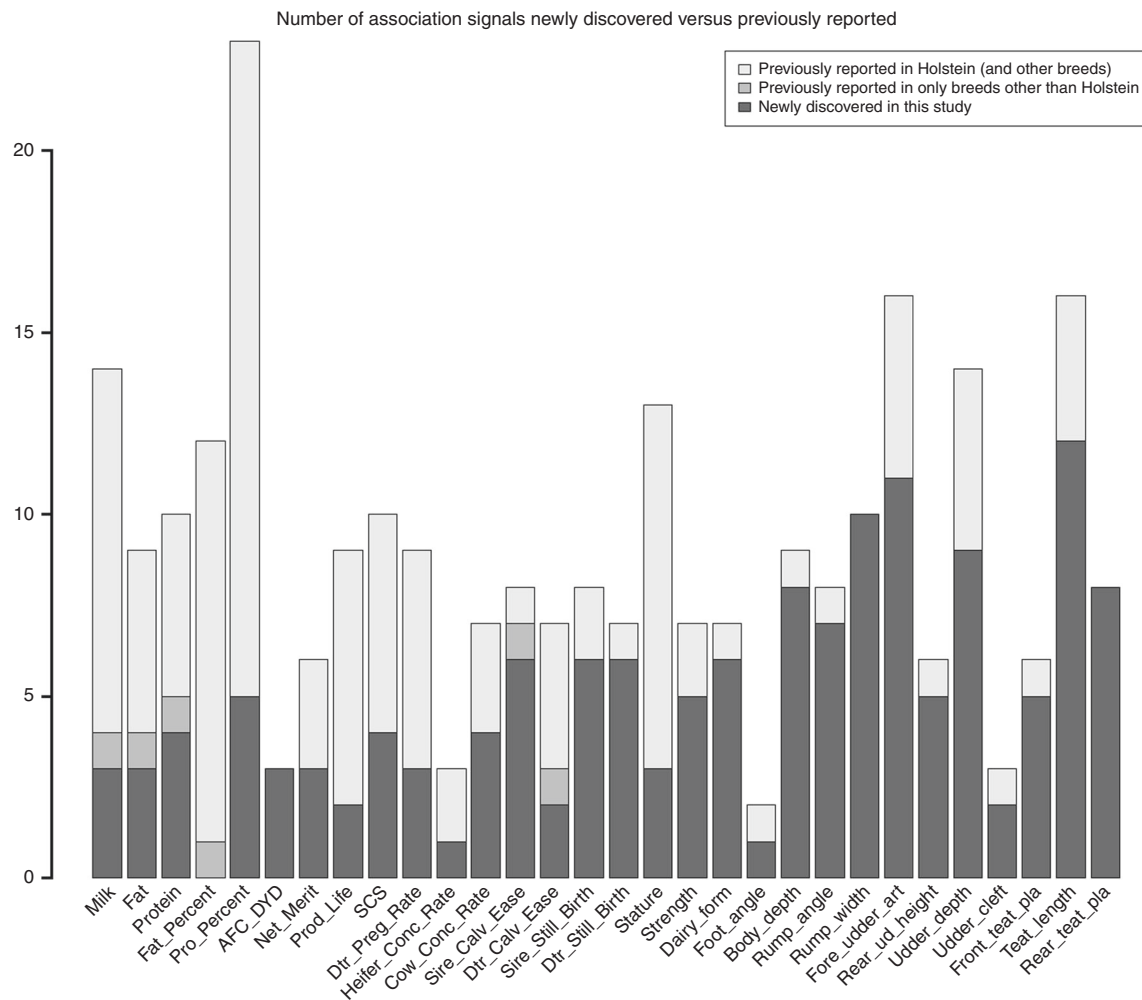


Fig. 1 Number of association signals newly discovered in our single-trait GWAS versus previously reported. There are in total 30 traits listed. Three leg traits were excluded since we found no associations passing genome-wide significance. DFB and final score were not listed because there were no matched traits in the Cattle QTLdb release 35

LD levels in the cattle genome²⁶, we observed distinctive enrichment patterns across these five categories (Fig. 6a). Using bootstrapping, we calculated 95% confidence intervals for the enrichment levels, showing significant enrichment of fine-mapped variants in CDS (4.52 \times) and 5' UTR (2.39 \times), but not in intron (0.93 \times) or 3' UTR (0.77 \times). We also analyzed a group of non-protein-coding genes but found significant depletion with $\widehat{E}_C = 3.2E-04$ (Supplemental Data 7), suggesting a lacking of functional impacts in these genes on dairy cattle traits.

We further investigated the enrichment of fine-mapped variants regarding their genomic locations and protein-coding effects (High, Moderate, Low or Modifier) predicted by SnpEff²³. When modeling these four categories, we found severe depletion of variants with high impact ($\widehat{E}_C = 2.51E-05$; Supplemental Data 8). This is strikingly different from a previous study on human complex traits and diseases that reported an enrichment of >100 for this category²⁵. As shown in Fig. 5b, we observed a significant enrichment in moderate-impact variants ($\widehat{E}_C = 8.7$; $P = 0.01$). Low-impact variants also showed an enrichment (2.0 \times), though it was not statistically significant (Fig. 6b). As expected, a minor depletion was seen in modifier variants (0.87 \times).

We also used constrained elements on the cattle genome to categorize variants into two groups (inside of or outside of constrained elements), as highly conserved DNA sequences may imply functional importance. As shown in Fig. 6c and

Supplemental Data 9, fine-mapped variants were significantly enriched in constrained elements (3.72 \times ; $P = 0.02$). When further categorizing variants into six groups based on both constrained elements and variant impacts (Moderate, Low or Modifier), we found the highest enrichment in moderate-impact variants inside constrained elements (25.56 \times ; $P = 0.005$). For the other categories, we observed no enrichment of fine-mapped variants (Fig. 6d and Supplemental Data 10).

When comparing different trait groups, we observed little difference in the pattern of enrichment regarding SnpEff-inferred effect impact (Fig. 7 and Supplemental Data 11). Moderate-impact variants had a clearly higher enrichment of being causal for production traits than for reproduction and type traits. We further used permutation to generate the null distribution of $E_C(\text{Production})/E_C(\text{Reproduction+Type})$ and showed that the difference was statistically significant ($P = 0.01$; Supplementary Fig. 4A). However, the enrichment for low-impact variants was similar between the three trait groups (Supplementary Fig. 4B).

Candidate genes. Based on the PPCs of variants after incorporation of SnpEff impact, we calculated PPC for each gene in each independent association signal. In total, there were 564 gene-trait association pairs with PPC >0.01 (Supplemental Data 12). Most of the genes had either a large (>0.95) or small PPC (<0.05) (Supplementary Fig. 5). We further obtained a short

