Biology Direct

RESEARCH

Open Access

Integrating multi-layered biological priors to improve genomic prediction accuracy in beef cattle

Zhida Zhao[1†], Qunhao Niu[1†], Jiayuan Wu[1], Tianyi Wu[1], Xueyuan Xie[1], Zezhao Wang[1], Lupei Zhang[1], Huijiang Gao[1], Xue Gao[1], Lingyang Xu[1*], Bo Zhu[1,2*] and Junya Li[1*]

Abstract

Background Integrating multi-layered information can enhance the accuracy of genomic prediction for complex traits. However, the improvement and application of effective strategies for genomic prediction (GP) using multi-omics data remains challenging.

Methods We generated 11 feature sets for sequencing variants from genomics, transcriptomics, metabolomics, and epigenetics data in beef cattle, then we assessed the contribution of functional variants using genomic restricted maximum likelihood (GREML). We next estimated and ranked variant scores for 43 economically important traits, and compared the prediction accuracy of the top and bottom sets using genomic best linear unbiased prediction (GBLUP) and BayesB model. In addition, we annotated the variants from GWAS with functional feature sets and performed enrichment analysis.

Results We observed significant enrichments for 32 functional categories in 11 feature sets. The evolutionary related sets (conservation regions and selection signatures) contributed significantly to heritability (31.78-fold and 14.48-fold enrichment), while metabolomics and transcriptomics showed low heritability enrichments. We observed a significant increase in prediction accuracy using the top feature set variants compared to whole-genome sequencing (WGS) data. The prediction accuracy based on the top 10% variant set showed an average increase of 11.6% and 7.54% using BayesB and GBLUP across traits, respectively. Notably, the greatest increase of 31.52% was obtained for spleen weight (SW) using BayesB. Also, we found that the top 10% of variants show strong enrichment with weight related QTLs based on the Cattle QTL database.

Conclusions Our findings suggest that integrating biological prior information from multiple layers can enhance our understanding of the genetic architecture underlying complex traits and further improve genomic prediction in beef cattle.

†Zhida Zhao and Qunhao Niu have contributed equally to this work.

*Correspondence:
Lingyang Xu
xulingyang@caas.cn
Bo Zhu
zhubo@caas.cn
Junya Li
lijunya@caas.cn
Full list of author information is available at the end of the article

Background

Genomic prediction (GP), an effective approach for enhancing selection and promoting breeding efficiency [1, 2], has been widely applied in the fields of plant and animal breeding [3, 4]. Many parametric and nonparametric statistical methods have been proposed to predict Genomic Estimated Breeding Values (GEBVs) [5]. GBLUP constructs a genetic relationship matrix

to facilitate the estimation of GEBVs [6]. The Bayesian alphabet assumes a priori that the variances of effects for many single nucleotide polymorphisms (SNPs) are zero, while the effect of SNPs follow Student's t distribution [7]. Such models formulate distinct hypotheses regarding the distribution of marker effects and their impact on genetic variation [8, 9]. GP has been primarily applied based on SNP arrays [10, 11]. With the cost of whole-genome sequencing (WGS) decreasing, the application of WGS for GP has been widely applied in farm animals [12, 13]. GP using WGS data may promote prediction accuracy because it covers more SNPs across the genome than SNP arrays [14, 15]. A previous simulation study showed the superiority of BayesB over GBLUP using WGS, while the accuracy of GEBV increased when compared to SNP array [16, 17]. GP based on WGS data, was more accurate when using BayesB than using GBLUP [18].

The use of WGS data for GP may be limited by the substantial presence of linkage disequilibrium (LD) and diverse genomic functional regions. This scenario reduces the signal-to-noise ratio when employing WGS data directly for GP without a biological prior [19–21], thus many studies have been conducted to incorporate genomic information into statistical models by controlling for LD and annotating variants based on different functional classes [22, 23]. Moreover, the availability of multi-omics information (e.g., genomics, transcriptomics, proteomics and metabolomics) bridges a vital link between genotypes and phenotypes which provides a biological prior for genomic prediction. The Genotype-Tissue Expression (GTEx) Project was initiated with the objective of collating genetic influences on gene expression in human tissues and facilitating an enhanced understanding of the dynamics of regulatory genetic variation by elucidating the molecular mechanisms underlying genetic correlations with complex diseases and traits [24]. Similarly, the Farm Animal Genotype-Tissue Expression (FarmGTEx) Project serves as an extensive public repository, facilitating the discovery of tissue-specific genetic regulatory variants and the prediction of molecular phenotypes in farm animals [25, 26]. Additionally, the Functional Annotation of Animal Genomes (FAANG) Project seeks to improve comprehension of genome functionality through comprehensive annotation efforts. This annotation information helps to refine the accuracy and sensitivity of genomic selection strategies for animal [27, 28].

Many approaches have been developed to estimate genomic values and improve genomic prediction accuracy using multi-omics information [28–30]. Ye et al. refined the GFBLUP model using transcriptomics information in *Drosophila* [31], revealing that significant variations identified through a transcriptome-wide association study (TWAS) contribute more effects than those identified through genome-wide association study (GWAS). Xu et al. applied Bayesian ridge regression (BR) to evaluate the accuracy of GP for complex traits by integrating transcriptomics, proteomics, and metabolomics data, suggesting that large biobanks could reliably and efficiently explore trait–disease associations using multi-omics data [32]. Additionally, Hu et al. designed a novel GP strategy called multilayered least absolute shrinkage and selection operator (MLLASSO) by integrating multi-omics data into a single model, their finding suggested MLLASSO can significantly improve the predictability of yield in rice [33]. Although these approaches have been successfully employed in many studies, the integration of disparate data types into comprehensive system-scale analyses represents a significant challenge [34].

In this study, we generated genomic feature sets for sequencing variants by integrating multi-layered biological priors in beef cattle, then we assessed the contribution of functional variants and estimated variant scores for 43 economically important traits. Further, we evaluated and compared the GP accuracy based on functional variants using GBLUP and BayesB (Fig. 1).

Methods

Dataset

The measurement of phenotypes and genotypes was performed as described in our previous studies [35, 36]. The phenotypic data were generated from 1577 Huaxi (derived from Chinese Simmental beef cattle), which were born between 2008 and 2020 from Ulgai, Xilingol League, and Inner Mongolia, China. After weaning, all individuals were moved to Jinweifuren Co., Ltd. for fattening under the same feeding and management conditions. All samples were genotyped by Illumina BovineHD SNP array. The SNP positions were determined based on the ARS-UCD1.2 reference genome, the SNP imputation was carried out using Run 8 of the 1000 Bull Genomes Project and 44 representative individuals from our studied population. After filtering variants with the threshold of MAF < 0.05 and $DR^2 < 0.8$, we retained a total of 10,213,925 autosome SNPs with an average DR^2 of 0.93. A total of 43 traits including carcass and beef cut traits were included in this study. Detailed information for traits is presented in (Supplementary file 1, Table S1). All of the traits were adjusted by gender, year, acid remove day, enter weight and enter day. The phenotypes were adjusted by the significant factors using the glm function in R.

