Epigenome-informed prioritization of bivalent chromatin SNPs enhances genomic prediction robustness: a proof-of-concept study in Pacific white shrimp (*Litopenaeus vannamei*)

Jia-Le Shi[1,2], Zhi-Fang Lu[1,2], Ming-Yi Sui[1,2], Miao-Miao Yin[1,2], Ming Mu[1,2], Dong-Dong Zhang[3], Zhen-Min Bao[1,2], Jing-Jie Hu[1,2], Qi-Fan Zeng[1,2,*], Zhi Ye[1,2,*]

[1]Key Laboratory of Tropical Aquatic Germplasm of Hainan Province, Sanya Oceanographic Institution, Ocean University of China, Sanya 572024, China

[2]MOE Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China

[3]School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication), School of Marine Biology and Fisheries, Hainan University, Sanya 572025, China

[*] Corresponding author: yezhi@ouc.edu.cn (Z.Y.); zengqifan@ouc.edu.cn (Q.Z.)

ABSTRACT

Genomic selection has become a powerful tool for accelerating genetic improvement in animal breeding, yet its broader application is often limited by high genotyping costs and reduced predictive reliability across populations. Functional genomic annotations offer a promising strategy to prioritize biologically meaningful variants and improve prediction robustness. Here, using Pacific white shrimp (*Litopenaeus vannamei*) as a model, we conducted a proof-of-concept study integrating whole-genome resequencing and phenotypic data from 972 individuals. High-resolution epigenomic maps were generated by profiling four histone marks (H3K4me1, H3K4me3, H3K27me3, and H3K27ac) across multiple embryonic stages and adult muscle tissue using CUT&Tag, and were then used to guide SNP prioritization for genomic prediction. Among the evaluated strategies, SNPs located in the muscle-specific bivalent promoter/enhancer (E6) state, which is characterized by the co-occurrence of active and repressive marks, consistently yielded the highest prediction accuracy. Notably, even at a moderate density (15k), E6-derived SNPs outperformed substantially larger genome-wide SNP sets. Importantly, in cross-population validation using an independent strain, the E6 SNP subset increased prediction accuracy by 47.62% (increasing from 0.21±0.05 to 0.31±0.04, P<0.05) relative

Uncorrected proof

to random subsets at equivalent density. These results demonstrate that epigenetic annotation–guided SNP prioritization can enhance prediction accuracy and robustness while reducing marker requirements, providing a practical framework for integrating functional genomic information into aquaculture breeding.

Keywords: Genomic selection; Histone modifications; Epigenetics; CUT&Tag; Functional annotation

INTRODUCTION

Genomic selection (GS) has fundamentally transformed animal breeding by accelerating genetic gain for complex traits, spanning from established livestock sectors such as pigs and cattle (Georges et al., 2019; Meuwissen et al., 2001) to emerging aquaculture species including fish, shrimp, and shellfish (Yáñez et al., 2023). By estimating breeding values using genome-wide molecular markers, GS circumvents the limitations of traditional pedigree-based selection. However, a critical bottleneck remains because the predictive accuracy of GS models often declines rapidly when applied across genetically divergent populations or across generations, largely due to the decay of linkage disequilibrium (LD) between markers and causal variants (Lund et al., 2014; Zheng et al., 2024). While whole-genome sequencing (WGS) data theoretically captures all causal mutations, studies have indicated that simply increasing marker density does not guarantee improved prediction accuracy (Yáñez et al., 2023; Yuan et al., 2025; Zhang et al., 2024). For instance, Zhuang et al. (2023) demonstrated in a commercial pig population that utilizing full WGS data did not significantly outperform single nucleotide polymorphism (SNP) chips, primarily due to the accumulation of "noisy" variants in non-functional genomic regions that dilute predictive signals.

To overcome these limitations, a paradigm shift is occurring from purely statistical modeling toward integrating "biological priors", specifically functional genomic annotations that pinpoint regulatory elements (MacLeod et al., 2016). Recent high-impact studies across diverse species have demonstrated that prioritizing variants within functional regions (e.g., promoters and enhancers) significantly enhances the robustness of genomic prediction (R. Ma et al., 2025; Yuan et al., 2025; Zheng et al., 2024). In cattle, Xiang et al. (2025) recently integrated extensive multi-omics annotations to develop a functional importance score, which improved the genomic prediction accuracy for climate resilience traits by approximately 11% across diverse populations. Similarly, W. Ma et al. (2025) developed a deep learning-based framework (DeepAnnotation) for pigs, showing that integrating functional annotations can effectively filter out noise

Uncorrected proof

and improve phenotype prediction. These findings suggest that identifying functional regulatory variants is key to breaking the LD dependency and achieving stable predictions across diverse genetic backgrounds. However, despite these encouraging breakthroughs, the integration of functional annotation into genomic prediction is still considered to be in its developmental stages (Alemu et al., 2025). Current applications are largely restricted to a few model livestock species and specific traits, and the optimal strategies for variant prioritization remain debated (W. Ma et al., 2025). To establish stable, reference-grade protocols that meet diverse breeding demands, it is essential to validate these functional genomics strategies across a wider spectrum of species and complex traits.

Among the various layers of functional information, epigenetic modifications, particularly histone post-translational modifications (PTMs), serve as dynamic indicators of the genomic regulatory landscape (Salem et al., 2024; Shi et al., 2025a; Zhao et al., 2021). Specific histone marks act as fundamental signatures for defining cis-regulatory elements (CREs); for instance, trimethylation of histone H3 at lysine 4 (H3K4me3) typically marks active promoters, whereas acetylation at lysine 27 (H3K27ac) distinguishes active enhancers from inactive ones. However, relying on individual marks provides only a partial view of regulatory complexity. By integrating profiles of multiple histone modifications (e.g., H3K4me1, H3K4me3, H3K27ac, and H3K27me3), the genome can be segmented into distinct "chromatin states" (e.g., active, repressed, or poised regions), providing a high-resolution map of functional elements (Ernst & Kellis, 2017; Pan et al., 2021; Salem et al., 2024). These chromatin states serve as powerful "biological priors" for genomic prediction, enabling the prioritization of SNPs located in biologically active regulatory regions while filtering out noise from non-functional sequences. Consequently, leveraging these multi-mark chromatin states offers a mechanism-based strategy to refine SNP selection and potentially enhance the predictive power of GS models.

In the aquaculture sector, concerted international initiatives like AQUA-FAANG (Functional annotation of animal genomes consortium) have begun to characterize functional regulatory landscapes for major European finfish species (Johnston et al., 2024), complemented by similar FAANG-associated efforts in the United States targeting key species such as rainbow trout (Salem et al., 2024). Emerging evidence from these projects indicates that functional genomic annotations can successfully assist in pinpointing causative variants underlying complex traits, such as disease resistance in sea bass (Johnston et al., 2024) and rainbow

Uncorrected proof

trout (Salem et al., 2024). However, despite this progress in map generation, the empirical application of functional genomic annotations into GS models remains unreported in aquaculture. Furthermore, compared to finfish, functional genomic resources for economically critical crustacean species, particularly the Pacific white shrimp (*Litopenaeus vannamei*), are severely lacking. While high-quality reference genomes for shrimp are now available, the comprehensive epigenetic maps required to filter "noisy" variants and guide robust genomic prediction are virtually nonexistent, limiting the translation of functional genomics into practical shrimp breeding.

To address this gap, this study presents a proof-of-concept framework for epigenetically informed genomic selection in *L. vannamei*. We generated high-resolution chromatin maps using CUT&Tag (Cleavage under targets and tagmentation) for four key histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac) across embryonic development and adult tissues. By integrating these functional annotations with whole-genome resequencing data from a breeding population, we systematically evaluated the predictive efficiency of different chromatin states. We specifically hypothesized that variants located in bivalent promoter/enhancer regions (E6 state) would offer superior predictive robustness compared to random genome-wide markers. Our findings demonstrate that prioritizing these functional SNPs significantly improves cross-population prediction accuracy, offering a cost-effective and biologically grounded strategy for precision breeding in aquaculture.

MATERIALS AND METHODS

Experimental populations

The three breeding populations analyzed in this study were derived from ongoing selective breeding programs conducted in our laboratory. Three domesticated breeding strains of *L. vannamei* maintained at a commercial hatchery (Hairen Aquatic Seed Industry Technology Co., Ltd., Hebei, China) were included. Each strain was produced from approximately 50 breeding pairs of broodstock through mass spawning, generating large mixed-family larval cohorts of approximately 5 000 000 offspring per strain. Individuals used in this study were randomly sampled from these cohorts for genotyping and phenotyping until they reached the desired age under rearing conditions described below. Due to the mass-spawning procedure, physical pedigree records (i.e., full- and half-sib family structure) were not available. Therefore, realized genetic relationships among individuals were estimated using a genomic relationship matrix derived from

Uncorrected proof

genome-wide SNP markers. In addition, a 110-day-old Hawaiian Primo (PLM) strain population from Haimao Seed & Breeding Technology Group Co., Ltd. (Guangdong, China) was included as an external, distantly related population for cross-population evaluation. This population originated from a single breeding family producing approximately 200 000 offspring, from which individuals were randomly sampled.

All shrimp were reared in controlled indoor recirculating aquaculture systems prior to sampling. Seawater was subjected to a standardized treatment process consisting of ozonation, ultraviolet sterilization, and fine mechanical filtration (1–5 μm). Each tank (effective volume 80–100 m³) was stocked at a density of 20–25 individuals/m³. During the rearing period, water temperature was maintained at 25–28°C, salinity at 28–31‰, and pH at 7.8–8.1. Dissolved oxygen levels were kept above 6 mg/L, while concentrations of ammonia-N and nitrite-N were controlled below 3 mg/L and 0.05 mg/L, respectively. To ensure stable environmental conditions, one-third to one-half of the tank volume was renewed daily. Shrimp were fed a high-protein formulated diet four times per day at a feeding rate of 12–15% of total biomass.

Phenotypic measurements and statistics

Body length and body weight were obtained using a digital caliper (±0.02 mm) and an electronic scale (±0.01 g), respectively. Summary statistics for each population are presented in **Table 1**.

Statistical analyses of the phenotypic data were performed using IBM SPSS Statistics v.25.0. To accurately assess phenotypic distributions, the normality of body length and body weight within each of the three populations was evaluated using the Shapiro-Wilk test alongside skewness calculations. Due to the non-normal distribution of body weight, the non-parametric Wilcoxon rank-sum test was employed to evaluate the statistical significance of phenotypic differences among the populations. P-value<0.05 was considered statistically significant, with extreme significance denoted at P<0.0001. Boxplots visualizing the phenotypic distributions were generated using the R package ggplot2 (Wickham, 2011). Furthermore, the correlation between body weight and body length was evaluated by linear regression, and coefficients of determination (R^2) were calculated in Microsoft Excel 2021.

Resequencing and SNP calling

Uncorrected proof

Genomic DNA was extracted from muscle tissue using the Tiangen Genomic DNA Extraction Kit (Beijing, China). Sequencing libraries were prepared with the VAHTS Universal DNA Library Prep Kit for Illumina V3® (Vazyme, China) and sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA) to generate 150 bp paired-end reads, targeting approximately 2 Gb of sequencing data per individual.

Raw sequencing reads were quality assessed and filtered using the NGS QC Toolkit v.2.3 (Patel & Jain, 2012). Reads containing adapter contamination,>10% ambiguous bases (N), or>40% bases with Phred quality<20 were removed. High-quality reads were aligned to the chromosome-level *L. vannamei* reference genome available through the Aquaculture Molecular Breeding Platform (AMBP; http://mgb.qnlm.ac/) (Q. Zeng et al., 2022) using BWA-MEM v.0.7.17 (Li, 2013). PCR duplicates were removed using the MarkDuplicates tool in GATK v.4.0 (McKenna et al., 2010), and alignment files were sorted and indexed with SAMtools v.1.17 (Li et al., 2009).

Variants were called using the GATK v.4.0 (McKenna et al., 2010) best-practice pipeline. The *HaplotypeCaller* was run in GVCF mode for each sample with default parameters, followed by CombineGVCFs and GenotypeGVCFs to produce a joint VCF (Variant call format) file containing all individuals. To ensure variant accuracy, only biallelic SNPs were retained using the GATK SelectVariants tool (flag –select-type-to-include SNP). Variants were filtered using the VariantFiltration function in GATK v.4.0, and loci with Quality by Depth (QD)<2.0, Fisher Strand Bias (FS)>60.0, Mapping Quality (MQ)<40.0, Mapping Quality Rank Sum (MQRankSum)< –12.5, Read Position Rank Sum (ReadPosRankSum)<-8.0, or Strand Odds Ratio (SOR)>3.0 were excluded from downstream analyses.

Genotypes were subsequently imputed using GLIMPSE v.2.0.0 (Rubinacci et al., 2021) with our in-house haplotype reference panel. SNP quality control was performed using Plink v.1.9 (Purcell et al., 2007): SNPs with minor allele frequency (MAF)<0.05 were excluded, yielding 2 978 750 SNPs. Hardy–Weinberg equilibrium (HWE) was tested (--hwe 0.0001), and loci with significant deviations (P<0.0001) were removed, resulting in 2 609 549 high-quality SNPs for downstream analysis. Sample- and locus-level missingness was assessed using PLINK (--missing), and variant statistics were visualized in R using ggplot2 (Wickham, 2011).

Uncorrected proof

Population genomics analyses

Population structure and genetic relatedness: Population structure was assessed using both principal component analysis (PCA) and model-based clustering. PCA was performed with Plink v.1.9 (--indep-pairwise 50 10 0.5; --pca 10) to remove linkage disequilibrium and extract the top ten principal components (PCs), which were used to visualize genetic stratification among populations. Model-based population structure analysis was conducted using ADMIXTURE v.1.3.0 (Alexander & Lange, 2011) with the number of assumed ancestral clusters (K) ranging from 2 to 30. Because the cross-validation (CV) error decreased gradually with increasing K without showing a clear minimum (**Supplementary Figure S1A**), we acknowledge that the selection of K=3 is a pragmatic choice informed by our prior knowledge of the three sampled breeding strains, rather than a purely data-driven statistical conclusion. Kinship among individuals was estimated using GEMMA v.0.98.5 (Zhou & Stephens, 2014) to construct the relatedness matrix. Visualization of PCA, ADMIXTURE ancestry coefficients, and kinship matrices was performed using the R packages ggplot2 (Wickham, 2011) and ComplexHeatmap (Gu, 2022). SNP density plots were generated with CMplot (Yin et al., 2021).

Estimation of genetic parameters: The narrow-sense heritability (h^2) of each trait was estimated using genome-wide SNP markers under a genomic linear mixed model (LMM) implemented in GCTA v.1.94.3 (Yang et al., 2011). The LMM was specified as:

$$y = Xb + Zu + \varepsilon (1)$$

where y is the vector of phenotypic observations, X is the design matrix for fixed effects, and b is the vector of fixed-effect coefficients. Because the three breeding populations were reared separately and environmental records were unavailable, population identity was included as the fixed effect to prevent confounding and inflation of additive genetic variance due to between-population divergence. Z is the incidence matrix relating records to individuals, and $u \sim N\left(0, G\sigma_g^2\right)$ represents the additive genetic effects, with G being the genomic relationship matrix (GRM) constructed from SNP genotypes using the command *--make-grm*. The residual term is denoted as $\varepsilon \sim N\left(0, I\sigma_e^2\right)$.

Variance components σ_g^2 (additive genetic variance) and σ_e^2 (residual variance) were estimated using restricted maximum likelihood (REML) with the parameters *--reml* and *--reml-alg 0*. The REML algorithm

Uncorrected proof

iteratively maximizes the restricted log-likelihood function to obtain optimal estimates of the variance components. The heritability of each trait was then calculated as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} (2)$$

Because population and environmental effects were confounded, this formulation estimates the within-population heritability, reflecting additive genetic variance after removing population mean differences. This approach ensures that the estimated h^2 values represent true intra-population genetic variation, rather than inflated values driven by between-population divergence.

Histone modification profiling and chromatin state annotation

CUT&Tag samples and experimental procedures: The CUT&Tag samples used in this study comprised specimens from 11 embryonic and larval developmental stages and adult muscle tissue. These 11 developmental stages were selected because embryogenesis is characterized by massive and highly dynamic gene expression regulation events. Profiling these stages allows us to comprehensively capture and annotate the landscape of *cis*-regulatory elements across the shrimp genome. In contrast, adult muscle tissue was chosen to provide highly trait-specific epigenetic annotations, as it is the primary tissue driving the targeted growth traits. Embryonic samples covered seven stages that were obtained from our previously published work. These stages included blastula (blas), gastrula (gast), limb bud embryo (lbe), larva in membrane (lim), nauplius I (N1), nauplius III (N3), and nauplius VI (N6) (Shi et al., 2025b). To extend the developmental coverage, additional CUT&Tag datasets were generated for four later laarval stages (zoea I, zoea III, mysis I, and mysis III). These newly generated datasets have been submitted to the NCBI BioProject database and will be made publicly available upon publication of this article.

