

Evolutionary persistence governs transmitted prediction across generations in hyper-diverse genomes

Shikai Liu

liushk@ouc.edu.cn

Ocean University of China <https://orcid.org/0000-0001-5777-489X>

Ben Yang

Ocean University of China

Deqi Sun

Ocean University of China

Mengmeng Xu

Ocean University of China

Chenyu Shi

Ocean University of China

Chao Guo

Ocean University of China

Xiaojie Han

Ocean University of China

Qi Li

Ocean University of China <https://orcid.org/0000-0002-3937-9324>

Article

Keywords:

Posted Date: July 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8780717/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

High nucleotide diversity and rapid linkage disequilibrium decay erode the stability of genomic predictions across generations in hyper-diverse genomes. Whether any variant classes can maintain associations that are sufficiently stable to support prediction across generations remains unresolved. Aquaculture breeding programs provide a unique experimental setting to quantify transmitted genetic effects through common-garden rearing and large-scale progeny validation. Leveraging this design in the Pacific oyster as a hyper-diverse model, we couple genome-wide variant discovery with cross-generational validation to quantify transmitted additive effects under minimized environmental variation. We uncover a fundamental decoupling between heritability and transmitted prediction accuracy. Rare variants in weak-LD regions explain substantial heritable variation, yet their effects are evolutionarily unstable and translate poorly into prediction accuracy across generations. In contrast, ancient variants with high allele frequencies and strong LD preserve persistent haplotype structure and dominate prediction accuracy. Prioritizing these evolutionarily persistent variants substantially improves accuracy relative to whole-genome and functionally annotated models and yields consistent gains when independently validated in representative fish and shrimp species. These findings establish an evolution-informed framework for genomic prediction in hyper-diverse genomes, in which prediction accuracy is governed by the evolutionary persistence of variants rather than by the amount of genetic variance they explain.

Introduction

Predicting the genetic potential of complex traits is a foundational goal across the life sciences, underpinning advances in human disease genetics, crop improvement, livestock breeding and conservation biology^{1,2}. Polygenic prediction provides a unifying statistical framework for these applications³. By aggregating small contributions across many loci, genomic data can be used to forecast phenotypes, select individuals for breeding and design efficient improvement program^{4,5}. In principle, as marker density and reference population sizes increase, prediction accuracy for additive genetic effects, which represent the heritable component that determines long term response to selection, should approach limits set by quantitative genetics theory. In practice, however, prediction accuracy for highly polygenic traits frequently plateaus well below these theoretical expectations⁶.

This performance gap is most acute in organisms characterized by hyper-diverse genomes, such as marine invertebrates and long-lived plants, where high nucleotide diversity and rapid decay of linkage disequilibrium conspire to decoupling between markers and causal variants^{7–9}. In these systems, recombination swiftly erodes local haplotype structure and leaves causal variants poorly tagged by nearby markers amid a vast excess of non-causal polymorphisms^{10,11}. Conventional whole-genome regression models are ill-suited to this landscape as they implicitly treat all markers as biologically equivalent and struggle to distinguish sparse causal variants from abundant neutral noise. This statistical assumption ignores the biological reality that genomic regions are functionally heterogeneous

and contribute unequally to phenotypic variation^{12,13}. Consequently, overcoming the noise barrier in these genomes requires moving beyond simple marker density to approaches that can explicitly prioritize variants based on their biological informativeness.

Functional and sequence-based annotations offer a principled route to address these limitations^{14,15}. By prioritizing variants with higher biological plausibility, annotation informed approaches aim to more effectively capture causal variants and suppress noise from non-causal polymorphisms. In humans, integrating regulatory and gene structural annotations, minor allele frequency (MAF) information and linkage disequilibrium (LD) based metrics into polygenic models has yielded significant gains in prediction accuracy for complex traits^{16–18}. However, the determinants of prediction performance remain unresolved in non-model species with highly polymorphic genomes. It is unclear whether prediction information is driven primarily by molecular function, such as coding versus noncoding regions, or by evolutionary context, defined by MAF and LD structure. These features reflect a variant's evolutionary persistence and embedding within the recombination landscape and may govern how stably its effect contributes to additive variance over time. Moreover, most existing studies evaluate prediction within a single cohort or generation, leaving it unresolved which classes of variants carry information about transmitted additive effects, defined as the component of genetic value that is stably passed to offspring and directly determines realized genetic gain.

Aquaculture breeding programs provide a unique and powerful experimental setting to address these challenges. The use of controlled crosses and structured pedigrees, integrated with standardized common-garden rearing, for the validation of parental genomic estimated breeding values (GEBV) against realized offspring phenotypes under minimized environmental noise^{19,20}. By isolating additive genetic signals from environmental noise, this framework enables the direct quantification of how specific variant classes contribute to the transmitted component of genetic effects, thereby surmounting the limitations of traditional within-cohort prediction. Many aquaculture species combine extremely high polymorphism with rapid linkage disequilibrium decay and strong selection on fitness traits^{21,22}. The Pacific oyster (*Crassostrea gigas*) exemplifies this combination, ranking among the most genetically diverse animals known²³. It serves as a stringent testbed for identifying general principles of variant prioritization, with insights likely to extend to a broad class of non-model species facing similar genomic challenges.

Here, we leverage the Pacific oyster to dissect the genomic features that govern the prediction of transmitted additive genetic effects in a hyper-diverse genome. We assembled a comprehensive suite of annotations that spans regulatory and gene-structural categories alongside minor allele frequency and linkage disequilibrium metrics which summarize the evolutionary and recombinational context of variants. By incorporating these features into polygenic prediction models, we evaluated how distinct genomic features contribute specifically to additive effects transmitted from parents to offspring. This progeny-based design allows us to measure directly how well different subsets of variants predict realized genetic gain. Such experimental validation is rarely feasible in human or wild population studies.

By synthesizing functional annotation, evolutionary information, and cross-generational validation, we aim to identify general principles that link variant properties to transmitted additive effects in complex and hyper-diverse genomes.

Results

An integrated framework to dissect transmitted genetic effects in hyper-diverse genomes

To resolve the genomic features that underpin the stable transmission of complex traits in highly polymorphic species, we implemented a multi-tiered experimental framework utilizing the Pacific oyster as a representative model (Fig. 1). *Vibrio* resistance was selected as the focal trait because it represents a quintessential complex phenotype governed by polygenic architecture²⁴. The experimental design was structured to isolate additive genetic signals from environmental noise through a large-scale common-garden culture²⁵. By rearing multiple breeding lineages under standardized conditions within a single marine environment, we minimized environmental confounding factors that typically obscure genetic contributions in wild or loosely controlled populations (Fig. 1a). Furthermore, we addressed the challenges posed by the rapid decay of linkage disequilibrium in this species by prioritizing high-density genome-wide variant discovery (Fig. 1b). This strategy was necessitated by the need to maintain sufficient resolution to tag causal variants across a genome characterized by exceptional nucleotide diversity and recombination rates, a landscape where sparse genotyping arrays often fail to capture the full spectrum of informative polymorphisms. A defining feature of this framework is the transition from traditional within-cohort prediction to a rigorous cross-generational validation. While standard genomic prediction models rely on statistical cross-validation, our approach utilizes a progeny test design to empirically measure the transmission of genetic value. By validating parental genomic estimated breeding values against the realized phenotypic performance of their offspring, we sought to directly quantify the transmitted component of additive variance (Fig. 1c). This component represents the portion of genetic merit that is stably inherited and determines the actual response to selection. This multi-tiered strategy provides a robust testbed for identifying the general principles that govern polygenic prediction in complex and hyper-diverse genomes.

Establishing a high-resolution phenotypic framework to capture precise resistance variation

Following a common-garden rearing phase, we subjected the experimental population to a standardized challenge to quantify variation in resistance components. Importantly, these breeding lineages had been cultured for more than 10 consecutive generations under a consistent common-environment framework²⁶, further minimizing environmental noise and potential transgenerational carry-over effects. A total of 2,124 oysters, comprising 787 individuals from selectively bred strains and 1,337 control

individuals, were subjected to a challenge experiment with *V. alginolyticus* (Supplementary Table 1). Following a 10-day challenge period, the selectively bred populations exhibited a significantly higher survival rate of 24.47% compared with 7.18% in the control population. Similarly, the mean survival time for the selectively bred oysters (5.53 days) was significantly greater than that of the control population (2.86 days) (Fig. 2). These results demonstrate substantial phenotypic divergence in resistance to *V. alginolyticus* infection among the populations, establishing a robust foundation for subsequent genetic investigations.

To capture resistance along a continuous axis rather than only at mortality endpoints, we quantified *Vibrio* load in the adductor muscle. Pathogen load provides a biologically informative measure of infection severity, as it quantifies pathogen replication within the host and reflects processes that occur before downstream survival outcomes^{27,28}. Based on the observed survival curves, 1,138 oysters spanning the full range of resistance phenotypes were selected for genome sequencing (Supplementary Fig. 1). Using unmapped reads from these genomes, we quantified *Vibrio* load in adductor muscle as a quantitative biomarker of infection severity²⁹. Selectively bred resistant populations carried substantially lower bacterial load than susceptible control populations, mirroring their survival differences (Fig. 2b). Moreover, *Vibrio* load was strongly and negatively correlated with survival time ($R = -0.702$, $P < 2.2 \times 10^{-16}$, Fig. 2c) and survival status ($R = -0.492$, $P < 2.2 \times 10^{-16}$; Fig. 2d). Together, these patterns indicate that *Vibrio* load provides a robust, biologically proximal biomarker that is closely aligned with conventional survival-based measures of disease resistance and is therefore well suited for genomic prediction of resistance traits.

High-resolution genomic profiling bridges the gap between markers and causal variants in a hyper-diverse landscape

Given that the decoupling between markers and causal variants occurs over exceptionally short physical distances in this species, a vast repertoire of polymorphisms is essential to ensure that functional signals are captured during genotyping. To achieve this high marker density without compromising population-scale throughput, we employed low-coverage whole-genome sequencing (lcWGS) integrated with a probabilistic genotype imputation framework³⁰. After stringent quality control and adapter trimming, we generated 2.77 TB of high-quality sequencing data. Read quality was consistently high, with mean Q20 and Q30 values of 98.41% and 96.12%, respectively. The mean data output per individual was 2.43 Gb and reads achieved an average mapping rate of 96.25% to the reference genome, resulting in an average sequencing depth of 3.41× across samples (Supplementary Table 2). Data uniformity was also high, with only 13 samples exhibiting depths below 1× (Supplementary Fig. 2).

Single nucleotide polymorphisms were called and imputed using STITCH³¹, yielding 12.59 million raw variants. To rigorously validate imputation accuracy at low-coverage, we selected 30 high-depth datasets and computationally down-sampled them to 1× coverage to serve as an internal gold standard through the same pipelines^{32,33}. This validation demonstrated high imputation accuracy even at 1× coverage,

with average genotype consistency of 0.958 and dosage R^2 of 0.909 across all variants. Accuracy remained robust for rare variants (minor allele frequency, MAF < 0.01), showing genotype consistency of 0.946 and imputation R^2 of 0.890. After filtering variants with MAF < 0.01 to ensure downstream reliability, 9.23 million SNPs were retained. Re-evaluation using this filtered set showed further improved accuracy on 1× down-sampled data (genotype consistency: 0.961, imputation R^2 : 0.914; Supplementary Fig. 3). These metrics demonstrate that our bioinformatic workflow successfully recovers high-density, research-grade genotypes from low-coverage data, enabling cost-effective genome-wide analysis in this hyper-diverse genome.

Annotation-informed modeling reveals the functional heterogeneity of the hyper-diverse genome

To move beyond the simplistic assumption that all markers are biologically equivalent, we next evaluated whether incorporating functional priors could sharpen prediction in this highly polymorphic genome. We compiled a panel of 48 functional annotations for SNPs using the Baseline-LD framework as a reference, covering gene structural regions, regulatory elements, and population-level measures including MAF and LD score³⁴ (Fig. 3a). This panel comprised three continuous and 45 binary annotations (Supplementary Table 3–4, Supplementary Fig. 4). Incorporating these annotations into the LDAK³⁵ framework markedly altered the genome-wide distribution of SNP effects (Fig. 3b–e). While the chromosomal locations of major signals were broadly conserved, the annotation-informed model exhibited a ~ 10-fold increase in mean absolute effect size and a variance increase of over one order of magnitude ($P < 2.2 \times 10^{-16}$), demonstrating a profound shift in effect-size distribution rather than random fluctuation (Supplementary Table 5). The modest rank concordance between the two models indicates that weighting does not merely scale effects uniformly (Supplementary Fig. 5). Instead, it redefines the genetic architecture by concentrating weight on biologically relevant variants, producing an effect-size landscape that is statistically and visually distinct from the unannotated baseline (Fig. 3f).