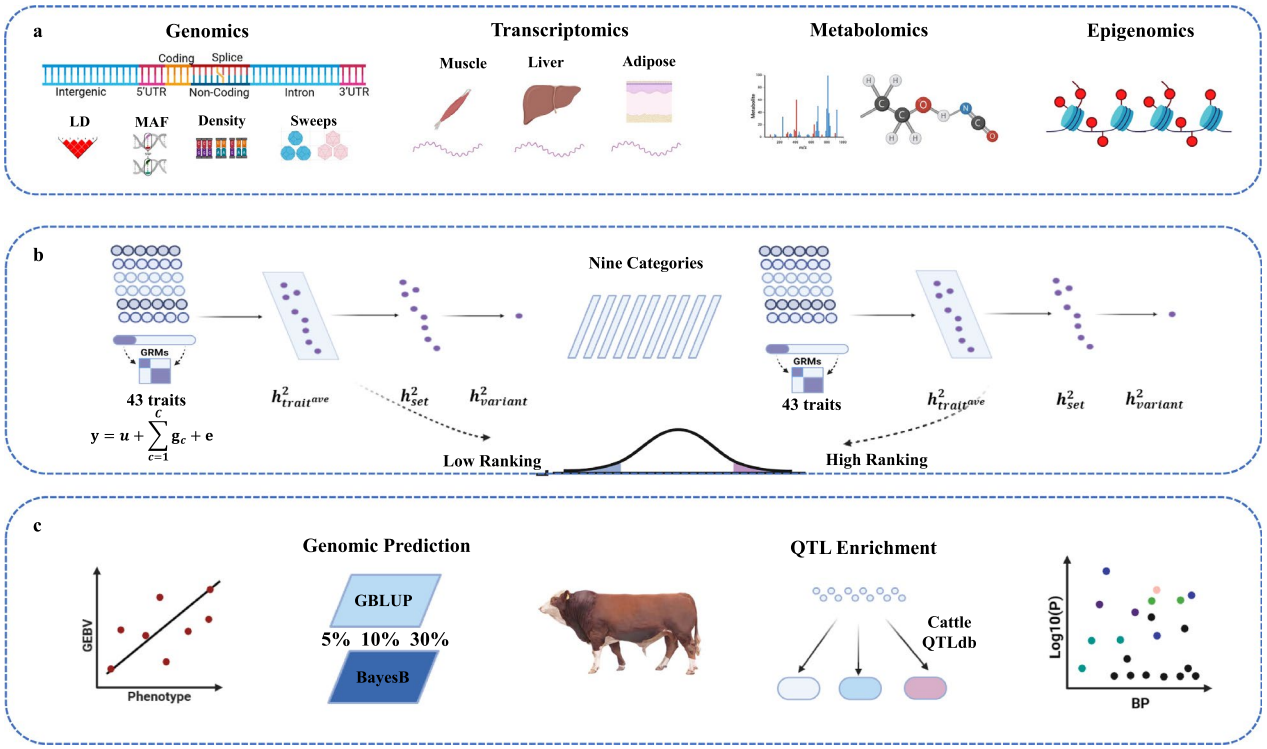

Fig. 1 Schematic overview of current study. **a** Data collection. We divided the full variants from WGS data into 11 feature sets (annotation, LD, allele frequency, variant density, p-variants, selection signature, conservation, eQTL, mQTL, OCR and HMRs) from genomics, transcriptomics, metabolomics, and epigenetics data. **b**. Calculate variants score. For each of the 43 traits, we estimated the variance explained by the random effects associated with each GRM using GREML. Each GREML analysis incorporated two random effects: one based on the targeted GRM and another based on the GRM derived from the remaining variants. We calculated the proportion of genetic variance attributed to the targeted GRM for each trait. To determine the per-variant heritability, we divided the explained variance by the number of variants in the set. Finally, we averaged this value across the 11 functional sets for each variant. **c**. Validation analysis. To assess the reasonableness of the scores, we established six thresholds: "top-5", "top-10", "top-30", "bottom-5", "bottom-10" and "bottom-30" We then compared the variance explained by each threshold with the accuracy of the genomic predictions. To ascertain whether there are pertinent QTL enrichments for our top variants, we conducted a QTL enrichment analysis using the Cattle QTL Database

Annotation

The SnpEff software was used to annotate and predict the effects of genetic variants, and the bovine genome annotation (ARS-UCD1.2) was downloaded from Ensembl [37, 38]. According to genome annotation information, the bovine genome was partitioned into six genomic classes, including 1) intergenic.regions, 2) intronic. regions, 3) geneend.regions, 4) cording.related.regions, 5) regulatory.related.regions, 6)3' untranslated regions (UTR) and 5' UTR.

LD, allele frequency and variant density

The levels of LD, allele frequency and variant density were divided based on three quartiles. We used the GCTA software to calculate the LD score in the surrounding 50 kb region, and then we used the LD score of each variant to bin all variants by quartile [39, 40]. Variants were unevenly distributed across the genome. VCFtools software was used to calculate the density within fixed 50 kb windows [41]. The allele frequency was divided by the minor allele frequency (MAF). The values of the quartiles were as follows: fourth quartile > third quartile > second quartile > first quartile.

Potential variants for production traits using WGS

We retrieved candidate variants related to body size and beef production traits in cattle, which are publicly available from sequence-based meta-GWAS with a larger number of animals from diverse cattle populations [42, 43]. Finally, 583,438 variants were used to define as the potential variants (p-variants).

Selection signature

The selscan software with a setting of the max-gap 800,000 bp was used to estimate the iHS score for autosomal SNPs [44]. The norm module of selscan was applied to normalize the iHS score, and single site values for iHS were averaged in nonoverlapping windows of 50 kb across the genome. Regions in the top 1% with the highest average |iHS| score and SNP numbers greater than 10 were regarded as candidate regions under positive selection.

Conserved sites

In this study, we used conserved sites to map gene regions that may be involved in basal metabolism. The conserved genome sites in cattle were transformed from humans. First, the conserved sites were based on conservation between 100 vertebrate species, and the Wiggle file was downloaded from UCSC. All the conserved sites in the human genome were 113,280,297, the sites were lifted over to the cattle genome (102,953,048) by LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver), and only the PhastCon score > 0.9 was chosen. Finally, a total of 192,825 variations were remained after merging the converted variants.

Meta-analysis of expression QTLs (eQTLs)

We conducted cis-eQTL mapping for three tissues (muscle, liver and adipose) according to our previous study [45]. The SNPs located within 1 Mb up/downstream of the transcription start sites (TSSs) were defined as potential cis-eQTLs [46]. In this study, each variant had an estimate of the effect and standard error (se), allowing for the three tissues to perform meta-analysis in this study by METAL software [47]. We obtained 240,683 variants as the eQTLs sites under threshold of 0.05 based on the false discovery rate(FDR) .

Meta-analysis of metabolic QTLs (mQTLs)

A total of 397 metabolites were extracted from 117 individuals. The polar metabolome extracts were analyzed using reversed-phase chromatographic separation with positive and negative ionization detection. The metabolome data were measured and corrected according to our previous study [48]. The metabolites were screened using three criteria: 1) We computed the pearson correlation coefficient between the traits and the metabolites. A total of 328 metabolites correlated with at least one trait were retained for subsequent analysis ($|r| \geq 0.25$). 2) The heritability of relevant metabolites was calculated, then we obtained 74 significant heritable metabolites ($0.1 < h^2 < 0.9$ and $P<0.05$). 3) GWAS analyses were performed based on the LMM for the 74 metabolites. We considered only the GWAS of 66 metabolites with inflation factors ranging from 0.98 to 1.02 for the meta-analysis in METAL software [47]. We obtained 386,995 candidate pleiotropic variants for metabolites with < *P*9.79E-08.

Epigenetic signals

Peak calling for muscle samples was performed using Genrich with the following parameters: -m 30, -j (ATAC-seq mode), -r (remove PCR duplicates), -e MT (to exclude mitochondrial chromosome), -q 0.05 (FDR-adjusted P-value). We obtained 495,903 variants as open chromatin region (OCR). For hypomethylated regions (HMRs) detection, we chose a 10 kb window size for muscle samples using the Methpipe software with the default parameters [49]. We obtained 762,835 variants as HMRs.

Variation score construction

The GREMLwas used to estimate the variance components. First, the different genomic relationship matrixes (GRMs) were made of the target variants and the remaining variants. Target variants were identified in 11 categories (conservation, annotation, selection signature, p-variants, LD, allele frequency, density, eQTLs, mQTLs, OCRs, and HMRs). Variants in these 11 categories refer to the target variants, whereas the remaining variants refer to the non-target variants. The variance components are estimated using a linear mixed model.

$$y^* = \sum_{c=1}^{C} g_c + e \tag{1}$$

where $\mathbf{y}^*$ is the vector of adjusted phenotypic values, $\mathbf{g}_{\boldsymbol{c}}$ is the vector of individual polygenic effects associated with annotation group$\boldsymbol{c}$, C is the total number of fitted annotation groups, and $\mathbf{e}$ is the random residual effect, which is assumed to follow a normal distribution of $\mathbf{e} \sim N\left(0, \sigma_e^2 I\right)$. where $\mathbf{g_c}$ is the genomic relationship matrix (GRM) computed using the variants present in category, which were calculated by Yang's method [50]. Then, the GCTA software was used to calculate the variance explained by random effects described for each GRM [39]. The mean heritability of the variation within target region for each trait was calculated, the partitioned heritability were estimated via the mean heritability divided by the number of variants [51]. To avoid LD heterogeneity along the genome, the LD level was chosen to adjust the score [52, 53] (Supplementary file 1, Table S2).

$$S_{\text{trait}} = \frac{\sum_1^{43} h_i^2}{43}$$

$$S_{\text{set}} = \frac{S_{\text{trait}}}{N}$$

$$\overline{h^2_{LD}} = \frac{n_{LD1} \times \overline{h^2_{LD1}} + n_{LD2} \times \overline{h^2_{LD2}} + n_{LD3} \times \overline{h^2_{LD3}} + n_{LD4} \times \overline{h^2_{LD4}}}{n_{LD1} + n_{LD2} + n_{LD3} + n_{LD4}} \quad (2)$$

$$\overline{\mathrm{S_{adj}}} = \mathrm{S_{set}} - \overline{h^2_{LD}}$$

$$S_{\mathrm{variant}} = \frac{\sum_1^{11} \overline{S_{\mathrm{adj}}}}{11}$$

$\mathrm{S_{trait}}$ is the heritability after pooling an average of 43 traits, and S_{set} is the average heritability of the variant within each category, N is the number of the functional sets. For the LD feature set, n_{LD1}, n_{LD2}, n_{LD3} and n_{LD4} are the numbers of variant members within the 1st, 2nd, 3rd and 4th LD score levels, respectively. $\overline{h^2_{LD1}}$, $\overline{h^2_{LD2}}$, $\overline{h^2_{LD3}}$ and $\overline{h^2_{LD4}}$ are the mean heritability of the 43 traits at the LD level. $\overline{S_{adj}}$ is the value after adjustment by LD. $S_{variant}$ is the mean heritability for 11 categories and is the combined score value of variance [51].

