Aquaculture Reports 45 (2025) 103088

Contents lists available at ScienceDirect

Aquaculture Reports

journal homepage: www.elsevier.com/locate/aqrep

Optimizing genotype imputation pipeline for low-coverage whole genome sequencing data in spotted sea bass and its application in genomic prediction

Chong Zhang [a], Yonghang Zhang [a], Pengyu Li [a], Cong Liu [a], Lingyu Wang [a], Yani Dong [a], Donglei Sun [a], Xin Qi [a], Haishen Wen [a], Kaiqiang Zhang [a], Shaosen Yang [b], Yun Li [a,*]

[a] *Key Laboratory of Mariculture, Ministry of Education (KLMME), Ocean University of China, Qingdao 266003, China*
[b] *Agro-Tech Extension Center of Guangdong Province, Guangzhou 510500, China*

ARTICLE INFO

Keywords:
Low-coverage whole genome sequencing
Genotype imputation
Reference panel
Lateolabrax maculatus
Genomic prediction

ABSTRACT

Genotype imputation following low-coverage whole genome sequencing (lcWGS) data offers a cost-effective approach for genotyping large populations, with significant potential to accelerate genomic selection in breeding programs. For spotted sea bass (*Lateolabrax maculatus*), genetic improvement is urgently required due to the degeneration of genetic traits and long generation intervals. However, the high costs associated with high-coverage WGS (hcWGS) for large populations have delayed breeding progress. To address this gap, the present study conducted a comprehensive evaluation of genotype imputation for lcWGS data down-sampled from 1107 individuals across four hcWGS datasets and aimed to develop an efficient imputation pipeline utilizing lcWGS data for spotted sea bass. Initially, 100data dataset was selected to preliminary assess the performance of various imputation pipelines. BEAGLE was excluded due to its lower accuracy and redundant computational requirements, while STITCH and GLIMPSE2 were retained for subsequent analyses. The effects of reference and target data on GLIMPSE2 imputation were then evaluated, identifying the optimal strategy for constructing the reference panel prioritizes population genetic diversity over sample size to maximizes imputation accuracy. It also highlighted the critical role of population structure, genetic relatedness and linkage disequilibrium (LD) level between reference and target data for imputation accuracy. Additionally, the imputation accuracy of STITCH and GLIMPSE2 was compared across three datasets, with GLIMPSE2 imputation using the optimal reference panel emerging as the most effective imputation pipeline for spotted sea bass. Finally, we demonstrated that lcWGS data combined with GLIMPSE2 imputation achieves predictive accuracy comparable to hcWGS data in genomic prediction. Our study presents an optimized workflow to impute lcWGS data in spotted sea bass and establishes the first publicly available reference panel with the highest known genetic diversity. This resource lays a crucial foundation for future genomic selection and breeding programs and serves as a valuable reference for genotype imputation in other aquaculture species.

1. Introduction

The rapid advancement of high-throughput genome sequencing technologies has spurred the development and application of diverse genotyping methods, including whole genome resequencing (WGS), single nucleotide polymorphism (SNP) arrays, and reduced representation sequencing (RRS), for population-level genetic variation studies (Peng et al., 2016; Zhang et al., 2023; Zhou et al., 2019). These methods have become fundamental tools in revealing the genetic architecture of complex traits through genome-wide association studies (GWAS) and accelerating genetic breeding programs via genomic selection (GS) (Georges et al., 2019; Gong et al., 2021; Visscher et al., 2017). While low- and medium-density SNP panels from arrays or RRS are cost-effective, WGS provides superior GWAS resolution and accuracy in identifying candidate loci due to its ultra-dense SNP coverage (Fernandes Garcia et al., 2022; Höglund et al., 2019). Similarly, accurate genomic prediction (GP) requires high-density SNP data across large cohorts (Iheshiulor et al., 2016; Tsai et al., 2017). However, the

[*] Corresponding author.
E-mail address: yunli0116@ouc.edu.cn (Y. Li).

https://doi.org/10.1016/j.aqrep.2025.103088
Received 26 February 2025; Received in revised form 27 July 2025; Accepted 2 September 2025
Available online 9 September 2025

prohibitive cost of high-coverage WGS genotyping for large populations remains a major barrier, limiting its widespread application in GWAS and GS for aquaculture breeding programs. Consequently, developing cost-effective strategies to obtain high-density genotype data has become an urgent priority.