We further assessed the prediction performance of these models to determine if resolving biological heterogeneity translates into superior forecasting of complex traits. Consistent with the effect-size repartitioning, annotation-informed models yielded substantial improvements in prediction accuracy across all three traits (Fig. 4a). Specifically, the annotation-informed models achieved prediction accuracies of 0.529 for *Vibrio* load, 0.383 for survival time, and 0.353 for survival status (Supplementary Fig. 6). To contextualize the performance of the annotation-informed models, we benchmarked them against several widely used genomic prediction methods including GBLUP, Bayes A, Bayes B, and Bayesian Lasso^{36,37}. We quantified the performance improvement of each model relative to the GBLUP baseline accuracy. While several Bayesian methods also showed improvements in prediction accuracy compared to GBLUP, the annotation-informed model demonstrated the most substantial relative gain across all traits (Fig. 4b–d). The magnitude of this relative improvement was highly trait dependent. The most pronounced gain was observed for the *Vibrio* load trait, with an improvement of 123.20% relative to

the GBLUP baseline (Fig. 4c). Survival status showed the next highest improvement at 46.07%, followed by survival time at 38.86% (Fig. 4b, d). This magnitude of improvement substantially exceeds the typical gains reported in previous studies integrating functional annotations in livestock^{38,39}, highlighting the unique leverage of annotation in species where rapid LD decay otherwise obscures causal variants. This trait-specific variation highlights that functional annotation utility depends on underlying genetic architecture, with the precisely measured *Vibrio* load phenotype exhibiting both highest absolute accuracy and greatest relative improvement.

Annotation enhances prediction transferability across populations

Effective cross-population prediction remains difficult because the statistical relationships between markers and causal variants are not preserved across differentiated groups. This instability arises from a combination of allele-frequency divergence, differences in linkage disequilibrium structure, and population-specific genetic architectures shaped by distinct demographic and selective histories⁴⁰. We assessed whether functional annotation could enhance prediction across genetically differentiated populations. Genetic diversity and population structure were analyzed in 1,138 individuals from four breeding populations using genome-wide SNP genotypes. Principal component analysis and neighbor-joining phylogenetic trees identified four distinct genetic clusters that were broadly consistent with meta-data, with only partial overlap between the RF and H1 populations (Fig. 5a, b). This well-defined population structure provided a solid foundation for assessing the accuracy of cross-population prediction (Supplementary Fig. 7 and Supplementary Fig. 8).

Using a leave-one-population-out cross-validation design, we trained models on three populations and predicted phenotypes in the remaining population⁴¹. Models without annotation information generally exhibited low and uniformly similar accuracies across all test sets (Fig. 5c-e). Moreover, their performance varied depending on trait type, for quantitative traits, accuracies ranked as H1 > RP > RF > H4, whereas for binary traits the ranking shifted to H1 > H4 > RF > RP. Importantly, these traditional models frequently showed sharp declines in prediction ability when applied across populations, in some cases even producing negative performance metrics. In contrast, annotation-informed models consistently achieved the highest prediction accuracy across all traits and held-out populations. Incorporating functional annotation information thus effectively mitigated the severe accuracy losses observed in cross-population prediction with traditional approaches, yielding more robust and transferable predictions across diverse genetic backgrounds.

Annotation-informed prediction benefits from functional annotation integration on a genome-wide variant landscape

To investigate how SNP density and genomic coverage affect prediction accuracy across different model types, we analyzed three datasets generated by representative genotyping methods: a low-density

SNP array (~ 24K SNPs)⁴², reduced-representation sequencing (ddRAD-seq, ~ 48K SNPs)⁴³, and a high-density SNP array (~ 122K SNPs)⁴⁴. Although the high-density chip provided substantially greater marker density than the other two, it still represented a set of selectively chosen loci rather than a comprehensive capture of genome-wide variation (Supplementary Fig. 9). Our results revealed a striking divergence in how different model architectures respond to increasing variant resolution. Traditional genomic prediction models, including GBLUP and Bayesian methods, were largely insensitive to SNP density. Prediction accuracy remained stable across the three datasets, with average accuracies of 0.281 for the low-density array, 0.276 for ddRAD-seq, and 0.265 for the high-density chip (Supplementary Table 6). This suggests that methods relying on coarse linkage disequilibrium saturate quickly⁵.

By contrast, annotation-informed models showed a strong dependence on a high-density, genome-wide variant landscape. When based on SNP sets from selective genotyping approaches (arrays or ddRAD-seq), their accuracy was comparable to or slightly below that of traditional models, likely because such datasets fail to capture a sufficient proportion of functional variants needed for annotation to be informative. However, when applied to genome-wide SNPs generated by lcWGS, annotation-informed models exhibited a pronounced improvement, achieving an average prediction accuracy of 0.422 (Fig. 6). These findings indicate that low- and medium-density SNP datasets cannot meaningfully enhance prediction through annotation weighting, whereas genome-wide variant sets provides the essential substrate for annotation-informed models to realize their full advantage¹⁵.

Empirical validation of transmitted effects through progeny test

To move beyond the constraints of statistical cross-validation and provide definitive evidence of realized genetic gain, we performed a rigorous progeny-testing analysis to quantify the prediction accuracy of transmitted additive genetic effects. The validation was based on the correlation between parental GEBV and the observed phenotypic performance of their offspring (Supplementary Fig. 10). Across all tested traits, annotation-informed models consistently achieved higher prediction accuracies than traditional genomic prediction approaches, with improvements ranging from 13.87% to 105.45% depending on the trait (Supplementary Table 7, 8). For example, in survival status, accuracy increased from 0.301 under GBLUP to 0.435 with the annotation-informed model, corresponding to a 44.52% relative improvement, and the genetic gain of the resistant population selected based on GEBV increased by 1.05-fold (Supplementary Table 9, 10).

Given that complex traits can be characterized by multiple phenotypic measurements representing different aspects of the same underlying trait⁴⁵ (Supplementary Fig. 11 and Supplementary Fig. 12), we assessed whether integrating these measures into a composite index could enhance prediction. We constructed an Aggregate Breeding Value (ABV) by combining GEBV for survival status, survival time, and *Vibrio* load into a single composite index⁴⁶ (Fig. 7a, b, Supplementary Fig. 13–17). Compared with single-trait GEBV, the ABV consistently achieved higher prediction accuracies across traits. For offspring survival rate, prediction accuracies for single-trait GEBV ranged from 0.378 (*Vibrio* load) to 0.435

(survival status), whereas the ABV reached 0.443, corresponding to relative gains of approximately 1.8–17.2% over the single-trait models. For offspring survival time, single-trait accuracies ranged from 0.339 to 0.385, while the ABV achieved 0.395, representing improvements of about 2.6–16.5%. These results indicate that aggregating multiple, complementary survival-related traits into a single ABV captures additional genetic effects and leads to modest but consistent improvements in genomic prediction accuracy.

The high fidelity of the ABV in capturing transmitted additive effects was further confirmed by the distinct phenotypic divergence observed among the selected lineages. Based on ABV values, breeding candidates were ranked, and three groups were selected for comparison: resistant, control, and susceptible (Supplementary Fig. 18). In the disease challenge test, the resistant group exhibited a survival rate of 51.16% and an average survival time of 5.46 days, compared with 37.35% and 4.75 days in the control group, and 23.97% and 4.29 days in the susceptible group (Fig. 7c, d). These results demonstrate that the ABV more effectively discriminates among individuals, leading to clearer separation of resistant and susceptible families in both survival rate and survival time.

Evolutionary context drives transmitted prediction and reveals a decoupling from heritability

To dissect the sources of prediction gain and quantify the contribution of individual annotation categories to transmitted additive genetic effects, we first fitted models that incorporated a single annotation while maintaining an identical progeny validation framework in which parental GEBV were regressed on observed progeny phenotypes. All single-annotation models increased prediction accuracy relative to the baseline model that lacked annotation, although the extent of improvement varied considerably among categories. The LD bins annotation produced the largest increase in accuracy and accounted for most of the improvement achieved in the full annotation model, which reached the mean prediction accuracy of 0.388. These results identify LD-related information as the dominant contributor to the additional prediction accuracy beyond the baseline model and suggest that the informativeness of variants is strongly shaped by their LD properties, as illustrated in Fig. 8a.

Given this strong contribution of LD-related annotations, we next asked whether the apparent performance of different annotation categories reflects intrinsic biological properties or is influenced by the number of SNPs they contain. Using the full annotation model, we recalculated prediction accuracy for SNPs assigned to individual functional categories. Across most categories, prediction accuracy increased with the number of SNPs, indicating that marker density influences the observed performance of annotation-defined subsets (Fig. 8b). Importantly, when considering MAF-bin and LD-bin annotations that were constructed to contain identical numbers of SNPs, we observed systematic stratification that persisted despite identical SNP numbers. Under this design, prediction accuracy exhibited stratification across bins, with consistently higher accuracy in bins characterized by higher LD or MAF. The highest bins systematically outperformed the lowest bins (Fig. 8c, d), indicating that LD and MAF capture

intrinsic prediction structure beyond differences in SNP count alone. To evaluate whether these trends were influenced by the LDAK weighting scheme, we repeated the bin-based analyses using GBLUP, which assumes equal SNP effects across the genome³⁷. Under GBLUP framework, the associations between accuracy and both LD and MAF remained detectable and followed the same directionality as in the annotation-informed analysis. Importantly, the rank ordering of bins was broadly consistent across the two frameworks, confirming that the observed enrichment of accuracy in high-LD and high-MAF regions is a robust biological feature rather than a model-specific artifact.

To reconcile these prediction accuracy patterns with the underlying genetic architecture, we estimated the additive genetic variance captured by each bin through a per-bin heritability analysis across the three resistance traits. This analysis revealed a striking contrast between the distribution of genetic variance and the distribution of prediction accuracy. For both MAF and LD binning schemes, the lowest bin yielded the highest absolute heritability estimates (Fig. 8e, f). Most notably, the lowest LD bin exhibited a mean heritability of approximately 0.710 across traits, yet this same subset conferred the lowest prediction accuracy in the corresponding validation. Conversely, higher bins displayed lower heritability estimates but sustained superior and more stable prediction accuracy.

These results demonstrate a fundamental decoupling of heritability and predictability in hyper-diverse genomes. While rare and weak-LD variants generate the bulk of phenotypic variance, a pattern consistent with population genetic theory where negative selection constrains large-effect mutations to low frequencies and weak linkage backgrounds^{34,47}, their effects are too evolutionarily unstable to support accurate transmitted prediction. Instead, prediction stems from the evolutionarily persistent scaffold provided by high-frequency and strong-LD variants. These ancient variants, although yielding less to heritability, maintain stable associations with causal haplotypes that persist through recombination and are thus robustly recovered across generations⁴⁸.

Prioritization of ancestral variants optimizes genomic prediction across diverse taxa

The fundamental decoupling between genomic regions that drive heritability and those that support prediction motivated us to investigate the evolutionary mechanisms underlying this divergence. To evaluate whether this contrast reflects differences in evolutionary persistence, we estimated allele ages using the GEVA genealogical framework⁴⁹ and examined their relationships with LD score and MAF. Both LD score and MAF displayed clear positive associations with inferred allele age (Fig. 9a–c). This relationship confirms that older variants are typically more frequent and embedded within regions of stronger LD. These empirical data directly support the hypothesis that evolutionary persistence acts as the primary determinant of prediction accuracy. Unlike young, deleterious variants that are rapidly purged or kept rare by purifying selection, ancient variants have survived extensive recombination and allelic turnover. Consequently, they retain stable linkage to causal genomic backgrounds and provide consistent information for forecasting transmitted additive effects.

To determine whether LD and MAF alone capture the main prediction accuracy contained in the broader annotation set, we fitted models that retained only LD bins and MAF bins annotations. Prediction accuracy under this simplified model was essentially indistinguishable from that of the full annotation model (Fig. 9d). This outcome indicates that LD and MAF jointly summarize most of the annotation-driven improvement and that additional functional annotations contribute comparatively little beyond these two axes of variation.