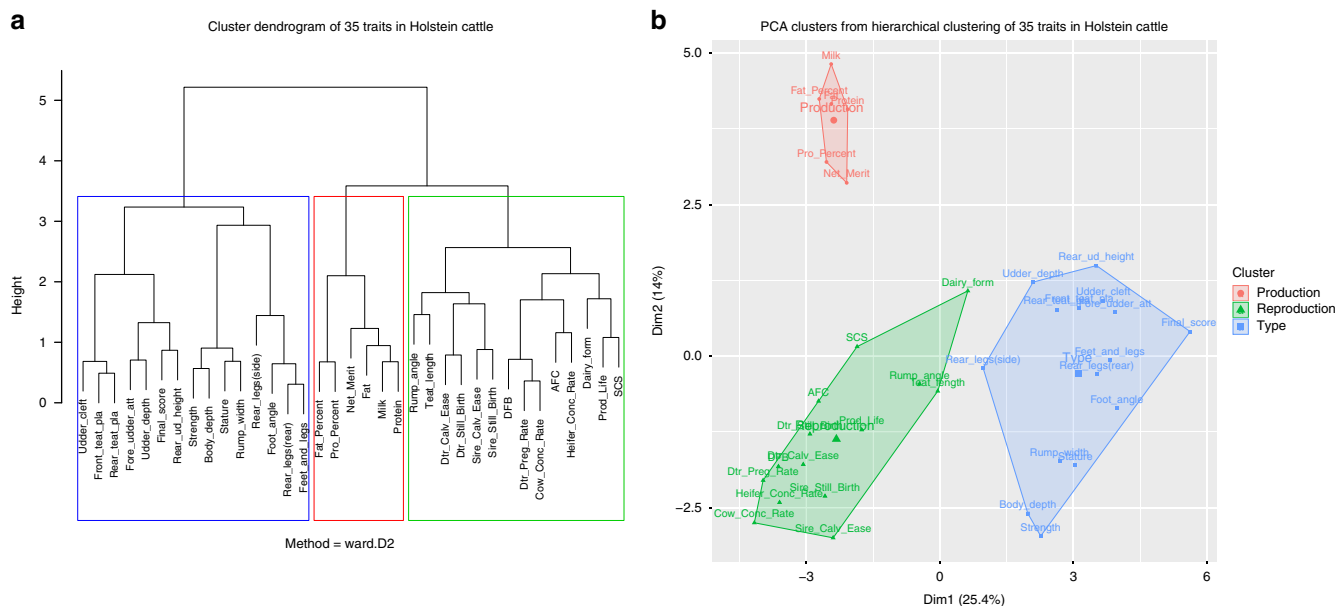


Fig. 2 Hierarchical clustering of 35 traits in Holstein cattle. **a** Cluster dendrogram. **b** PCA clusters

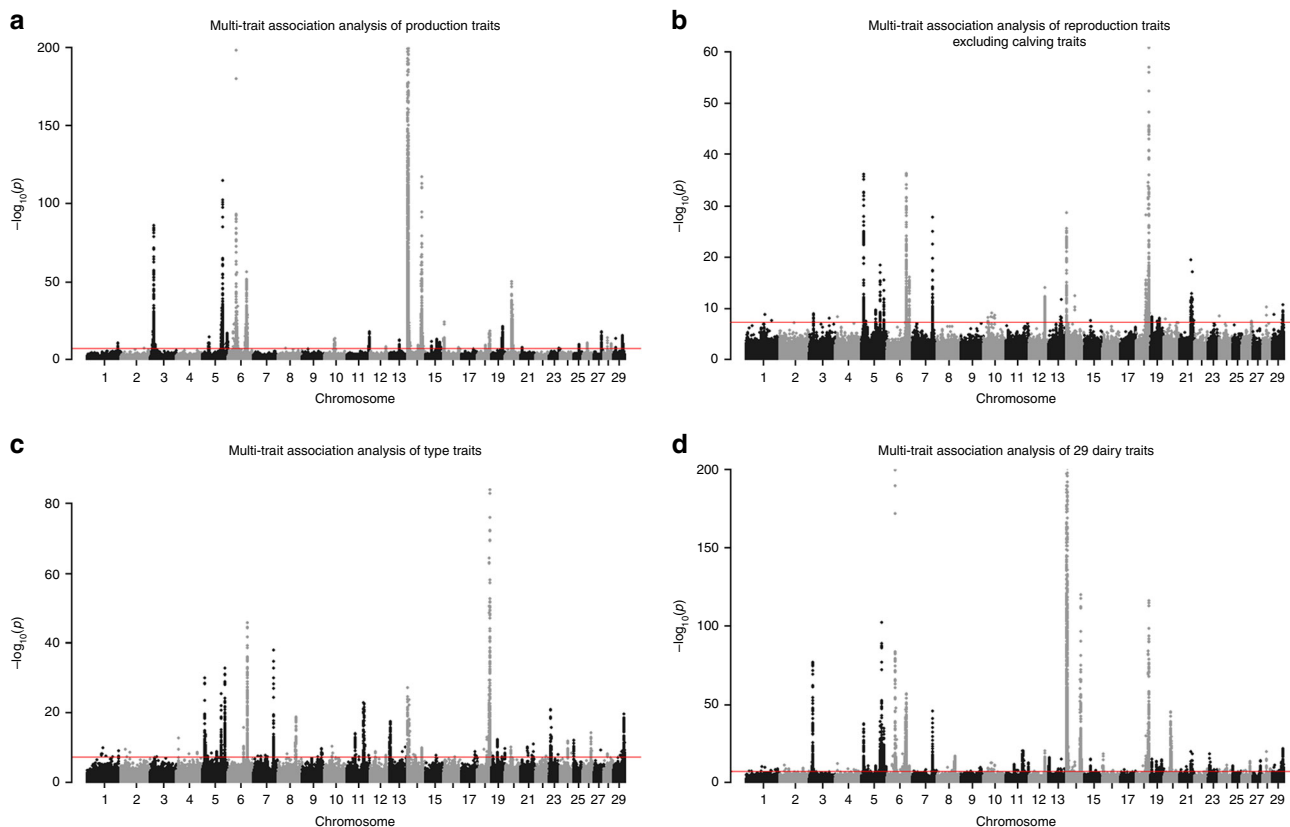


Fig. 3 Manhattan plots for multi-trait association analyses. **a** Production traits. **b** Reproduction traits, excluding four calving traits (calving ease and stillbirth traits). **c** Type traits. **d** All 29 dairy traits, excluding DFB, net merit, and four calving traits

list of the most promising candidates by applying conservative criteria: PPC >0.9 if a gene is associated with only one trait and PPC >0.5 for all traits if a gene affects multiple traits.

This short list had 69 unique genes including both previously reported genes and newly discovered ones for cattle traits (Table 2). For example, *ABCG2* and *DGAT1* are known to affect milk production in dairy cattle^{27,28}. The *ARRDC3* gene has been

associated with body confirmation traits and calving traits in beef and dairy cattle^{21,29,30}. Our fine-mapping study also revealed novel gene/association combinations for dairy traits. A previous study reported that the *ABCC9* gene was associated with fat yield, protein yield, and calving to first service interval in Holstein cattle³¹. In our study, we found a pleiotropic effect of this gene on body type traits (fore udder attachment and udder depth), milk

production (milk and protein yields), and daughter pregnancy rate, with a PPC of almost 1 for all the associated traits. In addition, we found that there were no common variants among the credible variant sets for these traits (Table 2), suggesting that *ABCC9* might have different causal mutations for the associated traits. *TMTC2* has been associated with teat length³⁰, and our fine-mapping showed that it had an effect on six type traits (including teat length, fore udder attachment, front teat placement, rear teat placement, rear udder height, and final score), with PPC being ≥ 0.95 for all those traits. Abo-Ismael et al. reported *CCND2* was associated with stature³⁰. Our fine-mapping results determined its association with four type traits (PPC > 0.95 for body depth, rump width, and stature). It is worth noting that our fine-mapping study not only discovered association of a

gene with a trait, but also provided the posterior probability of being causal for a gene.

Candidate variants. Because our stringent quality control filtering during and after imputation removed many variants (~20%, mostly intergenic with some genic), fine-mapping of the QTL regions to single-variant resolution could not always be achieved. Nevertheless, we obtained 95% credible variant set for each independent signal and merged them into one table. This resulted in a total of 1582 unique variants (Supplemental Data 13). We generated a short list of those variants with a moderate impact on protein coding and PPC > 0.2 (Table 3). Among the list, some variants have been previously reported, e.g., Chr6:38027010 in *ABCG2*²⁷ and Chr26:21144708 in *SCD*³². We also found other promising candidate variants, e.g., Chr7:93244933 in *ARRDC3* with an average PPC of 0.608 on 9 traits, Chr8:83581466 in *PTH1* with an average PPC of 0.68 on two type traits (body depth and strength), Chr1:69673871 in *KALRN* with an average PPC of 0.46 on two reproduction traits (cow conception rate and daughter pregnancy rate), Chr17:70276788 in *CHEK2* with an average PPC of 0.39 on two calving traits (sire calving ease and daughter calving ease).