Two independent biological replicates were performed for each histone modification under identical experimental conditions to ensure reliability and reproducibility. Approximately 200 embryos from each developmental stage or 100 mg of aseptically dissected adult dorsal muscle tissue were used per replicate. Because standard cell dissociation protocols are not fully optimized for *L. vannamei* tissues, we employed tissue-specific nuclei extraction methods. For embryonic samples, following Shi et al. (2025b), embryos were immersed in 300 μL of pre-chilled NE1 Buffer (mix 1 mL 1 mol/L HEPES-KOH pH 7.9, 500 μL 1 mol/L KCl, 12.5 μL 2 mol/L spermidine, 500 μL 10% Triton X-100, and 10 mL glycerol in 38 mL ddH_2O,

Zoological Research

and add 1 Roche Complete Protease Inhibitor EDTA-Free tablet) and homogenized on ice using sterile surgical scissors. An additional 500 μL of pre-chilled NE1 Buffer was added, and the homogenate was thoroughly pipetted to generate a single-cell suspension. After a 5-min incubation on ice, the suspension was filtered through a 40 μm strainer. The filtrate was centrifuged at 1 300×g for 4 min, and the resulting pellet was washed 2–3 times with 500 μL of Wash Buffer before being resuspended in 100 μL of Wash Buffer (mix 1 mL 1 mol/L HEPES pH 7.5, 1.5 mL 5 mol/L NaCl, 12.5 μL 2 mol/L spermidine, bring the final volume to 50 mL with ddH_2O, and add 1 Roche Complete Protease Inhibitor EDTA-Free tablet). For adult muscle samples, following Qi et al. (2025), the tissue block was placed in a 1.5 mL centrifuge tube with NE1 buffer, quickly clipped, and transferred into a Dounce homogenizer (Merck, Darmstadt, Germany). The tissue was smoothly ground using pestles A and B sequentially, then filtered through 70 μm and 40 μm cell strainers. The suspension was centrifuged to remove the supernatant, and the nuclear pellet was washed 1–2 times and resuspended in Wash buffer.

Following nuclei isolation, ~50 000 nuclei per sample were subjected to downstream CUT&Tag experiments. The CUT&Tag methodology is based on the in situ tethering principles originally described by Kaya-Okur et al. (2019). Libraries were constructed strictly following the manufacturer's instructions of the Hyperactive Universal CUT&Tag Kit for Illumina (Vazyme, China; Cat. #TD903). Information on the four histone modification antibodies (H3K4me1, H3K4me3, H3K27me3, and H3K27ac) and the control IgG is provided in **Supplementary Table S1**. Sequencing was conducted on an Illumina NovaSeq 6000 platform (Novogene, Tianjin, China) with an average depth of ~1 Gb per sample.

CUT&Tag data analysis was performed using the standardized pipeline established in our previous study (Shi et al., 2025b). Specifically, raw reads were quality-controlled using NGS QC Toolkit v.2.3 (Patel & Jain, 2012) by removing adapter sequences, reads containing>10% ambiguous bases (N), or reads with>50% of bases having Phred quality scores≤20 (summary in **Supplementary Table S2**). Clean reads were then aligned to the *L. vannamei* reference genome using BWA v.0.7.17 (Li, 2013; Zeng et al., 2022). Abnormally paired fragments were filtered using the *awk* command to retain valid fragments between 100–1000 bp. Peak calling was then conducted using the remaining fragments with MACS2 v.2.2.7.1 (Feng et al., 2012) using the following parameters: macs2 callpeak -t A.bam -c IgG.bam -f BAMPE -g 1.66e9 --keep-dup auto -q 0.05. Valid peaks were annotated using the R packages GenomicFeatures (Lawrence et al., 2013) and

Uncorrected proof

ChIPseeker (G. Yu et al., 2015).

Chromatin state modeling: Integration of multiple histone modification datasets enables the identification of chromatin states such as active enhancers, active transcription start sites, and silencers, thereby providing high-resolution functional annotation of the genome (Ernst & Kellis, 2017). In this study, CUT&Tag datasets for four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac) from 11 embryonic and larval developmental stages, as well as adult muscle tissue of *L. vannamei*, were used for chromatin state identification.

Chromatin state analysis was performed following the standardized pipeline described in our previous study (Shi et al., 2025b) using ChromHMM v.1.26 (Ernst & Kellis, 2017). The *L. vannamei* genome was divided into non-overlapping 200 bp bins, and enrichment signals for each histone modification were binarized into presence/absence states using the BinarizeBam module with parameters -gzip -b 200 -f 0 -g 0 -p 0.0001. Model training and chromatin segmentation were conducted using the LearnModel module with parameters -gzip -d 0.001 -color 129,0,0 -p 20 -i chrhmm. ChromHMM models with varying state numbers (5–15) were compared based on emission probabilities, enrichment patterns, and genomic coverage. Models with fewer than eight states failed to distinguish biologically meaningful features, whereas those with more than eight states introduced redundant categories (**Supplementary Figure S2A**). Therefore, an eight-state model was selected as the optimal balance between biological interpretability and model complexity. Each chromatin state was subsequently annotated based on combinatorial histone mark patterns, representing distinct functional elements such as active promoters, strong/weak enhancers, repressed or quiescent regions.

Conventional genomic selection (GS) analyses

In conventional genomic estimated breeding value (GEBV) analysis, random SNP sampling was conducted after LD filtering to reduce marker redundancy and ensure approximate independence among loci. High-LD loci were pruned using Plink v.1.9 with the parameter --indep-pairwise 50 10 0.5, resulting in 1 081 117 independent SNPs for downstream GS analysis.

To identify the optimal GS model for growth traits in *L. vannamei*, both genomic best linear unbiased prediction (GBLUP) and Bayesian regression approaches were evaluated. For GBLUP, the following linear

Uncorrected proof

mixed model was applied:

$$y = Xb + Zu + (3)$$

where y is the vector of phenotypic records, b is the vector of fixed effects including the population group, and u is the vector of random additive genetic effects with $u \sim N(0, G\sigma_u^2)$. G is the genomic relationship matrix (GRM) which was constructed following VanRaden (2008):

$$G = \frac{ZZ^T}{2\sum p_i(1-p_i)} (4)$$

where Z is the matrix of centered genotypes and p_i is the allele frequency of the ith SNP. Variance components were estimated using the restricted maximum likelihood (REML) algorithm implemented in the R package sommer (Covarrubias-Pazaran, 2016).

In addition to GBLUP, three Bayesian regression models (BayesA, BayesB, and BayesC) were implemented using the BGLR package (Pérez & de los Campos, 2014). The general model was defined as:

$$y = Xb + \sum_{j=1}^{m} z_j \alpha_j + (5)$$

where z_j is the genotype covariate for SNP j, and α_j is its effect. The three Bayesian methods differ in their prior assumptions: BayesA assigns marker-specific variances, BayesB introduces a probability (π) that a SNP has zero effect, and BayesC estimates π as a hyperparameter (Pérez & de los Campos, 2014; X. Wang et al., 2018). Prediction accuracies derived from each model were compared to determine the best-performing approach.

To assess the effect of marker density on prediction accuracy, SNP subsets ranging from 0.1k to 30k were randomly sampled from the LD-filtered dataset (the 1 081 117 SNPs described above) with the --thin-count parameter in Plink v.1.9. For each density, five independent random samplings were performed, and five-fold cross-validation was repeated five times, resulting in 25 replicates per density. To evaluate the impact of training sample size, random subsets of 100, 250, 500, 750, and 972 individuals were used. At a fixed marker density of 10k, loci were randomly sampled five times, and prediction accuracy was evaluated using the BayesA model across 20 rounds of five-fold cross-validation. In five-fold cross-validation, all individuals were randomly divided into five non-overlapping subsets, with one subset used as the validation set and the remaining four as the training set in each iteration until every individual had been used once for validation.

Uncorrected proof

Prediction accuracy was defined as the Pearson correlation between predicted GEBVs and observed phenotypes in the validation sets, adjusted by dividing by the square root of trait heritability to account for the influence of finite heritability on predictive performance.

Functional annotation–guided genomic selection

Because it remains unclear which histone modifications (in which tissues) have the greatest impact on genomic selection (GS), we performed a series of stepwise evaluations to comprehensively assess the effects of functional genomic annotations derived from histone modifications on GS performance. Given that histone modifications and chromatin states are dynamically regulated across embryonic development and different tissues, we first used the combined set of histone modification peaks from all tissues to ensure comprehensive coverage of functional genomic regions. Based on this union set, we evaluated the SNP filtering effects of individual histone modifications and compared them with results obtained from muscle tissue alone, which is expected to be more directly related to growth traits.

In addition, because SNPs located in distinct chromatin states may exert different regulatory influences on downstream genes, chromatin state annotations derived from all histone modification data were further applied. Specifically, we evaluated and compared the predictive performance of SNPs located within different chromatin states in embryonic stages and adult muscle tissue separately. The detailed analytical procedures are described below.

SNP filtering for genomic prediction based on histone modification peaks: SNPs were filtered based on histone modification peak regions identified from CUT&Tag narrowPeak files. Peak coordinates were extracted using Linux command-line tools (cut and awk), and SNP–peak overlaps were identified with the R package data.table, followed by deduplication. To ensure robustness, only SNPs located within peaks that were consistently detected in both biological replicates were retained. Functional annotation was performed with the R packages GenomicFeatures (Lawrence et al., 2013) and ChIPseeker (G. Yu et al., 2015), with a *flankDistance* of 5 kb.

From these peak-associated SNPs, five random subsets at different densities were sampled for each histone mark (H3K4me1, H3K4me3, H3K27me3, H3K27ac) and their union set (*All-Histone*) using the --thin-count

Uncorrected proof

parameter in Plink v.1.9. Parallel subsets were generated from muscle-specific peaks. Genomic prediction was primarily performed using the BayesA model, with predictive performance evaluated by five rounds of five-fold cross-validation. Prediction accuracies were compared with conventional GS (random LD-filtered SNPs), and accuracy was calculated as described in Section 2.3. At equal marker densities, statistical significance was tested using the Wilcoxon rank-sum test ($^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$), and results were visualized using ggplot2 (Wickham, 2011).

Chromatin state–based SNP prioritization: Chromatin state profiles for each embryonic stage and adult muscle tissue were generated using the eight-state ChromHMM model described in Section 2.2.3. SNPs passing quality control (prior to LD pruning) were intersected with state-specific genomic segments by first extracting chromatin-state coordinates with Linux command-line tools (cut and awk) and then identifying SNP–state overlaps using the R package data.table. SNPs falling within each chromatin state were retained. For embryonic stages, SNPs belonging to the same chromatin state across all 11 stages were merged to generate embryonic state-specific SNP sets. For adult muscle, SNPs were grouped according to the muscle-specific chromatin-state segmentation.

For each chromatin state, SNP subsets of varying marker densities were sampled using the --thin-count function in Plink v.1.9, with five replicates per density. Predictive performance was evaluated using the same five-round five-fold cross-validation procedure described in Section 2.4.1. Because muscle tissue is expected to be more directly related to growth traits, chromatin-state–filtered SNPs from adult muscle were analyzed separately.

To assess whether improvements were dependent on the statistical model, the chromatin state showing the highest predictive enhancement (state E6 in adult muscle) was additionally evaluated using GBLUP, BayesB, and BayesC, in addition to BayesA.

Cross-population validation of predictive performance

As shown in Section 3.5, SNPs within the muscle bivalent promoter/enhancer (E6) state significantly improved prediction accuracy. To further evaluate their stability and generalizability across populations, the distantly related PLM population (see **Table 1** for sample size and trait statistics) was used as an external

Uncorrected proof

validation set. The same number of SNPs located in the muscle E6 state were compared with randomly sampled SNPs of equal density.

Before genomic prediction, phenotypes in the reference population were adjusted for population group effects to avoid confounding between population divergence and additive genetic signals. For model training, SNPs from the reference population were partitioned into five subsets for five-fold cross-validation, and the entire procedure was repeated five times, resulting in 25 training iterations in total. In each iteration, four folds were used to estimate marker effects, and the remaining fold served as the internal validation set. After model fitting, the estimated marker effects were directly applied to the genotypes of PLM individuals to compute uncalibrated GEBVs, thereby evaluating the transferability of SNP effects without phenotype-based re-estimation in the target population.

Prediction accuracy in the PLM population was calculated as the Pearson correlation between predicted GEBVs and the observed phenotypes, and was further adjusted by dividing by the square root of the trait heritability to account for finite heritability constraints. The same number of E6-state SNPs was compared against randomly sampled SNPs of equal density to evaluate the specificity of functional annotation–guided SNP selection.

Genome-wide association study (GWAS) of E6 SNPs

To further characterize the functional relevance of SNPs located in the muscle bivalent promoter/enhancer (E6) state and explore their underlying genetic mechanisms, GWAS was conducted using GEMMA v.0.98.1 (Zhou & Stephens, 2014). Group identity was included as a covariate to account for differences among populations. GWAS was performed for body length using a univariate linear mixed model (LMM) to appropriately control for population structure and individual relatedness.

A standardized genomic relationship matrix (GRM) was first computed with the -gk 2 option in GEMMA according to the following formula:

$$G = \frac{1}{M} SS^T (6)$$

where *M* represents the total number of SNP markers, and S denotes the centered and standardized genotype

Zoological Research

matrix. The resulting GRM reflects pairwise genetic similarity among individuals and is equivalent to the kinship matrix used in mixed linear model-based GWAS. Subsequently, a univariate LMM was fitted for each SNP using the -lmm 1 option in GEMMA:

$$y = W\alpha + x\beta + u + \varepsilon (7)$$

where y is the phenotype vector (body length); W is the design matrix for fixed effects, including group identity and the top principal components (PC1 and PC2); α is the corresponding vector of fixed-effect coefficients; x is the genotype vector of the SNP being tested; and β is the SNP effect size. The random polygenic effect is modeled as $u \sim N\left(0, K\sigma_g^2\right)$, where K is the GRM. The residual errors follow $\varepsilon \sim N\left(0, I\sigma_\varepsilon^2\right)$. SNP significance was assessed using the Wald test.

Complete GWAS results are provided in **Supplementary Table S3**. SNP annotation was performed using GenomicFeatures (Lawrence et al., 2013) and ChIPseeker (G. Yu et al., 2015), followed by Gene Ontology (GO) enrichment analysis using the R package clusterProfiler (G. Yu et al., 2012) to identify biological processes and molecular functions enriched among significant SNPs.

To further refine candidate variants and genes associated with growth traits, RNA-seq data from *L. vannamei* muscle were incorporated to provide supporting transcriptomic evidence. Raw reads were downloaded from NCBI SRA, with accession numbers SRR6466352, SRR6466355, and SRR6466357 for fast-growth performance (FG) individuals, and SRR6466336, SRR6466339, and SRR6466341 for slow-growth performance (SG) individuals (Santos et al., 2021). Quality control and filtering, including automatic adapter removal, trimming of low-quality bases, and minimum-length filtering, were performed using SeqyClean v.1.10.09 (https://github.com/ibest/seqyclean), following the preprocessing strategy described in Santos et al. (2021). Reads were further filtered by requiring an average Phred quality≥23, end-base quality≥30, and a minimum read length of 65 bp, consistent with the original workflow. Clean reads were subsequently mapped to the same chromosome-level reference genome used in this study for whole-genome resequencing and CUT&Tag. Read quantification and normalization were performed as in the original study to ensure methodological consistency. Differential transcript abundance between the FG and SG groups was assessed with the Wilcoxon rank-sum test (P<0.05), and RNA level visualization was conducted with GraphPad Prism v.8.0.2.

Uncorrected proof

RESULTS

Population characteristics and SNP discovery

Within each of the three populations, body length exhibited an approximately normal distribution (W>0.990, absolute skewness<0.1; P>0.05 for groups 1 and 3, P=0.027 for group 2). However, consistent with traits scaling volumetrically, body weight displayed a visible right-skewed distribution and significantly deviated from normality in all groups (W=0.974–0.989, skewness=0.335–0.415, all P<0.05) (**Figure 1A**). Furthermore, highly significant differences in both traits were confirmed among the three populations of the same age (P<0.0001), as indicated by asterisks in **Figure 1B**. Linear regression further revealed a strong correlation between body length and body weight (R^2=0.918) **(Supplementary Figure S3**).

Resequencing of the 972 individuals identified 14 276 173 SNPs after imputation. Following minor allele frequency filtering (MAF<0.05), 2 978 750 SNPs remained (**Figure 1C**), and HWE filtering yielded 2 609 549 high-quality SNPs for downstream GS analysis (**Figure 1D**). Heritability estimates for body weight and body length were 0.29±0.07 and 0.44±0.08, respectively. Given the higher heritability of body length, it was selected as the representative growth trait for subsequent GS analyses.

BayesA model achieved the highest prediction accuracy for the growth trait in *L. vannamei*

Principal component analysis (PCA) revealed population structure among the samples (**Figure 2A**). Consistently, ADMIXTURE analysis illustrated distinct ancestry components with varying degrees of admixture (**Supplementary Figure S1B**), reflecting their divergent breeding histories. The kinship heatmap exhibited a block-like pattern of within-population genetic similarity (**Figure 2B**). Together, these analyses indicated the presence of population stratification, although Groups 1 and 3 exhibited substantial genetic overlap. To prevent genomic prediction models from being confounded by baseline phenotypic differences among the commercial strains, we conservatively retained population origin as a fixed effect, despite incomplete genetic isolation.

After quality control, 2 609 549 SNPs were retained and were evenly distributed across the genome (**Figure 2C**), consistent with the expectation that most SNPs arise from genetic drift and lack strong regional bias (Santure & Garant, 2018). Following LD filtering, 1 081 117 SNPs remained for GS model training, were

Uncorrected proof

also evenly distributed (**Figure 2D**).

The predictive performance of four GS models (GBLUP, BayesA, BayesB, and BayesC) was compared. Prediction accuracy increased with SNP number across all models, but gains diminished at higher densities, showing only marginal increases beyond 10k SNPs (**Figure 2E; Supplementary Table S4**). BayesA consistently outperformed the other models under multiple densities (**Figure 2E; Supplementary Table S4**) and was therefore adopted for subsequent analyses.