We next investigated whether a targeted SNPs selection strategy based on LD and MAF could enhance prediction beyond what was achievable with all annotated variants. An optimized SNP subset was generated by selecting variants from the highest LD and highest MAF bins. When the annotation-informed model was fitted using only this subset, accuracy increased from 0.388 to 0.426, corresponding to a relative improvement of 9.79%. This improvement shows that prioritizing variants characterized by both strong LD and high MAF can yield higher prediction value than using the full complement of annotated SNPs (Fig. 9e). To evaluate whether this improvement depended on annotation modeling, we applied the same high-LD and high-MAF SNP subset to the model without annotation. In this setting, restricting analysis to the optimized SNPs increased prediction accuracy by 23.68% relative to the model using all SNPs. Control subsets that matched the optimized subset in size but were distributed uniformly across the genome achieved considerably lower accuracy. These comparisons demonstrate that the gain arises from intrinsic information carried by variants with high LD and high MAF rather than from a reduction in marker number alone.

Finally, we evaluated the cross-taxa applicability of the optimized SNP prioritization strategy using independent disease-resistance datasets from two additional aquaculture species. In large yellow croaker (*Larimichthys crocea*) dataset⁵⁰, a species with genomic architecture comparable to that of the oyster, the optimized SNP subset achieved an 81.35% increase in prediction accuracy relative to the raw model (Fig. 9f). Similarly, in Pacific white shrimp (*Litopenaeus vannamei*)⁵¹, the same strategy resulted in a more moderate but consistent improvement of 18.76% in predictive performance (Fig. 9g). Despite differences in genome organization, marker density, and trait architecture between the two species, both independent validations showed concordant gains in the accuracy of parental GEBV. These results show that prioritizing ancient variants with high LD and high MAF concentrates stable, transmitted additive effects and consistently improves genomic prediction across diverse taxa.

Discussion

This study integrates dense genome-wide variation, functional and evolutionary annotation, and cross-generational progeny test to dissect the mechanisms governing polygenic prediction in a genome characterized by high polymorphism and rapid linkage disequilibrium decay. Analysis of resistance to *Vibrio* infection in the Pacific oyster shows that substantially more prediction accuracy can be recovered when genome-wide variation is comprehensively sampled and phenotypes closely reflect underlying biological processes. The resulting gains reveal a fundamental decoupling between variants that contribute most to heritability and those that support accurate, transmitted prediction.

Trait definition fundamentally dictates the resolving power of estimated genetic effects. Survival status and survival time compress multiple components of the host-pathogen interaction and are sensitive to stochastic variation. By contrast, *Vibrio* load provides a biologically proximal and continuous readout of pathogen burden, displays strong correlations with survival traits, and supports the highest absolute accuracy among all phenotypes. Annotation-informed models achieved their largest relative improvements for this trait. These patterns illustrate that phenotypes positioned closer to causal mechanisms sharpen the genotype-phenotype mapping and reveal variants whose effects are sufficiently stable to inform long-term selection.

The genomic setting in which these variants are evaluated is equally decisive. The Pacific oyster, representing a broad class of highly polymorphic marine invertebrates, combines extreme nucleotide diversity with limited haplotype structure and rapid LD decay. In this context, traditional whole-genome regression methods saturate quickly with increasing SNP density because their performance is governed by coarse LD patterns. In contrast, annotation-informed models display qualitatively different behavior as their advantage becomes apparent only when applied to genome-wide datasets generated by whole-genome sequencing. Conversely, sparse marker panels provide too limited a representation of the genome for annotation to improve accuracy⁵². This result defines a simple design principle for genomic prediction in such genomes, which indicates that dense and relatively unbiased sampling of variation is a prerequisite for functional or evolutionary annotation to improve accuracy.

Incorporating annotation fundamentally reshapes the distribution of genetic effects. Weighting variants by gene-structural, regulatory, and population-genetic features yields effect-size landscapes that are visually and statistically distinct from uninformed models and yields consistent gains in prediction for all resistance traits. Notably, these gains extend beyond within-population cross-validation. In cross-population prediction, where models trained in one set of breeding populations are applied to another, unannotated models frequently lose accuracy or become anti-prediction, whereas annotation-informed models retain positive and often substantially higher accuracies. This behavior indicates that annotation helps concentrate weight on variants whose effects are more stable across backgrounds that differ in minor allele frequency spectra and recombination histories, thereby supporting more transferable prediction in structured breeding program.

A key implication of our findings is that functional annotations contribute far less to transmitted prediction than population-genetic descriptors such as LD and MAF, despite their direct biological relevance. This result does not imply that functional categories lack causal variants. Instead, it suggests that the effect from functional variants in hyper-diverse genomes is fundamentally eroded by recombination and allelic turnover. First, extreme polymorphism inflates the number of non-causal variants within coding and regulatory regions, which dilutes enrichment and diminishes the statistical contrast between functional and non-functional sites. Second, rapid LD decay disrupts the physical linkage between functional annotations and causal alleles, preventing functional categories from maintaining consistent predictive value across generations. Consequently, molecular annotations

describe biological mechanisms, whereas population-genetic features describe evolutionary persistence, and the latter property is the most relevant to long-term predictability.

These observations also clarify why heritability and predictability become decoupled in highly polymorphic genomes. Rare and recently arisen variants, which are often enriched for larger effect sizes under purifying selection, may contribute disproportionately to heritability^{34,47}. Yet their weak LD, short evolutionary lifespan, and rapid allelic turnover render them unreliable for forecasting genetic response across generations. In contrast, ancient variants that segregate at higher frequencies and reside within stronger LD environments contribute modest effects but generate stable, transmitted accuracy that persist through recombination. As a result, variants that explain substantial heritability are not necessarily those that best predict long-term selection response. This conceptual separation reframes the architecture of complex traits in hyper-diverse genomes and motivates an evolutionary definition of marker informativeness grounded in persistence rather than effect size.

This evolutionary model leads naturally to a simple and general prioritization strategy. Restricting annotation models to LD and MAF captures nearly all improvement provided by the full annotation catalogue, indicating that these axes of variation summarize the most relevant biological information. Moreover, selecting variants from the highest LD and highest MAF bins yields higher prediction accuracy than using all annotated SNPs in both annotation-informed and unannotated frameworks. Control subsets of identical size do not reproduce this improvement, confirming that prediction gains arise from the intrinsic information content of evolutionarily persistent variants rather than from marker thinning. The same strategy yields a large accuracy improvement in independent teleost and crustacean species, suggesting that prioritizing ancient variants in strong LD is a general principle applicable beyond oysters to other highly polymorphic genomes.

In conclusion, these results refine the role of annotation in genomic prediction. Molecular annotations offer biological insight, but population-genetic descriptors determine the stability of prediction accuracy across generations. For applications that aim to forecast genetic potential and realized response to selection in complex, highly polymorphic genomes, dense variant discovery and precise characterization of evolutionary context emerge as central design priorities. By aligning phenotypic resolution, genome-wide variant discovery, and annotation with the evolutionary histories of segregating variants, this work establishes an evolution-informed framework for genomic prediction. In highly polymorphic genomes, evolutionary persistence emerges as the defining property of marker informativeness, reconciling the decoupling between heritability and predictability and providing a practical foundation for stable, long-term genomic prediction.

Methods

Establishment and common-garden rearing of the reference cohort

The reference cohort used in this study comprised four distinct lineages derived from a long-term breeding program maintained under a consistent common-environment regime for over ten consecutive generations^{26,53}. Among these, two lineages served as control groups, while the other two strains, selected for *Vibrio* resistance, were derived from our oyster *Vibrio*-resistant breeding program^{43,54}. To minimize environmental confounding and transgenerational carry-over effects, all lineages were propagated during the 2022 spawning season using a synchronized artificial fertilization and larval rearing protocol. Mature oysters with well-developed gonads were selected, and gonadal cells were extracted using capillary tubes. These cells were smeared onto glass slides for microscopic examination to determine the sex of each individual. Subsequently, the soft tissue of selected individuals was dissected to collect eggs and sperm, which were placed into separate 10 L containers. For artificial insemination, sperm were added incrementally to the egg containers, with the quantity adjusted to ensure that 5–10 sperm surrounded each egg under microscopic observation. Fertilization was monitored, and approximately 15–20 minutes after sperm addition, 30% of the eggs displayed the presence of a first polar body, indicating successful fertilization. The fertilized eggs were transferred to filtered seawater for incubation. After 24 hours, D-shaped larvae were isolated using a 300-mesh sieve and transferred to new containers with filtered seawater. The water temperature was maintained at 23–25°C and the salinity at with daily water replacement using seawater filtered through a sand filter and non-woven polypropylene fabric. Larvae were fed with *Isochrysis galbana* three times daily. Once the larvae developed to the umbo stage, *Platymonas* sp. was supplemented to their diet. When approximately 30% of the pediveliger larvae exhibited eyespot, scallop shells were introduced to provide substrates for larval settlement. Following a two-week settlement period, the spat was transitioned to a standardized outdoor pond system for initial acclimation. When the spat reached a shell height of 2–3mm, they were transferred to oyster raft culture system located in Sanggou Bay, Shandong Province, China (37°06'N, 122°31'E) for nature culture⁵⁴. This multi-stage common-garden design, spanning from hatchery to open-sea culture, ensured that all phenotypic variations observed during the subsequent challenge could be attributed to additive genetic differences rather than environmental heterogeneity.

Pathogen challenge trial

In May 2023, oysters were acclimated to laboratory conditions for two weeks to adapt to the environment prior to the *Vibrio* challenge experiment. The *Vibrio* strain Cg5, previously stored in glycerol at -80°C, was revived by culturing on marine agar plates⁵⁵. Single colonies were inoculated into 2216E liquid medium and incubated at 28°C for 12 hours until the culture reached the logarithmic growth phase. Bacterial cells were harvested by centrifugation (5000×g, 10 min, 4°C) and washed three times with sterile seawater. The pellet was then resuspended in sterile seawater to achieve a final concentration of 5×10^8 CFU/mL. Oysters were anesthetized with a 50 g/L magnesium chloride solution for 4 hours. Following anesthesia, 100 µL of the bacterial suspension was injected into the adductor muscle of each oyster. Post-injection, oysters were monitored every two hours for mortality. Oysters that were unable to close their shells were recorded as dead. Mortality rates were calculated throughout the

10-day challenge experiment. Adductor muscle tissue from dead individuals was collected and preserved in 95% ethanol at -20°C. At the conclusion of the experiment, adductor muscles from surviving oysters were similarly sampled and stored for subsequent analyses.

DNA Extraction and Sequencing Library Construction

Approximately 50 mg of oyster adductor muscle tissue was excised and placed in a sterile 1.5 mL centrifuge tube. A total of 600 µL of proteinase K lysis buffer (300 µg/mL) was added, and the tissue was homogenized using magnetic beads for 1 minute. The homogenized tissue was then digested overnight at 37°C. Following digestion, DNA was extracted using the nucleic acid extraction kit based on magnetic bead adsorption. The extracted DNA was dissolved in double-distilled water and assessed for quality and concentration using Qubit 4.0 Fluorometer and agarose gel electrophoresis. High-quality genomic DNA was enzymatically fragmented via Tn5 transposase-mediated tagmentation to generate libraries⁵⁶. Unique dual-index (UDI) adapters were then incorporated into the fragments through PCR amplification with indexed primers, followed by purification to remove primer dimers. The constructed libraries were initially quantified using a Qubit 4.0 Fluorometer, and the fragment size distribution was analyzed using an Agilent 2100 Bioanalyzer. Libraries with the desired fragment size underwent quantitative PCR (qPCR) for precise quantification. Whole-genome sequencing was conducted with 3× depth using the Illumina NovaSeq 6000 platform, following the “paired-end 150” sequencing strategy.