To bridge this gap, low-cost genotyping strategies coupled with genotype imputation have thus emerged as a cost-effective solution to obtain high-density genotype data across large populations at minimal expense (Huang et al., 2012; Tsai et al., 2017). Genotype imputation typically involves two key steps: (1) constructing haplotype reference panels (HRPs) using high-density genotyping data from represented individuals, and (2) inferring and imputing missing genotypes or ungenotyped markers in low-density SNP panels (Browning and Browning, 2009); (Davies et al., 2016); (Zhang et al., 2022). These approaches maximize the utility of genomic data by imputing low-density SNP panels to high-density SNP data, even up to WGS level, and have been successfully applied in several breeding programs for economically important livestock and crop species (Fernandes Júnior et al., 2021; Hayes et al., 2012; Hickey et al., 2012). Despite their success, several challenges in these imputation approaches remain unresolved. For example, common SNP arrays are more prone to bias in capturing genetic variation and are limited in detecting novel variants compared to WGS, thereby constraining imputation accuracy (Lachance and Tishkoff, 2013; Zhang et al., 2023). Additionally, high-quality HRPs for comprehensive genome-wide imputation are often unavailable for non-human species, necessitating the construction of HRPs from high-coverage WGS data of the same or closely related populations, which can be leveraged to significantly improve imputation accuracy (Hayward et al., 2019; Ji et al., 2019; Sargolzaei et al., 2014). Therefore, array-based genotype imputation in aquaculture species has primarily focused on Atlantic salmon (*Salmo salar*) and Nile tilapia (*Oreochromis niloticus*), both globally significant breeding species with high-quality commercial SNP arrays and well-established breeding programs (Fernandes Garcia et al., 2022; Tsai et al., 2017; Tsairidou et al., 2020; Yoshida and Yáñez, 2021). The lack of SNP arrays and established pedigree populations for most aquaculture species continues to hinder the widespread application of array-based genotype imputation in breeding programs (Zhang et al., 2021).

Given these challenges, low-coverage whole genome sequencing (lcWGS) has emerged as a promising, low-cost alternative for the imputation of complete genotypes (Pasaniuc et al., 2012). Compared with SNP array and RRS strategy, lcWGS maximizes coverage breadth at the expense of sequencing depth, capturing more comprehensive genetic variation of whole genome, including population-specific variants (Lou et al., 2021). The use of lcWGS provides greater power for GWAS in detecting associated signals compared to SNP arrays (Alicia et al., 2021; Arthur et al., 2019; Gilly et al., 2016). Moreover, genotype data generated by lcWGS can be further imputed to WGS level using genotype imputation strategies, which mainly including two categories: those relying on genotype reference panels and reference-free approaches (Zhang et al., 2021; Zhang et al., 2022). For instance, using a high-quality reference panel phased by Beagle v5.4, lcWGS data imputed with GLIMPSE2 achieved an average concordance rate greater than 0.99 in cattle (Zhang et al., 2023). Similarly, STITCH imputation, a reference-free imputation method based solely on genetic sites information, achieved a genotype concordance rate above 0.99 and was identified as the optimal imputation strategy in rabbits (Wang et al., 2022). In recent years, lcWGS-based genotype imputation has been increasingly applied to aquatic species, including large yellow croaker (*Larimichthys crocea*), Russian sturgeon (*Acipenser gueldenstaedtii*), Pacific oyster (*Crassostrea gigas*) and Scallops (Song et al., 2024; Wang et al., 2025; Yang et al., 2024; Zhang et al., 2021). However, most studies have focused exclusively on either reference panel-based or reference-free approaches, and systematic comparisons between these strategies remain in their infancy. Given the limited availability of large-scale reference panels in aquaculture species, it is essential to investigate whether constructing reference panels from a small set of high-coverage WGS (hcWGS) datasets can enhance imputation accuracy compared to reference-free methods.