The enrichment of each category is quantified using the ratio of *EST* to *EXP*. *EST* denotes the estimated total heritability associated with the category, normalized by the estimated SNP heritability, which represents the proportion of heritability attributable to SNPs within that specific category. In contrast, *EXP* signifies the expected contribution of the category to overall SNP heritability, which was calculated by the ratio of the number of SNPs within the category to the total number of SNPs analyzed [54].

Finally, the combined variant score among the 11 feature sets was obtained. The top 5%, 10% and 30% and bottom 5%, 10% and 30% ranked variants were selected as the "top-5", "top-10", "top-30", "bottom-5", "bottom-10" and "bottom-30", respectively.

GBLUP model

$$y^* = Xg + e \quad (3)$$

where $\mathbf{y}^*$ is the vector of adjusted phenotypic values, $\mathbf{X}$ is the design matrix selected by different thresholds (the top 5%, 10%, and 30% and bottom 5%, 10%, and 30%) relating additive genetic values to the phenotype, $\mathbf{g}$ is the vector of genomic values captured by the genetic markers linked to target variants, which follow a normal distribution $g \sim \mathbf{N}\left(0, \sigma_g^2 \mathbf{G}\right)$. $\mathbf{G}$ was calculated using Yang's method, and $\mathbf{e}$ is a vector of random residual effects. The GBLUP model is implemented in the GCTA software [39].

BayesB model

In this study, we chose the BayesB model to predict the individual phenotypes, the equation:

$$y^* = \sum_j^k z_j a_j \delta_j + e \quad (4)$$

where y^* is the adjusted phenotype, z_j is the vector of genotype (0,1,2) across animals for the SNP, a_j is the allele substitution effect for the SNP, and δ_j is an indicator of whether the SNP was included ($\delta_j = 1$) or excluded ($\delta_j = 0$) in the model for a given Markov chain Monte Carlo (MCMC) iteration. The BayesB model is implemented in the GCTB software [55].

Evaluation of prediction performance

To avoid the impact of the number of variants on the accuracy of genomic prediction, we performed subsampling on WGS data by simulating a variant set containing 10% variants. The autosomal variants were sorted based on their base pair (BP) positions and then divided into bins, each containing 10 variants. From each bin, one variant was randomly selected to compose the variant panel [56, 57]. We then performed genomic prediction with simulated panels using the GBLUP model.

The accuracy of the predictions was assessed using a five-fold cross-validation method with five repetitions. Genomic prediction accuracy ($Acc = cor\left(y^*, GEBV\right)$) was determined by calculating the Pearson correlation coefficient between adjusted phenotypic values and GEBVs separately for each of the five-fold cross-validation replicates.

The GWAS significant regions mapping and QTLs enrichment

The GWAS was performed by using a mixed linear model-based association analysis in GCTA software [39]. The mixed linear model was used for GWAS:

$$y^* = bx + g + e \quad (5)$$

where $\mathbf{y}^*$ is the adjusted phenotype, $\mathbf{b}$ is the additive effect (fixed effect) of the candidate SNP to be tested for association, x is the SNP genotype indicator variable coded as 0, 1 or 2, $\mathbf{g}$ is the random effect and accumulated effect of all SNPs, $\mathbf{g} \sim N\left(0, \sigma^2 \mathbf{G}\right)$, and $\mathbf{e}$ is a vector of random residual effects. FDR was used to determine the threshold values for single-trait GWAS.

QTL enrichment analysis was carried out using the GALLO package [58] by comparing the number of annotated QTLs within candidate regions to the total number of QTLs in the Cattle QTL database [59].

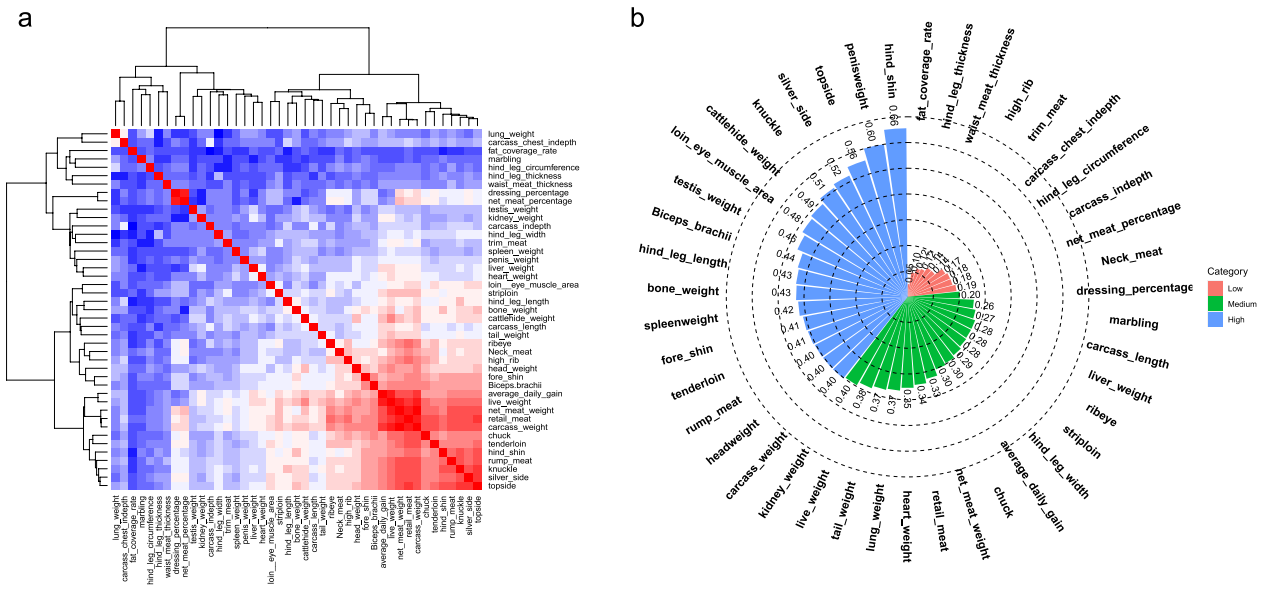

Fig. 2 The phenotype and genetic parameters of carcass traits. **a** The phenotype correlation between the 43 carcass traits is presented in the form of a color-coded box plot. The color of each box represents a positive correlation (red) or a negative correlation (blue). **b** The heritability of the 43 traits is presented in a similar format. Low heritability (0~0.2), medium heritability (0.2~.4), and high heritability (0.4~1)

Table 1 Merged and original functional annotation of sequence variants

Merged set name	Original annotation set name	Number	Sum
UTR	3_prime_UTR_variant	33,021	43,207
	5_prime_UTR_variant	1669	
	5_prime_UTR_premature_start_codon_gain_variant	8517	
intergenic	intergenic_region	6,132,394	6,132,394
geneend	downstream_gene_variant	388,028	863,401
	upstream_gene_variant	475,373	
intron	intron_variant	3,096,627	3,096,627
regulatory.related	splice_acceptor_variant	110	11,534
	splice_donor_variant	164	
	splice_region_variant	6222	
	non_coding_transcript_exon_variant	5038	
coding.related	synonymous	43,651	66,762
	missense_variant	22,680	
	initiator_codon_variant	23	
	start_lost	66	
	stop_gained	279	
	stop_lost	41	
	stop_retained_variant	22	
Total			10,213,925

Results

Summary statistics and genetic parameter estimations of 43 traits

The detailed summary statistics of the 43 traits are presented in (Supplementary file 1, Table S1), including the mean, median, coefficient of variation (CV), and standard deviation (SD). As expected, strong phenotypic correlations were observed between traits (Fig. 2a). The heritability estimates from the GREML are displayed in Supplementary file 1, Table S1. Seventeen traits showed high heritability, 16 showed medium heritability, 10 showed low heritability. Hind shin (HS) had

Table 2 The summary of the functional annotation sets in this study

Omics	Partitions (Number of animals)	Targeted variant sets (no. of variants)
Genomic	Conserved 100 species (NA)	Bovine genome sites lifted over from human sites with PhastCon score > 0.9 calculated using genomes of 100 vertebrate species (192825)
	Annotation (NA)	SnpEff was used to annotate the variants which was annotated as UTR (43,207), intergenic (6,132,394), geneend (863,401), intron (3,096,627), regulatory.related (11,534) and coding.related (66,762)
	Selection signature (44)	Regions at the top 1% with the highest average |iHS| score and SNP numbers greater than 10 were regarded as candidate regions under positive selection (109,325)
	P-variants (NA)	Variants have been identified in cattle: those relating to body size and beef production
	LD (1577)	First quartile (2,553,483), second quartile (2,553,481), third quartile (2,553,486), and fourth quartile (2,553,475)
	Freq (1577)	First quartile (2,557,044), second quartile (2,555,826), third quartile (2,550,339), and fourth quartile (2,550,716)
	Variant density (1577)	First quartile (2,564,482), second quartile (2,564,733), third quartile (2,538,073), and fourth quartile (2,546,637)
Metabolomics	mQTLs (117)	mQTLs with meta-analysis P < (0.05/10,213,925) from 66 types of metabolites (386,995)
Transcriptomics	eQTLs(227 muscle, 120 liver and 117 adipose)	eQTLs with meta-analysis (0.05) from 3 types of tissues (240,683)
Epigenomics	OCR (10)	Peak calling for muscle samples was performed using Genrich with -m 30, -j (ATAC-seq mode), -r (remove PCR duplicates), -e MT (to exclude mitochondrial chromosome), -q 0.05 (FDR-adjusted P-value) (495,903)
	HMRs (10)	chose a 10 kb window size with the default parameters (762,835)

The LD score indicates the linkage disequilibrium between pairwise variants in the surrounding 50-kb region. For the three quartiles, fourth quartile scores > third quartile > second quartile > first quartile. NA indicates sample from public data

the highest heritability (0.66), while fat cover percentage (FCR) showed the lowest heritability (0.05) (Fig. 2b).