Genotype

Raw sequencing data were processed using fastp (version 0.23.2)⁵⁷ to perform quality control and remove adapter sequences. The cleaned reads were aligned to the reference genome using BWA-MEM algorithm in BWA (version 0.7.17)⁵⁸ following the creation of alignment index with “bwa index” command. The resulting SAM files were converted to compressed BAM format, sorted by chromosomal coordinates using SAMtools (version 1.9)⁵⁹, and subsequently indexed to facilitate downstream analysis and storage optimization. PCR duplicates were removed using Picard MarkDuplicates (version 2.26.0), followed by BAM file re-indexing. Genotype calling was performed using GATK pipeline⁶⁰, while genotype imputation was carried out with STITCH (version 1.6.6), which employs a Hidden Markov Model (HMM) coupled with Expectation-Maximization (EM) algorithm to estimate ancestral haplotypes in the population³¹. This probabilistic framework enables individual genotype determination based on inferred haplotype patterns, generating an initial VCF output. To assess imputation accuracy, the original 10× sequencing data were down-sampled to 1× coverage to simulate low-depth genotypes. The down-sampled datasets were then processed with STITCH for genotype imputation, and the imputed genotypes were evaluated against the corresponding high-depth (10×) genotypes using the squared correlation (R^2) and genotype concordance metrics. The VCF file underwent additional refinement through Beagle (version 5.4) for genotype refinement and haplotype phasing⁶¹. Finally, PLINK (version

1.90b6.26) was utilized to filter variants with minor allele frequency (MAF) below 0.01, ensuring high-quality SNP retention for subsequent studies⁶².

Vibrio load quantification

To evaluate the physiological resistance of oysters to *Vibrio* infection, the bacterial load of *V. alginolyticus* in adductor muscle tissues was quantified using metagenomic sequencing. Unmapped reads were extracted from host-aligned BAM files using SAMtools (version 1.9)⁵⁹ by selecting sequences with a SAM flag of 4 and converting them to FASTQ format. The reference genome for *V. alginolyticus* (NCBI Assembly: GCF_023650915.1) was retrieved and indexed with BWA (version 0.7.17)⁵⁸, and the unmapped reads were subsequently aligned to this pathogen genome using the BWA-MEM algorithm. The relative pathogen load was calculated as follows:

$$load = \frac{n}{N}$$

where n represents the number of reads aligned to the *V. alginolyticus* genome, and N denotes the total number of sequencing reads.

Polygenic prediction model

To ensure normality assumptions for downstream statistical analyses, phenotypic data underwent rigorous preprocessing. Continuous traits, including survival time and *Vibrio* load, were transformed using the Box-Cox method to address right-skewed distributions⁶³. The optimal power parameter (λ) was estimated via maximum likelihood using the MASS package in R. Additionally, survival status was redefined as a semi-quantitative trait to enhance analytical power. Individuals were categorized into two groups based on the population median survival time: resistant (coded as 1) for those surviving beyond the median and susceptible (coded as 0) for those falling below the threshold.

Polygenic prediction analyses were conducted using LDAK with two models: a standard Elastic Net model and an annotation-informed Elastic Net model⁶⁴. SNPs were annotated with reference to the Baseline LD annotation data for the human genome³⁴, which encompassed the following elements: ten minor allele frequency (MAF) bins represented as binary annotations; ten linkage disequilibrium (LD) score bins encoded as binary annotations, along with continuous LD-related annotations by LDAK weights, and gene-based annotations generated using ANNOVAR from GFF files⁶⁵, as well as promoter sequence annotations derived from ATAC-seq data⁶⁶. Protein-coding sequences were further categorized according to codon degeneracy (0d, 2d, 3d, and 4d), yielding four binary annotations. Furthermore, nucleotide diversity and heterozygosity were calculated for each SNP and incorporated as continuous annotations.

The classical genomic selection model, GBLUP, was employed as the baseline model³⁷. A genomic relationship matrix (G matrix) for 1,138 individuals and 9.23 million SNPs was constructed using GCTA⁶⁷, followed by GBLUP implementation via the R package rrBLUP⁶⁸. Additionally, classical Bayesian models, including Bayes A, Bayes B, and Bayesian Lasso, were also used for comparison³⁶.

A standard 10-fold cross-validation with five repetitions was conducted to evaluate the prediction accuracy of the models. The accuracy was measured as the mean Pearson correlation coefficient (r) between the observed phenotypes and the genetic estimated breeding values (GEBV).

Cross populations prediction

Genetic diversity and population structure were analyzed across 1,138 individuals representing four distinct breeding populations. The SNPs were initially filtered through linkage disequilibrium-based pruning using PLINK (version 1.90b6.26) with parameters: 50-SNPs sliding window, 5-SNPs step size, and pairwise r^2 threshold of 0.2. This process retained 2,150,621 SNPs with minimal allelic correlation for downstream population structure analyses.

Principal component analysis (PCA) was conducted using GCTA (v1.93.2beta) with the pruned SNP dataset to visualize genetic clustering⁶⁷. Genetic distance matrices were generated from the pruned SNP dataset using VCF2Dis with default settings⁶⁹. Subsequently, neighbor-joining (NJ) trees were constructed via the FateME2.0 (<http://www.atgc-montpellier.fr/fastme/>)⁷⁰. Final tree visualization was performed in Interactive Tree of Life (<https://itol.embl.de>) with customized branch coloring. Population structure was further assessed using ADMIXTURE (v1.3.0) to estimate individual ancestry proportions based on the pruned SNP dataset.⁷¹

Cross-population genomic prediction accuracy was evaluated via a leave-one-population-out cross-validation framework. Prediction performance was quantified by the Pearson correlation coefficient (r) between estimated breeding values and observed phenotypic values across all pairwise population comparisons.

The impact of SNP density on genomic prediction

To evaluate the prediction accuracy of different SNP densities derived from various genotyping methods, we conducted a systematic comparison using three established genotyping approaches, including a commercial Axiom Affymetrix 57K Oyster Array⁴², a 48K SNP panel derived from double digest Restriction-site Associated DNA sequencing⁴³, and a high-density 190K Pacific Oyster Axiom Genotyping Array⁴⁴. For SNP array, SNPs and their flanking regions were aligned to the reference genome (GCA_902806645.1)⁷², and the physical positions of the SNPs on the reference genome were identified. Variants within 10 bp upstream and downstream of each SNP were extracted for genotyping. Density-

matched SNP subsets were extracted and used to construct genomic selection models. Prediction accuracy was assessed via cross-validation.

Progeny Test

Genomic prediction accuracy was empirically validated through a progeny challenge trial specifically designed to assess disease resistance traits. For this validation, breeding cohorts were established by mating selected broodstock. Parent selection was guided by an Aggregate Breeding Value (ABV) index, which we computed as a composite score integrating estimated genomic breeding values for three key traits: survival time, survival status, and *Vibrio* load. To calculate the ABV, we first converted the GEBV for each individual trait into percentile ranks using a non-parametric ranking system. The transformation applied was:

$$R_i(X) = \frac{\text{rank}(x_i)}{N}$$

where N represents the population size, and ranks were assigned according to the ordering of GEBV. The ranking direction for each trait was aligned with the desired breeding objective: survival time and status were ranked in ascending order of GEBV (higher GEBV values received higher ranks) to prioritize extended survival duration and enhanced viability, while *Vibrio* load was ranked in descending order of GEBV (lower GEBV values indicating reduced pathogen burden received higher ranks). The composite ABV index for each individual was then calculated by averaging these standardized percentile ranks:

$$ABV_I = \frac{R(\text{Time}_i) + \text{inverted } R(\text{Load}_i) + R(\text{Status}_i)}{3}$$

where $R(\text{Time})$ and $R(\text{Status})$ are the percentile ranks from ascending ranking, and inverted $R(\text{Load})$ is the percentile rank from descending ranking of *Vibrio load* GEBV. This averaging approach implicitly applies an equal weight (+ 1/3) to the beneficial ranked values from each trait, while the non-parametric ranking approach eliminated scale dependency between traits and preserved ordinal relationships without assuming specific parametric distributions for the GEBV. Subsequent progeny validation analyses evaluated the correlation between parent GEBV (including the composite ABV and individual trait GEBV) and offspring phenotypes.

A total of 44 full-sib families were derived from 37 dams and 41 sires. These families were categorized into three groups based on their ABVs: the top 20% (9 families) were designated as resistant, the bottom 20% (9 families) as susceptible, and the remaining 60% (26 families) served as control. In May 2024, a *Vibrio* challenge trial was conducted to test family-level resistance. Progeny performance was measured using the survival rate within each family and the mean survival time of all individuals in the family. The prediction accuracy of parental aggregate breeding values (ABVs) was then assessed by applying

Pearson correlation analysis to correlate progeny survival rates and mean survival times with the sum of parental ABVs ($H_{\text{Dam}} + H_{\text{Sire}}$)⁷³.

Inference of variations age

The age of variants was inferred using the Genealogical Estimation of Variant Age (GEVA) framework,⁴⁹ which estimates allele age from shared haplotype segments in a coalescent-based setting. For each chromosome, phased VCF files were converted into the binary haplotype, marker and site files required by GEVA. Age estimation was performed for SNPs using GEVA's default pipeline that combines mutation- and recombination-based clocks and integrates information across concordant and discordant haplotype pairs to obtain a composite posterior distribution of allele age. The resulting age estimates were post-processed and quality-filtered with the R utilities provided in the GEVA package.

External dataset for cross-species validation

To independently assess the generalizability of the optimized SNP prioritization strategy developed in this study, we incorporated two external datasets from aquaculture species. The first dataset was derived from large yellow croaker (*Larimichthys crocea*), generated in previous studies⁵⁰, it comprises individuals phenotyped for resistance to *Cryptocaryon irritans* and genotyped using a 55K SNP array. After standard genotype quality control, 480 individuals from the 2018 challenge experiment and 276 individuals from the 2020 challenge experiment were retained and used as the reference population. Using this reference panel, GEBV for the parental generation were computed under GBLUP model and the optimized prediction model proposed in this study. The resulting parental GEBV were then regressed against the observed phenotypic performance of their offspring in subsequent challenge tests. The second dataset consisted of Pacific white shrimp (*Litopenaeus vannamei*) from a commercial breeding population. This dataset included survival phenotypes following *Vibrio parahaemolyticus* (AHPND) infection for 735 individuals, among which 447 individuals were genotyped using a 45K SNP array and served as the reference population⁵¹. Genomic prediction models were trained using the reference population, and the performance of prioritized high-LD and high-MAF SNP subsets was evaluated by comparing parental GEBV with the survival performance of their progeny in an independent challenge experiment. This cross-species progeny-validation framework provided an independent evaluation of the transmitted additive genetic effects captured by the optimized SNP subsets.

Declarations

Competing interests

The authors declare no competing interests.

Author contributions

S.L. conceived and supervised the project, secured funding, and finalized manuscript revisions. B.Y. led the experimental design, performed data analysis, and drafted the manuscript. B.Y., C.G., D.S. and Q.L. established the oyster populations. Sample collections were conducted by B.Y., M.X., D.S., X.H., C.S., and C.G. All authors contributed to manuscript review and critical editing and approved the final version.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32341060), National Key Research and Development Program of China (2022YFD2400301), the Key Research and Development Program of Shandong Province (2023GJLJRC-058) the Fundamental Research Funds for the Central Universities (202541002) and the earmarked fund for the Agriculture Research System of China (CARS-49). We acknowledge the support of the High-Performance Biological Supercomputing Center at the Ocean University of China for this research. We also thank Professor Luan Sheng (Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences) for providing the *Litopenaeus vannamei* disease-resistance dataset, and Professor Xu Peng (Xiamen University) for providing the *Larimichthys crocea* dataset used for cross-species validation.

Data availability

The Raw sequencing data are available from NCBI SRA under BioProject accessions PRJNA1244556.

Code availability

Codes used for the analysis are available at <https://github.com/YangBen98/G2P>.