Prediction accuracy and stability also improved with increasing training sample size, reaching a plateau at ~750 individuals (**Supplementary Figure S4; Supplementary Table S5**). This indicates that the 972 samples used in this study were sufficient to ensure reliable GS analyses and provided a robust foundation for downstream comparisons.

Differential effects of histone modification–based SNP subsets on genomic selection

CUT&Tag data from 11 embryonic stages and adult muscle tissue of *L. vannamei* were analyzed for four histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3) together with an IgG control to identify SNPs with potential regulatory functions. Sequencing alignment rates exceeded 86% (mostly>90%) (**Supplementary Table S6**), indicating high data quality. Further quality assessments, including fragment size distribution, fraction of reads in peaks (FRiP), Pearson correlation coefficients for the biological replicates, peak signal intensity heatmaps, and the genomic feature distribution of histone modification peaks (**Supplementary Figure S5–S9; Supplementary Table S7–8**), also confirmed that both the newly generated embryonic-stage samples and the adult muscle samples produced reliable CUT&Tag profiles (Kaya-Okur et al., 2019; Qi et al., 2025; Shi et al., 2025b). The number of histone modification peaks varied substantially across developmental stages and tissues (ranging from 212 to 33 694), reflecting their dynamic regulatory activity, whereas the high concordance between biological replicates demonstrated the robustness of the CUT&Tag signals. The number of peaks detected at each developmental stage or in adult muscle tissue, together with the corresponding numbers of SNPs located within each peak set, are summarized in **Supplementary Table S9–10**.

When merging peak regions from all embryonic stages and adult muscle tissue, histone modification signals

Uncorrected proof

505 collectively covered ~65.34% of annotated coding genes (**Supplementary Figure S10**), indicating
506 widespread regulatory element occupancy across the genome had been identified. In total, 306 487 SNPs
507 (11.47% of all high-quality SNPs) were located within histone modification peaks and were broadly
508 distributed across the genome (**Figure 3A**). Among these, H3K4me3 peaks contained the largest proportion
509 (7.34%), followed by H3K4me1 (4.72%), H3K27me3 (2.15%), and H3K27ac (1.34%) (**Table 2**).
510
511 To test their utility for GS, SNP subsets of varying densities (0.5k–30k) were sampled from each histone
512 modification region and their union (All-Histone) and compared with random SNPs using BayesA and five-
513 fold cross-validation. Prediction accuracy increased with SNP density, with the rate of improvement
514 diminishing substantially beyond 10k. At low densities (0.5k–5k), SNPs from H3K4me1 and H3K4me3
515 showed higher accuracy (4.00%–18.75%) and greater stability than random SNPs, although differences were
516 not always significant (**Figure 3B; Supplementary Table S11**). By contrast, SNPs from H3K27me3 and
517 H3K27ac often underperformed, with H3K27me3 accuracy significantly reduced at 20k density. Combining
518 all histone modifications did not improve prediction, likely because broad inclusion diluted tissue- or stage-
519 specific signals. These results suggest that different histone modifications provide distinct functional
520 annotation information, which can differentially influence the prediction accuracy of GS when used for SNP
521 selection.
522
523 Because muscle is directly related to growth, we further examined muscle-specific regulatory regions,
524 identifying 70 480 SNPs (2.70% of all) across the four histone marks (**Figure 3C; Table 2**). The H3K4me1
525 subset was small (163 SNPs), but at 0.1k density, its prediction accuracy (0.11±0.02) exceeded that of
526 random SNPs (0.08±0.05) by 37.5%, with improved stability (**Supplementary Table S12**). Across muscle
527 histone marks, accuracy increased with density, yielding diminishing returns beyond ~10k. At low densities
528 (0.5k–2.5k), H3K4me3 subsets consistently outperformed random SNPs (4%–12.5%), while H3K27ac
529 subsets generally underperformed. Notably, H3K27me3 SNPs from muscle showed a trend of improved
530 accuracy (**Figure 3D; Supplementary Table S12**), in contrast to their performance when embryonic data
531 were included.
532
533 Overall, histone modification–based SNP selection showed the potential to improve the predictive accuracy
534 and stability of GS models under low marker densities, though effects varied by tissue and modification.

Uncorrected proof

While many subsets performed comparably to random SNPs, H3K4me3 consistently provided modest but stable gains at low densities, supporting its role as a key regulatory mark. Furthermore, when evaluated at equivalent marker densities, using muscle-specific information tended to produce greater relative improvements over random SNPs compared to using the union of all tissues. These findings suggest that integrating histone modification data to refine functional SNP annotations may further enhance GS accuracy.

Functional annotation of embryonic chromatin states reveals SNP subsets with distinct predictive power

The GS results based on individual histone modifications indicated that SNPs located in different functional regions exert distinct predictive effects. To obtain higher-resolution functional information, we integrated four histone modification datasets (H3K4me1, H3K4me3, H3K27me3, and H3K27ac) from embryonic stages and adult muscle of *L. vannamei* and identified eight chromatin states using ChromHMM (Ernst & Kellis, 2017). These included: flanking bivalent transcription start site (TSS)/enhancer 1 (BivFInk1), quiescent/low expression regions (Quies), weak repressed Polycomb regions (ReprPCWk), repressed Polycomb regions (ReprPC), weak enhancer regions (EnhWk), flanking bivalent TSS/enhancer 2 (BivFInk2), weak transcription regions (TxWk), and flanking TSS downstream regions (TssFlnkD) (**Figure 4A**). As histone modification effects appeared tissue-specific, chromatin states from embryonic stages and adult muscle were analyzed separately.

In embryos, the number of chromatin state sites ranged from 11 313 to 113 769, with SNP coverage from 13 291 to 1 912 007 (**Supplementary Table S13-14**), reflecting dynamic changes in chromatin accessibility and its relationship with genetic variations. The quiescent/low expression state (Quies, E2), which lacks histone modification marks and likely represents inactive regions, encompassed the largest number of SNPs (90.10%). In contrast, SNPs were far fewer in active states such as TxWk (E7), BivFInk1/2 (E1/E6), and TssFlnkD (E8) (**Table 3**).

To evaluate predictive performance, SNP subsets (0.5k–30k, five replicates per density) were sampled from each chromatin state and assessed with the BayesA model using five rounds of five-fold cross-validation. SNPs in BivFlnk1 (E1), EnhWk (E5), BivFlnk2 (E6), and TssFlnkD (E8) improved prediction accuracy by 4.00–11.11% over random SNPs at low-to-moderate densities (1k–2.5k) (**Figure 4B; Supplementary Table**

Uncorrected proof

565 **S15**). However, the overall improvement from embryonic chromatin states was modest, likely because
566 embryonic regulatory landscapes differ from those driving adult growth phenotypes.
567
568 Notably, SNPs from repressive or weak transcriptional states reduced predictive performance. In ReprPCWk
569 (E3), ReprPC (E4), and TxWk (E7), accuracy decreased by 2.86–12.00% compared with random SNPs at
570 2.5k–15k densities (**Figure 4B; Supplementary Table S15**), suggesting that such regions contribute little
571 to GS predictions and may even introduce noise.
572
573 **Predictive superiority and cross-population robustness of muscle chromatin state–based SNPs**
574 Analysis of chromatin states in adult shrimp muscle revealed 19 763–105 594 peaks across different states,
575 reflecting substantial variations among functional regions such as promoters, enhancers, and repressive
576 domains. Corresponding SNP coverage ranged from 15 073 to 1 689 212, with the quiescent state (E2) again
577 encompassing the largest number of SNPs (**Table 4; Supplementary Table S14**).
578
579 Genomic prediction analyses revealed that SNPs within the muscle bivalent promoter/enhancer state (E6)
580 consistently exhibited the strongest predictive performance across all marker densities. At low densities
581 (0.5k–2.5k), E6 SNPs improved prediction accuracy by 12.00–44.44% compared with random LD-filtered
582 SNPs. At a density of 5k, E6 SNPs achieved accuracy comparable to that obtained using 15k randomly
583 sampled SNPs. At 15k density, E6 SNPs produced an additional 25.71% improvement in accuracy relative
584 to random SNPs (**Figure 4C; Table 5**).
585
586 To further assess the robustness of the chromatin state–based improvement, three additional genomic
587 prediction models (BayesB, BayesC, and GBLUP) were applied to the muscle E6 SNP set. Consistent with
588 the BayesA results, E6 SNPs yielded higher predictive accuracy across all marker densities (**Supplementary**
589 **Figure S11**), indicating that the improvement in prediction accuracy was not dependent on the prior
590 assumptions of the BayesA model but rather stemmed from the unique biological relevance of the E6
591 chromatin state. These results demonstrate that E6 SNPs offer strong predictive power and meaningful
592 biological interpretability under both low- and moderate-density scenarios.
593
594 Furthermore, we sought to confirm that the predictive superiority of the bivalent state is robust to alternative

Uncorrected proof

595 ChromHMM parameterizations. We extracted the corresponding bivalent state SNPs from both 7-state and
596 9-state models generated for the adult muscle tissue and evaluated their genomic prediction accuracies. As
597 shown in **Supplementary Figure S2C**, SNPs residing in the bivalent state consistently outperformed
598 randomly sampled SNPs across different marker densities, regardless of whether a 7-state, 8-state, or 9-state
599 model was utilized. This confirms that the predictive advantage is driven by the underlying biological
600 relevance of bivalent chromatin regions rather than the specific model parameterization.
601
602 In cross-population validation, using the PLM population as an independent validation set, the prediction
603 accuracy of E6 SNPs was significantly improved by 47.62% compared with random SNPs (from 0.21±0.05
604 (95% CI: 0.166–0.254) to 0.31±0.04 (95% CI: 0.275–0.345), $P<0.05$) (**Figure 4D, 4E**), highlighting their
605 excellent generalizability and practical value across populations.
606
607 Notably, SNPs within certain chromatin state regions exerted a negative impact on predictive performance.
608 For example, in the quiescent state (E2), weak/strong repressive regions (ReprPCWk, E3; ReprPC, E4), and
609 weak transcriptional activity regions (TxWk, E7), prediction accuracy decreased by 2.86%–12.00%
610 compared with random SNPs under 2.5k–15k marker densities (**Figure 4C; Table 5**), suggesting that regions
611 with low activity, repressive function, or weak transcriptional activity may contribute little, or even introduce
612 noise, to GS prediction.
613
614 In summary, SNPs from the E6 (bivalent promoter/enhancer) region exhibited not only high predictive
615 performance within the focal population but also strong robustness in cross-population validation,
616 underscoring their potential as functional genetic markers for application in molecular breeding.
617
618 **Muscle bivalent promoter/enhancer (E6) state SNPs associated genes are significantly enriched in**
619 **growth-related pathways**
620 Given the significant advantage of SNPs within the muscle bivalent promoter/enhancer (E6) state in genomic
621 prediction, we further evaluated their direct association with body length to identify potential key regulatory
622 elements, functional genes, and related pathways. The results showed that the heritability of this trait was
623 0.44, indicating a moderately high level and strong potential for genetic improvement.
624

Uncorrected proof

A total of 15 073 high-quality SNPs located within the E6 state were included in the GWAS, and these SNPs showed a relatively even distribution across chromosomes (**Figure 5A**). In total, 751 SNPs showed significant associations with body length (**Figure 5B**). GO enrichment analysis indicated that the significant SNPs were predominantly enriched in biological processes related to muscle cell differentiation and development, chromatin modification, biosynthesis, gene expression regulation, and metabolic processes (**Figure 5C**). Enrichment of muscle development pathways underscores their central roles in muscle growth and functional maintenance (Braun & Gautel, 2011). The enrichment of chromatin modification pathways revealed the bridging role of epigenetic mechanisms between genetic variation and trait expression (Zhang et al., 2020). Meanwhile, the enrichment of metabolism-related pathways suggests that these SNPs may indirectly regulate growth performance by affecting cellular metabolism and energy supply (Stitt et al., 2010).

Integrating GWAS signals with gene annotation further identified several strong candidate genes associated with growth and their putative regulatory SNPs. For example, XM_027356510.1 (*MyoD1-like*) encodes a master transcription factor essential for myogenic determination (Vo et al., 2023). XM_027359869.1 (*ARPC4-like*) encodes a core subunit of the Arp2/3 complex, an actin polymerization factor involved in muscle fiber formation (Deng et al., 2015). XM_027375747.1 (*COMP/TSP-5-like*) belongs to the thrombospondin (TSP) family and regulates the development of the muscle–tendon junction (Chanana et al., 2007). Significant SNPs adjacent to these genes showed strong associations with body length in GWAS (**Figure 5D–F**), suggesting their important roles in growth regulation. Moreover, transcriptome analysis confirmed that these genes were widely expressed in muscle tissue, with significantly higher expression in the fast-growth (FG) group compared to the slow-growth (SG) group (**Figure 5G–I**), further supporting their functional roles in growth regulation.

In summary, functional SNPs within the E6 state not only performed well in genomic selection prediction but also revealed a series of growth-related genes and pathways through GWAS, confirming their potential value as functional markers for molecular breeding.

DISCUSSION

The use of functional genomic annotation to enhance genomic selection (GS) accuracy has long been considered a promising direction (Andersen et al., 2025; Johnston et al., 2024; R. Ma et al., 2025; Salem et

Uncorrected proof

al., 2024). However, practical applications remain scarce in aquaculture species, largely due to the lack of systematic and high-resolution functional annotations. This study provides one of the first systematic demonstrations of integrating histone modification–based functional annotations into genomic selection for shrimp growth traits. By leveraging CUT&Tag-derived high-resolution maps of four histone marks (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), we identified and annotated regulatory elements, and used them to guide SNP selection for GS of growth traits. Our findings demonstrate that epigenetic annotation–based SNP subsets improved prediction accuracy and stability, with statistically significant gains in multiple conditions, particularly SNPs located in the muscle-specific bivalent promoter/enhancer (E6) state. These SNPs not only achieved comparable or superior predictive accuracy under reduced marker densities, but also exhibited strong cross-population generalizability. This strategy provides a feasible path to lowering genotyping costs while maintaining or enhancing GS performance.

Genetic architecture and the need for functional SNP prioritization

The heritability estimates for body length (0.44) and body weight (0.29) indicated moderate genetic control, consistent with previous reports in aquatic species (Jerry et al., 2022; Yoshida et al., 2019; Y. Yu et al., 2019). Among the GS models tested, BayesA achieved the highest accuracy, corroborating earlier findings that it better captures polygenic architectures by allowing heterogeneous effect distributions (Daetwyler et al., 2010; Meuwissen et al., 2001).

Traditional GS relies on dense genome-wide SNP coverage, but the inclusion of numerous non-informative markers dilutes the effects of causal variants and rapidly diminishes the marginal benefits of increasing SNP density (Wray et al., 2013; Yáñez et al., 2023; Zenger et al., 2019). In this study, the gains in prediction accuracy diminished significantly beyond 10k SNPs, highlighting the potential of low-density panels optimized by functional annotation for cost-effective breeding applications.

Our results confirmed that SNPs within active regulatory regions such as promoters (H3K4me3, H3K27ac) and enhancers (H3K4me1) conferred moderate accuracy improvements at low densities, consistent with their role in transcriptional regulation. However, the gains from single histone marks were limited, as each modification reflects only a single dimension of chromatin regulation (Lennartsson & Ekwall, 2009; Peterson & Laniel, 2004; Zhang et al., 2020). In contrast, ChromHMM-based chromatin state annotation

Uncorrected proof

685 integrates multiple marks to partition the genome into biologically meaningful states, thereby improving
686 regulatory element identification and predictive utility.
687
688 **Tissue-specific chromatin states and GS performance**
689 GS accuracy is strongly influenced by the genetic relatedness between training and validation populations
690 and often declines in cross-population scenarios (Yáñez et al., 2023; Zenger et al., 2019). Identifying
691 biologically stable, functionally relevant SNPs is therefore essential. Previous studies have attempted to
692 incorporate functional annotations (e.g., genomic structural features) or to up-weight SNPs identified as
693 significant in GWAS, but these approaches generally rely on simplified functional partitions or limited prior
694 knowledge, which restricts their ability to comprehensively capture key regulatory variants (Lu et al., 2016;
695 Xu & Taylor, 2009; Yuan et al., 2025). Our analyses revealed that embryonic chromatin state–based SNPs
696 provided only modest improvements, likely due to developmental-stage differences from adult phenotypes.
697 In contrast, muscle chromatin state SNPs, particularly those in the E6 bivalent promoter/enhancer state,
698 exhibited remarkable predictive performance. At low-to-moderate densities, E6 SNPs achieved accuracy
699 comparable to or exceeding that of 15k random SNPs, and in the independent PLM population, they
700 improved prediction accuracy by 47.62%. This highlights the dual value of E6 SNPs in enhancing both cost-
701 efficiency and cross-population robustness. These results emphasize the importance of tissue-specific
702 regulatory information for identifying functional SNPs that enhance both cost-efficiency and cross-
703 population robustness.
704
705 **Biological relevance of E6 SNPs**
706 The superior performance of E6-state SNPs reflects both the intrinsic regulatory properties of bivalent
707 chromatin and the direct functional relevance of variants residing in these regions. Bivalent
708 promoter/enhancer elements marked simultaneously by H3K4me3 and H3K27me3 are well known to
709 maintain key developmental genes in a poised but responsive state, enabling rapid transcriptional activation
710 under environmental or physiological cues (Voigt et al., 2013; Z. Zeng et al., 2019). This chromatin
711 flexibility facilitates dynamic responses to stressors such as temperature fluctuations, hypoxia, and pathogen
712 exposure, conditions frequently encountered in aquaculture and known to influence growth performance
713 (Burgents et al., 2005; J. Wang et al., 2019; Z. Wang et al., 2019). Consistent with this regulatory potential,
714 our GWAS of 15 073 E6-state SNPs identified 751 significant variants associated with body length, and the

Uncorrected proof

associated genes were enriched in pathways related to muscle development, chromatin regulation, transcriptional control, and metabolic processes (**Figure 5C**), all of which are core biological functions underlying growth (Braun & Gautel, 2011; Zhang et al., 2020; Stitt et al., 2010). Moreover, major candidates such as MyoD1-like, ARPC4-like, and TSP-5-like not only carried significant E6 variants but also showed higher transcript level in fast-growing individuals (**Figure 5G–I**), providing strong evidence that E6-state SNPs capture regulatory polymorphisms influencing core growth genes.