Discussion

In this study, we performed GWAS for 35 production, reproduction, and type traits in dairy cattle with a uniquely large data set, and then fine-mapped the GWAS signals to single-gene resolution. With the fast computing method that we developed (BFMAP), we attempted to find causal effects in hundreds of loci each of which contained thousands of variants. We also investigated the functional enrichment patterns of several functional annotation data available in the cattle genome, and incorporated useful functional information into the final fine-mapping. In sum, we provided not only a credible candidate gene list for follow-up functional validation, but also a unique resource that can be easily employed by future functional enrichment studies.

In the single-trait GWAS, we found many association signals that have not been discovered (Fig. 1), clearly demonstrating the benefits of using large dairy cattle data for GWAS of complex

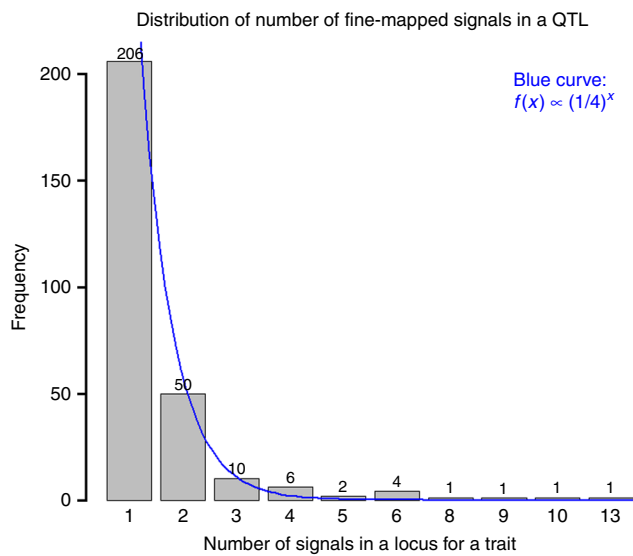


Fig. 4 Distribution of number of fine-mapped signals in a candidate locus for a trait. Signals were filtered by a significance threshold of $5E-7$

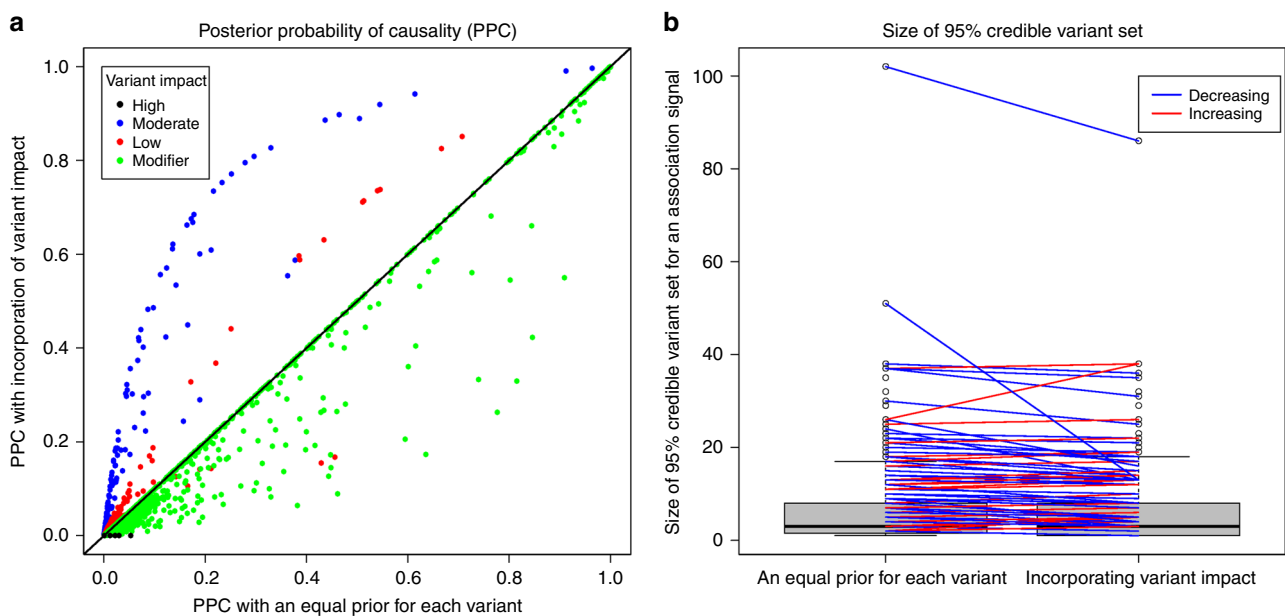


Fig. 5 Effect of incorporation of SnpEff-inferred impact on fine-mapping performance. **a** PPC with incorporation of SnpEff impacts versus PPC with an equal prior for each variant. **b** Size of 95% credible variant set generally decreased after incorporation of SnpEff-inferred impact

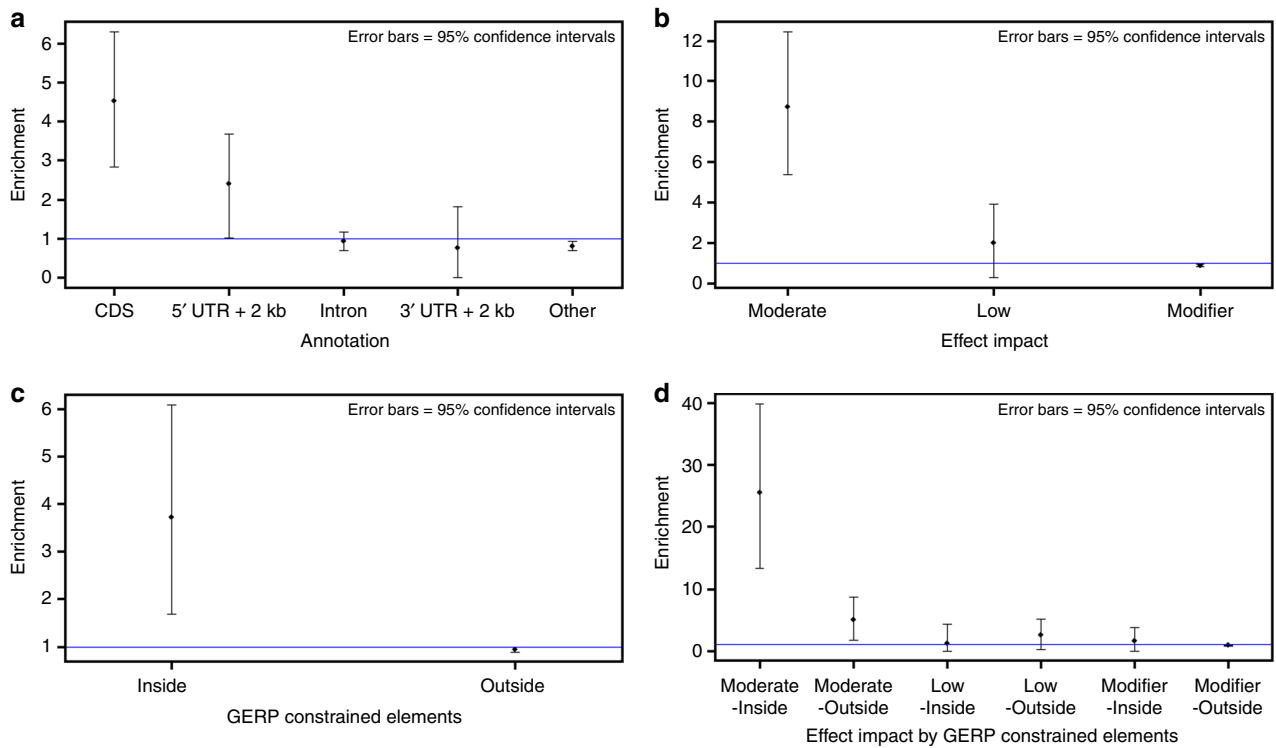


Fig. 6 Enrichment of fine-mapped variants across various functional annotations. **a** Locations of variants regarding protein-coding genes. **b** SnpEff predicted impact. **c** GERP-constrained elements. **d** SnpEff impact and GERP-constrained elements