The category of variants

In this study, we divided the full variants from WGS data into 11 feature sets (annotation, LD, allele frequency, variant density, p-variants, selection signature, conversation, eQTL, mQTL, OCR and HMRs). First, we annotated a total of 10,213,925 variants based on the ARS-UCD1.2 reference genome (Table 1). Among them, the majority of variants were located in intergenic and intronic regions, while only 0.65% were found in coding regions. The p-variant set contains 583,438 variants.

Selection signature set were selected based on the top 1% of regions with the highest |iHS| values (109,325). Different sets of variants were also divided according to the distribution of LD score (LD1, LD2, LD3, and LD4), MAF (Freq1, Freq2, Freq3, and Freq4) and variant density (Density1, Density2, Density3, and Density4) based on the quartile approach. In addition, conserved sites were selected according to a PhastCon score > 0.9 (Table 2). Finally, a total of 192,825 variants remained in our subsequent analysis.

For the transcriptomic data, a meta-analysis was subsequently carried out to identify the candidate variants that influence the expression of genes. We obtained 240,683 variants as eQTL sites with a threshold (FDR < 0.05) (Supplementary file 2**,** Fig. S1a).

For the metabolomic analysis, metabolites with |Correlation coefficients| > 0.25 for at least one trait were retained. Heritability analysis was then performed on the remaining 328 metabolites with strict criteria ($0.1 < h2 < 0.9$ and $P < 0.05$) (Supplementary file 2**,** Fig. S1b). Finally, 254 categories remained for subsequent analysis. GWAS was performed based on 74 metabolites, and we obtained 66 metabolites for meta-analysis after filtering based on the inflation factor. We identified 386,995 mQTL variants with a threshold of 0.05/10,213,925 (Supplementary file 2**,** Fig. S1c). For epigenomics data, a total of 495,903 OCRs were detected in muscle tissue by Genrich. We identified 762,835 HMRs in muscle tissue using Methpipe with a 10 kb window size.

The partitioned heritability estimation based on feature sets

We calculated the partitioned heritability for 43 traits to assess the contribution of each feature set to traits (Table 3). Regulatory-related regions accounted for 6.22% of S_{trait} while representing only 0.11% of genome variants. Our analysis showed a decrease in the square of the average heritability S_{trait} from the first to fourth quartile, indicating local LD and variant density influence variants effect. P-variants explained approximately 9.35% of the average genetic variance and constituted 5.71% of WGS variants. Conservation and selection signature variants accounted for 18.87% and 4.64% of genetic variance,

Table 3 The partitioned heritability enrichment for the functional annotation sets

Category	Averaged heritability, %	Min /Max heritability, %	Number	Genome fraction, %	Averaged Enrichment_ratio	Min /Max Enrichment_ ratio
OCR_target	9.54	1.87/ 22.47	495,903	4.86	6.37	1.25/ 15.02
Conservation_target	18.87	2.02/ 52.62	192,825	1.89	31.78	3.38/ 88.57
coding.related	1.95	0.47/ 7.9	66,762	0.65	9.40	2.27/ 38.08
Density1	19.48	0.99/ 57.07	2,564,482	25.11	2.37	0.12/ 6.94
Density2	5.68	1.20/ 22.25	2,564,733	25.11	0.69	0.16/ 2.69
Density3	4.26	0.98/ 15.17	2,538,073	24.85	0.52	0.12/ 1.85
Density4	3.33	0.64/ 13.98	2,546,637	24.93	0.41	0.08/ 1.72
eQTLs_target	2.74	0.49/ 9.39	240,683	2.36	3.78	0.67/ 12.97
eQTLs_rest	28.04	2.04/ 57.14	9,973,242	97.64	0.93	0.07/ 1.90
Freq1	8.82	1.42/ 39.34	2,557,044	25.03	1.11	0.18/ 4.95
Freq2	8.23	0.81/ 22.14	2,555,826	25.02	1.04	0.10/ 2.81
Freq3	8.00	1.02/ 37.54	2,550,339	24.97	1.01	0.13/4.74
Freq4	6.55	1.02/ 17.74	2,550,716	24.97	0.83	0.13/ 2.23
Geneend	2.04	0.54/ 6.61	863,401	8.45	0.76	0.21/ 2.47
SLSN_target	4.64	0.61/ 11.59	109,325	1.07	14.48	1.93/ 36.17
Intergenic	11.28	0.58/ 42.19	6,132,394	60.04	0.59	0.03/ 2.21
Intron	6.73	0.61/ 34.33	3,096,627	30.32	0.70	0.06/ 3.57
LD1	16.14	2.05/ 51.85	2,553,483	25.00	1.92	0.24/ 6.29
LD2	8.00	1.41/ 25.35	2,553,481	25.00	0.96	0.17/ 2.47
LD3	5.37	0.94/ 30.83	2,553,486	25.00	0.64	0.11/ 3.65
LD4	3.98	0.57/ 20.52	2,553,475	25.00	0.48	0.07/ 2.48
OCR_rest	21.25	1.86/ 52.03	9,718,022	95.14	0.73	0.06/ 1.79
Conservation_rest	12.55	1.81/ 35.52	10,021,100	98.11	0.41	0.05/ 1/16
SLSN_rest	25.31	2.09/ 50.70	10,104,600	98.93	0.85	0.07/ 1.70
P-variant_rest	21.53	2.14/ 48.88	9,630,487	94.29	0.74	0.08/ 1.68
HMRs_rest	23.49	1.96/ 56.12	9,451,090	92.53	0.82	0.07/ 1.96
P-variant_target	9.35	1.02/ 21.86	583,438	5.71	5.31	0.58/ 12.43
Regulatory.related	6.20	0.72/34.58	11,534	0.11	172.96	20.11/ 964.67
mQTLs_target	5.84	1.22/ 16.29	386,995	3.79	5.03	1.06/ 14.03
mQTLs_rest	24.84	1.94/ 55.60	9,826,930	96.21	0.84	0.07/ 1.88
UTR	3.51	0.51/ 18.06	43,207	0.42	26.15	3.80/ 134.57
HMRs_target	7.43	1.32/ 21.26	762,835	7.47	3.22	0.57/9.21

SLSN represents selection signature

respectively, despite representing only 1.89% and 1.07% of genome variants. Variants derived from transcriptomics and metabolomics data explained 2.74% and 5.84% of genetic variance. Epigenetically, HMRs and OCRs explained 7.43% and 9.54% of genetic variance. Overall, functional annotation sets contributed significantly to heritability (Fig. 3a), with regulatory-related SNPs (11,534) showing the greatest enrichment (172.96-fold), and coding regions providing a per-SNP predictability enrichment of 9.40. Smaller SNP counts in conservation and selection signatures yielded substantial contributions (31.78-fold and 14.48-fold enrichment). Metabolomics data provided greater heritability enrichment (5.03-fold) than transcriptomics (3.78-fold), while in the epigenetics category, OCRs outperformed HMRs with a 6.37-fold enrichment versus 3.22-fold. After correcting for LD categories, the ranking of variant sets from per-variant heritability showed highly correlated ($R^2=0.91$) with that of unranking variant sets(Fig. 3b).

Performance of GP based on different sets in the GBLUP model

We examined the heritability and compared the prediction accuracy of different variant sets and WGS data using GBLUP (Supplementary file 1, Table S3 and S4).

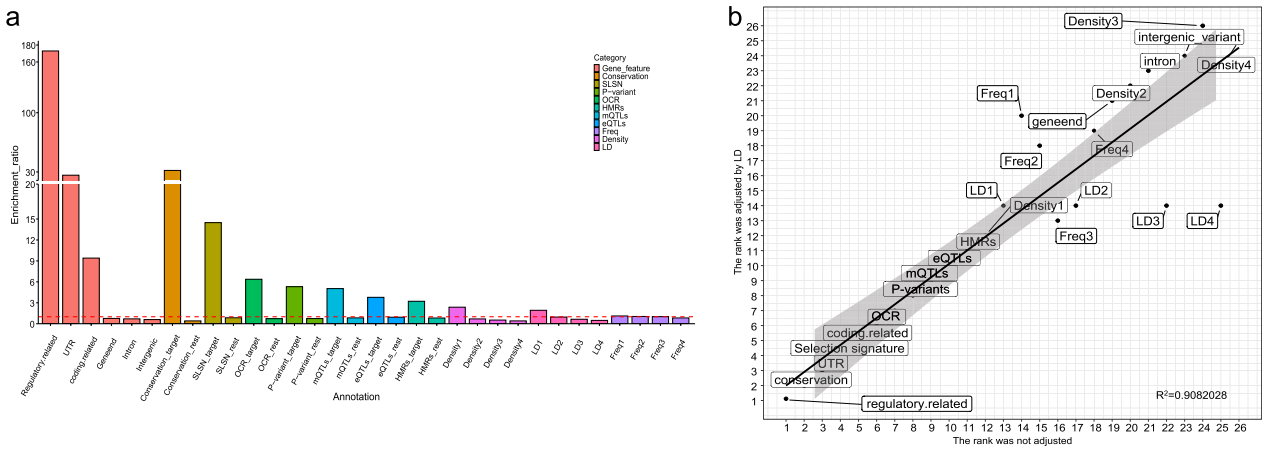

Fig. 3 The ranking analysis of 11 feature sets. **a** The heritability enrichment ratio of functional variant sets averaged across beef cattle. SLSN represents selection signature **b** The correlation between the ranking of variant set based on the LD adjusted per-variant $\overline{S_{adj}}$ (y-axis) and the ranking of variant set based on the unadjusted (observed) per-variant S_{set} (x-axis)