While these results highlight the biological importance of E6 regions, it is critical to distinguish between statistical association and true biological causation. Given the extent of linkage disequilibrium (LD) in the genome, it is possible that a portion of these prioritized E6 SNPs simply act as effective tags for nearby ungenotyped causal variants. However, several lines of evidence in our study strongly suggest that the E6 subset is genuinely enriched for true causal regulatory variants rather than merely capturing population-specific LD proxies. First, as detailed above, the remarkably high proportion of significant associations in our targeted GWAS and their specific enrichment in core growth-related pathways point toward direct functional roles. Second, because LD patterns and marker-causal phase typically diverge across distinct populations, the substantial predictive improvement (47.62%) maintained by E6 SNPs in the genetically distant PLM strain indicates a strong reduction in LD dependency. Regardless of whether they are directly causal or act via tight LD, prioritizing SNPs based on these functional annotations successfully concentrates the predictive signals.

Importantly, the predictive advantage of E6-state SNPs was model-independent. Across four genomic prediction models, specifically BayesA, BayesB, BayesC and GBLUP, E6 SNPs consistently outperformed randomly selected SNPs across all marker densities, indicating that their improvement does not rely on the statistical assumptions of any single model. Instead, the enhanced accuracy stems from the underlying biological informativeness of E6-state regulatory elements, highlighting the robustness of chromatin-state–guided SNP prioritization across analytical frameworks and its practical utility for breeding applications.

Building upon this mechanistic and empirical foundation, we propose an epigenome-guided SNP prioritization framework (**Figure 6**), which integrates multiple histone modification profiles to define key regulatory chromatin states and systematically enrich SNPs within active promoters, enhancers, and bivalent

Uncorrected proof

regions. By focusing on genomic segments with high regulatory potential, this framework ensures effective coverage of core functional variants while substantially reducing marker density requirements, providing a biologically grounded and cross-population–robust strategy for enhancing genomic prediction in aquaculture breeding.

Broader implications for livestock genomic prediction

Although validated here in shrimp, our findings support a broader principle that epigenome-informed regulatory SNP prioritization can enhance genomic prediction across farmed animal species. In livestock species like pigs and cattle, large-scale functional annotation projects (e.g., FAANG) have revealed that causal variants for complex traits are disproportionately enriched in regulatory regions rather than coding sequences (Pan et al., 2021; Xiang et al., 2025). Our results extend this understanding by highlighting the specific utility of bivalent domains (marked by H3K4me3 and H3K27me3). Biologically, these domains maintain developmental genes in a "poised" state, allowing for rapid activation in response to environmental stimuli (Salem et al., 2024; Voigt et al., 2013), a mechanism that is evolutionarily conserved across metazoans (Boyle et al., 2014; Fu et al., 2025).

In livestock production, animals frequently face variable stressors such as heat stress in dairy cows or pathogen exposure in poultry. We propose that genetic variations within these bivalent regions may capture key genotype-by-environment (GxE) interactions that are often missed by neutral markers. This hypothesis is strongly supported by recent findings in cattle, where integrating functional annotations significantly enhanced genomic prediction for climate resilience traits like heat tolerance (Xiang et al., 2025). Therefore, the strategy of prioritizing bivalent SNPs could be particularly valuable for developing robust breeding lines in livestock that maintain high performance across diverse rearing environments. Importantly, the predictive utility of epigenetically prioritized SNPs is likely to be trait- and tissue-dependent, as functional elements identified in the most relevant tissues (e.g., muscle for growth, adipose for fat deposition and immune-related tissues for disease resistance) have been shown to yield the greatest benefit in genomic prediction models (R. Ma et al., 2025; Yuan et al., 2025; Pan et al., 2021; Xiang et al., 2025). Consistent with this view, our ongoing work on pigmentation traits in fish indicates that skin-derived regulatory annotations outperform those from other tissues, underscoring the importance of context-specific epigenomic prioritization.

Uncorrected proof

Overall, this epigenome-informed prioritization strategy offers a practical route toward designing low-density, functionally enriched SNP panels that reduce genotyping costs while maintaining prediction accuracy across generations and populations. As functional genomic resources continue to expand in livestock species, integrating chromatin-state annotations into GS frameworks may provide an effective and scalable solution for improving breeding efficiency in large-scale animal agriculture.

Limitations and future directions

Although this study generated high-resolution histone modification maps and demonstrated their application to GS, several aspects remain open for future improvement. In this study, we prioritized four histone marks (H3K4me1, H3K4me3, H3K27me3, and H3K27ac) and one adult tissue (muscle), as these modifications are well-established indicators of promoter and enhancer activity and are closely linked to growth regulation. Although this design provides a robust foundation for functional annotation, it may not fully capture the broader regulatory complexity across diverse tissues and developmental stages. Expanding to additional histone marks (e.g., H3K36me3, H3K9me3) and other functional layers such as ATAC-seq or Hi-C would further refine functional annotations. Second, while our prioritized SNPs showed strong predictive and cross-population performance, experimental validation of their causal roles will be an important next step. Finally, the general applicability of this strategy to other traits (such as disease resistance) and aquaculture species remains to be systematically assessed.

Additionally, while our internal cross-validation demonstrated the superiority of E6-state SNPs, we acknowledge that screening multiple chromatin states to identify the optimal feature set may introduce a degree of selection bias, potentially inflating internal accuracy estimates. However, this theoretical limitation is strongly mitigated by our experimental design of utilizing the independent PLM population as a true, untouched external validation set. The substantial 47.62% predictive improvement observed in the completely independent PLM strain confirms the robust and transferable predictive power of the E6 subset, demonstrating that our findings are driven by underlying biological mechanisms rather than statistical overfitting.

Future work should therefore focus on integrating multi-omics datasets, extending analyses to diverse tissues and developmental stages, and testing the transferability of this approach across species and traits. To

Uncorrected proof

translate these findings into practice, the prioritized E6-state SNPs can be directly utilized to design cost-effective, low-density genotyping platforms (e.g., targeted amplicon sequencing) or spiked into existing commercial SNP arrays to balance trait-specific precision with overall genome-wide coverage. Their subsequent validation in commercial breeding programs to quantify realized genetic gains will further accelerate the transition toward functionally informed GS in aquaculture.

CONCLUSIONS

In summary, this study establishes the first chromatin state–guided framework for SNP prioritization in GS of shrimp. By integrating multiple histone modification signals, we identified functional SNP subsets that improve predictive accuracy, reduce dependence on high-density genotyping, and enhance cross-population stability. The strong performance of E6 SNPs underscores the central role of bivalent promoter/enhancer regions in growth regulation and environmental adaptability. This work provides a theoretical and methodological basis for building "epigenetically informed" marker systems, advancing molecular breeding in aquaculture from traditional structural markers toward functional, biologically grounded markers. Future work should focus on validating these prioritized SNP panels in real-world breeding programs.

ETHICS DECLARATIONS

Protocols for animal use in this study were approved by the Institutional Animal Care and Use Committee (IACUC) of Ocean University of China.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The resequencing datasets analyzed in this study were generated as part of ongoing selective breeding programs in our laboratory and are not publicly available due to their proprietary nature. These data may be made available from the corresponding author upon reasonable request. The CUT&Tag datasets used in this study include two categories: (i) seven embryonic developmental stages (blastula, gastrula, limb bud embryo, larva in membrane, nauplius I, nauplius III, and nauplius VI), which are publicly available from our previous publication (NCBI BioProject accession: PRJNA1224423) (Shi et al., 2025b); and (ii) newly generated adult

Uncorrected proof

muscle tissue data, which have been deposited in Figshare (https://figshare.com/) with the link (https://figshare.com/s/a9b187bdd1f6fe2cc523). In addition, building upon these datasets, we generated CUT&Tag data for four later larval developmental stages, specifically zoea I, zoea III, mysis I, and mysis III. These newly generated datasets have also been submitted to the NCBI BioProject database under accession number PRJNA1304883. Meanwhile, all newly generated datasets in this study have been deposited in the Genome Sequence Archive at the National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/gsa) under accession number PRJCA064872, and the Science Data Bank (CSTR:31253.11.sciencedb.j00139.00315). All newly generated datasets will be made publicly available upon publication of this article. A comprehensive summary of all datasets used in this study, including newly sequenced data and those downloaded from previous studies, is provided in **Supplementary Table S16**.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was funded by the Project of Sanya Yazhou Bay Science and Technology City (SCKJ-JYRC-2023-62), National Natural Science Foundation of China (32200667), and Shandong Provincial Special Funds for Taishan Scholars (tsqn202306104).

AUTHORS' CONTRIBUTIONS

J.S. and Z.Y. conceived and designed the study. J.S. performed the methodology development, investigation, formal analysis, data curation, software implementation, validation, and prepared the original draft. Z.L., M.S., and M.Y. contributed to methodology and investigation. M.M. and D.Z. assisted with visualization and manuscript revision. Z.B. provided supervision, project administration, and resources. Q.Z. and J.H. contributed resources and participated in manuscript editing. Z.Y. supervised the project, acquired funding, and led the writing, review, and editing process. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

Uncorrected proof

We acknowledge the support of the High-Performance Biological Supercomputing Center at the Ocean University of China for this research.

During the preparation of this work the authors used ChatGPT v.5.2 (OpenAI) for assistance in language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Alemu SW, Lopdell TJ, Trevarton AJ, et al. 2025. Comparison of genomic prediction accuracies in dairy cattle lactation traits using five classes of functional variants versus generic SNP. *Genetics Selection Evolution,* **57**(1): 20.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**(1): 246.

Andersen LK, Thompson NF, Abernathy JW, et al. 2025. Advancing genetic improvement in the omics era: Status and priorities for United States aquaculture. *BMC Genomics*, **26**(1): 155.

Boyle AP, Araya CL, Brdlik C, et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**(7515): 453–456.

Braun T, Gautel M. 2011. Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nature Reviews Molecular Cell Biology*, **12**(6): 349–361.

Burgents JE, Burnett KG, Burnett LE. 2005. Effects of hypoxia and hypercapnic hypoxia on the localization and the elimination of *Vibrio campbellii* in *Litopenaeus vannamei*, the Pacific white shrimp. *The Biological Bulletin*, **208**(3): 159–168.

Chanana B, Graf R, Koledachkina T, et al. 2007. αPS2 integrin-mediated muscle attachment in Drosophila requires the ECM protein Thrombospondin. *Mechanisms of Development*, **124**(6): 463–475.

Covarrubias-Pazaran G. 2016. Genome-assisted prediction of quantitative traits using the R package sommer. *PLOS ONE*, **11**(6): e0156744.

Daetwyler HD, Pong-Wong R, Villanueva B, et al. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, **185**(3): 1021–1031.

Deng S, Bothe I, Baylies MK. 2015. The formin diaphanous regulates myoblast fusion through actin polymerization and Arp2/3 regulation. *PLOS Genetics*, **11**(8): e1005381.

Uncorrected proof

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, **12**(12): 2478–2492.

Feng J, Liu T, Qin B, et al. 2012. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, **7**(9): 1728–1740.

Fu X, Mo S, Buendia A, et al. 2025. A foundation model of transcription across human cell types. *Nature*, **637**(8047): 965–973.

Georges M, Charlier C, Hayes B. 2019. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*, **20**(3): 135–156.

Gu Z. 2022. Complex heatmap visualization. *iMeta*, **1**(3): e43.

Jerry DR, Jones DB, Lillehammer M, et al. 2022. Predicted strong genetic gains from the application of genomic selection to improve growth related traits in barramundi (*Lates calcarifer*). *Aquaculture*, **549**: 737761.

Johnston IA, Kent MP, Boudinot P, et al. 2024. Advancing fish breeding in aquaculture through genome functional annotation. *Aquaculture*, **583**: 740589.

Kaya-Okur HS, Wu SJ, Codomo CA, et al. 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, **10**(1): 1930.

Lawrence M, Huber W, Pagès H, et al. 2013. Software for computing and annotating genomic ranges. *PLOS Computational Biology*, **9**(8): e1003118.

Lennartsson A, Ekwall K. 2009. Histone modification patterns and epigenetic codes. *Biochimica et Biophysica Acta (BBA) – General Subjects*, **1790**(9): 863–868.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint, arXiv*:1303.3997.

Li H, Handsaker B, Wysoker A, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16): 2078–2079.

Lu Q, Yao X, Hu Y, et al. 2016. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics*, **32**(4): 542–548.

Lund MS, Su G, Janss L, et al. 2014. Genomic evaluation of cattle in a multi-breed context. *Livestock Science*, **166**: 101–110.

Ma R, Kuang R, Zhang J, et al. 2025. Annotation and assessment of functional variants in livestock through epigenomic data. *Journal of Genetics and Genomics*, **52**(12): 1588–1599.

Uncorrected proof

925 Ma W, Zheng W, Qin S, et al. 2025. DeepAnnotation: A novel interpretable deep learning–based genomic
926 selection model that integrates comprehensive functional annotations. *GigaScience*, **14**: giaf083.
927 MacLeod I, Bowman P, Vander Jagt C, et al. 2016. Exploiting biological priors and sequence variants
928 enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*, **17**(1): 144.
929 McKenna A, Hanna M, Banks E, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for
930 analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9): 1297–1303.
931 Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense
932 marker maps. *Genetics*, **157**(4): 1819–1829.
933 Pan Z, Yao Y, Yin H, et al. 2021. Pig genome functional annotation enhances the biological interpretation
934 of complex traits and human disease. *Nature Communications*, **12**(1): 5848.
935 Patel RK, Jain M. 2012. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data.
936 *PLOS ONE*, **7**(2): e30619.
937 Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package.
938 *Genetics*, **198**(2): 483–495.
939 Peterson CL, Laniel M-A. 2004. Histones and histone modifications. *Current Biology*, **14**(14): R546–R551.
940 Purcell S, Neale B, Todd-Brown K, et al. 2007. PLINK: A tool set for whole-genome association and
941 population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3): 559–575.
942 Qi Z, Liu J, Shi J, et al. 2025. Integrated transcriptomic and epigenomic analysis reveals mechanisms
943 underlying melanotic spot formation in red tilapia (*Oreochromis spp.*). *International Journal of Molecular*
944 *Sciences*, **26**(9): 4370.
945 Rubinacci S, Ribeiro DM, Hofmeister RJ, et al. 2021. Efficient phasing and imputation of low-coverage
946 sequencing data using large reference panels. *Nature Genetics*, **53**(1): 120–126.
947 Salem M, Al-Tobasei R, Ali A, et al. 2024. Functional annotation of regulatory elements in rainbow trout
948 uncovers roles of the epigenome in genetic selection and genome evolution. *GigaScience*, **13**: giae092.
949 Santos CA, Andrade SCS, Teixeira AK, et al. 2021. Transcriptome differential expression analysis reveals
950 the activated genes in *Litopenaeus vannamei* shrimp families of superior growth performance. *Aquaculture*,
951 **531**: 735871.
952 Santure AW, Garant D. 2018. Wild GWAS–association mapping in natural populations. *Molecular Ecology*
953 *Resources*, **18**(4): 729–738.
954 Shi J, Liu J, Wei Z, et al. 2025a. Epigenetic regulation and maternal-to-zygotic transition in dwarf surf clam

Uncorrected proof

(*Mulinia lateralis*): Insights from chromatin state profiling and transcriptomics. *Zoological Research*, **46**(5): 1009–1028.

Shi J, Qi Z, Yin M, et al. 2025b. Dynamic epigenomic landscape and gene regulatory networks during embryonic development in Pacific white shrimp (*Litopenaeus vannamei*) as revealed by histone modification profiling using CUT&Tag. *Epigenetics & Chromatin*, **18**(1): 50.

Stitt M, Sulpice R, Keurentjes J. 2010. Metabolic networks: How to identify key components in the regulation of metabolism and growth. *Plant Physiology*, **152**(2): 428–444.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**: 4414–4423.

Vo L, Schmidtke MW, Rosa-Junior NTD, et al. 2023. Cardiolipin metabolism regulates expression of muscle transcription factor MyoD1 and muscle development. *Journal of Biological Chemistry*, **299**(3): 102978.

Voigt P, Tee W-W, Reinberg D. 2013. A double take on bivalent promoters. *Genes & Development*, **27**(12): 1318–1338.

Wang J, Huang Y, Xu K, et al. 2019. White spot syndrome virus (WSSV) infection impacts intestinal microbiota composition and function in *Litopenaeus vannamei*. *Fish & Shellfish Immunology*, **84**: 130–137.

Wang X, Xu Y, Hu Z, et al. 2018. Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal*, **6**(4): 330–340.

Wang Z, Qu Y, Yan M, et al. 2019. Physiological responses of Pacific white shrimp *Litopenaeus vannamei* to temperature fluctuation in low-salinity water. *Frontiers in Physiology*, **10**: 1025.