Spotted sea bass (*Lateolabrax maculatus*) is a promising candidate for aquaculture in China due to its significant market demand and potential for genetic improvement (Zhang et al., 2023). Considering the higher cost of WGS and the absence of SNP arrays, genotype imputation using lcWGS presents a cost-effective genotyping solution for large populations of spotted sea bass. Therefore, establishing an efficient imputation pipeline based on lcWGS is essential for leveraging genomic resources and facilitating selective breeding at minimal expense. In this study, genotype imputation was performed using lcWGS data down-sampled from 1107 hcWGS data form four datasets, and their sequencing depth, linkage disequilibrium and population structure were captured. Initially, 100data was selected to conduct a preliminary comparison of imputation accuracy across various pipelines. Subsequently, we thoroughly evaluated the impact of reference and target data on GLIMPSE2 imputation, and the first reference panel of spotted sea bass was constructed by combining all resequencing data and publishing with open access. Additionally, a systematic comparison was then made between two specific imputation pipelines: one relying on a haplotype reference panel and one not requiring reference panel. Finally, the feasibility of genotype imputation in genomic selection was assessed by comparing the accuracies of GP using hcWGS and imputed lcWGS data. Our study provided the optimal imputation pipeline for largescale lcWGS data, demonstrating the potential of lcWGS for genomic selection in spotted sea bass. This work provides a valuable reference for lcWGS-based studies in other aquaculture species.

2. Materials and methods

2.1. Sample collection and whole genome resequencing

In this study, a total of 1107 spotted sea bass samples with high-coverage WGS data were collected. Specifically, 1007 samples were sourced from three local fish farms in Dongying (DY), Tangshan (TS), and Yantai (YT), China. This included 301 one-year-old fish from DY population, collected from natural populations in the Yellow Sea and Bohai Sea, 213 five-year-old broodstock from TS population, and 493 two-year-old fish from YT population, derived from northern and southern cultivated populations. Growth traits, including total length (TL) and body weight (BW), were measured for each individual, and pectoral fin samples were stored in anhydrous ethanol for DNA extraction. Genomic DNA was extracted with TIANamp Genomic DNA Kit (TIANGEN, Beijing, China). WGS libraries for 513 samples from DY and TS populations were constructed and sequenced via BGISEQ-500 platform to generate paired-end 100 bp reads, following the protocols described previously (Fang et al., 2018). For YT population, WGS libraries were prepared using NEBNext® UltraTM DNA Library Prep Kit and sequenced on the DNBSEQ-T7 platform to generate paired-end 150 bp reads. Additionally, we downloaded 100 accessions, published in recent genome resequencing studies (Chen et al., 2023), from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (PRJNA701455). These libraries were constructed using Illumina DNA preparation kits, and 150-bp paired-end reads were generated on the Illumina HiSeq 4000 platform. Therefore, four WGS datasets, referred to as 100data, DY, TS and YT, were used in this study.