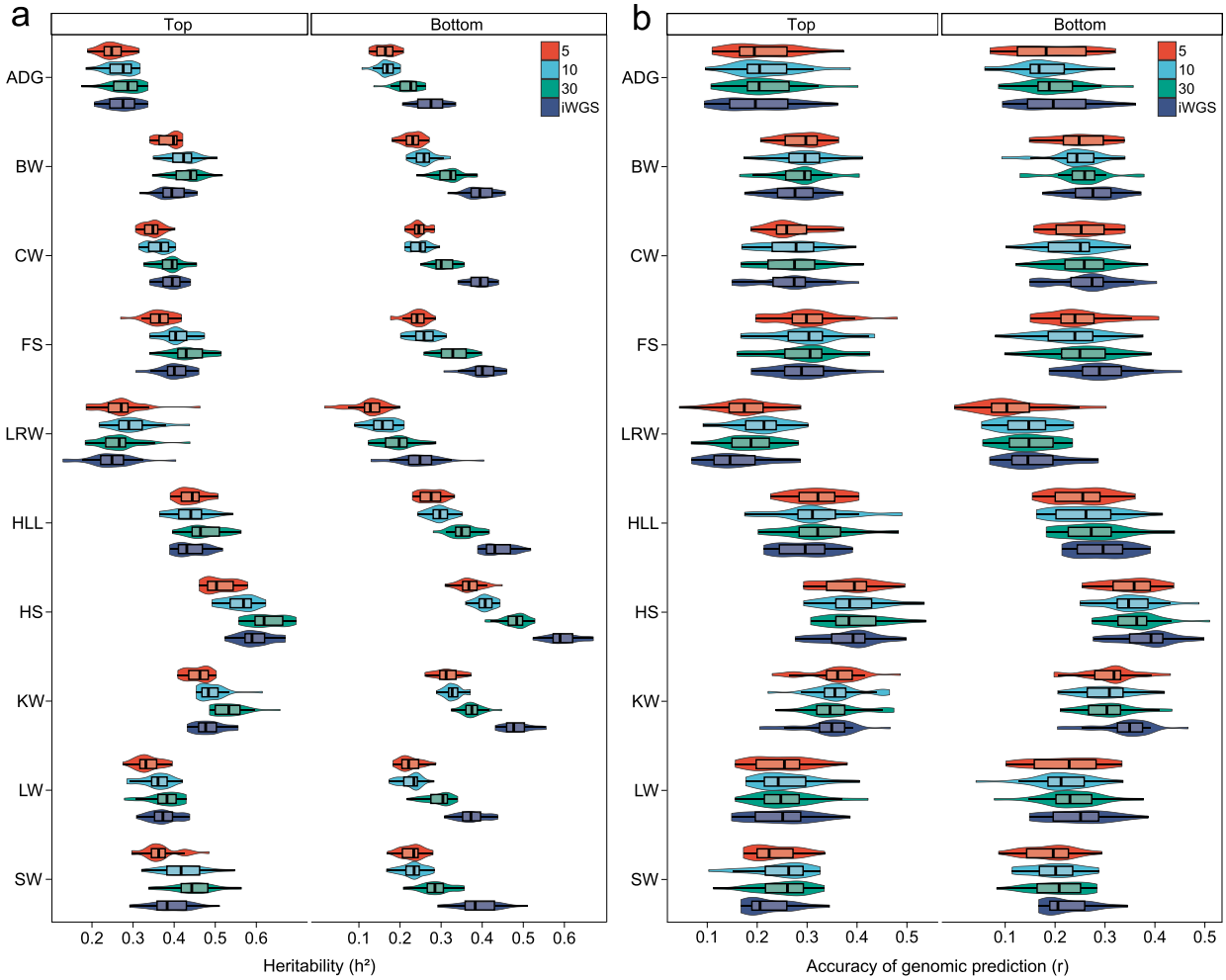

Fig. 4 Genomic prediction based on different thresholds using the GBLUP model **a** The heritability among 6 thresholds and WGS data in the GBLUP model **b** The accuracy of genomic prediction among 6 thresholds and WGS data in GBLUP model

The top sets performed significantly better than the bottom sets. We found that the top sets (averages of 0.3740, 0.4026, and 0.4267 for top-5, top-10, and top-30, respectively) had significantly higher heritability estimates than the bottom sets (averages of 0.2420, 0.2578, and 0.3164 for bottom-5, bottom-10, and bottom-30, respectively) (Fig. 4a), and the top sets (averages of 0.2793, 0.2865, and 0.2841 for top-5, top-10, and top-30, respectively) achieved higher accuracies than the bottom sets (averages of 0.2376, 0.2364, and 0.2487 for bottom-5, bottom-10, and bottom-30, respectively).

Across all sets and WGS data, the heritability of the top-30 exceeded that of GBLUP, with an average increase of 6.54%. Specifically, SW, knuckle (KK), and bone weight (BW) increased by 13.76%, 10.01%, and 9.36%, respectively. For the top-10 set, the LRW trait exhibited 20.67% higher heritability compared to WGS data. Regarding prediction accuracy, the top-10 set significantly outperformed the WGS data, with an average improvement of 7.54%. Notably, the LRW, SW, and hind leg length (HLL) accuracies increased by 29.40%, 11.30%, and 7.29%, respectively (Fig. 4b).

In this study, we randomly selected the 10% variants from the WGS data and compared with the full WGS data, we observed enhancements in prediction accuracy (4.18%, 2.67%, and 0.78%) for the GBLUP model for average daily gain (ADG), carcass weight (CW), and live weight (LW), respectively. The top-10 set demonstrated higher accuracy (4.18%, 2.76%, and 7.95%) compared to the randomly selected variants in ADG, LW, and KK (Supplementary file 2, Fig. S2).

Performance of GP based on different sets using the BayesB model

In this study, we found that the top-5 set achieved higher accuracy compared to the bottom-5 (average