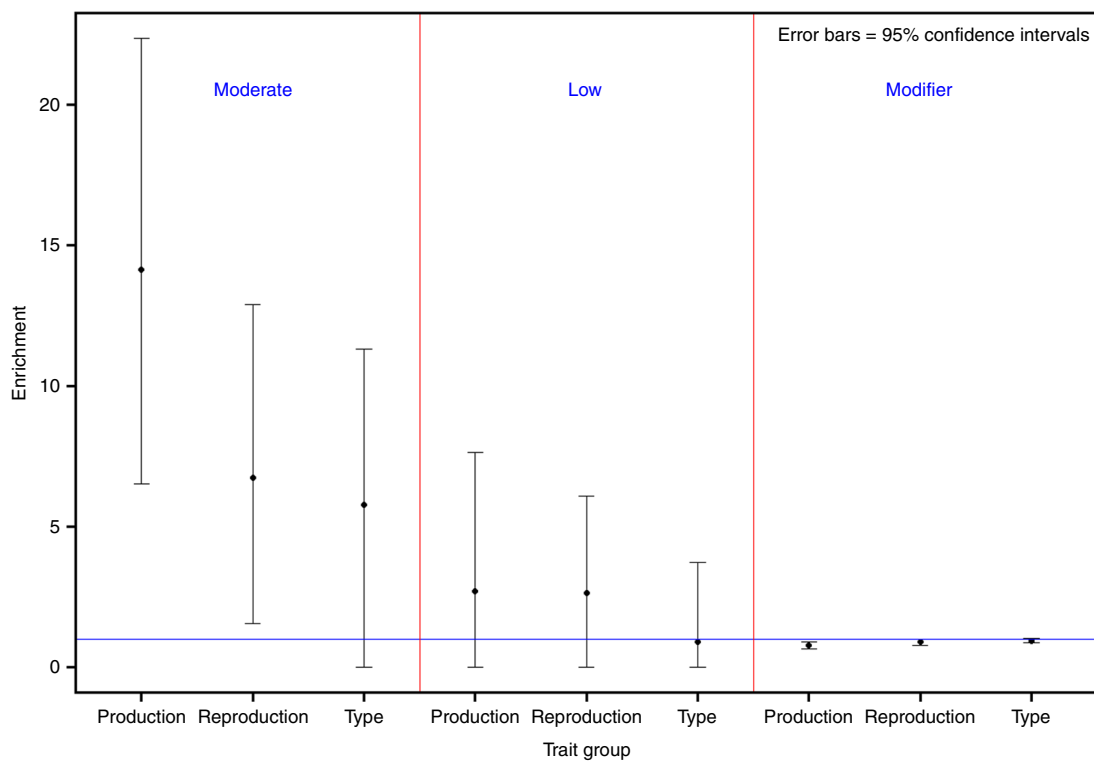


Fig. 7 Enrichment estimates for SnpEff predicted impact by three groups of traits

traits. Reliabilities of deregressed PTAs were modeled for most of the traits. For the traits with small variation of reliability, we observed similar results for the models with and without reliability; e.g., QTLs found when not modeling reliability were largely

the same as those by incorporating reliability for fat percentage and daughter pregnancy rate (Supplementary Fig. 6). Interestingly, we observed some deflations in the GWAS of production traits, which could be due to the large QTL effects on these traits

Table 2 Candidate genes with high posterior probability of causality

Gene	Traits	Gene PPC	Minimal p-values
ABCG2	Fat Fat_Percent Milk Net_Merit Pro_Percent Protein	0.85-1.00	1.1E-09-1.5E-221
TMTC2	Final_score Fore_udder_att Front_teat_pla Rear_teat_pla Rear_ud_height Teat_length	0.95-1.00	1.2E-09-4.9E-26
ARRDC3	Dtr_Calv_Ease Rear_ud_height Sire_Calv_Ease Strength Teat_length Udder_depth	0.56-0.91	8.4E-09-2.7E-15
ABCC9	Dairy_form Dtr_Preg_Rate Fore_udder_att Milk Protein Udder_depth	0.999-1.00	4.4E-07-2.6E-21
DGAT1	Milk Net_Merit Pro_Percent Protein SCS	0.99-1.00	1.5E-21-2.0E-260
VPS13B	Fat_Percent Milk Pro_Percent Rear_ud_height Udder_cleft	0.97-1.00	1.5E-07-1.5E-76
ZNF613	Body_depth Net_Merit Sire_Still_Birth Stature Strength	0.61-0.84	2.2E-14-8.9E-37
CCND2	Body_depth Rump_width Stature Strength	0.71-1.00	2.4E-19-4.5E-26
MGST1	Fat Fat_Percent Milk Pro_Percent	0.999-1.00	7.1E-21-2.4E-75
FGF6	Body_depth Rump_width Stature Strength	0.76-1.00	1.1E-07-3.9E-21
CCDC88C	DFB_PTA Dairy_form Rear_ud_height	0.89-1.00	2.7E-10-2.9E-22
LOC751788	Dairy_form Final_score	0.92 0.96	4.2E-09 1.4E-11
SCD	Fat Fat_Percent	1.00 1.00	9.7E-13 4.6E-10
MKL1	Milk Protein	1.00 1.00	2.0E-14 2.9E-10
SYT8	Final_score Foot_angle	1.00 0.998	1.6E-10 1.4E-09
LOC782261	Milk Net_Merit	0.92 0.61	6.5E-09 3.8E-10
CHEK2	Dtr_Calv_Ease Sire_Calv_Ease	0.65 0.67	1.9E-12 3.8E-07
C8H9orf3	Final_score Rump_width	1.00 0.63	1.2E-09 3.5E-09
GC	Cow_Conc_Rate Udder_depth	1.00 0.69	8.5E-08 1.8E-09
KALRN	Cow_Conc_Rate Dtr_Preg_Rate	0.54 0.92	1.4E-07 2.8E-08
CSN1S1	Pro_Percent Protein	0.999 1.00	1.2E-14 8.7E-14
SCAPER	Fore_udder_att Front_teat_pla	0.999 0.77	5.3E-08 3.3E-08
TCP11	Stature Udder_depth	1.00 0.97	5.9E-15 3.1E-08
PAEP	Fat_Percent Protein	0.996 0.84	2.0E-11 1.1E-07
ANKFN1	Rump_width SCS	0.98 0.87	1.5E-09 3.8E-07
NADSYN1	Dtr_Preg_Rate Stature	0.65 0.99	2.9E-08 3.3E-07
LOC100852273	Final_score Fore_udder_att	0.995 0.97	3.2E-09 8.4E-09
RAB6A	Milk Pro_Percent	0.79 0.72	2.1E-13 3.5E-13
LOC107132925	Fore_udder_att Udder_depth	0.75 0.999	1.6E-15 9.7E-18
POLD1	Foot_angle Protein	0.98 0.99	3.8E-12 4.7E-13
RAB11FIP2	Front_teat_pla Rear_teat_pla	0.83 0.64	2.5E-10 1.5E-07
MGMT	Rump_angle	1	4.15E-11
BOSTAUV1R417	Sire_Still_Birth	1	1.64E-16
SLCS0A1	Pro_Percent	1	2.48E-11
RNF217	Pro_Percent	1	2.29E-09
LOC104974054	Rump_angle	1	3.19E-15
HSD17B12	Fat_Percent	1	9.24E-10
LOC104975270	Fore_udder_att	1	3.69E-11
LOC104972568	Sire_Calv_Ease	1	4.13E-10
ADGRV1	Sire_Calv_Ease	1	4.93E-10
CD276	Dtr_Preg_Rate	1	3.86E-11
TTC28	Dtr_Calv_Ease	1	4.98E-10
LSP1	Udder_depth	1	2.95E-12
VEPH1	Udder_cleft	0.999	3.49E-07
TIGAR	Prod_Life	0.999	9.64E-17
CCDC57	Fat	0.999	2.12E-09
GON4L	Protein	0.998	1.45E-10
FASN	Fat_Percent	0.998	7.47E-10
COLEC12	Rump_angle	0.997	1.05E-08
C6	SCS	0.997	3.95E-08
MYH10	Udder_depth	0.996	1.71E-09
GPAT4	Fat_Percent	0.995	3.93E-11
EXOC6B	Teat_length	0.992	1.09E-09
ABO	Pro_Percent	0.988	4.37E-11
LOC619012	Sire_Still_Birth	0.988	5.09E-09
MRGPRG	Sire_Calv_Ease	0.987	3.63E-07
FSTL1	Stature	0.985	2.13E-08
SFTPD	Pro_Percent	0.985	3.92E-10
SLC24A2	Rump_angle	0.973	5.17E-09
ESR1	Dtr_Calv_Ease	0.971	2.29E-11
LDLR	SCS	0.965	2.85E-08
TBC1D22A	Pro_Percent	0.947	3.73E-14
PTCH1	Body_depth	0.941	7.46E-09
LOC101903327	Prod_Life	0.936	9.30E-06
FAM98B	Stature	0.93	5.08E-08
VWA2	Teat_length	0.929	7.82E-06
LOC786966	Pro_Percent	0.919	1.33E-08
MROH9	Rear_teat_pla	0.908	1.27E-08