References

1. Ruan, Y., et al.: Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022)
2. Daetwyler, H.D., Calus, M.P., Pong-Wong, R., de Los Campos, G., Hickey, J.M.: Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics.* **193**, 347–365 (2013)
3. Ma, Y., Zhou, X.: Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends Genet.* **37**, 995–1011 (2021)
4. Wu, Y., et al.: Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell.* **186**, 2313–2328 (2023)

5. Meuwissen, T.H., Hayes, B.J., Goddard, M.: Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. **157**, 1819–1829 (2001)
6. Chatterjee, N., et al.: Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013)
7. Neale, D.B., Savolainen, O.: Association genetics of complex traits in conifers. *Trends Plant. Sci.* **9**, 325–330 (2004)
8. Butler, J.B., et al.: Patterns of genomic diversity and linkage disequilibrium across the disjunct range of the Australian forest tree *Eucalyptus globulus*. *Tree Genet. Genomes*. **18**, 28 (2022)
9. Plough, L.V.: Genetic load in marine animals: a review. *Curr. Zool.* **62**, 567–579 (2016)
10. de los Campos, G., Sorensen, D., Gianola, D.: Genomic heritability: what is it? *PLOS Genet.* **11**, e1005048 (2015)
11. Wray, N.R., et al.: Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013)
12. Yang, J., et al.: Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010)
13. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., Smoller, J.W.: Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019)
14. Zeng, J., Visscher, P.M.: Harnessing functional annotation to improve the accuracy and transferability of polygenic scores. *Nat. Rev. Genet.* **26**, 805–806 (2025)
15. Zheng, Z., et al.: Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet.* **56**, 767–777 (2024)
16. Zhang, Q., Privé, F., Vilhjálmsson, B., Speed, D.: Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021)
17. Hu, Y., et al.: Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Comput. Biol.* **13**, e1005589 (2017)
18. Márquez-Luna, C., et al.: Incorporating functional priors improves polygenic prediction accuracy in UK biobank and 23andMe data sets. *Nat. Commun.* **12**, 6052 (2021)
19. Liu, M., et al.: Assessing accuracy of genomic breeding values of selection candidates under biosecurity restrictions by progeny testing in Chinese shrimp *Fenneropenaeus chinensis*. *Aquaculture*. **566**, 739181 (2023)
20. Houston, R.D., et al.: Harnessing genomics to fast-track genetic improvement in aquaculture. *Nat. Rev. Genet.* **21**, 389–409 (2020)
21. Hollenbeck, C.M., Johnston, I.A.: Genomic tools and selective breeding in molluscs. *Front. Genet.* **9**, 253 (2018)
22. Liu, S., et al.: Haplotype-resolved genomes provide insights into the origins and functional significance of genome diversity in bivalves. *Cell. Rep.* **44**, 115697 (2025)

23. Zhang, G., et al.: The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*. **490**, 49–54 (2012)
24. Yang, B., et al.: Genome-wide association study toward efficient selection breeding of resistance to *Vibrio alginolyticus* in Pacific oyster, *Crassostrea gigas*. *Aquaculture*. **548**, 737592 (2022)
25. de Villemereuil, P., Gaggiotti, O.E., Mouterde, M.: Till-Bottraud, I. Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity*. **116**, 249–254 (2016)
26. Li, Q., Wang, Q., Liu, S., Kong, L.: Selection response and realized heritability for growth in three stocks of the Pacific oyster *Crassostrea gigas*. *Fish. Sci.* **77**, 643–648 (2011)
27. Medzhitov, R., Schneider, D.S., Soares, M.: P. Disease tolerance as a defense strategy. *Science*. **335**, 936–941 (2012)
28. Cunnington, A.J.: The importance of pathogen load. *PLOS Pathog.* **11**, e1004563 (2015)
29. De Lorgeril, J., et al.: Immune-suppression by OsHV-1 viral infection causes fatal bacteraemia in Pacific oysters. *Nat. Commun.* **9**, 4215 (2018)
30. Yang, B., Li, Y., Li, Q., Liu, S.: High-throughput and cost-effective genotyping by low-coverage whole genome sequencing with genotype imputation in Pacific oyster, *Crassostrea gigas*. *Aquaculture*. **591**, 741134 (2024)
31. Davies, R.W., Flint, J., Myers, S., Mott, R.: Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016)
32. Liu, S., et al.: Utilizing non-invasive prenatal test sequencing data for human genetic investigation. *Cell. Genomics*. **4**, 100669 (2024)
33. Browning, B.L., Browning, S.R.: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009)
34. Gazal, S., et al.: Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017)
35. Speed, D., Hemani, G., Johnson, M.R., Balding, D.J.: Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012)
36. Pérez, P., de los Campos, G.: Genome-wide regression and prediction with the BGLR statistical package. *Genetics*. **198**, 483–495 (2014)
37. VanRaden, P.M.: Efficient methods to compute genomic predictions. *J. Dairy. Sci.* **91**, 4414–4423 (2008)
38. Yuan, C., et al.: Evaluation of genomic selection models using whole genome sequence data and functional annotation in Belgian Blue cattle. *Genet. Sel. Evol.* **57**, 10 (2025)
39. MacLeod, I.M., et al.: Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genom.* **17**, 144 (2016)
40. Wientjes, Y.C.J., Veerkamp, R.F., Calus, M.P.: L. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*. **193**, 621–631 (2013)

41. Lotterhos, K.E.: Principles in experimental design for evaluating genomic forecasts. *Methods Ecol. Evol.* **15**, 1466–1482 (2024)
42. Gutierrez, A.P., et al.: Development of a medium density combined-species SNP array for Pacific and European oysters (*Crassostrea gigas* and *Ostrea edulis*). *G3 Genes/Genomes/Genet.* **7**, 2209–2218 (2017)
43. Yang, B., et al.: Genomic selection accelerates genetic improvement of resistance to *Vibriosis* in the Pacific oyster, *Crassostrea gigas*. *Aquaculture*. **584**, 740679 (2024)
44. Qi, H., et al.: Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*). *PLOS One*. **12**, e0174007 (2017)
45. Mbebi, A.J., Mercado, F., Hobby, D., Tong, H., Nikoloski, Z.: Advances in multi-trait genomic prediction approaches: classification, comparative analysis, and perspectives. *Briefings Bioinf.* **26**, bbaf211 (2025)
46. Hazel, L.N.: The genetic basis for constructing selection indexes. *Genetics*. **28**, 476–490 (1943)
47. Zeng, J., et al.: Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018)
48. O'Connor, L.J., et al.: Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019)
49. Albers, P.K., McVean, G.: Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biol.* **18**, e3000586 (2020)
50. Wang, J., et al.: Effects of artificial mating on genomic selection of resistance against *Cryptocaryon irritans* in large yellow croaker. *Aquaculture*. **561**, 738617 (2022)
51. Liu, M., et al.: Genomic selection of resistance to acute hepatopancreatic necrosis disease in *Litopenaeus vannamei*. *Aquaculture*. **596**, 741828 (2025)
52. Zhang, H., et al.: A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nat. Genet.* **55**, 1757–1768 (2023)
53. Liang, Y., et al.: Hybridization between Haida 1 and orange-shell line of the Pacific oyster reveals high heterosis in survival. *Aquaculture*. **551**, 737945 (2022)
54. Yang, B., Sun, D., Xu, C., Li, Q., Liu, S.: Comprehensive performance evaluation of *Vibrio* resistance strains from selective breeding in the Pacific oyster, *Crassostrea gigas*. *Aquaculture*. **612**, 743179 (2026)
55. Yang, B., et al.: Identification of *Vibrio alginolyticus* as a causative pathogen associated with mass summer mortality of the Pacific oyster (*Crassostrea gigas*) in China. *Aquaculture*. **535**, 736363 (2021)
56. Yang, B., Li, Y., Xu, M., Li, Q., Liu, S.: Developing efficient and cost-effective NGS sequencing libraries for high-throughput genotyping by whole-genome resequencing in aquaculture. *Aquaculture*. **613**, 743431 (2026)

57. Chen, S., Zhou, Y., Chen, Y., Gu, J.: fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018)
58. Li, H., Durbin, R.: Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009)
59. Li, H., et al.: The sequence alignment/map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009)
60. McKenna, A., et al.: The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010)
61. Browning, B.L., Zhou, Y., Browning, S.R.: A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018)
62. Purcell, S., et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007)
63. Box, G.E., Cox, D.R.: An analysis of transformations. *J. R Stat. Soc. B*. **26**, 211–243 (1964)
64. Hof, J.P., Speed, D.: LDAK-KVIK performs fast and powerful mixed-model association analysis of quantitative and binary phenotypes. *Nat. Genet.* **57**, 2116–2123 (2025)
65. Wang, K., Li, M., Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010)
66. Wan, K., et al.: Profiling of chromatin accessibility and regulatory elements toward understanding the mechanisms underlying growth regulation in the Pacific oyster (*Crassostrea gigas*). *Aquaculture*. **595**, 741531 (2025)
67. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M.: GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011)
68. Endelman, J.B.: Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant. Genome* **4**, (2011)
69. Xu, L., et al.: VCF2Dis: an ultra-fast and efficient tool to calculate pairwise genetic distance and construct population phylogeny from VCF files. *GigaScience* **14**, giaf032 (2025)
70. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015)
71. Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009)
72. Peñaloza, C., et al.: A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. *GigaScience* **10**, giab020 (2021)
73. Vallejo, R.L., et al.: Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. *Genet. Sel. Evol.* **49**, 17 (2017)

Figures

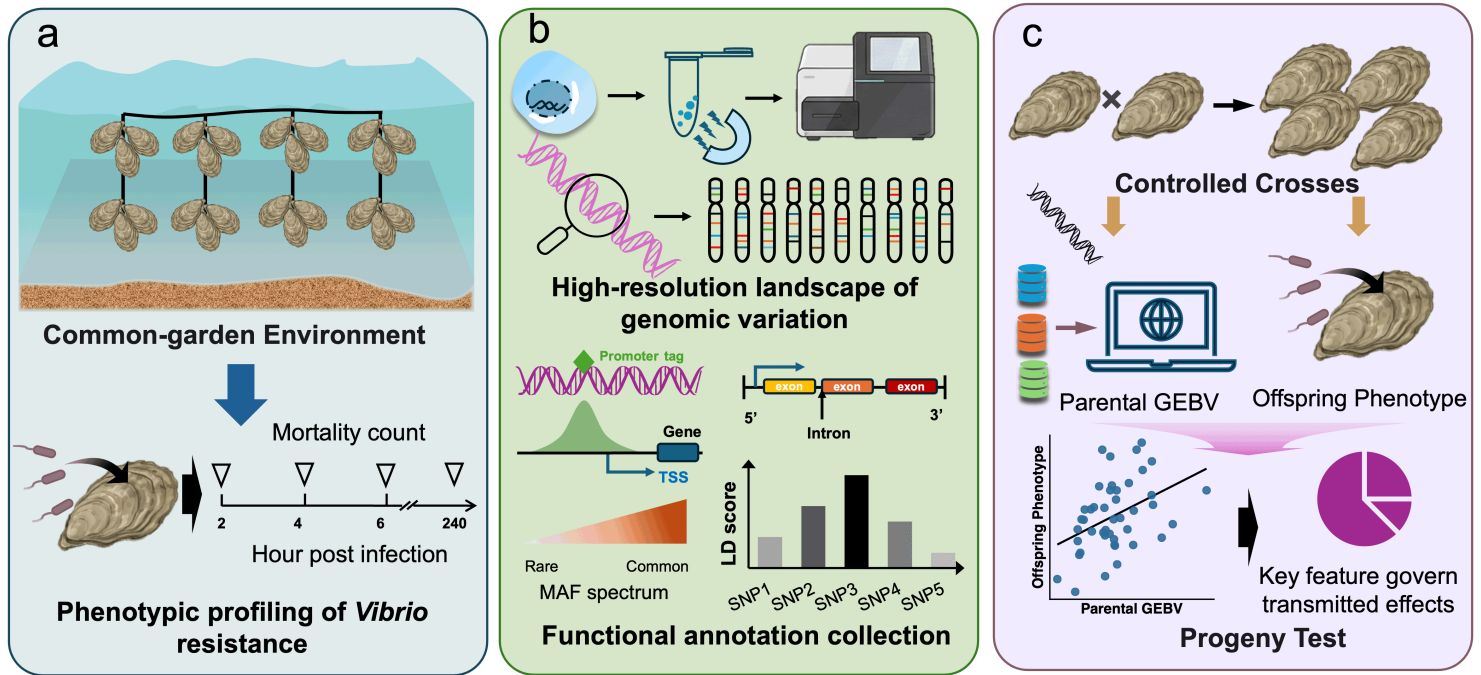


Figure 1

Integrated experimental and analytical framework for quantifying transmitted additive effects. a, Experimental common-garden rearing and *Vibrio* challenge. Diverse oyster lineages were reared under standardized marine conditions in a large-scale common-garden experiment, followed by a controlled challenge with *V. alginolyticus* to quantify phenotypic variation in disease resistance while minimizing environmental confounding. b, High-resolution genomic profiling and integration of biological priors. Comprehensive genome-wide variant discovery was achieved through low-coverage whole-genome sequencing (lcWGS) and genotype imputation to ensure sufficient genomic coverage against rapid linkage disequilibrium decay. These high-density variants were subsequently integrated with a panel of functional and evolutionary annotations to prioritize biologically informative markers from the vast pool of polymorphisms. c, Empirical validation of transmitted genetic merit through progeny test. Parental genomic estimated breeding values (GEBV) were validated against the realized phenotypic performance of their offspring in a cross-generational design, enabling the direct quantification of the transmitted genetic effects.