2.2. Variant calling and quality control

All raw sequence data were filtered by Fastp v0.20.0 (Chen et al., 2018) and subsequently aligned to the *L. maculatus* reference genome (JAYMHB000000000) using the BWA-MEM algorithm in BWA v0.7.17 (Li and Durbin, 2010) with default parameters. The resulting Sequence Alignment Map (SAM) files were converted into Binary Alignment Map (BAM) files, which were then indexed and sorted using SAMtools v1.17

(Li et al., 2009). Duplicate reads were identified and excluded by picard v1.90 (http://broadinstitute.github.io/picard/). Finally, SNP calling and joint genotyping were performed with GATK v4.1.8 (McKenna et al., 2010), followed by hard filtering with QD $\geq$ 2.0 || FS $\leq$ 60.0 || SOR $>$ 3.0 || RMS mapping quality $\geq$ 40.0 || MQRankSum $\geq$ -12.5 || ReadPosRankSum $\geq$ -8.0. After variant calling, SNPs with a missing rate above 5 %, minor allele frequency (MAF) below 5 %, and non-biallelic SNPs were excluded using BCFtools v1.20 to obtain the final SNP database.

2.3. Sequences depth, linkage disequilibrium (LD) and population structure analysis

First, the sequencing depth of each sample was captured using Mosdepth v0.2.5 (Pedersen and Quinlan, 2018) for subsequent down-sampling process. The LD coefficient (r^2) for each dataset was calculated using the PopLDdecay v3.41 package with the parameters of "-MaxDist 300 kb" (Zhang et al., 2019). To better understand the population structure of all samples, principal component analysis (PCA) was performed using Plink v1.9 (Purcell et al., 2007). Additionally, the genetic groups of each sample were further investigated using Admixture v1.3.0 (Alexander et al., 2009), with the number of clusters (*K*) ranging from 2 to 7. The *K* value with the smallest CV error was assumed to be the optimal population stratification number, and individuals with *q* values of genetic components greater than 50 % were assigned to their corresponding population. Finally, genetic relatedness of population was generated using GEMMA v0.98.1 (Zhou and Stephens, 2012) and visualized with a heatmap using the hist function in R.

2.4. Evaluation of the accuracy of different genotype imputation pipelines

To preliminary access the accuracy of different genotype imputation pipelines, we selected the 100data as test dataset for three categories of imputation pipelines: (1) STITCH (v1.6.6) imputation for BAM files, based solely on SNP site information without a reference panel (Davies et al., 2016), (2) GLIMPSE2 (v2.0.0) imputation for BAM files relying on a reference panel (Rubinacci et al., 2023), and (3) BEAGLE (v4.1) imputation for VCF files using a reference panel (Browning et al., 2018). First, all 1007 samples of DY, TS and YT datasets were included to construct a reference panel using SHAPEIT5 following default parameters (Hofmeister et al., 2023). Then, to investigate the impact of sequencing depth on genotype imputation accuracy, we randomly down-sampled paired-reads from the BAM files (average coverage is 15.9 ×) of 100 samples to lcWGS data with depths of 1 ×, 3 × and 5 × using DownsampleSam in Picard tools. STITCH and GLIMPSE2 accepted down-sampled BAM files as input, while BEAGLE requires VCF files. SNP calling and quality control for the down-sampled BAM files followed the same procedures described in Section 2.2 "Variant Calling and Quality Control". Additionally, the pilot study evaluated the impact of the K value (number of ancestral haplotypes) on imputation accuracy, finding that a K value of 25 is optimal for STITCH imputation in this study, while GLIMPSE2 and BEAGLE were used with default parameters. Considering imputation efficiency and the robustness of imputation performance across chromosomes, three chromosomes (chr1, chr8, and chr24) of lcWGS data were chosen for the comparison of imputation accuracy. Two metrics were introduced to evaluate the imputation accuracy, including genotype concordance (GC) and the squared Pearson correlation coefficient of genotype dosage (R^2) between the imputed lcWGS data and corresponding hcWGS data (Browning et al., 2018). The calculation formulas for GC are defined as follows:

$$\mathrm{GC} = \frac{(\mathrm{mRef/Ref} + \mathrm{mRef/Alt} + \mathrm{mAlt/Alt})}{(\mathrm{xRef/Ref} + \mathrm{xRef/Alt} + \mathrm{xAlt/Alt} + \mathrm{mRef/Ref} + \mathrm{mRef/Alt} + \mathrm{mAlt/Alt})}$$

where **m** means number of matches between imputed and observed genotype, **x** means number of mismatches between imputed and observe genotype.