Wickham H. 2011. Ggplot2. Wiley Interdisciplinary Reviews: *Computational Statistics*, **3**(2): 180–185.

Wray NR, Yang J, Hayes BJ, et al. 2013. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, **14**(7): 507–515.

Xiang R, Breen E, Bolormaa S, et al. 2025. Integrating extensive functional annotations and multiomics of cattle enhances climate resilience prediction and mapping. *Proceedings of the National Academy of Sciences*, **122**(49): e2514736122.

Xu Z, Taylor JA. 2009. SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Research*, **37**(Suppl 2): W600–W605.

Yáñez JM, Barría A, López ME, et al. 2023. Genome-wide association and genomic selection in aquaculture. *Reviews in Aquaculture*, **15**(2): 645–675.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis.

Uncorrected proof

985 *The American Journal of Human Genetics*, **88**(1): 76–82.
986 Yin L, Zhang H, Tang Z, et al. 2021. rMVP: A memory-efficient, visualization-enhanced, and parallel-
987 accelerated tool for genome-wide association study. Genomics, *Proteomics & Bioinformatics*, **19**(4): 619–
988 628.
989 Yoshida GM, Lhorente JP, Correa K, et al. 2019. Genome-wide association study and cost-efficient genomic
990 predictions for growth and fillet yield in Nile tilapia (*Oreochromis niloticus*). *G3: Genes, Genomes, Genetics*,
991 **9**(8): 2597–2607.
992 Yu G, Wang L-G, Han Y, et al. 2012. ClusterProfiler: An R package for comparing biological themes among
993 gene clusters. *Omics: A Journal of Integrative Biology*, **16**(5): 284–287.
994 Yu G, Wang L-G, He Q-Y. 2015. ChIPseeker: An R/Bioconductor package for ChIP peak annotation,
995 comparison and visualization. *Bioinformatics*, **31**(14): 2382–2383.
996 Yu Y, Wang Q, Zhang Q, et al. 2019. Genome scan for genomic regions and genes associated with growth
997 trait in Pacific white shrimp *Litopenaeus vannamei*. *Marine Biotechnology*, **21**(3): 374–383.
998 Yuan C, Gillon A, Gualdrón Duarte JL, et al. 2025. Evaluation of genomic selection models using whole
999 genome sequence data and functional annotation in Belgian Blue cattle. *Genetics Selection Evolution*, **57**(1):
1000 10.
1001 Zeng Q, Zhao B, Wang H, et al. 2022. Aquaculture Molecular Breeding Platform (AMBP): A comprehensive
1002 web server for genotype imputation and genetic analysis in aquaculture. *Nucleic Acids Research*, **50**(W1):
1003 W66–W74.
1004 Zeng Z, Zhang W, Marand AP, et al. 2019. Cold stress induces enhanced chromatin accessibility and bivalent
1005 histone modifications H3K4me3 and H3K27me3 of active genes in potato. *Genome Biology*, **20**(1): 123.
1006 Zenger KR, Khatkar MS, Jones DB, et al. 2019. Genomic selection in aquaculture: Application, limitations
1007 and opportunities with special reference to marine shrimp and pearl oysters. *Frontiers in Genetics*, **9**: 693.
1008 Zhang Y, Sun Z, Jia J, et al. 2020. Overview of histone modification. *Histone Mutations and Cancer*: 1–16.
1009 Zhang Y, Zhuang Z, Liu Y, et al. 2024. Genomic prediction based on preselected single-nucleotide
1010 polymorphisms from genome-wide association study and imputed whole-genome sequence data annotation
1011 for growth traits in Duroc pigs. *Evolutionary Applications*, **17**(2): e13651.
1012 Zhao Y, Hou Y, Xu Y, et al. 2021. A compendium and comparative epigenomics analysis of cis-regulatory
1013 elements in the pig genome. *Nature Communications*, **12**(1): 2217.
1014 Zheng Z, Liu S, Sidorenko J, et al. 2024. Leveraging functional genomic annotations and genome coverage

Uncorrected proof

to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics*, **56**(5): 767–777.

Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, **11**(4): 407–409.

Zhuang Z, Wu J, Qiu Y, et al. 2023. Improving the accuracy of genomic prediction for meat quality traits using whole genome sequence data in pigs. *Journal of Animal Science and Biotechnology*, **14**(1): 67.

uncorrected proof

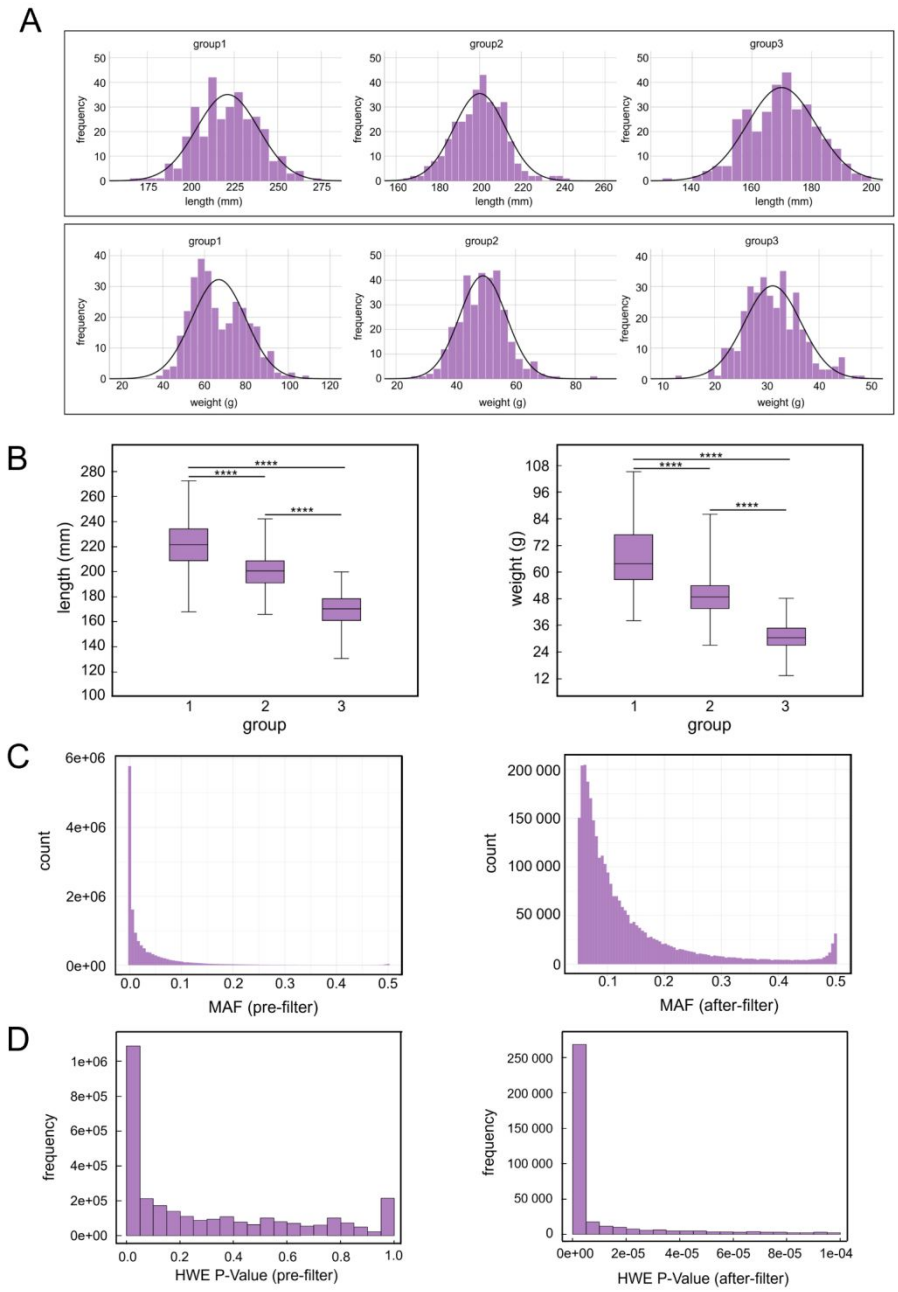


Figure 1. Phenotypic distributions and genotypic quality control for growth traits in selective breeding populations of *Litopenaeus vannamei*

A: Frequency distributions of body length (top row) and body weight (bottom row) within each of the three populations. B: Boxplots of body length and body weight for the three populations in the breeding program. Asterisks denote statistically significant differences between groups (Wilcoxon rank-sum test, **** $P<0.0001$). C: Allele frequency distributions before (left) and after (right) filtering for minor allele frequency (MAF)>0.05. D: Distributions of Hardy–Weinberg equilibrium (HWE) *P*-values before (left) and after (right)

Zoological Research

filtering.

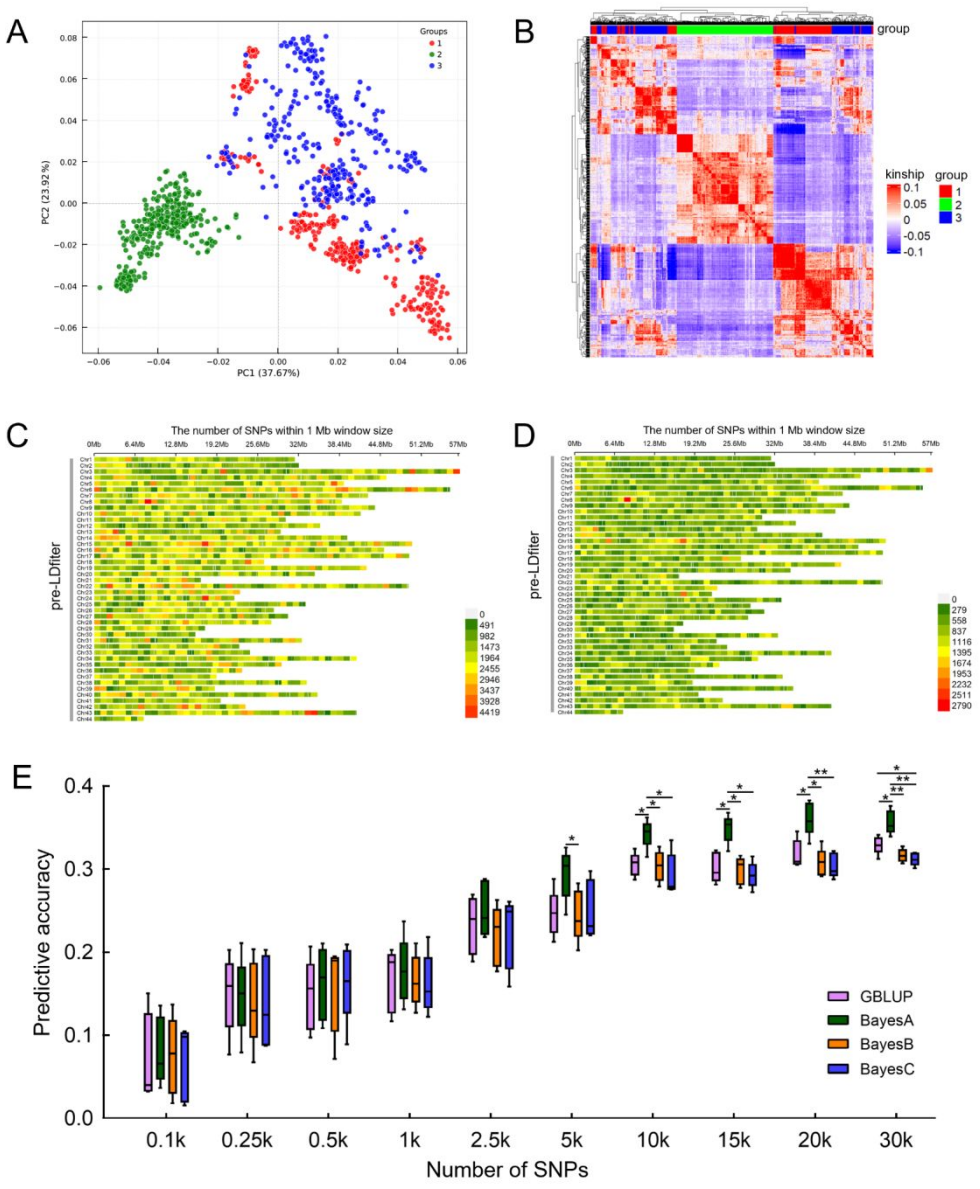


Figure 2. Effects of SNP density and training population size on genomic selection (GS) accuracy for body length in *Litopenaeus vannamei*

A: Principal component analysis (PCA) based on genome-wide SNPs, showing clear population stratification along PC1 and PC2. B: Heatmap of the genomic relationship matrix illustrating pairwise relatedness among individuals. C–D: SNP density plots across chromosomes before and after linkage disequilibrium (LD) pruning, showing even genomic coverage after pruning. E: Boxplots of predictive accuracy of genomic estimated breeding values (GEBVs) across four GS models (GBLUP, BayesA, BayesB,

Zoological Research

and BayesC) at different SNP densities. For each density, SNPs were randomly sampled five times and assessed using fivefold cross-validation repeated five times. Accuracy increased with SNP density, showing only marginal improvements beyond 10k SNPs, with BayesA consistently achieving the highest accuracy.

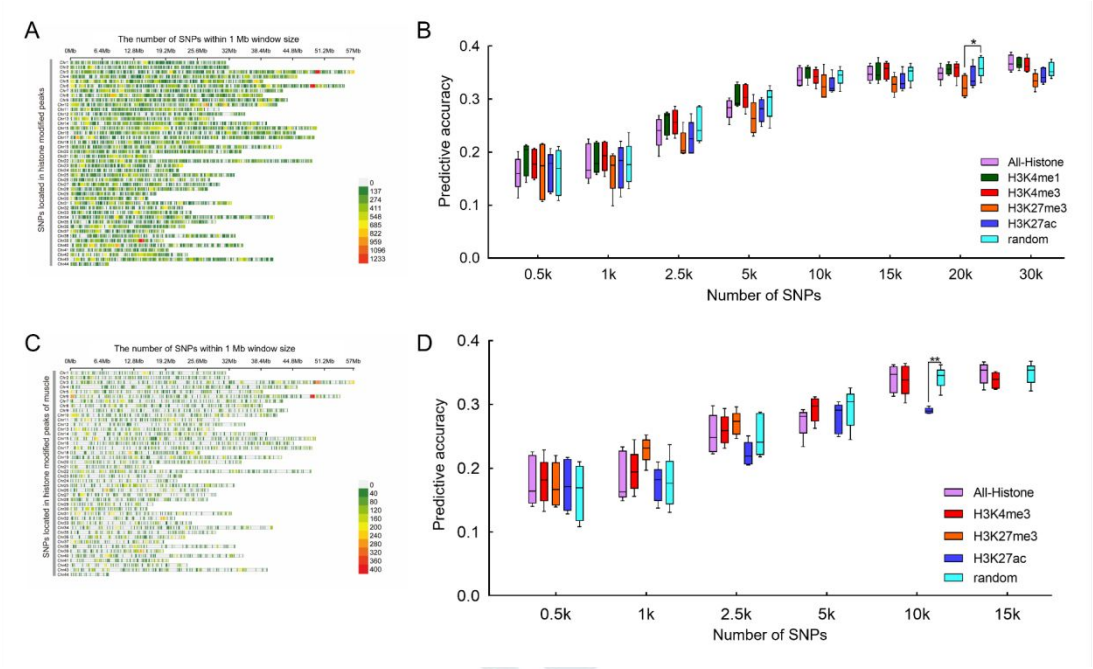


Figure 3. Genomic distribution and predictive performance of histone modification–based SNPs in genomic selection (GS) for growth traits of *Litopenaeus vannamei*

A: Genomic distribution of SNPs located within histone modification peaks from embryonic stages and adult muscle, showing broad and even coverage. B: Predictive accuracy of SNP subsets from histone modification peaks of embryonic and muscle tissues at varying marker densities, compared with randomly selected SNPs. For each density, SNPs were sampled five times, and predictive accuracy was assessed with BayesA using five rounds of fivefold cross-validation (25 replicates per density). C: Genomic distribution of SNPs within histone modification peaks from adult muscle tissue, showing uniform chromosomal coverage. D: Predictive accuracy of SNP subsets from muscle histone modification peaks compared with random SNPs at the same density. Evaluation procedure was as in (B). Note that H3K4me1 subset is absent from panel D because it contained only 163 SNPs in adult muscle, restricting its evaluation to a maximum density of 0.1k (details provided in **Supplementary Table S12**). Significance levels: $^{*}P<0.05$, $^{**}P<0.01$.