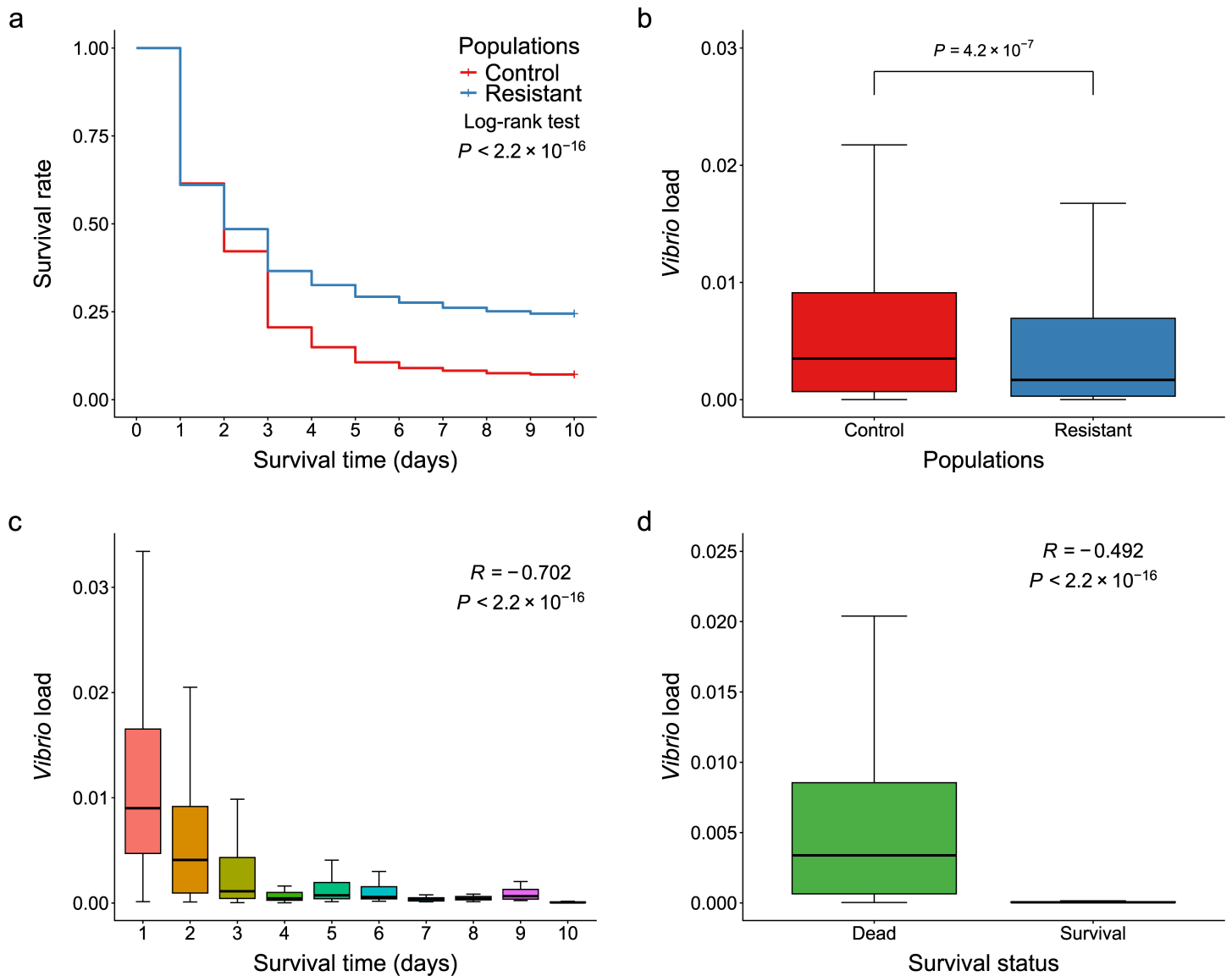


Figure 2

Phenotypic divergence in *Vibrio* resistance and associations between bacterial load and survival traits. a, Survival curves of resistant and control populations. Kaplan-Meier survival analysis comparing selectively bred resistant oysters and unselected controls during *V. alginolyticus* challenge. Resistant oysters show significantly higher survival rate across the 10-day trial (log-rank test, $P < 2.2 \times 10^{-16}$). b, Comparison of *Vibrio* load between populations. Distribution of bacterial load in resistant versus control oysters. Resistant oysters exhibit significantly lower bacterial load ($P = 7.8 \times 10^{-7}$), consistent with enhanced resistance. c, Negative association between *Vibrio* load and survival time. Boxplots showing *Vibri*oload across individuals grouped by survival time. Bacterial load declines sharply with increasing survival duration (Spearman's $r = -0.702$, $P < 2.2 \times 10^{-16}$), indicating that load is a quantitative biomarker of resistance. d, Differences in *Vibrio* load between survival status groups. Oysters that died during the challenge exhibit substantially higher bacterial load than survivors (Spearman's $r = -0.492$, $P < 2.2 \times 10^{-16}$), further supporting pathogen burden as a robust indicator of disease outcome.

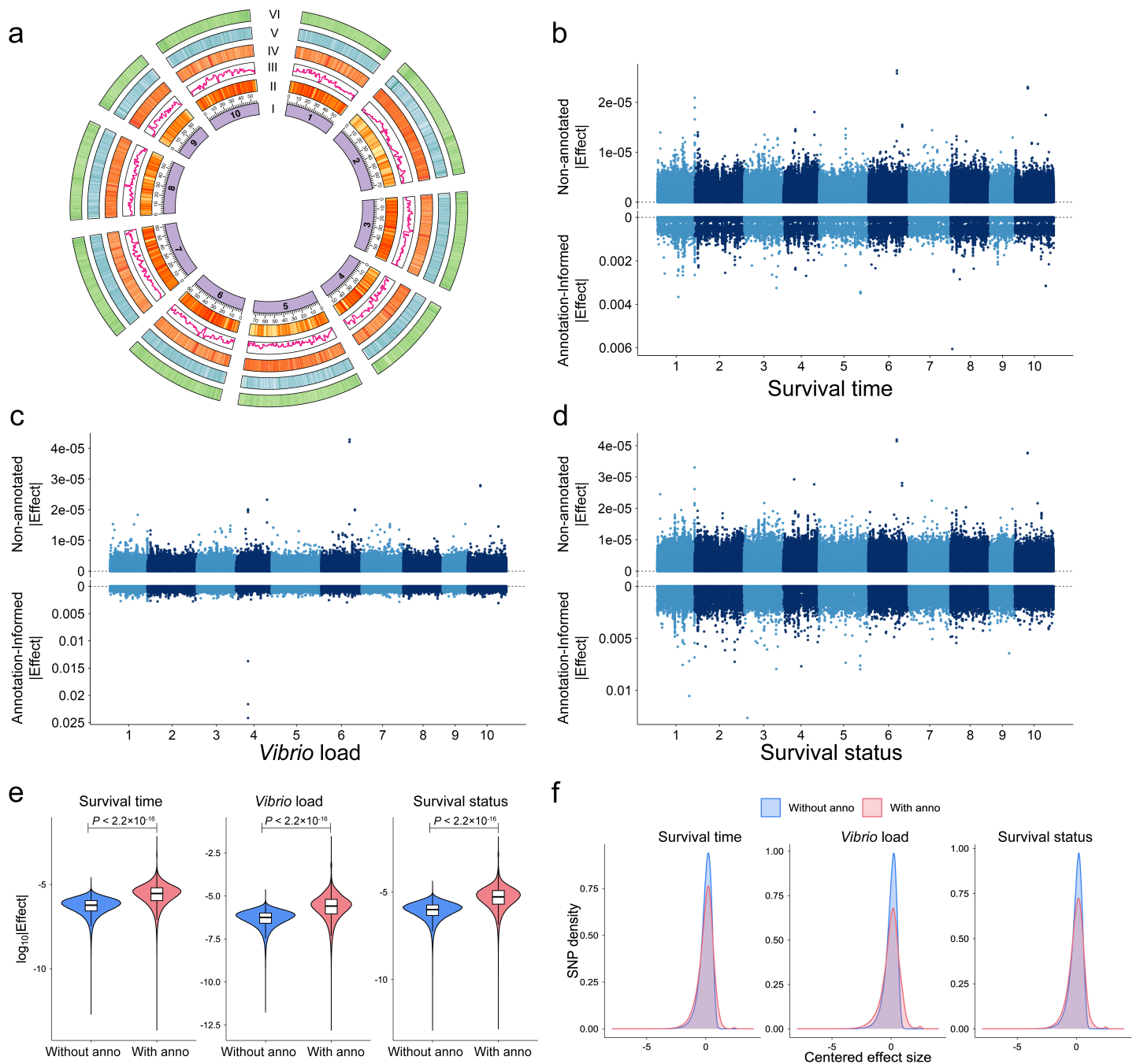


Figure 3

Annotation-informed models reshape the genome-wide distribution of SNP effects. **a**, Circos overview of genome-wide SNP architecture and genomic features across the *C. gigas* genome. From inner to outer rings: (I) chromosomal scaffolds, (II) SNPs density, (III) LD score distribution, (IV) transposable element (TE) density, (V) gene density and (VI) GC content. **b**, Effect-size landscape for survival time trait with and without annotation information. Mirror Manhattan plot showing the absolute SNPs effect sizes estimated by models without annotation (top) and with annotation (bottom). Annotation-informed weighting markedly increases effect-size magnitude and reshapes the distribution of influential loci

across the genome. c, Effect-size landscape for *Vibrio* load trait with and without annotation information. d, Effect-size landscape for survival status trait with and without annotation information. e, Pairwise comparison of SNPs effect sizes between annotation-informed and non-annotated models ($P < 2.2 \times 10^{-16}$). f, Distribution of scaled SNP effect sizes across models. Density distributions of scaled effect sizes after mean-centering. Non-annotated model effects (blue) show a narrow, low-variance distribution, whereas annotation-informed effects (red) exhibit broader spread and heavier tails, consistent with increased resolution in distinguishing informative variants from background noise.