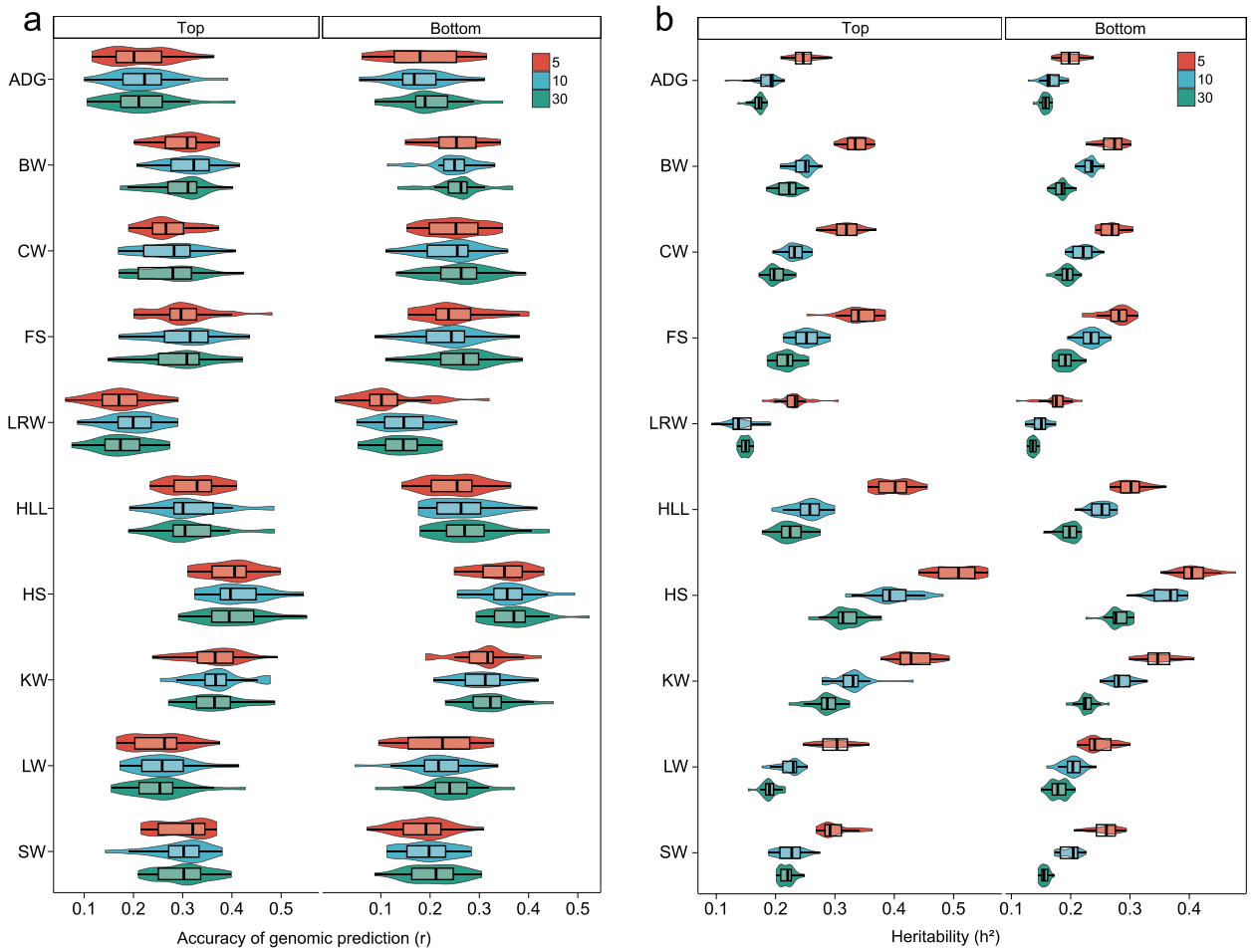

Fig. 5 Genomic prediction based on different thresholds using the BayesB model **a** The accuracy of genomic prediction among 6 thresholds in BayesB model. **b** The heritability among 6 thresholds in the BayesB model

~22.43%) using BayesB, the top-10 set outperformed the bottom-10 set (average ~24.75%), and the top-30 set surpassed the bottom-30 set (average ~ 14.50%). These findings are consistent with the performance observed in the GBLUP model (Fig. 5a). Furthermore, our analysis indicated that the top-10 demonstrated superior predictive accuracy when utilizing both the GBLUP and BayesB models. Specifically, we observed that the top-10 outperformed the top-5 by an average of approximately 2.88% and surpassed the top-30 by an average of approximately 2.83% (Supplementary file 1, Table S3 and S4).

Also, the heritability estimates of the studied traits using the BayesB were lower than GBLUP (Fig. 5b**,** Supplementary file 2, Fig. S3a), and BayesB significantly outperformed GBLUP in terms of genomic prediction accuracy. We found a significant increase in the prediction accuracy using BayesB. For instance, the prediction accuracy for the SW trait using the top-30 set showed an increase of 34.48% over the WGS data, and it increased by 31.52% for the SW using the top-10 set. Compared with the other methods, BayesB exhibited more improvements based on the top-10 set than others. For example, LRW, BW, and KK achieved the highest improvements in accuracy (24.90%, 12.95%, and 9.31%, respectively). Compared to GBLUP, BayesB showed higher predicted accuracy based on the top feature sets for all traits except LRW (Supplementary file 2, Fig. S3b). Particularly, the SW showed an approximately 18.17% increase based on the top-10 set compared to GBLUP, while BayesB model improved by 3.85% over GBLUP on average. The unbiasedness of genomic predictions (from ~0.84 to ~1.17) based on different thresholds were also estimated using the GBLUP and BayesB models (Supplementary file 2, Fig. S4).

Overlap of GWAS variants with functional features and QTLs enrichment analysis

The prioritized variants in the variant score were mapped to significant regions from the GWAS results for the HS and LRW. The top variant with an FDR of 3.3E-04 in the candidate region (BTA 6: 36,816,554-37,883,636 bp) was associated with HS, exhibiting a strong LD with nearby variants. Subsequently, we annotated the variants in this region with functional feature sets and observed that several variants overlapped with selection signatures, conservation, p-variants, and OCRs (Fig. 6a). For LRW, a QTL (BTA 14:25,906,554-26,883,636) was annotated with variants from OCRs (51.65%), mQTLs (25.87%), HMRs (10.39%) and conservation regions (3.87%) (Fig. 6b).

We carried out the overlap analysis between the top-10 variant set and the functional annotation set, revealing significant correlations between the occurrence ratio of the functional annotation set and their enrichment folds ($R^2 = 0.56$). Variants within the functional annotation set

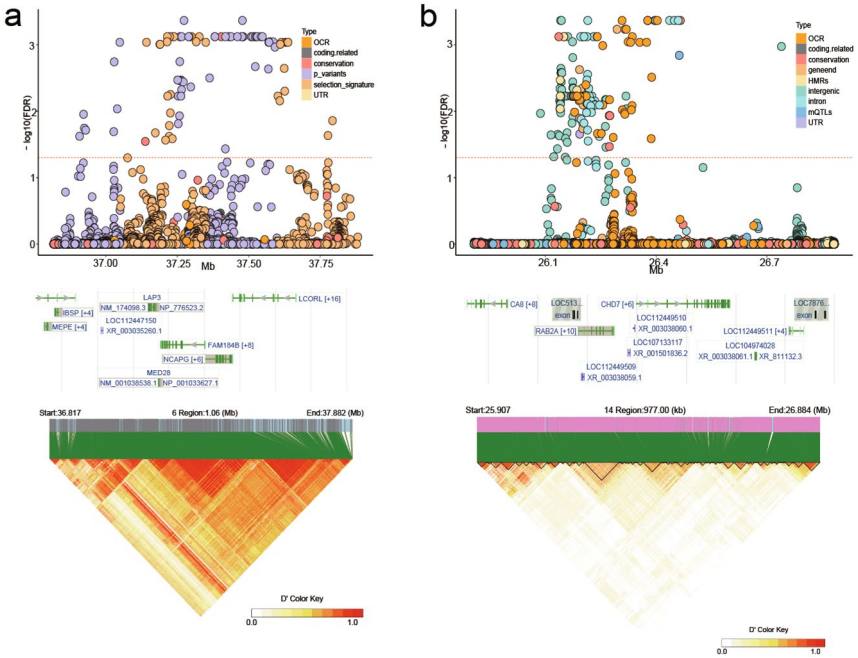

Fig. 6 Integrative GWAS analysis of two traits. **a** Results of the GWAS of the HS trait; the region plot of BTA 6. The colors of the variants are based on their LD with the most significant variants. **b** GWAS results for the FZ trait; the region plot of BTA 14. The colors of the variants are based on their LD with the most significant variants

that exhibited high enrichment were more frequently represented in the top-10 set. Notably, all variants categorized as regulatory-related, conservation, selection signatures, and coding-related were identified within the top-10 set. Furthermore, our findings indicated that variants within the lower enrichment-fold functional annotation set, specifically those in LD4, Freq4, and Density4 categories, occurred less frequently compared to other sets (Supplementary file 1, Table S6). Based on the Cattle QTL database [59], the top-10 set of variants was predominantly enriched with the production, meat, and carcass categories of QTLs, representing 19.15% and 21.96% of these categories, respectively (Supplementary file 2, Fig. S5a). Further enrichment analysis revealed a significant association of the top-10 set with weight related QTLs, such as metabolic body weight, average daily gain, carcass weight, and longissimus muscle area (Supplementary 2, Fig. S5b). Notably, these QTLs showed a prominent enrichment in meat color and conformation traits, demonstrating the highest richness factor among the analyzed traits.

Discussion

In recent years, numerous GP approaches have emerged for farm animals and plants [60, 61]. To improve the prediction accuracy, many methods have been proposed by integrating biological priors from multi-layered information [62, 63]. In cattle, a previous study performed GP using the BayesRC model by integrating independent biological priors, their findings revealed that BayesRC was more effective than BayesR in identifying candidate causal variants and predicting milk traits[64]. Liang et al. utilized transcriptomic data as a T matrix and combined them with wmssGBLUP and reported that the transcriptome data has the potential to improve genomic predictions [65]. However, a comprehensive methodology for GP by integrating this multi-layered information has not yet been fully developed [66]. To address this issue, we constructed 11 feature sets from multi-omics data and evaluated the contributions of functional variants to 43 economically important traits in beef cattle.

Multi-omics analyses are useful to characterize the regulatory regions and annotated mammalian genomes [27, 67–70]. Our study integrated multi-omics data, including comprehensive evolutionary, selection signal, transcriptomic, metabolomic and epigenetic data, based on the FAETH framework [51]. By estimating the variance explained by each feature set using GREML, we selected functional variant sets based on their ranking scores. This strategy can help to improve the accuracy of genomic prediction, and understand the genetic architecture underlying complex traits.

Our findings underscore the significance of incorporating functional annotation into genomic analyses [71, 72]. Zeng et al. revealed that genomic loci displaying conservation across a wider array of species were more prone to containing variants correlated with a heightened enrichment of heritability [73]. In humans, conserved regions of the genome have been found to markedly enhance the estimation of trait heritability [74], a phenomenon often attributed to the concentration of functional elements within these conserved regions [75]. Our findings revealed that the selection signature set was capable of elucidating a significant portion of the genetic variation observed in studied traits, which can contribute larger effect for GP compared with WGS data as reported by Kemper et al. [76]. Furthermore, our analysis of the set ranking result confirmed that the LD heterogeneity of variants have a substantial impact on trait heritability, which is consistent with recent evidence for the strong influence of LD properties on complex traits [19, 53].