Zoological Research

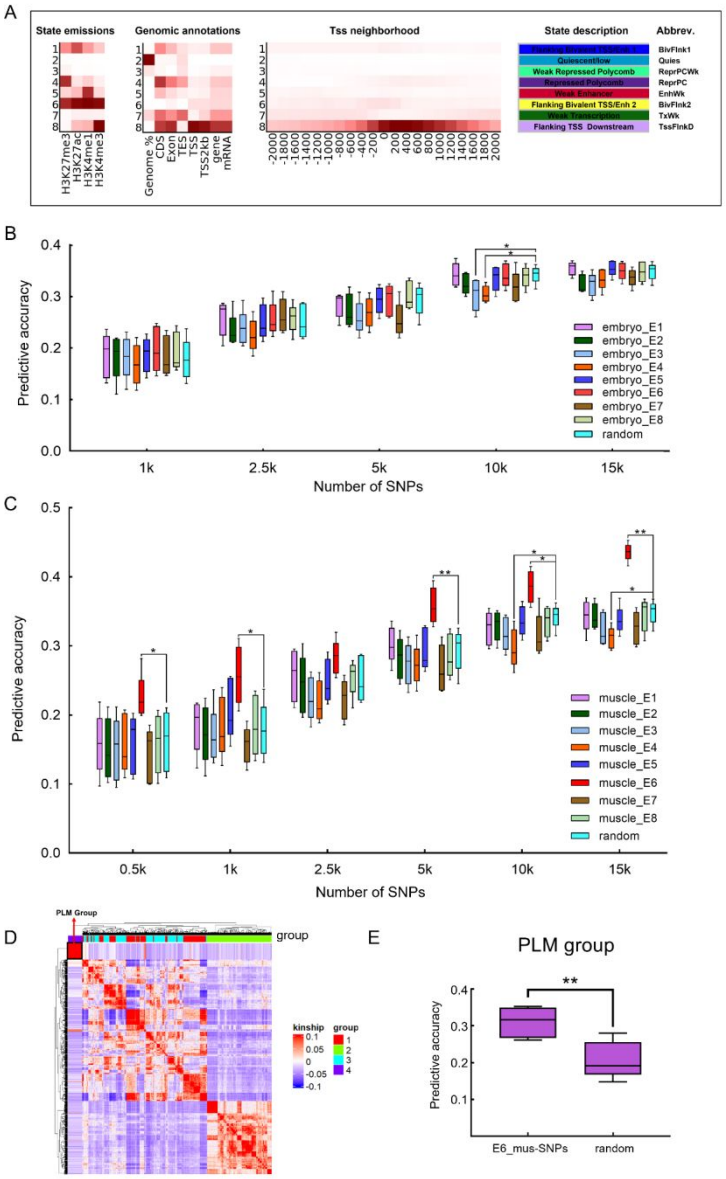


Figure 4. Chromatin state–based SNP annotation reveals tissue-specific predictive performance and cross-population robustness in genomic selection (GS) for *Litopenaeus vannamei* growth traits

A: Chromatin state annotation based on H3K4me1, H3K4me3, H3K27me3, and H3K27ac marks from the embryonic and muscle tissues. The left panel shows the emission parameters of the ChromHMM 8-state model, where color intensity represents the probability of each histone modification occurring within a given state. The middle panel depicts state enrichments across external genomic annotations, with color gradients indicating enrichment levels. The right panel illustrates fold enrichment of chromatin states within±2 kb of transcription start sites (TSSs) at 200 bp resolution, with deeper red reflecting stronger enrichment. The state descriptions and abbreviations are summarized in the right panel. B: Predictive accuracy of SNPs located in embryonic chromatin states across varying marker densities compared with randomly sampled SNPs. C:

Zoological Research

Predictive accuracy of SNPs located in muscle chromatin states. SNPs within the muscle bivalent promoter/enhancer state (E6) consistently outperformed random SNPs across densities, achieving significantly higher predictive accuracy. D: Kinship matrix showing genetic divergence between the training population and the independent PLM validation population. E: Cross-population validation of muscle E6 SNPs in the PLM population. Predictive accuracy was compared with equal-sized sets of random SNPs. Models were trained with fivefold cross-validation in the reference population, with PLM as the validation set. E6 SNPs improved accuracy by 47.62% (from 0.21±0.05 (95% CI: 0.166–0.254) to 0.31±0.04 (95% CI: 0.275–0.345)). Significance levels: $^{*}P<0.05$, $^{**}P<0.01$.

uncorrected proof

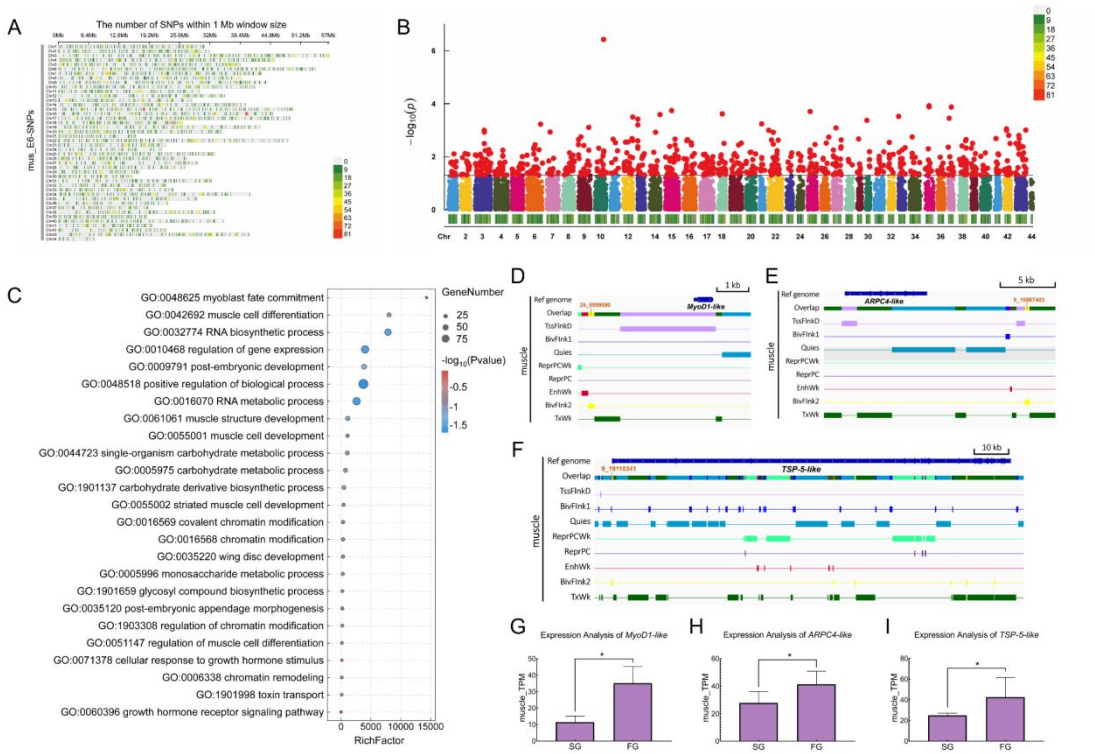


Figure 5. Genome-wide association analysis (GWAS) of SNPs located in the muscle bivalent promoter/enhancer (E6) state

A: Genomic distribution of 15 073 high-quality E6 SNPs across chromosomes, showing broad and even coverage. B: Manhattan plot of GWAS for body length, with the horizontal line indicating the significance threshold (P<0.05). Red dots represent significant SNPs, and chromosome density is shown at the bottom. C: GO functional enrichment of genes annotated from significant E6 SNPs. Bubble size represents the number of enriched genes, and color intensity reflects statistical significance (P<0.05). Results indicate enrichment in multiple growth-related pathways. D–F: The ChromHMM IGV visualization results of loci 25_8999590, 22_15987403, and 9_19110341, all of which are located within the E6 state region. G–I: RNA levels of *MyoD1-like*, *ARPC4-like*, and *TSP-5-like* in muscle tissues of fast-growth (FG) and slow-growth (SG) groups, with significantly higher level in FG shrimp (P<0.05).

Uncorrected

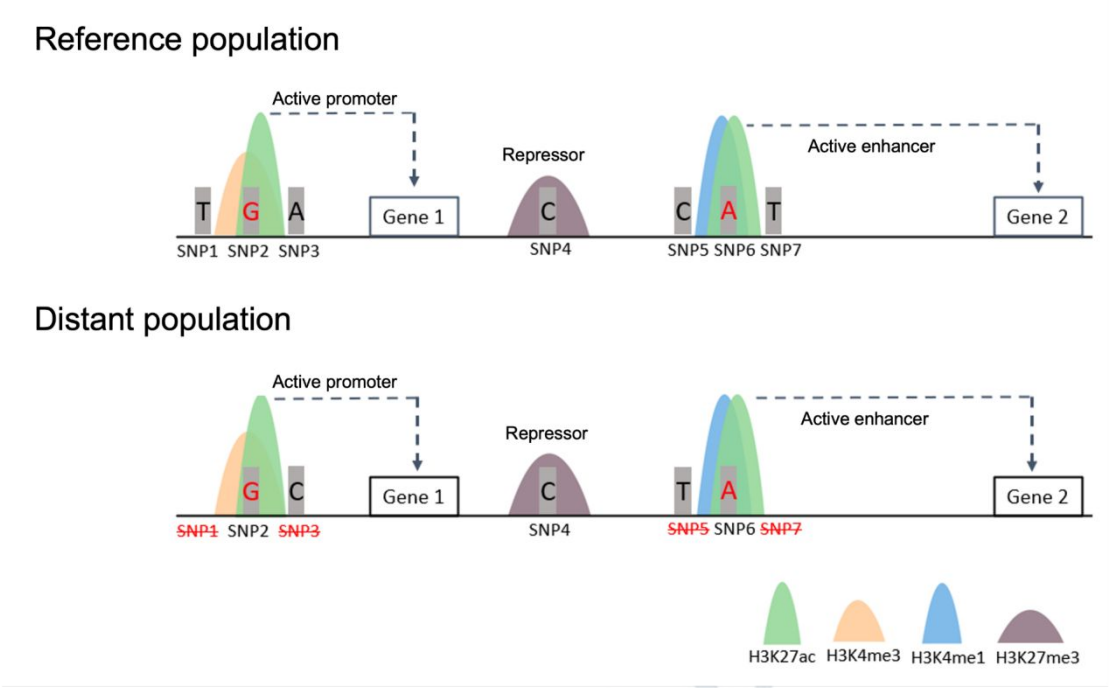


Figure 6. Conceptual framework of histone modification–guided SNP prioritization for improving genomic prediction and cross-population robustness. By integrating chromatin states defined by histone modifications (H3K27ac, H3K4me3, H3K4me1, and H3K27me3), active promoters, enhancers, repressive regions, and bivalent elements can be precisely annotated. SNPs located within these regulatory regions are prioritized as functionally relevant markers, ensuring effective coverage of key regulatory elements. This strategy not only improves the accuracy of genomic selection within populations but also enhances prediction stability and robustness across genetically distant populations, as a higher proportion of phenotype-associated causal SNPs are retained.

uncorrected

1100 Table 1. Phenotypic statistics of growth traits in *L. vannamei.*

Group	Sample size	Length (mm) (Mean±SEM)	Weight (g) (Mean±SEM)
1	313	221.14±1.01	66.79±0.73
2	331	200.18±0.68	49.27±0.43
3	328	169.95±0.64	31.18±0.30
PLM	73	133.82±0.10	18.90±0.44

1101

1102 Table 2. Number and proportion of SNPs located within histone modification peaks.

Stage	Histone modification	Total SNPs	SNPs located in peaks	Proportion
Embryo development stages and adult muscle tissue	H3K4me1	2 609 549	123 174	4.72%
	H3K4me3		191 517	7.34%
	H3K27me3		56 172	2.15%
	H3K27ac		34 878	1.34%
	All-Histone		306 487	11.47%
Adult muscle tissue	H3K4me1	2 609 549	163	0.01%
	H3K4me3		62 192	2.38%
	H3K27me3		3 263	0.13%
	H3K27ac		11 381	0.44%
	All-Histone		70 480	2.70%

1103

uncorrected proof

1104 Table 3. Number and proportion of SNPs located within different chromatin state regions in embryonic
1105 tissues.

Chromatin state		Total SNPs	SNPs in chromatin state regions	Proportion
E1	BivFlnk1	2 609 549	201 729	7.73%
E2	Quies		2 351 162	90.10%
E3	ReprPCWk		828 478	31.75%
E4	ReprPC		238 452	9.14%
E5	EnhWk		277 111	10.62%
E6	BivFlnk2		107 027	4.10%
E7	TxWk		1 080 973	41.42%
E8	TssFlnkD		312 303	11.97%

1106

1107 Table 4. Number and proportion of SNPs located within different chromatin state regions of adult shrimp
1108 muscle tissue.

Chromatin state		Total SNPs	SNPs in chromatin state regions	Proportion
E1	BivFlnk1	2 609 549	44 299	1.70%
E2	Quies		1 689 212	64.73%
E3	ReprPCWk		207 586	7.95%
E4	ReprPC		39 606	1.52%
E5	EnhWk		17 290	0.66%
E6	BivFlnk2		15 073	0.58%
E7	TxWk		467 392	17.91%
E8	TssFlnkD		129 889	4.98%

1109

1110

uncorrected proof

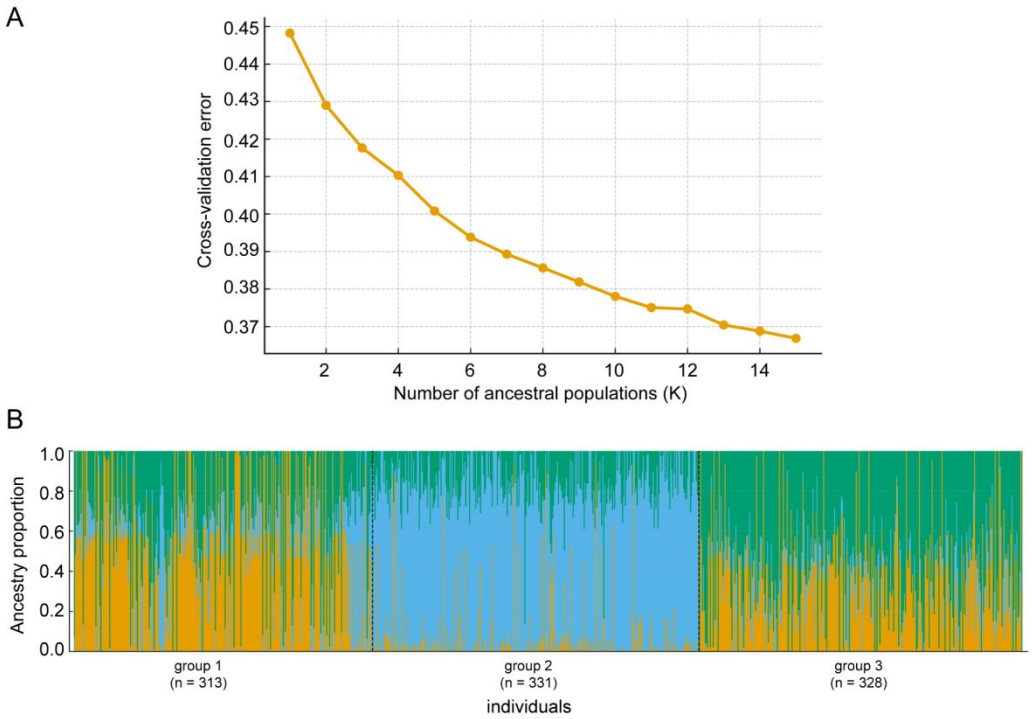
Table 5. Predictive accuracy of SNPs located in adult muscle chromatin states across varying marker densities compared with randomly sampled SNPs.

Density	Random	E1	E2	E3	E4	E5	E6	E7	E8
0.5k	0.16±0.04	0.16±0.04 (0%)	0.15±0.04 (–6.25%)	0.15±0.05 (–6.25%)	0.16±0.04 (0%)	0.16±0.04 (0%)	0.23±0.03* (+43.75%)	0.14±0.04 (–12.5%)	0.16±0.04 (0%)
1k	0.18±0.04	0.19±0.04 (+5.56%)	0.17±0.04 (–5.56%)	0.17±0.04 (–5.56%)	0.18±0.04 (0%)	0.21±0.04 (+16.67%)	0.26±0.04* (+44.44%)	0.16±0.03 (–11.11%)	0.18±0.04 (0%)
2.5k	0.25±0.03	0.26±0.04 (+3.85%)	0.24±0.04 (–4.00%)	0.22±0.03 (–12.00%)	0.22±0.03 (–12.00%)	0.25±0.03 (0%)	0.28±0.03 (+12.00%)	0.22±0.03 (–12.00%)	0.26±0.03 (+4.00%)
5k	0.29±0.03	0.3±0.03 (+3.45%)	0.28±0.03 (–3.45%)	0.27±0.03 (–6.90%)	0.27±0.03 (–6.90%)	0.29±0.03 (0%)	0.36±0.03** (+24.14%)	0.27±0.03 (–6.90%)	0.29±0.03 (0%)
10k	0.34±0.02	0.33±0.02 (–2.94%)	0.33±0.02 (–2.94%)	0.32±0.02 (–5.88%)	0.3±0.03* (–11.76%)	0.34±0.02 (0%)	0.39±0.02* (+14.71%)	0.32±0.03 (–5.88%)	0.34±0.02 (0%)
15k	0.35±0.02	0.35±0.02 (0%)	0.34±0.02 (–2.86%)	0.32±0.02 (–8.57%)	0.31±0.02* (–11.43%)	0.34±0.02 (–2.86%)	0.44±0.01** (+25.71%)	0.33±0.02 (–5.71%)	0.35±0.02 (0%)

Note: Breeding value prediction accuracy is presented as mean±SD. Percent values in parentheses indicate the relative change in accuracy (increase "+" or decrease "–") compared with randomly sampled SNPs at the same SNP density. The chromatin states (E1–E8) were identified from the combined chromatin-state peaks detected across muscle tissue of adult shrimp. Statistical significance: * $P<0.05$, ** $P<0.01$.