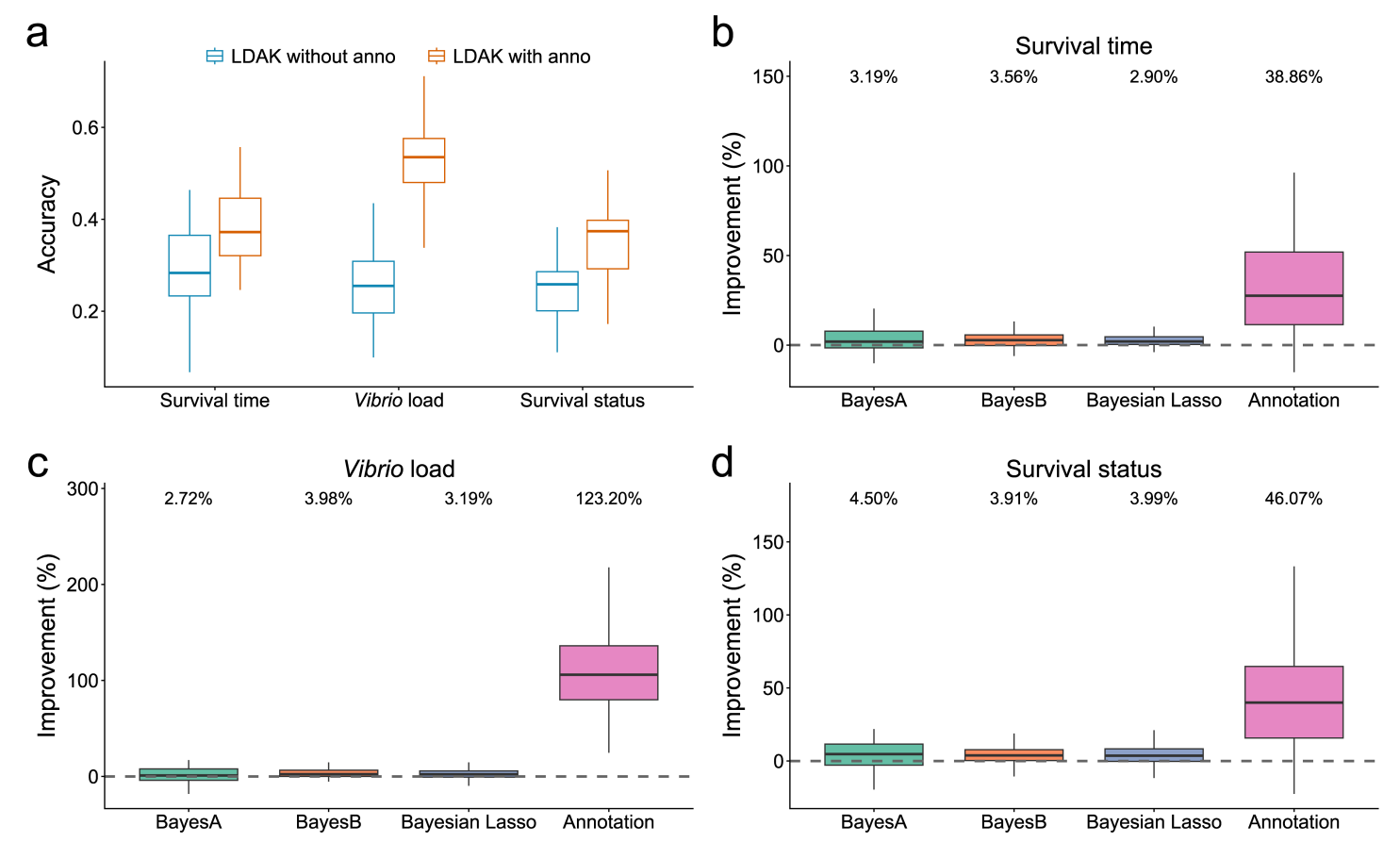


Figure 4

Annotation-informed models improve genomic prediction accuracy across traits. (a). Genomic prediction performance of LDAK models with and without annotation information for survival time, *Vibrio* load, and survival status. Incorporating functional annotations consistently increases prediction accuracy relative to unannotated models across all traits. The relative improvement in accuracy (percent increase compared with GBLUP) for survival time (b), *Vibrioload* (c), and survival status (d) when using Bayes A, Bayes B, Bayesian Lasso, and the annotation-informed model.

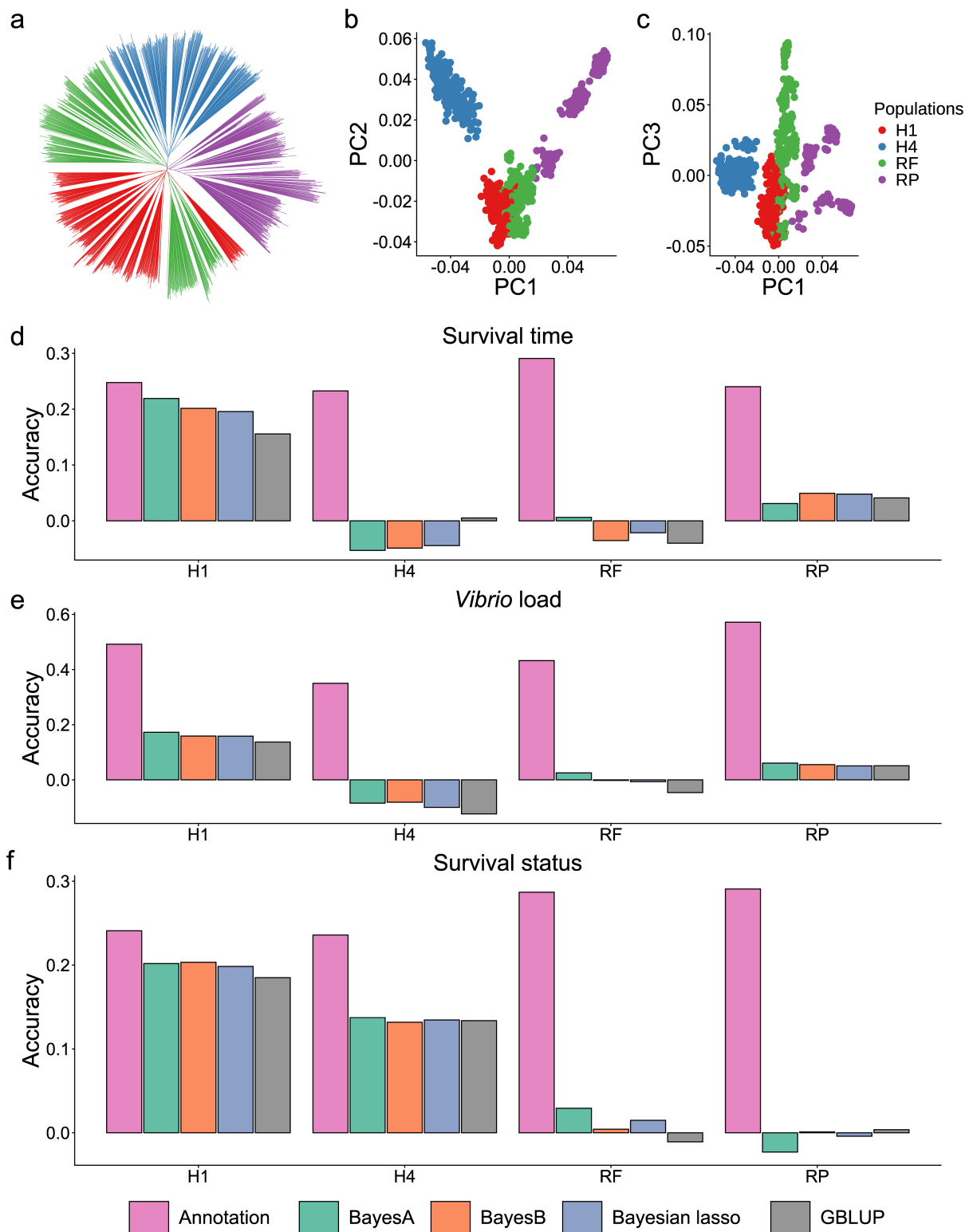


Figure 5

Population structure and cross-population genomic prediction performance. a, A neighbor-joining tree depicting relationships among the four breeding populations (H1, H4, RF and RP), showing well-defined genetic clusters with limited overlap. b and c, Principal component analyses based on genome-wide SNPs, with PC1–PC2 and PC1–PC3, further illustrating the strong genetic differentiation among populations. d–f, Cross-population genomic prediction accuracy for survival time (d), *Vibrio* load (e) and survival status (f).

survival status (f) under a leave-one-population-out framework, where models were trained on three populations and validated in the held-out population.

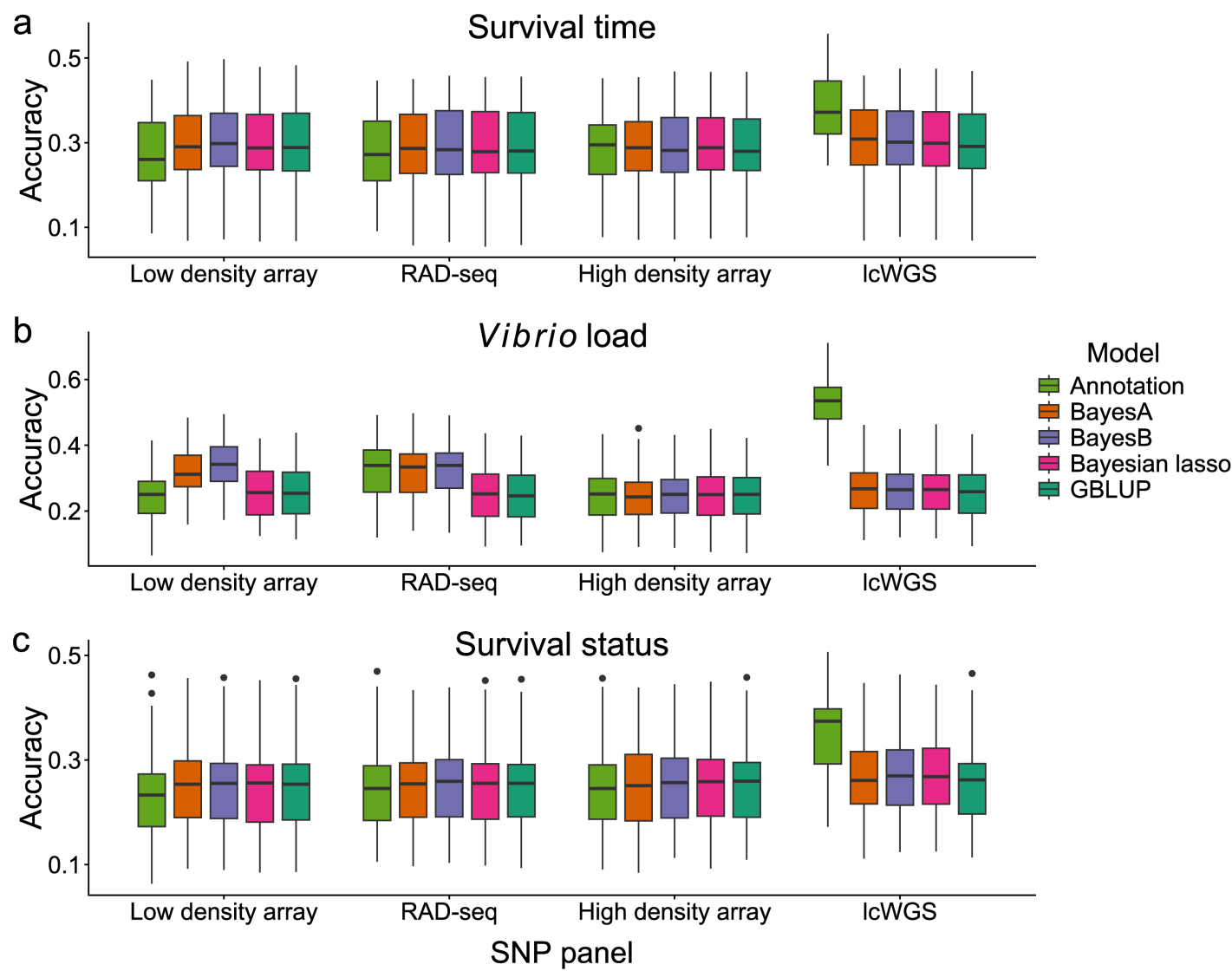


Figure 6

Effects of SNPs density on genomic prediction accuracy across models and traits.

a–c, Prediction accuracies for survival time (a), *Vibrio* load (b) and survival status (c) across four SNP panels: a low-density SNP array (~27K), a ddRAD-seq panel (~48K), a high-density SNP array (~190K) and the full genome-wide variant set (~9.23M) from lcWGS.