including the *DGAT1* gene. Minor inflations were observed in GWAS for calving traits (i.e., calving ease and stillbirth) and final score (Supplementary Fig. 1). Although there were many sporadic variants passing the threshold of genome-wide significance ($P < 5E-8$), we could still locate a few credible GWAS peaks where there were a cluster of significant variants.

Initially, our fine-mapping discovered as many as 19 signals in a candidate region for a trait, as we applied a variant inclusion threshold accounting for only the effective number of independent variants (m_{eff}) at the locus-by-trait level. We also noticed that there were more locus-by-trait association pairs with multiple signals than with only one signal. By examining those with

Table 3 Missense variants with largest posterior probabilities of causality (>0.2)

Variant	Gene	MAF	Average PPC	Traits
7:93244933	<i>ARRDC3</i>	0.099	0.608	Body_depth Dtr_Calv_Ease Net_Merit Prod_Life Rear_ud_height Sire_Calv_Ease Strength Teat_length Udder_depth
6:38027010	<i>ABCG2</i>	0.015	0.87	Fat Fat_Percent Milk Net_Merit Pro_Percent Protein
8:83581466	<i>PTCH1</i>	0.027	0.678	Body_depth Strength
26:21144708	<i>SCD</i>	0.253	0.571	Fat Fat_Percent
1:69673871	<i>KALRN</i>	0.105	0.462	Cow_Conc_Rate Dtr_Preg_Rate
19:7521843	<i>ANKFN1</i>	0.217	0.446	Rump_width SCS
29:50290087	<i>SYT8</i>	0.388	0.438	Final_score Foot_angle
29:50286107	<i>TNNI2</i>	0.203	0.436	Rump_width Stature
29:50289940	<i>SYT8</i>	0.387	0.399	Final_score Foot_angle
17:70276788	<i>CHEK2</i>	0.088	0.388	Dtr_Calv_Ease Sire_Calv_Ease
18:57017616	<i>POLD1</i>	0.103	0.291	Foot_angle Protein
8:83044210	<i>FANCC</i>	0.116	0.252	Rear_teat_pla Udder_depth
14:1321450	<i>LOC782261</i>	0.207	0.206	Milk Net_Merit
14:2072259	<i>LOC786966</i>	0.090	0.919	Pro_Percent
18:44378414	<i>CHST8</i>	0.120	0.889	DFB_PTA
5:118244695	<i>TBC1D22A</i>	0.177	0.676	Pro_Percent
5:30259026	<i>NCKAP5L</i>	0.252	0.611	Teat_length
3:15464749	<i>GBA</i>	0.063	0.601	Milk
3:20189903	<i>ADAMTSL4</i>	0.075	0.571	Dairy_form
11:104232298	<i>ABO</i>	0.309	0.449	Pro_Percent
19:51319797	<i>CCDC57</i>	0.350	0.423	Fat
18:61020273	<i>ZNF331</i>	0.038	0.322	Dairy_form
19:51319759	<i>CCDC57</i>	0.350	0.304	Fat
8:85147150	<i>LOC101906801</i>	0.117	0.302	Strength
13:58716308	<i>C13H20orf85</i>	0.116	0.297	Fore_udder_att
11:104232319	<i>ABO</i>	0.309	0.223	Pro_Percent
14:66328304	<i>SPAG1</i>	0.119	0.222	SCS

multiple signals, we found the models often contained a strong signal and several much weaker ones. Those weak signals might result from imperfect model fitting of the lead variants in other signals, instead of being true positives. Nevertheless, filtering out these weak signals with genome-wide significance levels did little harm to the discovery of strong ones.

The enrichment results for SnpEff-inferred variant impact in our study were very different from those reported in human studies²⁵. The differences among the four categories in the human study are more distinctive than ours. This is consistent with our anticipation that high LD in cattle genome makes such enrichment difficult to detect. In addition, high-impact variants generally have a lower frequency than other variants and are thus harder to impute in cattle where the number of reference sequences is small and the original genotype data are of moderate density. Nevertheless, we found a considerable enrichment of candidate causal effects in moderate-impact variants. Incorporation of this enrichment into fine-mapping facilitated the discovery of more candidate causal variants (Fig. 5). The discovery of biologically meaningful enrichment patterns will be valuable for the development of new methods to incorporate functional information into fine-mapping and genomic prediction.

Different functional annotations are often related, so we analyzed the enrichment of each functional annotation separately. Although single-annotation analysis does not resolve confounding of multiple annotations, the enrichment estimates can still provide informative priors for fine-mapping. We analyzed various functional annotations by single-annotation enrichment analysis and determined the ones that provide highly differential priors. LDSC may be able to dissect heritability enrichment between multiple functional annotations, and BFMAP can incorporate these outputs into fine-mapping simultaneously.

It is widely acknowledged that population structure and relatedness need to be properly accounted for in GWAS via a linear mixed model³³ or a linear model with principal components extracted from genomic relationship matrix³⁴. Similarly, we need to account for population structure and relatedness in fine-mapping analyses as proposed in our BFMAP. However, existing fine-mapping approaches have not fully addressed this issue. For instance, BIMBAM models only intercept and SNPs of interest, and thus only works for independent samples³⁵. BayesFM can include principal components as covariates, but it does not have a random component to fully account for relatedness³⁶. piMASS applies Bayesian variable selection regression to modeling genome-wide variants to control for population structure and relatedness, but it uses a Markov chain Monte Carlo (MCMC) algorithm that is computationally impractical for large studies³⁷. CAVIARBF, PAINTOR, and FINEMAP³⁸, use summary test statistics and are approximately equivalent to BIMBAM. In theory, these methods work for independent samples by using summary statistics from linear model analyses and genotype correlations between variants. Further studies are warranted to investigate how these summary-statistics methods perform for structured or related samples when using summary statistics from linear mixed models.

Using BFMAP, we pinpointed some promising candidate genes for economically important traits in dairy cattle. It is promising to validate those genes with high posterior probability of causality (Table 2) in future functional studies. In addition, with our new method of functional enrichment analysis in BFMAP, our fine-mapping result of hundreds of QTLs (Supplemental Data 6) can be readily used to estimate enrichments of causal effects for additional functional annotation data. Thus, we provided an easy-to-use enrichment analysis resource to test the functional

annotations that are being generated by the on-going FAANG and related projects for cattle¹⁸.