Several studies had showed that the incorporation of transcriptomic and metabolomic information can help to elucidate the genetic basis for complex traits[77–79]. Intermediate QTLs such as eQTLs and mQTLs have consistently demonstrated their significance in contributing to the regulation of complex traits [80, 81]. These intermediate QTLs act as crucial mediators, connecting genetic variants to phenotypic outcomes [82, 83]. In this study, we found that the prediction accuracy using metabolomics information surpasses that of transcriptomic information, this could be explained by that the complex gene expression patterns of the diverse tissue sampled in different developmental stages [84–86]. Our analysis further revealed the pivotal role of intermediate QTLs, emphasizing their importance in the genetic architecture of complex traits.

Different models may influence the predictive accuracy of GEBVs [10, 18, 87]. Using the GBLUP model, Xiang et al. predicted GEBVs by calculating the variant scores in dairy cattle, and their findings suggested that genomic prediction using high-ranking variants was more accurate than genomic prediction using low-ranking variants in most scenarios [51]. In this study, we found that the top variant sets showed higher prediction accuracy than the bottom sets using the BayesB and GBLUP, which is consistent with the findings of a previous study by Xiang et al. [51].The difference between GBLUP and BayesB was mainly caused by the different assumptions regarding variation effects. Compared with

the GBLUP model, the performance of Bayesian models is superior when the studied trait are controlled by multiple QTLs [88, 89].

Our study suggested that pinpointing the functional variants on the top feature set may contribute larger effect for the genetic architecture underlying complex traits. Our approach provided a comprehensive framework for GP from multi-layered biological priors, and refined our understanding of complex trait genetics architecture. New approaches from machine learning and deep learning should advance analysis strategies for incorporating multi-layered biological datasets and promote genetic gains in animal breeding programs [66, 90, 91]. Overall, integrating multi-omics data from new approaches can further facilitate investigations of the functional impacts of variants and improve genomic prediction accuracy in farm animals.

Conclusions

Our study revealed that pinpointing the effect of variants on the top feature set can enhance our understanding of the genetic architecture underlying complex traits. Genomic selection by integrating multi-layered biological priors can improve prediction accuracy for important traits in cattle.

Abbreviations

GP	Genomic prediction
GBLUP	Genomic best linear unbiased prediction
GEBV	Genomic estimated breeding value
SNP	Single nucleotide polymorphism
MCMC	Markov chain Monte Carlo
WGS	Whole-genome sequencing
LD	Linkage disequilibrium
TWAS	Transcriptome-wide association study
GWAS	Genome-wide association study
SD	Standard deviation
HMRs	Hypomethylated regions
MAF	Minor allele frequency
TSSs	Transcription start sites
REML	Restricted maximum likelihood
BP	Base pair
QTL	Quantitative trait locus
BLUP	Best linear unbiased prediction

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13062-024-00574-y.

Additional file 1.

Additional file 2.

Acknowledgements

The authors would like to thank the staff at the experimental unit in Beijing and Ulgai for caring for animals and collecting biological samples.

Author contributions

JYL, LYX and BZ conceived and designed the study. ZDZ and QHN performed the statistical analyses. ZDZ and LYX wrote the paper. ZDZ, JYW, TYW, XYX, XG, and HJG participated in the data analyses. LPZ and LYX participated in the design of the study and contributed to the acquisition of data. All authors read and approved the final manuscript.

Funding

This study was supported by the National Key R&D Program of China (2022YFF1000600, 2023YFD1300105-020), the National Natural Science Foundation of China (31972554, 32272843), the Agricultural Science and Technology Innovation Program in the Chinese Academy of Agricultural Sciences (ASTIP-IAS-TS-16, ASTIP-IAS03), the National Beef Cattle Industrial Technology System (CARS-37), and the Inner Mongolia Autonomous Region Seed Industry Science and Technology Innovation Major Demonstration "Announce the List and Take-charge" Project (2022JBGS0018). LYX was supported by the Elite Youth Program in the Chinese Academy of Agricultural Sciences. BZ was supported by the Hohhot Science and Technology Innovation Talent Project (2022RC-5).

Availability of data and materials

The datasets used during the current study are available from the corresponding authors upon reasonable request.

Declarations

Ethics approval and consent to participatel

All animals were treated following the guidelines for experimental animals which were established by the Council of China. The study involving the use of tissue samples was approved by the ethics committee of the Science Research Department of the Institute of Animal Science, Chinese Academy of Agricultural Sciences under IAS2020-48.

Consent for publication

All authors have approved the final manuscript.

Competing interests

The authors declare no conflict of interest.

Author details

[1]Key Laboratory of Animal Genetics Breeding and Reproduction, Ministry of Agriculture and Rural Affairs, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China. [2]Northern Agriculture and Livestock Husbandry Technology Innovation Center, Hohhot 010010, China.

Received: 27 September 2024 Accepted: 2 December 2024
Published online: 31 December 2024

References

1. Marshall DM. Breed differences and genetic parameters for body composition traits in beef cattle. J Anim Sci. 1994;72(10):2745–55.
2. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic Selection in Dairy Cattle: The USDA Experience. Annu Rev Anim Biosci. 2017;5:309–27.
3. Werner CR, Gaynor RC, Sargent DJ, Lillo A, Gorjanc G, Hickey JM. Genomic selection strategies for clonally propagated crops. Theor Appl Genet. 2023;136(4):74.
4. Meuwissen T, Hayes B, Goddard M. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 2013;1:221–37.
5. Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JC. Applied animal genomics: results from the field. Annu Rev Anim Biosci. 2014;2:105–39.
6. Clark SA, van der Werf J. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. Methods Mol Biol. 2013;1019:321–30.
7. Gianola D. de los Campos G, Hill WG, Manfredi E, Fernando R: Additive genetic variability and the Bayesian alphabet. Genetics. 2009;183(1):347–63.
8. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.

9. Zhang Q, Zhang Q, Jensen J. Association Studies and Genomic Prediction for Genetic Improvements in Agriculture. Front Plant Sci. 2022;13: 904230.
10. Gao H, Su G, Janss L, Zhang Y, Lund MS. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. J Dairy Sci. 2013;96(7):4678–87.
11. Powell RL, Norman HD. Major advances in genetic evaluation techniques. J Dairy Sci. 2006;89(4):1337–48.
12. Zhu D, Zhao Y, Zhang R, Wu H, Cai G, Wu Z, Wang Y, Hu X. Genomic prediction based on selective linkage disequilibrium pruning of low-coverage whole-genome sequence variants in a pure Duroc population. Genet Sel Evol. 2023;55(1):72.
13. Song H, Ye S, Jiang Y, Zhang Z, Zhang Q, Ding X. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. Genet Sel Evol. 2019;51(1):58.
14. van Binsbergen R, Calus MP, Bink MC, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2015;47(1):71.
15. Heidaritabar M, Calus MP, Megens HJ, Vereijken A, Groenen MA, Bastiaansen JW. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. J Anim Breed Genet. 2016;133(3):167–79.
16. Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol. 2009;41(1):2.
17. Meuwissen T, Goddard M. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics. 2010;185(2):623–31.
18. de Los CG, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193(2):327–45.
19. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91(6):1011–21.
20. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. Nat Genet. 2017;49(9):1304–10.
21. Yengo L, Vedantam S, Marouli E, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610(7933):704–12.
22. Ren D, Cai X, Lin Q, Ye H, Teng J, Li J, Ding X, Zhang Z. Impact of linkage disequilibrium heterogeneity along the genome on genomic prediction and heritability estimation. Genet Sel Evol. 2022;54(1):47.
23. Xu L, Gao N, Wang Z, Xu L, Liu Y, Chen Y, Xu L, Gao X, Zhang L, Gao H, et al. Incorporating Genome Annotation Into Genomic Prediction for Carcass Traits in Chinese Simmental Beef Cattle. Front Genet. 2020;11:481.
24. Consortium GT: Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015, 348(6235):648–660.
25. Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, Li B, Xiang R, Chamberlain AJ, Pairo-Castineira E, et al. A multi-tissue atlas of regulatory variants in cattle. Nat Genet. 2022;54(9):1438–47.
26. Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, Bai L, Cai Z, Zhao B, Li X, et al. A compendium of genetic regulatory effects across pig tissues. Nat Genet. 2024;56(1):112–23.
27. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16(1):57.
28. Giuffra E, Tuggle CK, Consortium F: Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annual Review of Animal Biosciences 2019, 7(1):65–88.
29. Pérez-Enciso M, Rincón JC, Legarra A: Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution 2015, 47(1):43.
30. Amariuta T, Siewert-Rocks K, Price AL. Modeling tissue co-regulation estimates tissue-specific contributions to disease. Nat Genet. 2023;55(9):1503–11.
31. Ye S, Li J, Zhang Z. Multi-omics-data-assisted genomic feature markers preselection improves the accuracy of genomic prediction. J Anim Sci Biotechnol. 2020;11(1):109.
32. Xu Y, Ritchie SC, Liang Y, Timmers P, Pietzner M, Lannelongue L, Lambert SA, Tahir UA, May-Wilson S, Foguet C, et al. An atlas of genetic scores to predict multi-omic traits. Nature. 2023;616(7955):123–31.
33. Hu X, Xie W, Wu C, Xu S. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol J. 2019;17(10):2011–20.
34. Earls JC, Rappaport N, Heath L, Wilmanski T, Magis AT, Schork NJ, Omenn GS, Lovejoy J, Hood L, Price ND: Multi-Omic Biological Age Estimation and Its Correlation With Wellness and Disease Phenotypes: A Longitudinal Study of 3,558 Individuals. J Gerontol A Biol Sci Med Sci 2019, 74(Suppl_1):S52-s60.
35. Niu Q, Zhang T, Xu L, Wang T, Wang Z, Zhu B, Zhang L, Gao H, Song J, Li J, et al. Integration of selection signatures and multi-trait GWAS reveals polygenic genetic architecture of carcass traits in beef cattle. Genomics. 2021;113(5):3325–36.
36. Zhu B, Niu H, Zhang W, Wang Z, Liang Y, Guan L, Guo P, Chen Y, Zhang L, Guo Y, et al. Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. BMC Genomics. 2017;18(1):464.
37. Shamimuzzaman M, Le Tourneau JJ, Unni DR, Diesh CM, Triant DA, Walsh AT, Tayal A, Conant GC, Hagen DE, Elsik CG. Bovine Genome Database: new annotation tools for a new reference genome. Nucleic Acids Res. 2020;48(D1):D676–81.
38. Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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). 2012;6(2):80–92.
39. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
40. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J. Schizophrenia Working Group of the Psychiatric Genomics C, Patterson N, Daly MJ, Price AL, Neale BM: LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291–5.
41. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
42. Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, Sahana G, Govignon-Gion A, Boitard S, Dolezal M, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet. 2018;50(3):362–7.
43. Sanchez MP, Tribout T, Kadri NK, Chitneedi PK, Maak S, Hoze C, Boussaha M, Croiseau P, Philippe R, Spengeler M, et al. Sequence-based GWAS meta-analyses for beef production traits. Genet Sel Evol. 2023;55(1):70.
44. Szpiech ZA, Novak TE, Bailey NP, Stevison LS. Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. Evol Lett. 2021;5(4):408–21.
45. Cai W, Zhang Y, Chang T, Wang Z, Zhu B, Chen Y, Gao X, Xu L, Zhang L, Gao H, et al. The eQTL colocalization and transcriptome-wide association study identify potentially causal genes responsible for economic traits in Simmental beef cattle. J Anim Sci Biotechnol. 2023;14(1):78.
46. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016;32(10):1479–85.
47. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.
48. Du L, Chang T, An B, Liang M, Deng T, Li K, Cao S, Du Y, Gao X, Xu L et al.: Transcriptomics and Lipid Metabolomics Analysis of Subcutaneous, Visceral, and Abdominal Adipose Tissues of Beef Cattle. Genes (Basel) 2022, 14(1).
49. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS ONE. 2013;8(12): e81148.
50. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9.
51. Xiang R, Berg IVD, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, Bolormaa S, Liu Z, Rochfort SJ, Reich CM, et al. Quantifying the

contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A. 2019;116(39):19398–408.
52. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. Schizophrenia Working Group of the Psychiatric Genomics C: LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291–5.
53. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ. the UC: Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017;49(7):986–92.
54. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat Genet. 2019;51(2):277–84.
55. Zeng J, de Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, Yap CX, Xue A, Sidorenko J, McRae AF, et al. Signatures of negative selection in the genetic architecture of human complex traits. Nat Genet. 2018;50(5):746–53.
56. Della Coletta R, Fernandes SB, Monnahan PJ, Mikel MA, Bohn MO, Lipka AE, Hirsch CN. Importance of genetic architecture in marker selection decisions for genomic prediction. Theor Appl Genet. 2023;136(11):220.
57. Jeong S, Kim J-Y, Kim N. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Sci Rep. 2020;10(1):19653.
58. Fonseca PAS, Suárez-Vega A, Marras G, Cánovas Á: GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. *Gigascience* 2020, 9(12).
59. Hu Z-L, Park CA, Reecy JM. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 2021;50(D1):D956–61.
60. Tang Z, Toneyan S, Koo PK. Current approaches to genomic deep learning struggle to fully capture human genetic variation. Nat Genet. 2023;55(12):2021–2.
61. Alemu A, Åstrand J, Montesinos-López OA, Isidro y Sánchez J, Fernández-Gónzalez J, Tadesse W, Vetukuri RR, Carlsson AS, Ceplitis A, Crossa J et al.: Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant* 2024, 17(4):552–578.
62. Young AI. Solving the missing heritability problem. PLoS Genet. 2019;15(6): e1008222.
63. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nat Rev Genet. 2013;14(2):139–49.
64. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.
65. Liang M, An B, Chang T, Deng T, Du L, Li K, Cao S, Du Y, Xu L, Zhang L, et al. Incorporating kernelized multi-omics data improves the accuracy of genomic prediction. J Anim Sci Biotechnol. 2022;13(1):103.
66. Zeng Y, Yin R, Luo M, Chen J, Pan Z, Lu Y, Yu W, Yang Y: Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Brief Bioinform* 2023, 24(2).
67. Peng S, Dahlgren AR, Donnelly CG, Hales EN, Petersen JL, Bellone RR, Kalbfleisch T, Finno CJ. Functional annotation of the animal genomes: An integrated annotation resource for the horse. PLoS Genet. 2023;19(3): e1010468.
68. Clark EL, Archibald AL, Daetwyler HD, Groenen MAM, Harrison PW, Houston RD, Kühn C, Lien S, Macqueen DJ, Reecy JM, et al. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. Genome Biol. 2020;21(1):285.
69. Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? Trends Genet. 2021;37(2):109–24.
70. Zhong Y, Perera MA, Gamazon ER. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. Am J Hum Genet. 2019;104(6):1097–115.
71. Koufariotis LT. Chen Y-PP, Stothard P, Hayes BJ: Variance explained by whole genome sequence variants in coding and regulatory genome annotations for six dairy traits. BMC Genomics. 2018;19(1):237.
72. Levenstien MA, Klein RJ. Predicting functionally important SNP classes based on negative selection. BMC Bioinformatics. 2011;12:26.
73. Zheng Z, Liu S, Sidorenko J, Wang Y, Lin T, Yengo L, Turley P, Ani A, Wang R, Nolte IM, et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nat Genet. 2024;56(5):767–77.
74. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228–35.
75. Hujoel MLA, Gazal S, Hormozdiari F, van de Geijn B, Price AL. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. Am J Hum Genet. 2019;104(4):611–24.
76. Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics. 2014;15(1):246.
77. Zou Z, Zhang C, Wang Q, Hou Z, Xiong Z, Kong F, Wang Q, Song J, Liu B, Liu B et al.: Translatome and transcriptome co-profiling reveals a role of TPRXs in human zygotic genome activation. *Science* 2022, 378(6615):abo7923.
78. Zhai Y, Yu H, An X, Zhang Z, Zhang M, Zhang S, Li Q, Li Z. Profiling the transcriptomic signatures and identifying the patterns of zygotic genome activation—a comparative analysis between early porcine embryos and their counterparts in other three mammalian species. BMC Genomics. 2022;23(1):772.
79. Zuo X, Chen M, Zhang X, Guo A, Cheng S, Zhang R. Transcriptomic and metabolomic analyses to study the key role by which Ralstonia insidiosa induces Listeria monocytogenes to form suspended aggregates. Front Microbiol. 2023;14:1,260,909.
80. VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017;49(1):32.
81. Khansefid M, Pryce JE, Bolormaa S, Chen Y, Millen CA, Chamberlain AJ, Vander Jagt CJ, Goddard ME. Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. BMC Genomics. 2018;19(1):793.
82. Xiang R, Fang L, Liu S, Macleod IM, Liu Z, Breen EJ, Gao Y, Liu GE, Tenesa A, Mason BA, et al. Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle. Cell Genomics. 2023;3(10): 100385.
83. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van 't Hof P, Mei H, van Dijk F, Westra H-J et al.: Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* 2017, 49(1):139–145.
84. Hu H, Campbell MT, Yeats TH, Zheng X, Runcie DE, Covarrubias-Pazaran G, Broeckling C, Yao L, Caffe-Treml M, Gutiérrez La et al.: Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. *Theoretical and Applied Genetics* 2021, 134(12):4043–4054.
85. Knoch D, Werner CR, Meyer RC, Riewe D, Abbadi A, Lücke S, Snowdon RJ, Altmann T. Multi-omics-based prediction of hybrid performance in canola. Theor Appl Genet. 2021;134(4):1147–65.
86. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet. 2019;20(3):135–56.
87. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92(2):433–43.
88. Burch KS, Hou K, Ding Y, Wang Y, Gazal S, Shi H, Pasaniuc B. Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. Am J Hum Genet. 2022;109(4):692–709.
89. Zhong S, Dekkers JCM, Fernando RL, Jannink J-L. Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics. 2009;182(1):355–64.
90. Li J, Zhao T, Guan D, Pan Z, Bai Z, Teng J, Zhang Z, Zheng Z, Zeng J, Zhou H, et al. Learning functional conservation between human and pig to decipher evolutionary mechanisms underlying gene expression and complex traits. Cell Genom. 2023;3(10): 100390.
91. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49: 107739.