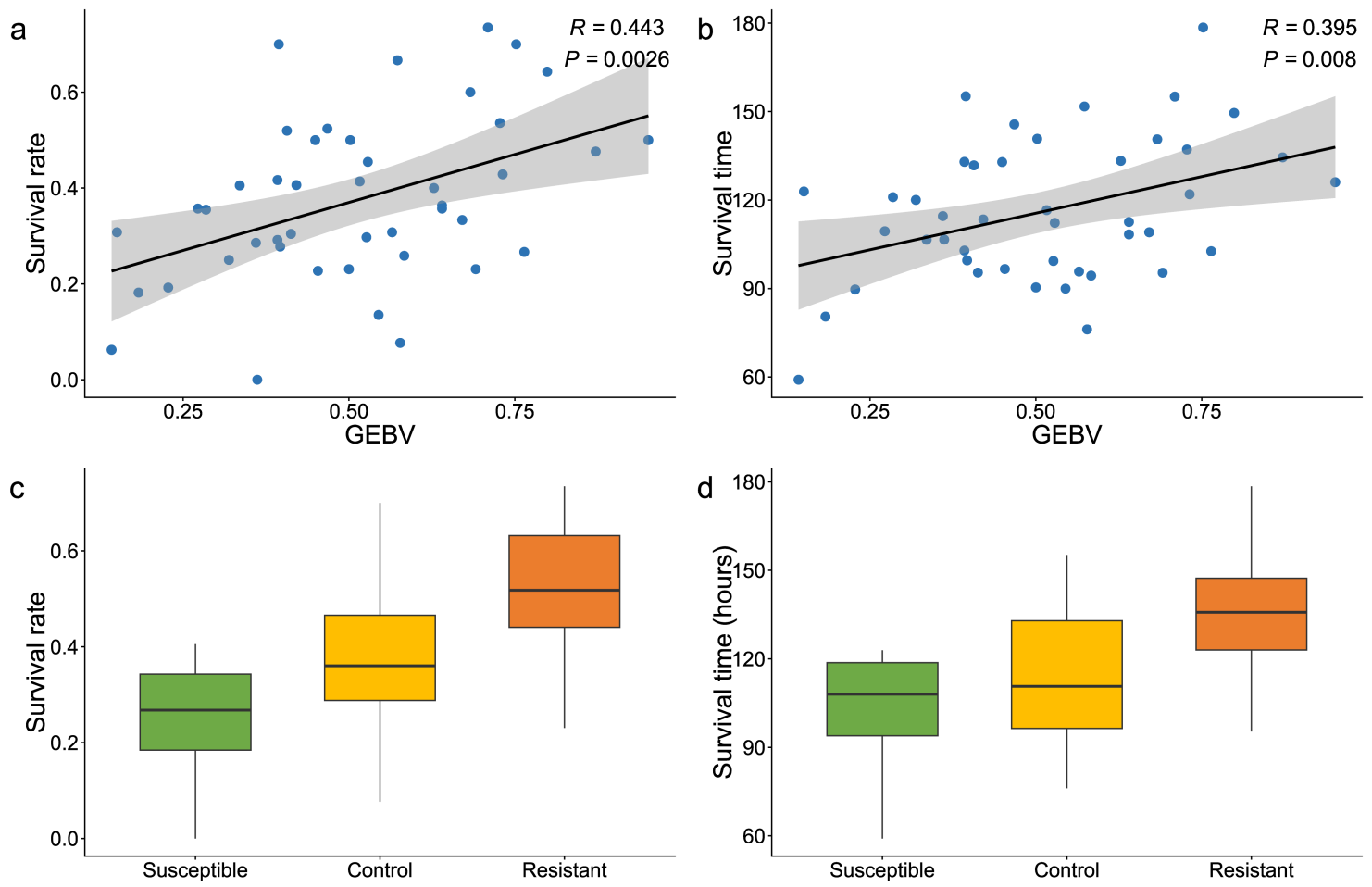


Figure 7

Progeny validation of annotation-informed genomic estimated breeding values (GEBV). Correlations between parental GEBV and offspring performance for survival rate (a) and survival time (b). Each point represents a full-sib family, with linear regression lines and 95% confidence intervals shown in grey. Offspring survival rate (c) and survival time (hours) (d) across families classified as susceptible, control, or resistant based on parental aggregate breeding values. Families ranked as resistant show markedly higher survival metrics compared with control and susceptible groups, confirming that annotation-informed GEBV effectively discriminate breeding candidates and translate into improved realized genetic gain.

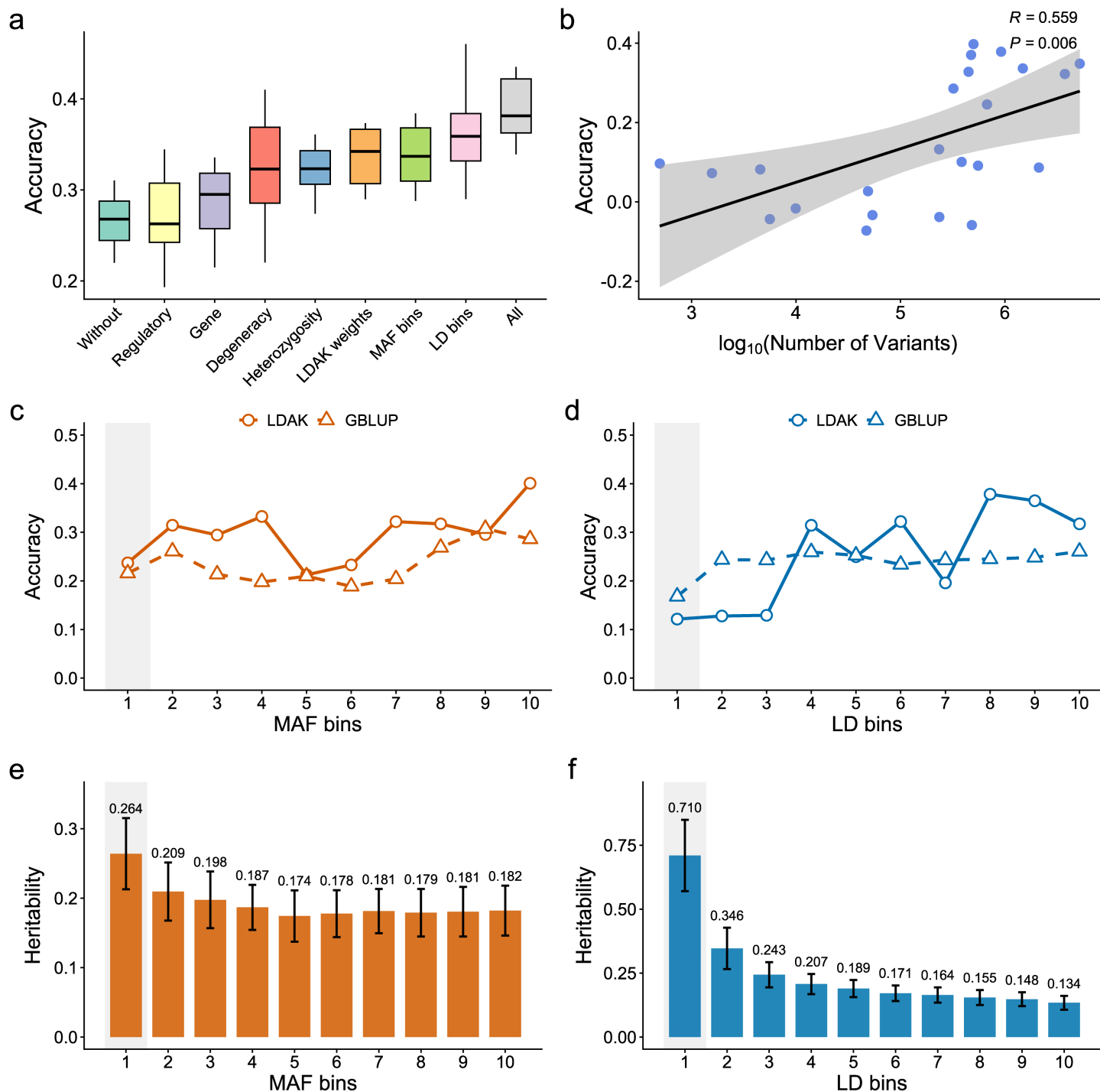


Figure 8

The LD and MAF annotations capture the majority of transmitted prediction accuracy. **a**, Prediction accuracy of models incorporating individual annotation categories, including regulatory features, gene-structural elements, codon degeneracy, heterozygosity, LDK weights, MAF bins, LD bins, and the full annotation set. **b**, Positive association between prediction accuracy and the number of SNPs contained in each annotation category, indicating that part of the apparent contribution of many functional annotations reflects SNP abundance. **c**, Prediction accuracy across MAF bins for the LDK (solid) and

GBLUP (dashed) model, the grey shaded region marks bin 1. d, Prediction accuracy across LD bins for the LDAK (solid) and GBLUP (dashed) models, the grey shaded region marks bin 1. e, Mean per-bin heritability (bars) across three traits for MAF bins with error bars ($\pm$ standard deviation across traits); bar labels indicate the mean value, the grey shaded region marks bin 1. f, Mean per-bin heritability (bars) across three traits for LD bins with error bars ($\pm$ standard deviation across traits); bar labels indicate the mean value, the grey shaded region marks bin 1.

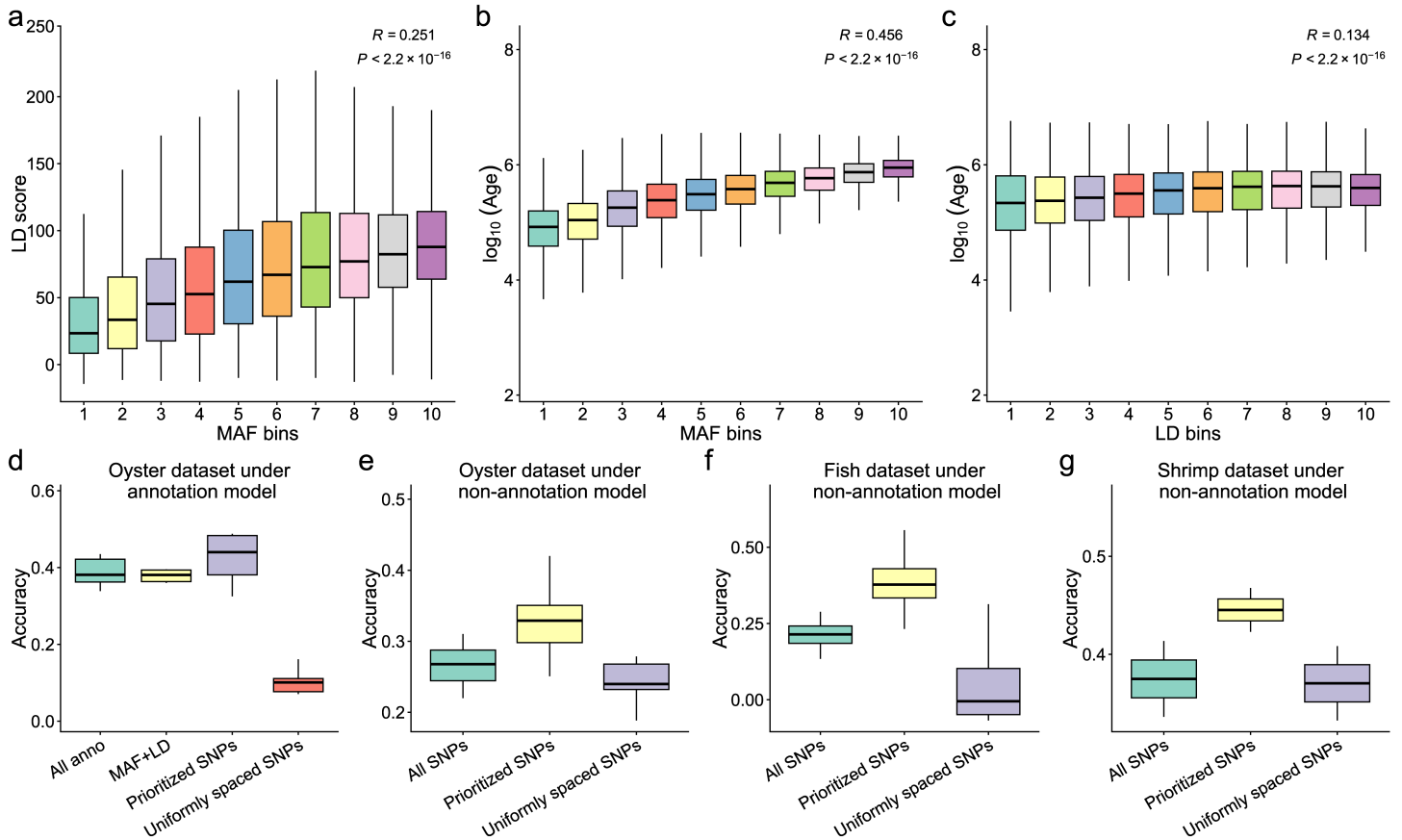


Figure 9

Ancient, high LD and high MAF variants drive transmitted additive prediction accuracy and enable optimized SNP selection. a–c, Relationships among MAF, LD score and estimated SNP age. LD scores increase with MAF (a), and SNP age increases with both MAF (b) and LD score (c), indicating that variants at higher MAF and in stronger LD tend to be evolutionarily older. d, Prediction accuracy of models using all annotations, only MAF and LD annotations, a prioritized SNP subset selected from the highest MAF and highest LD bins, and size-matched uniformly spaced SNP subsets. e, Performance of the prioritized SNP subset in models without annotation information, compared with corresponding models using all SNPs and uniformly spaced SNP subsets. f, Cross-species evaluation in an independent large yellow croaker dataset shows that prioritizing SNPs with high MAF and high LD markedly improves transmitted prediction relative to models using all SNPs or uniformly spaced SNP subsets. g, Consistent improvement is observed in an independent Pacific white shrimp dataset, demonstrating that the high-LD and high-MAF SNP prioritization strategy robustly enhances transmitted prediction across species.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables.xlsx](#)
- [SupplementaryMaterialsSL260204.docx](#)