Methods

Genotype and phenotype data. Genotype data have been described in more details previously⁵. Here we provide a brief summary. SNP and insertion-deletion (InDel) calls (sequence variants) from Run 5 of the 1000 Bull Genomes Project⁶ were released in July 2015. After stringent quality control edits and removal of intergenic and intronic SNPs, 3,148,506 sequence variants were retained for 444 Holstein bulls. The sequence variants and high-density SNP genotypes of 312K markers for 26,949 progeny-tested Holstein bulls (and 21 Holstein cows) were combined by imputation using the FindHap software (version 3)³⁹. Finally, we had genotypes of 3,148,506 sequence variants for 27,214 Holstein bulls (179 bulls had both sequence and high-density genotypes) and 21 cows. Imputation quality from FindHap was assessed with 404 of the sequenced animals as the reference population and 40 randomly selected animals for validation. The sequence genotypes of the validation animals were reduced to high-density SNP genotypes and then imputed back to sequence variants. The average imputation accuracy was 96.7% for the 3,148,506 variants⁵. After excluding high-density SNPs, we found an average accuracy of 96.4% for the newly imputed sequence variants. Chromosome-specific imputation accuracy was >95% for all autosomes except Chromosome 12.

All of the 27,214 Holstein bulls used in this study had highly reliable (average reliability >71% across traits) PTAs for 35 production, reproduction, and type traits (Table 1). Transmitting ability is basically the additive genetic values of cattle. Reliability quantifies the amount of information available in a PTA and measures its accuracy⁴⁰. Deregressed PTAs were used as phenotype in all our analyses, which excludes parent information and reduces the dependence among animals⁴¹. Because each of the bulls had many phenotyped daughters, their PTAs were generally of high reliability, even for low-heritability traits (Table 1). The trait definitions are shown in Table 1 and Supplementary Note 1. We categorized the 35 traits into three groups, i.e., production, reproduction, and body type, based on a clustering analysis.

Single-trait GWAS. The software MMAP¹⁹ was used for all single-trait GWAS analyses (<https://mmap.github.io/>). Basically, MMAP efficiently implements a mixed-model approach for association tests that is similar to GEMMA⁴² but different from EMMAX⁴³ that is, variance component is estimated uniquely for each marker. We used the following model

$$y = \mu + Xb + g + e \text{ with } g \sim N(0, \sigma_g^2 G) \text{ and } e \sim N(0, \sigma_e^2 R), \quad (1)$$

where y is deregressed PTAs, μ is global mean, X is genotype of a candidate variant (coded as 0, 1 or 2) and b is its effect, g is a polygenic effect accounting for population structure, and e is residual. The genomic relationship matrix (G)⁴⁴ was built using ~312K high-density SNP markers (filtered by MAF >1%). R is a diagonal matrix ($R_{ii} = 1/r^2 - 1$), which is used to model differential reliability among animals.

We disregarded variants on the X chromosome. We also filtered out variants with an MAF of <1% or failing Hardy-Weinberg Equilibrium test ($p < 1E-6$). After QC, there were ~2.7 million variants to be tested for association. We used a genome-wide significance level of $P < 5E-8$. QTLs were located by finding GWAS peaks where there were a cluster of significant variants. We used a custom Perl script to find all GWAS peaks and further examined each of the peaks based on the Manhattan plots to filter out suspicious ones (i.e., sporadic significant variants). Subsequently, we determined a total of 286 QTLs (Supplemental Data 2) that were further analyzed in fine-mapping studies.

To find which ones are novel among the 286 QTLs, we compared our result with Cattle QTLdb (release 35 published on April 29, 2018) that contains 113,256 QTLs/associations from 848 publications²⁰. To ensure correct physical positions of QTLs on UMD 3.1, we first extracted the rs identifiers (rs#) of flanking SNPs for each term from the Cattle QTLdb data, and then used the identifiers to find flanking SNPs' positions on UMD 3.1 in the Ensembl genome variation database. These SNP positions were used as QTL positions. This procedure can rule out QTL terms whose physical positions are inaccurately converted from genetic maps. The Cattle QTLdb release 35 covers 599 different traits, in which we extracted those with the (almost) same definition as the 35 traits in our study (Supplemental Data 14). For each of the QTLs that we detected, we determined that it is either previously reported if it is within ± 500 kb of any QTL/association for the same trait(s) in the Cattle QTLdb or is newly discovered otherwise (Supplemental Data 2).

Multi-trait association analysis. Following a previous study²¹, our multi-trait association tests were based on a chi-square statistic with multiple degrees of freedom. For each variant, the chi-square statistic for the multi-trait association test was calculated by:

$$\text{Multi-trait } \chi_{df=n}^2 = \mathbf{t}_i' \mathbf{V}^{-1} \mathbf{t}_i,$$

where \mathbf{t}_i is a $n \times 1$ vector of the signed t -values of variant i for n traits, and \mathbf{V} is an $n \times n$ correlation matrix for the n traits which is calculated using signed t -values of

genome-wide variants. In our analysis, the signed t -values were obtained from single-trait GWAS for 2,619,418 variants passing QC, and the correlations between traits were calculated using all the variants. To test the robustness of the estimated correlation using all sequence variants⁴⁵, we also computed the correlation matrix using two variant subsets obtained by selecting every 10th and every 100th variant. The three variant sets produced similar correlation estimations (Supplementary Fig. 7).

We performed hierarchical clustering based on the absolute correlation coefficients, and then did multi-trait association analysis for each of the three resulting clusters of traits as shown in Fig. 2. Specifically, we excluded net merit and days to first breeding (DFB) in production and reproduction clusters, respectively, because these traits are linear combinations of other traits and the number of bulls for DFB was much smaller compared to other traits. We also excluded the four calving traits to avoid sporadic significant variants. Additionally, all the traits except for the six traits aforementioned were analyzed as a whole in a separate multi-trait association test.

Bayesian Fine-MAPping (BFMAP). We developed the following Bayesian model for fine-mapping:

$$\begin{aligned} y &= Xb + Za + g + e \\ b &\sim N(0, \varphi \sigma_e^2 I) \\ a &\sim N(0, \gamma \sigma_e^2 I) \\ g &\sim N(0, \eta \sigma_e^2 I) \\ e &\sim N(0, \sigma_e^2 R) \\ P(\sigma_e^2) &\propto 1/\sigma_e^2 \end{aligned} \quad (2)$$

where y is a phenotype vector of size n for a complex trait, b is a vector of covariate (other than genomic variants) effects and X is corresponding design matrix, a is a vector of variant effects and Z is corresponding genotype coding matrix (e.g., genotype coding for additive, dominance, or imprinting effects⁴⁶), g is a vector of polygenic effect for controlling population structure, G is a corresponding variance structure matrix (e.g., genomic relationship matrix), and e is the residual with variance structure R for modeling reliability or accuracy of phenotypic records as in model (1). The common variance component (σ_e^2) is given by a non-informative Jeffrey's prior. Other variance parameters (φ, γ , and η) are treated as known. Generally, we can set φ to a large value (e.g., 1E8) to make a act like fixed effects. A genomic variant is usually considered to have a small but noticeable effect, so we can set γ at 0.01 or 0.04^{47,48}. When Za only accounts for a tiny proportion of phenotypic variance (this is true when modeling variants from a small genomic region), we can set η based on the heritability (h^2) by $\eta = h^2/(1-h^2)$. In practice, we can instead use heritability estimate (\hat{h}^2) in the null model without variants to determine η .