uncorrected proof

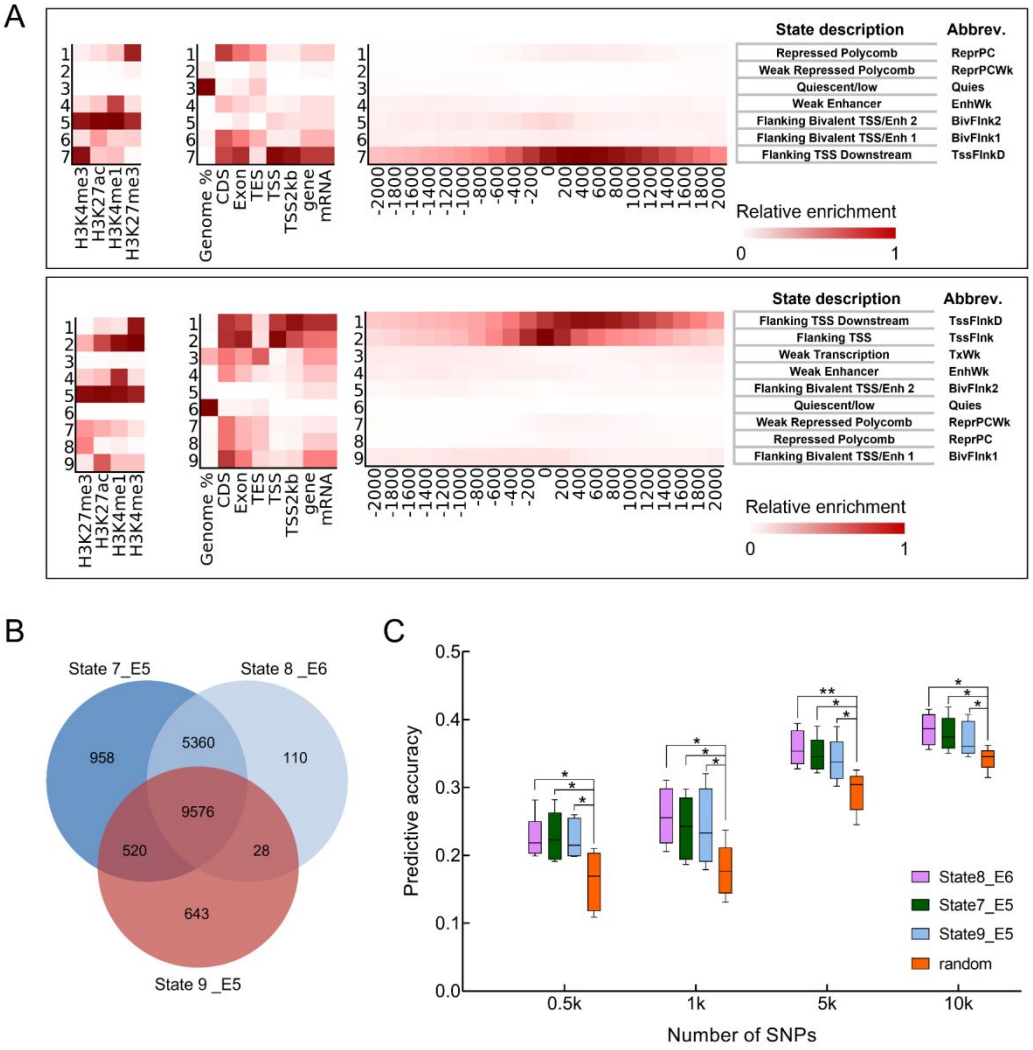
Supplementary Materials



Supplementary Figure S1. Model-based population structure analysis of *Litopenaeus vannamei* populations inferred using ADMIXTURE

A: Cross-validation (CV) error across different numbers of ancestral populations (K=2–15). The CV error decreases gradually with increasing K, without showing a clear minimum. B: Ancestry proportion bar plot for K=3. Each vertical bar represents an individual, and different colors denote the estimated ancestry proportions from distinct genetic components. Individuals are grouped according to their sampling populations (group 1: n=313; group 2: n=331; group 3: n=328), showing clear genetic differentiation and varying degrees of admixture among groups.

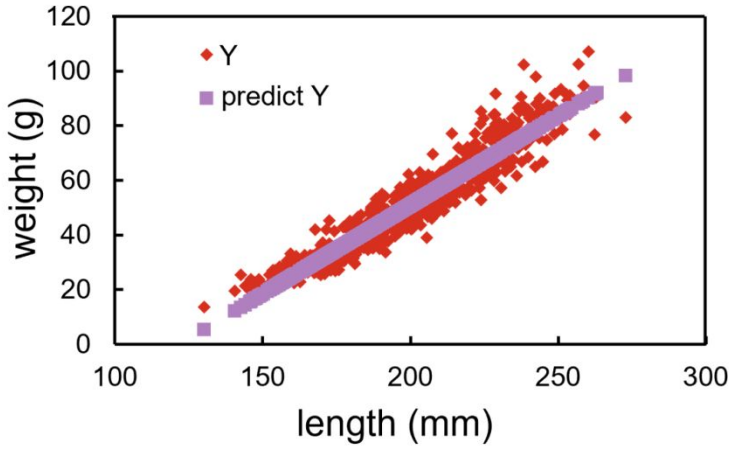
uncorrect



Supplementary Figure S2. Model comparison and robustness evaluation of ChromHMM parameterizations

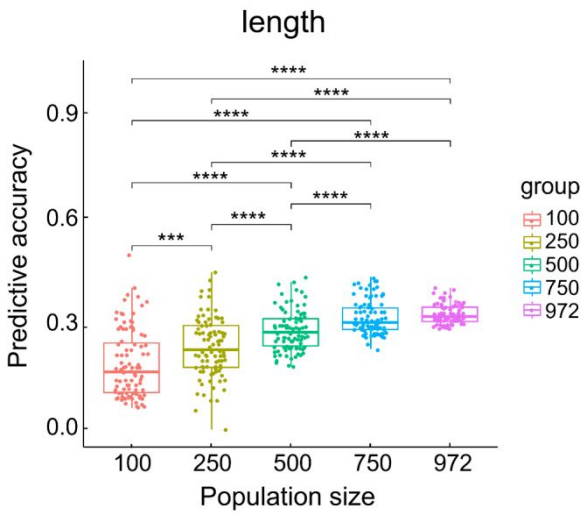
A: Emission probabilities and genomic enrichment heatmaps for 7-state (top) and 9-state (bottom) ChromHMM models. The heatmaps illustrate the emission probabilities for histone modifications (H3K4me3, H3K27ac, H3K4me1, and H3K27me3), genomic coverage, and enrichment relative to genomic features (e.g., TSS, TES, and mRNA). The 7-state model lacks the resolution to distinguish the weak transcription state, while the 9-state model introduces redundant flanking TSS categories. B: Venn diagram showing the overlap of bivalent promoter/enhancer state loci across different models. Comparison of bivalent regions identified in the 7-state (State 7_E5), 8-state (State 8_E6), and 9-state (State 9_E5) models. The 8-state model (k=8) captures the majority of functional loci while avoiding the artificial fragmentation and loss of regions observed in the 9-state model. C: Robustness of genomic prediction accuracy across alternative model parameterizations. Predictive accuracies were compared between SNPs residing in the bivalent states of 7, 8, and 9-state models versus randomly sampled SNPs at varying marker densities (0.5k to 10k). SNPs within bivalent regions consistently outperformed random sets regardless of the model chosen ($^{*}P<0.05$, $^{**}P<0.01$), confirming that the predictive superiority is biologically driven and robust to model selection.

Zoological Research



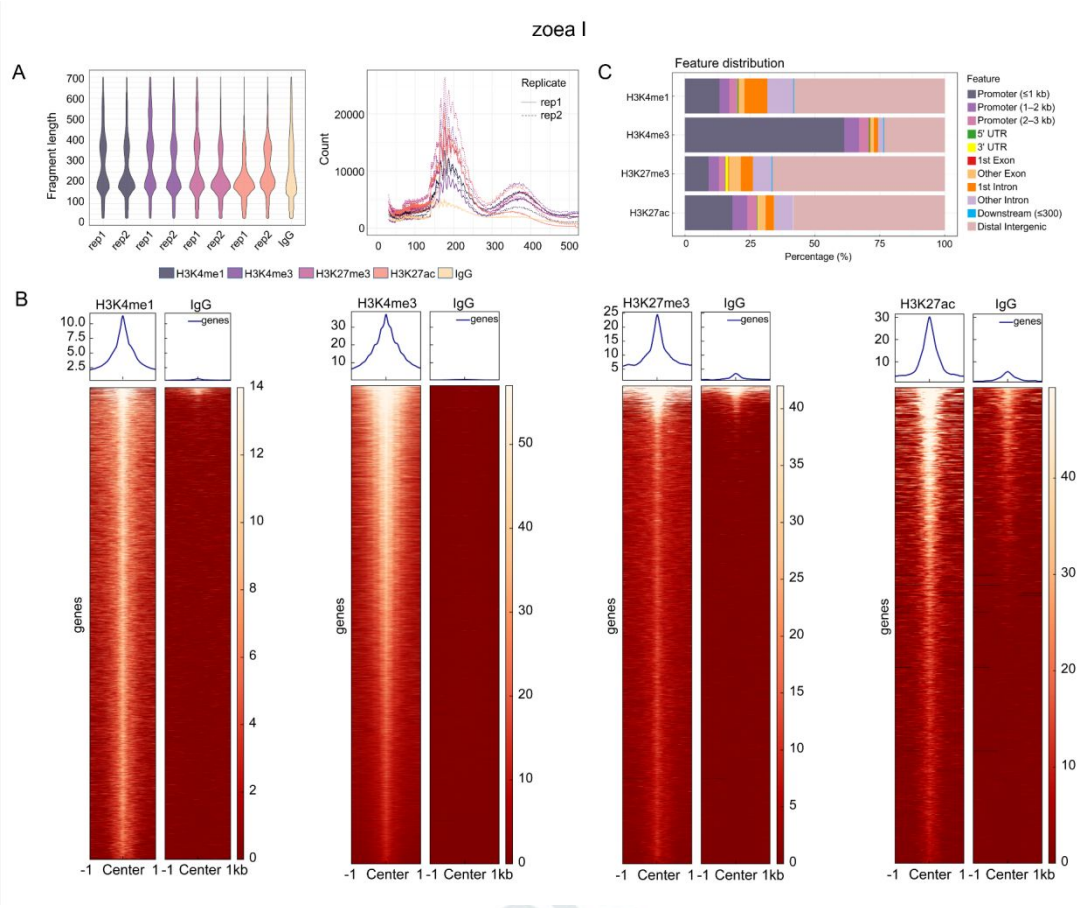
Supplementary Figure S3. Linear regression of body length on body weight. Red dots represent observed values, and the purple line indicates the fitted regression. A strong positive correlation was observed (R^2=0.918).

uncorrected pr



Supplementary Figure S4. Effect of training population size on prediction accuracy at a fixed density of 10k SNPs, evaluated using BayesA. For each size, SNPs were randomly sampled five times and assessed using 20 repetitions of cross-validation, withholding 20% of individuals for validation set in each iteration. Accuracy improved with larger training sets (*** P<0.001, **** P<0.0001) and plateaued beyond ~750 individuals.

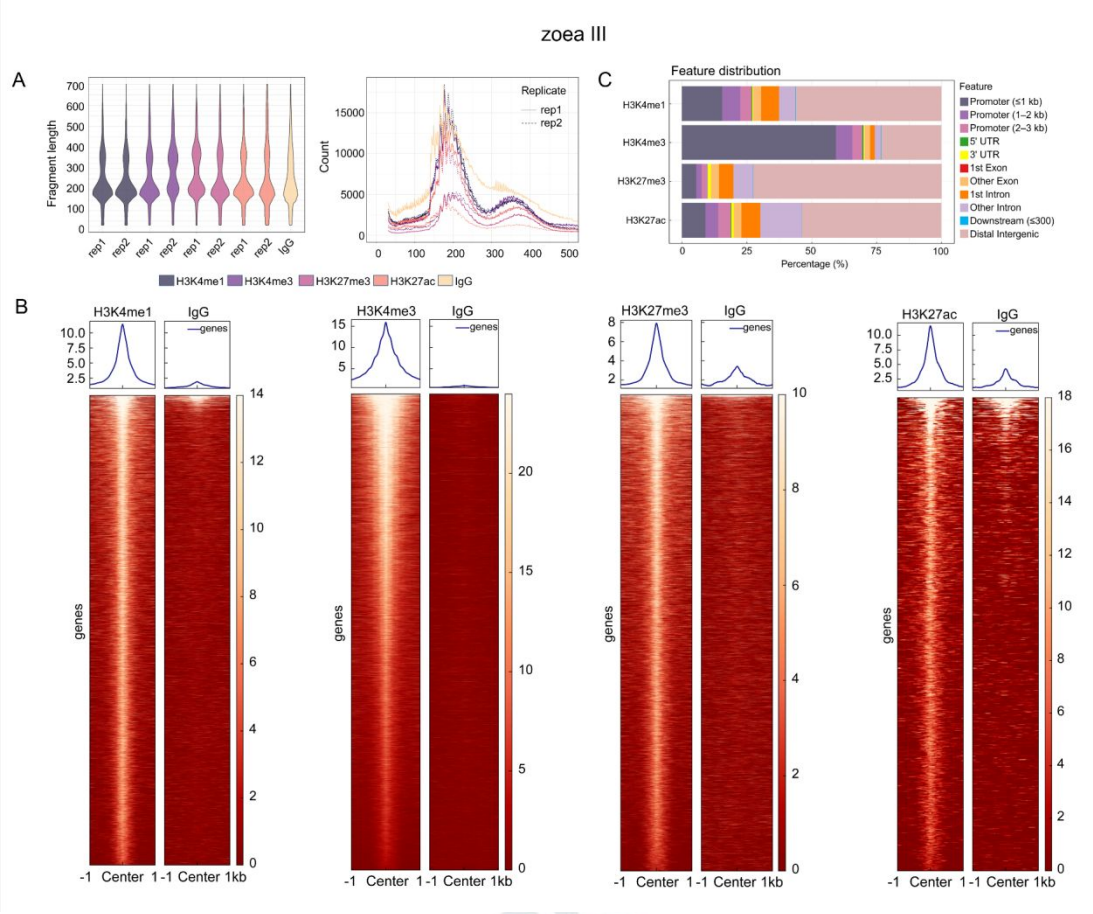
uncorrected proof



Supplementary Figure S5. Quality assessment of CUT&Tag data at the zoea I stage

A: Fragment size distribution. Plots showing the proportion (left) and raw counts (right) of DNA fragments across different sizes. Histone modification samples display strong enrichment of mononucleosome-sized fragments, whereas the IgG control samples exhibit a random fragment size distribution. B: Peak signal enrichment. Heatmaps (bottom) and line plots (top) showing CUT&Tag signal enrichment within ±1 kb of peak summits for the four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), compared with the random signal distribution observed in the IgG control. C: Distribution of peaks across genomic features.

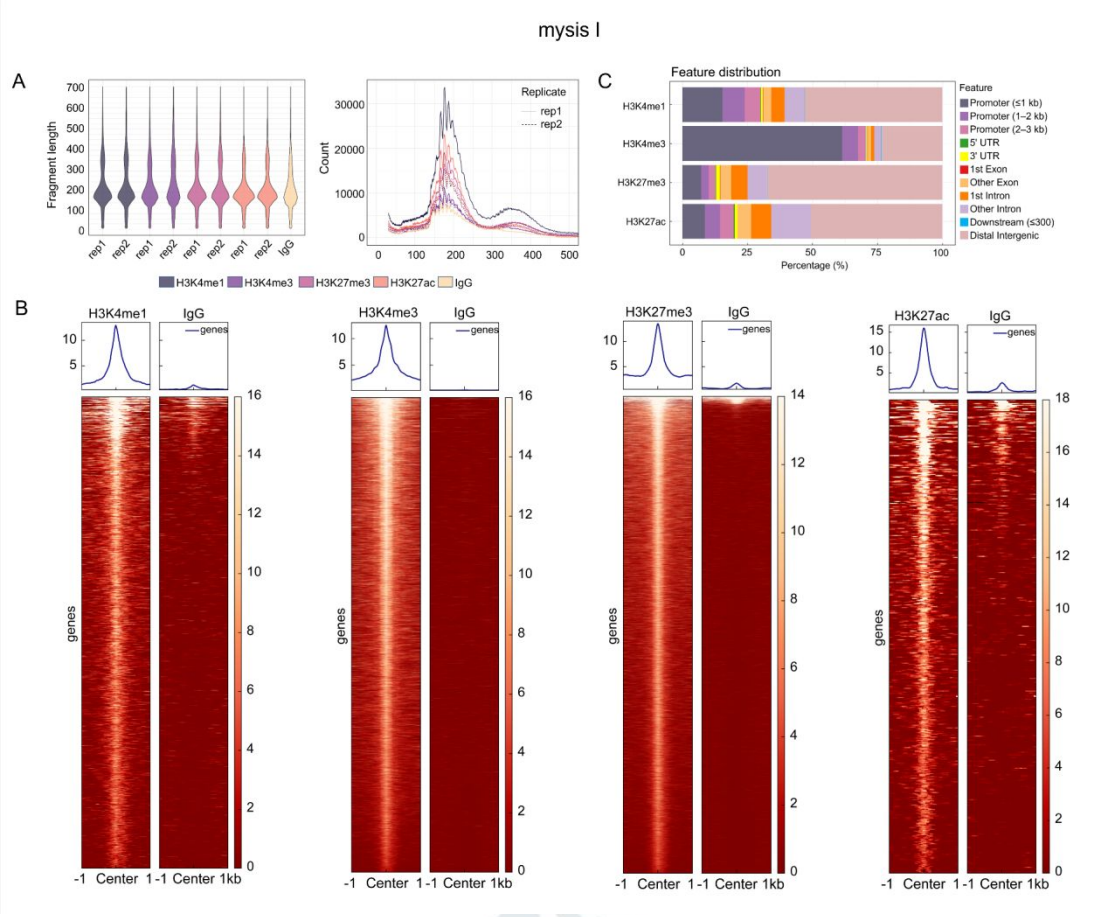
uncorrected



Supplementary Figure S6. Quality assessment of CUT&Tag data at the zoea Ⅲ stage

A: Fragment size distribution. Plots showing the proportion (left) and raw counts (right) of DNA fragments across different sizes. Histone modification samples display strong enrichment of mononucleosome-sized fragments, whereas the IgG control samples exhibit a random fragment size distribution. B: Peak signal enrichment. Heatmaps (bottom) and line plots (top) showing CUT&Tag signal enrichment within ±1 kb of peak summits for the four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), compared with the random signal distribution observed in the IgG control. C: Distribution of peaks across genomic features.