We can easily compute $P(D|M)$ (data D , and model M regarding variant inclusion) by integrating out σ_e^2 based on model (2). To allow easy calculation, we use a linear transformation to the model (Supplemental Note 2). We can further obtain the null distribution of Bayes factors ($H_0: a = 0$) in model (2) by an extension of the results by Zhou and Guan⁴⁸ (Supplemental Note 3). Based on the null distribution, scaled Bayes factor⁴⁸ and corresponding p -value can be computed for our model.

We seek to identify independent association signals within a QTL region and to assign a posterior probability of causality (PPC) to each variant with fine-mapping. Following the method by Huang et al.¹⁰, our fine-mapping approach includes three steps: forward selection⁴⁹ to add independent signals in the model, repositioning signals, and generating credible variant set for each signal. Although our approach uses the same framework as Huang et al.¹⁰, there are a few notable differences (Supplemental Data 15). While they only provided some R scripts for disease data, we provide a fast, general-purpose software tool for fine-mapping analysis of complex traits.

We set $\varphi = \gamma = 1E8$ in model (2) for fine-mapping, which enables easy calculation of p -value for a newly added variant conditional on variants already in model (Supplemental Note 4). We use a Bonferroni-corrected threshold⁴⁹ as stopping criterion in forward selection; that is, forward selection stops when $(2\log BF + 1) < 2\log m_{\text{eff}}$, where m_{eff} is the effective number of independent variants calculated using the method by Li and Ji⁵⁰. Suppose that we select p independent signals in forward selection and determine a set of lead variants (S_i) for the p signals after repositioning. Then, for signal i with lead variant (l_i), we have a variant set (S_i) containing variants that have substantial LD with l_i but weak LD with lead variants in other signals $S_j / \{ l_j \}$. Accordingly, we can compute PPC of variant j (v_{ij}) in S_i conditioning on $S_i / \{ l_i \}$:

$$P(M_i = v_{ij} | y, X, Z, S_i \setminus \{ l_i \}) = \frac{P(y | X, Z, M_i = v_{ij}, S_i \setminus \{ l_i \}) P(M_i = v_{ij})}{\sum_j P(y | X, Z, M_i = v_{ij}, S_i \setminus \{ l_i \}) P(M_i = v_{ij})} \quad (3)$$

Where $M_i = v_{ij}$ denotes that the causal variant in signal i is variant j in S_i (i.e. v_{ij}). Weóó can easily get a credible variant set passing a given confidence level (e.g., 95%) for a signal, by sorting variants in a descending order of PPC and including

them in the set from top to bottom. We can also calculate PPC of a gene by summing up PPCs of all variants within the gene.

In the study by Huang, et al.¹⁰, an equal prior for each variant was used; that is, $P(M_i = v_{ij}) = 1 \forall v_{ij} \in S_i$. Here, we propose a method to apply differential prior probabilities by integrating functional annotation, following a previous study on adjusting significance threshold based on functional annotation in GWAS²⁵. With our fine-mapping procedure, it is usually safe to assume one and only one causal variant in each independent signal. For a functional annotation with several categories, we denote the probability of a causal variant being in category C as p_C and the probability of a non-causal variant being of category C as q_C . We can accordingly obtain:

$$P(M_i = v_{ij}) = P(c_{ij} | M_i = v_{ij}) \prod_{j \neq i} P(c_{ij} | M_i \neq v_{ij}) = p_{c_{ij}} \prod_{j \neq i} q_{c_{ij}} \quad (4)$$

where c_{ij} denotes the category of variant j in S_i (i.e. v_{ij}).

We estimate q_C with the genome-wide frequencies of the categories²⁵. To estimate p_C , we can use all available independent signals (M_i):

$$L(\{p_C\} | y, Z) \propto \prod_i P(M_i, y, Z | \{p_C, q_C\}) \\ \propto \prod_i \prod_j P(y | X, Z, M_i = v_{ij}) P(M_i = v_{ij} | \{p_C, q_C\}) \quad (5)$$

When the signals identified in fine-mapping are independent of each other, we can get:

$$P(y | X, Z, M_i = v_{ij}) \approx P(y | X, Z, M_i = v_{ij}, S_i \setminus \{i\}) \quad (6)$$

Taking Eqs. 4 and 6 into Eq. 5, we obtain a likelihood function regarding $\{p_C\}$ and then get their maximum likelihood estimates (MLEs), $\{\hat{p}_C\}$. By taking the estimates of $\{p_C, q_C\}$ and Eq. 4 to Eq. 3, we get updated PPCs with incorporation of function annotation, which is actually an empirical Bayes approach.

When setting an equal prior for each variant, we find:

$$P(M_i = v_{ij} | y, X, Z, S_i \setminus \{i\}) \propto P(y | X, Z, M_i = v_{ij}, S_i \setminus \{i\}) \quad (7)$$

Thus, to estimate $\{p_C\}$ by Eq. 5, we can use PPCs from the computation assuming an equal prior for each variant. Accordingly, incorporation of functional annotation includes three steps: computing PPCs given an equal prior for each variant, estimating $\{q_C\}$ with the genome-wide frequencies of the categories and estimating $\{p_C\}$ with these PPCs, and updating PPCs with $\{p_C, q_C\}$. These features make our approach easier to use compared with PAINTOR¹³ and CAVIARBF¹².

Fine-mapping of dairy cattle traits. Genomic regions for fine-mapping were determined by lead variants in single-trait and multi-trait GWAS results. We first determined a minimal region that covered each lead variants (either in single- or multi-trait QTLs), and then extended it 1 Mb upstream and downstream, resulting in a ≥ 2 Mb candidate region for fine-mapping. The 1-Mb extension allowed the region to cover most variants that have an LD r^2 of >0.3 with the lead variants²⁶.

We obtained a total of 125 loci from single- and multi-trait GWAS results (Supplemental Data 16). Three loci without enough high-density SNPs were removed to ensure imputation quality, thus leaving 122 loci for fine-mapping. A total of 57 loci were associated with more than one trait. Fine-mapping was performed for individual traits, and these 122 loci represented 282 locus-by-trait pairs for 32 traits (three leg traits were excluded for lack of significance). When fine-mapping identified multiple signals in a candidate locus, we kept the strongest one and filtered the rest. The effective number of independent tests was 54,403 for the 282 locus-by-trait pairs (Supplemental Data 17). Considering that our effective number estimates were already conservative⁵¹, we used $5E-7$ ($<0.05/54,403$) as the significance threshold. Subsequently, we found 434 association signals (Supplemental Data 5).

We found that the locus-by-trait association pairs with more than three signals identified were mostly from still birth and final score (Supplemental Data 5). We also noticed slight inflation of GWAS results of these two traits (Supplementary Fig. 1). Therefore, we removed the 16 QTLs with >3 fine-mapped signals from all following analyses. We further removed 15 signals whose variant set had ≤ 10 variants of distinct genotypes, as a small cluster of highly linked variants could indicate inaccurate imputation. Additionally, if there were multiple QTL on a chromosome for a trait, all lead variants in these loci were modeled jointly in fine-mapping. Accordingly, 13 association signals whose lead variant had a p -value $> 5E-7$ were removed. After all these edits, we determined a total of 308 association signals (Supplemental Data 6).

Besides assuming an equal prior for each variant, we further applied differential prior probabilities based on SnpEff-inferred impacts²³. Since using Eq. 5 requires independent association signals, we removed all the association signals for protein, cow conception rate, rear teat placement, udder depth and strength, because they have high correlation ($r^2 > 0.5$) with other traits. We also removed another six association signals, since these signals have a substantial LD with another signal (measured by LD r^2 between lead variants >0.25). These edits reduced the number of association signals from 308 to 249. We estimated $\{p_C, q_C\}$ for variant impact categories based on the 249 association signals, and updated PPCs for all 308 signals by integrating the estimated functional enrichment. Effect impact-incorporated PPCs were used for determining candidate variants or genes. When

computing PPC of a gene, all variants within its 2-kb upstream and downstream ranges were included.

Enrichment analysis in BFMAP. Our enrichment analysis was based on our 249 fine-mapped association signals to estimate p_C (the probability of a causal variant being in category C) and q_C (the probability of a non-causal variant being in category C). The two probabilities can be estimated using the models described in BFMAP. The enrichment for category C is defined as $E_C = p_C/q_C$ ²⁵, for which a value larger than one indicates that candidate causal variants are more enriched in category C than across the whole genome. Functional annotations investigated included locations of variants regarding protein-coding genes, effect impact inferred by SnpEff²³, and constrained elements predicted by GERP²⁴. Confidence intervals of the enrichment estimates were derived by percentile bootstrap as in ref. ²⁵. The association signals were resampled 1,000 times to calculate the confidence intervals. We removed very small categories (like HIGH in SnpEff-inferred effect impacts) in bootstrapping to avoid non-convergence of the maximum likelihood estimation.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The GWAS summary statistics of all 36 dairy traits have been made publically available through Figshare (<https://figshare.com/s/ea726fa95a5bac158ac1>). The reference sequence data have been described and published previously by the 1000 bull genome project, and the NCBI Sequence Read Archive accession codes are, SRP039339, SRR1293227, SRR1262614-SRR1262659, SRR1188706, SRR1262533, SRR1262536, SRR1262538, SRR1262539, SRR1262660-SRR1262788 and SRR1262789-SRR1262846. The original genotype data are owned by third parties and maintained by the Council on Dairy Cattle Breeding (CDCB). A request to CDCB is necessary for getting data access on research, which may be sent to: João Dürr, CDCB Chief Executive Officer (joao.durr@cdcb.us). All other relevant data are available in the manuscript, Supporting Information files, and from the corresponding author upon request.

Code availability

BFMAP: <https://jiang18.github.io/bfmap/>

MMAP: <https://mmap.github.io/>

Cattle constrained elements: ftp://ftp.ensembl.org/pub/release-90/bed/ensembl-compara/68_eutherian_mammals_gerp_constrained_elements/gerp_constrained_elements_bos_taurus.bed.gz

Cattle genome annotation: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000003055.6_Bos_taurus_UMD_3.1.1/GCF_000003055.6_Bos_taurus_UMD_3.1.1_genomic.gff.gz

Cattle QTLdb: <https://www.animalgenome.org/cgi-bin/QTLdb/BT/index>

Cattle genome variation: ftp://ftp.ensembl.org/pub/release-89/variation/gvf/bos_taurus/

Received: 27 September 2018 Accepted: 29 April 2019

Published online: 18 June 2019

References

- Wiggans, G., Misztal, I. & Van Vleck, L. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. *J. Dairy Sci.* **71**, 54–69 (1988).
- VanRaden, P. Invited review: selection on net merit to improve lifetime profit. *J. Dairy Sci.* **87**, 3125–3131 (2004).
- Brotherstone, S. & Goddard, M. Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philos Trans. R. Soc. Lond. B Biol. Sci.* **360**, 1479–1488 (2005).
- van Binsbergen, R. et al. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* **46**, 41 (2014).
- VanRaden, P. M., Tooker, M. E., O’Connell, J. R., Cole, J. B. & Bickhart, D. M. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* **49**, 32 (2017).
- Daetwyler, H. D. et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* **46**, 858–865 (2014).
- Pausch, H. et al. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. *Genet. Sel. Evol.* **49**, 24 (2017).
- Hayes, B. J. & Daetwyler, H. D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu. Rev. Anim. Biosci.* **7**, 89–102 (2018).
- Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337 (2015).
- Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

11. Kim, E. S. & Kirkpatrick, B. W. Linkage disequilibrium in the North American Holstein population. *Anim. Genet.* **40**, 279–288 (2009).
12. Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J. Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics* **204**, 933–958 (2016).
13. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
14. Van Binsbergen, R. et al. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* **47**, 71 (2015).
15. MacLeod, I. et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genom.* **17**, 144 (2016).
16. Veerkamp, R. F., Bouwman, A. C., Schrooten, C. & Calus, M. P. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. *Genet. Sel. Evol.* **48**, 95 (2016).
17. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).
18. Andersson, L. et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
19. O’Connell, J. R. in 63th Annual Meeting of The American Society of Human Genetics.
20. Hu, Z. L., Park, C. A. & Reecy, J. M. Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res.* **44**, D827–D833 (2016).
21. Bolormaa, S. et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet.* **10**, e1004198 (2014).
22. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229 (2018).
23. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly. (Austin)* **6**, 80–92 (2012).
24. Cooper, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901–913 (2005).
25. Sveinbjornsson, G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
26. Bohmanova, J., Sargolzaei, M. & Schenkel, F. S. Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genom.* **11**, 421 (2010).
27. Cohen-Zinder, M. et al. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Res* **15**, 936–944 (2005).
28. Grisart, B. et al. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl Acad. Sci. USA* **101**, 2398–2403 (2004).
29. Saatchi, M., Schnabel, R. D., Taylor, J. F. & Garrick, D. J. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC Genom.* **15**, 442 (2014).
30. Abo-Ismael, M. K. et al. Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. *Genet. Sel. Evol.* **49**, 82 (2017).
31. Nayeri, S. et al. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet* **17**, 75 (2016).
32. Pegolo, S. et al. Effects of candidate gene polymorphisms on the detailed fatty acids profile determined by gas chromatography in bovine milk. *J. Dairy Sci.* **99**, 4558–4573 (2016).
33. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203 (2006).
34. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904 (2006).
35. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).
36. Fang, M. & Georges, M. BayesFM: a software program to fine-map multiple causative variants in GWAS identified risk loci. *bioRxiv*, 067801, <https://doi.org/10.1101/067801> (2016).
37. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).
38. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
39. VanRaden, P. M. findhap.f90: Find haplotypes and impute genotypes using multiple chip sets and sequence data, <https://aipl.arsusda.gov/software/findhap> (2016).
40. VanRaden, P. M. & Wiggans, G. R. Derivation, calculation, and use of national animal model information. *J. Dairy Sci.* **74**, 2737–2746 (1991).
41. Garrick, D. J., Taylor, J. F. & Fernando, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **41**, 55 (2009).
42. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
43. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
44. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
45. Zhu, X. et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96**, 21–36 (2015).
46. Jiang, J. et al. Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genom.* **18**, 425 (2017).
47. Chen, W. et al. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719–736 (2015).
48. Zhou, Q. & Guan, Y. On the null distribution of Bayes factors in linear regression. *J. Am. Stat. Assoc.* **113**, 1362–1371 (2017).
49. Foster, D. P. & George, E. I. The risk inflation criterion for multiple-regression. *Ann. Stat.* **22**, 1947–1975 (1994).
50. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Hered. (Edinb.)* **95**, 221–227 (2005).
51. Hendricks, A. E., Dupuis, J., Logue, M. W., Myers, R. H. & Lunetta, K. L. Correction for multiple testing in a gene region. *Eur. J. Hum. Genet.* **22**, 414–418 (2014).

Acknowledgements

We thank the Council on Dairy Cattle Breeding (CDCB) for the access to the genotype data. We thank the 1000 Bull Genomes Project for providing genome references for sequence imputation. This work was supported in part by AFRI grant number 2016-67015-24886 and 2018-67015-28128 from the USDA National Institute of Food and Agriculture (NIFA) and BARD grant number US-4997-17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund. JBC and PMV were also supported by appropriated projects 1265-31000-096-00, Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information, and 8042-31000-104-00, Enhancing Genetic Merit of Ruminants Through Genome Selection and Analysis, of the Agricultural Research Service of the United States Department of Agriculture. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The USDA is an equal opportunity provider and employer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

J.J., Y.D., and L.M. conceived and designed the experiments. P.M.V. and J.B.C. provided genotype and phenotype data, J.J. and E.F. performed computational and statistical analyses. J.J., E.F., Y.D., and L.M. wrote the paper. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s42003-019-0454-y>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019