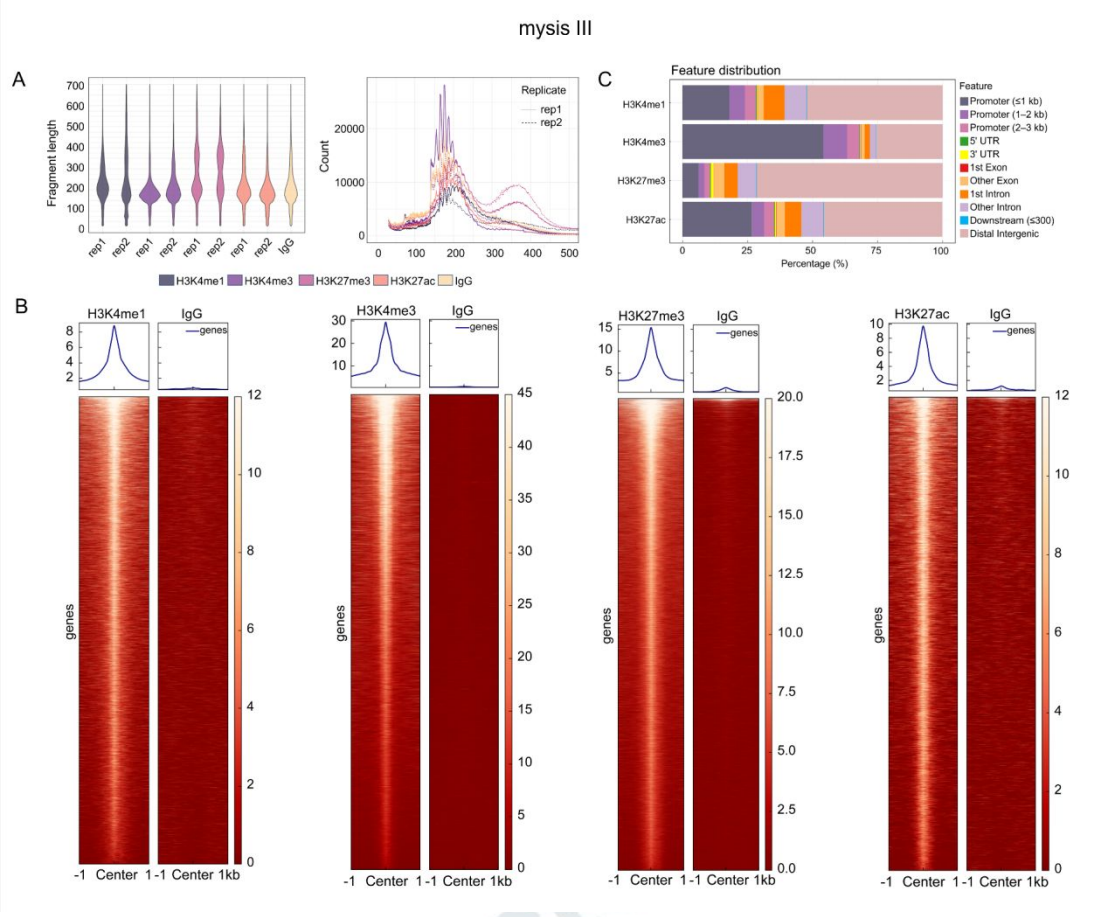
uncorrected



Supplementary Figure S7. Quality assessment of CUT&Tag data at the mysis Ⅰ stage

A: Fragment size distribution. Plots showing the proportion (left) and raw counts (right) of DNA fragments across different sizes. Histone modification samples display strong enrichment of mononucleosome-sized fragments, whereas the IgG control samples exhibit a random fragment size distribution. B: Peak signal enrichment. Heatmaps (bottom) and line plots (top) showing CUT&Tag signal enrichment within ±1 kb of peak summits for the four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), compared with the random signal distribution observed in the IgG control. C: Distribution of peaks across genomic features.

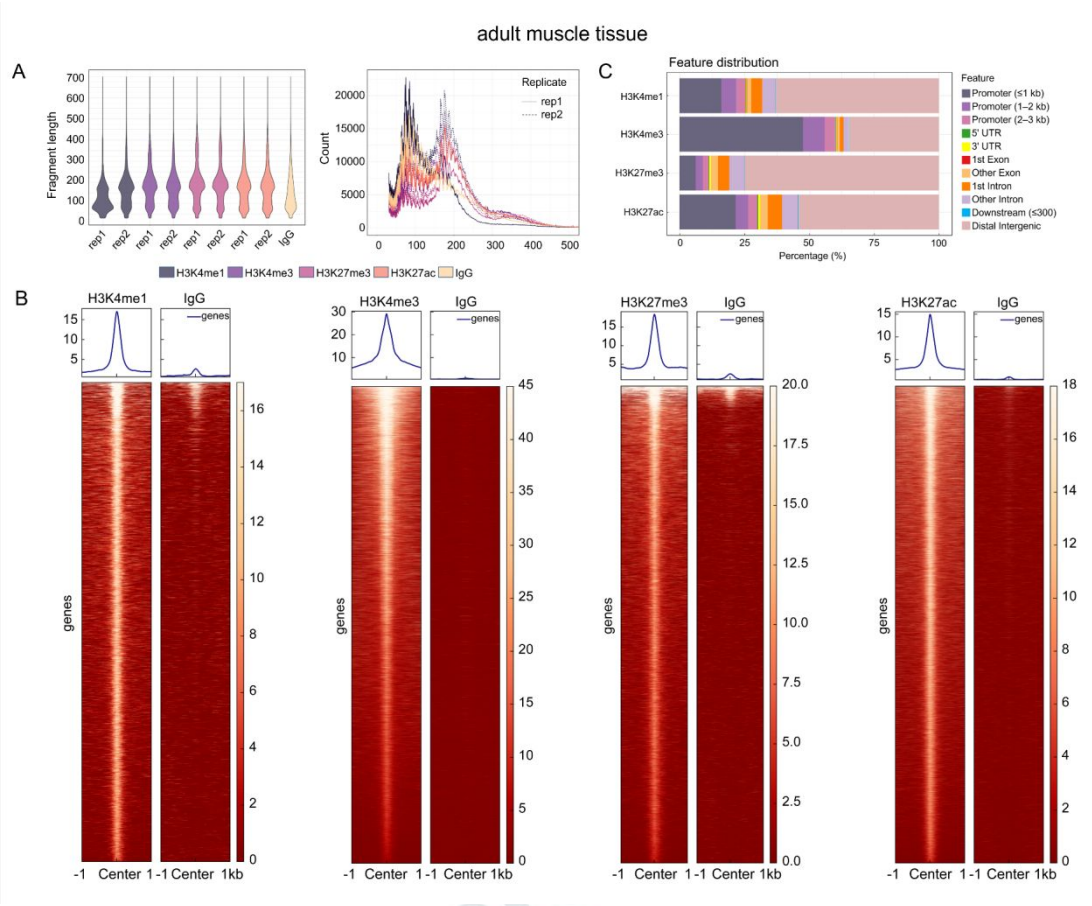
Zoological Research



Supplementary Figure S8. Quality assessment of CUT&Tag data at the mysis Ⅲ stage

A: Fragment size distribution. Plots showing the proportion (left) and raw counts (right) of DNA fragments across different sizes. Histone modification samples display strong enrichment of mononucleosome-sized fragments, whereas the IgG control samples exhibit a random fragment size distribution. B: Peak signal enrichment. Heatmaps (bottom) and line plots (top) showing CUT&Tag signal enrichment within ±1 kb of peak summits for the four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), compared with the random signal distribution observed in the IgG control. C: Distribution of peaks across genomic features.

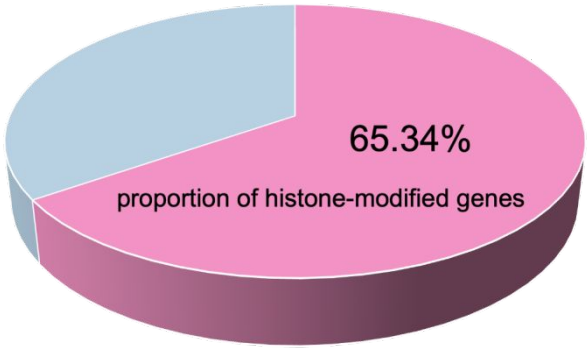
Zoological Research



Supplementary Figure S9. Quality assessment of CUT&Tag data at the adult muscle tissue

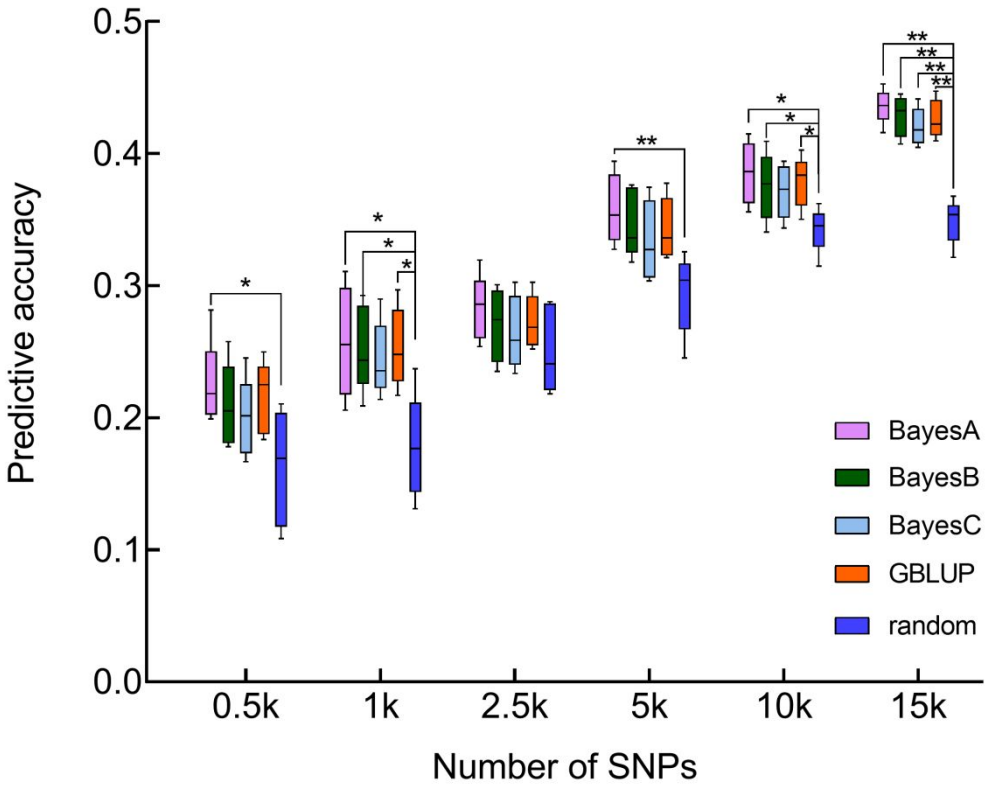
A: Fragment size distribution. Plots showing the proportion (left) and raw counts (right) of DNA fragments across different sizes. Histone modification samples display strong enrichment of mononucleosome-sized fragments, whereas the IgG control samples exhibit a random fragment size distribution. B: Peak signal enrichment. Heatmaps (bottom) and line plots (top) showing CUT&Tag signal enrichment within ±1 kb of peak summits for the four histone modifications (H3K4me1, H3K4me3, H3K27me3, and H3K27ac), compared with the random signal distribution observed in the IgG control. C: Distribution of peaks across genomic features.

Zoological Research



Supplementary Figure S10. Gene annotation of CUT&Tag peaks during embryonic development of *Litopenaeus vannamei*. CUT&Tag peaks from embryonic stages overlapped with 65.34% of annotated genes in the genome, indicating widespread association of histone modifications with coding regions and suggesting abundant cis-regulatory information during development.

uncorrected proof



Supplementary Figure S11. Validation of SNPs from the muscle bivalent promoter/enhancer state (E6) across different genomic selection (GS) models. Four GS models (BayesA, BayesB, BayesC, and GBLUP) were used to evaluate the predictive accuracy of SNPs located in the muscle E6 chromatin state for body length across five SNP densities (0.5k–15k). For each density, SNPs were randomly sampled five times and assessed using fivefold cross-validation repeated five times. Across all models, SNPs from the E6 state consistently outperformed random SNPs. Significance levels: $^{*}P<0.05$, $^{**}P<0.01$.

uncorrected

Supplementary Table S1. Antibodies used in this study.

Supplementary Table S2. Quality control results of CUT&Tag datasets in embryonic tissues and adult muscle tissues.

Supplementary Table S3. GWAS results of SNPs within the muscle bivalent promoter/enhancer (E6) state.

Supplementary Table S4. Predictive accuracy of different statistical models across varying marker densities.

Supplementary Table S5. Predictive accuracy of different sample sizes.

Supplementary Table S6. Alignment rate (%) of clean CUT&Tag reads to the reference genome.

Supplementary Table S7. Fraction of Reads in Peaks (FRiP, %) of mapped fragments for each CUT&Tag sample.

Supplementary Table S8. Replicate reproducibility.

Supplementary Table S9. Number of histone modification peaks located within embryonic tissues and adult muscle tissues.

Supplementary Table S10. Number of SNPs located within histone modification peaks in embryonic tissues and adult muscle tissues.

Supplementary Table S11. Predictive accuracy of SNP subsets from histone modification peaks in embryonic and muscle tissues at varying marker densities.

Supplementary Table S12. Predictive accuracy of SNP subsets from histone modification peaks in adult muscle tissue at varying marker densities.

Supplementary Table S13. Number of peaks located within different chromatin state regions in embryonic tissues and adult muscle tissues.

Supplementary Table S14. Number of SNPs located within different chromatin state regions in embryonic tissues and adult muscle tissues.

Supplementary Table S15. Predictive accuracy of SNPs located in embryonic chromatin states across varying marker densities compared with randomly sampled SNPs.

Uncorrected proof

Supplementary Table S16. Summary of all datasets used in this study (newly sequenced and downloaded).

表观基因组信息辅助优选的二价染色质可提升凡纳滨对虾（*Litopenaeus vannamei*）基因组预测稳健性

史佳乐[1,2]，鲁志芳[1,2]，隋明益[1,2]，尹苗苗[1,2]，牟铭[1,2]，张冬冬[3]，包振民[1,2]，胡景杰[1,2]，曾启繁[1,2,*]，叶质[1,2,*]

[1]中国海洋大学三亚海洋研究院，海南省热带水产种质重点实验室，三亚 572024，中国

[2]中国海洋大学海洋生命学院，海洋遗传与育种教育部重点实验室，青岛 266003，中国

[3]海南大学三亚南繁研究院，海洋生物与水产学院，三亚 572025，中国

*通讯作者：叶质 (yezhi@ouc.edu.cn)；曾启繁 (zengqifan@ouc.edu.cn)

摘要：基因组选择（GS）已成为加速动物遗传改良的重要工具，但其广泛应用仍受到基因分型成本较高以及跨群体预测可靠性下降等问题的限制。利用功能基因组学注释信息来优先选择具有生物学意义的 SNPs，是提高基因组预测稳健性的一种极具前景的策略。本研究以凡纳滨对虾（*Litopenaeus vannamei*）为模型进行了一项概念验证研究，整合了来自 972 个个体的全基因组重测序数据与表型数据。通过利用 CUT&Tag 技术，对多个胚胎发育阶段以及成体肌肉组织中的四种核心组蛋白修饰（H3K4me1、H3K4me3、H3K27me3 和 H3K27ac）进行了高分辨率表观基因组图谱绘制，并基于这些表观组信息指导 SNP 优先筛选以用于基因组预测。在评估的多种筛选策略中，位于肌肉特异性二价启动子/增强子区域（E6 状态）的 SNPs 始终表现出最高的预测准确率。该状态以活化性与抑制性组蛋白修饰共存为特征。值得注意的是，即使在中等标记密度（15k）条件下，来源于 E6 区域的 SNP 集合仍显著优于规模更大的全基因组随机 SNP 集合。更重要的是，在利用独立对虾品系进行的跨群体验证中，与同等密度的随机 SNP 组合相比，E6 状态 SNP 子集将基因组预测准确度显著提升了 47.62%（由 0.21±0.05 提升至 0.31±0.04，$P<0.05$）。上述结果表明，基于表观遗传注释指导的 SNP 优先选择策略，可以在大幅降低分子标记需求量的同时，提高基因组预测的准确性和跨群体稳健性，为功能基因组信息在水产动物育种中的整合应用提供了一种切实可行的方法学框架。

关键词：基因组选择；组蛋白修饰；表观遗传学；CUT&Tag；功能注释

uncorrected proof