The calculation formulas for R^2 are defined as follows:

$$\mathrm{R}^2 = \left(\frac{\left(\sum_{\{i=1\}}^{n} (\mathrm{lg}_i - \bar{l}g)(hg_i - \bar{h}g \quad) \right)}{\sqrt{\sum_{\{i=1\}}^{n} (\mathrm{lg_i} - \bar{\mathrm{l}}\mathrm{g})^2 * \sum_{\{i=1\}}^{n} (\mathrm{hg_i} - \mathrm{hg})^2}} \right)^2$$

where genotypes were coded as 0, 1, or 2, representing the number of the minor allele; $\boldsymbol{lg_i}$ is the imputed genotype for individual **i** in lcWGS data, and $\bar{l}g$ denotes mean imputed genotype value across all individuals; $\boldsymbol{hg_i}$ and $\bar{h}g$ were the observed true genotypes and mean value of the observed genotypes calling by hcWGS data, and **n** is the number of all individuals used for genotype imputation.

2.5. Effect of reference and target data on imputation accuracy

To investigate the impact of reference data on GLIMPSE2 imputation, the YT dataset, characterized by heterogeneous population structure, was selected as test data for genotype imputation using three different reference panel construction strategies varying in population genetic diversity and sample size. In detail, SHAPEIT5 was used to construct following reference panel: (1) The reference panel called "DY+TS" was constructed by integrating 613 samples from 100data, DY and TS datasets. (2) The reference panel called "YT" was constructed from YT samples only. To avoid overestimation of imputation accuracy caused by overlap between the imputation samples and the reference panel construction samples, we employed a five-fold cross-validation approach. In this approach, 80 % of the samples from each population, selected based on the optimal YT population stratification, were used to construct the reference panel, while the remaining 20 % were designated as the imputation dataset. This approach ensured genetic diversity in the reference panel and minimized the risk of accuracy overestimation. This procedure was repeated five times until all YT samples were imputed. (3) The comprehensive "ALL" reference panel, combining the previous two strategies, was formulated through 613 samples from 100data, DY and TS datasets, together with 80 % of the YT samples (394). The remaining 20 % of YT samples served as the target data for imputation. The five-fold cross-validation procedure was repeated until all YT samples were imputed. Additionally, to investigate the impact of target data on GLIMPSE2 imputation, imputation accuracy was compared across DY, TS, and YT datasets, which vary in population structure, genetic relatedness, LD level and sample size, using the optimal reference panel. All BAM sequence files of DY, TS and YT datasets were randomly down-sampled to varying lcWGS levels of 0.5 ×, 1 ×, 2 ×, 3 ×, and 5 × using Picard. Three chromosomes (chr1, chr8, and chr24) of lcWGS data were selected for GLIMPSE2 imputation and imputation accuracy was assessed by calculating the GC and R^2 between the imputed and true genotypes.

2.6. Comparison of imputation accuracy between GLIMPSE2 and STITCH

Given the critical impact of sample size on the accuracy of STITCH imputation, which is highly sensitive to imputation samples to estimate ancestral haplotype for inferring missing genotypes. For example, STITCH imputation accuracy for Pacific oyster exhibited an increasing trend with the increase of sample size, and the accuracy tended to stabilize after the sample size reached 300 (Davies et al., 2016; Yang et al., 2024). Therefore, the lower accuracy of STITCH relative to GLIMPSE2 observed in the 100 test samples aligns with methodological expectations. This result shouldn't be interpreted as definitive evidence of GLIMPSE2's superiority, and we couldn't determine GLIMPSE2 as an optimal approach rather than STITCH. To address this critical dependency and determine the optimal genotype imputation pipeline for an adequate sample size, we systematically compared the imputation accuracy of GLIMPSE2 and STITCH for DY, TS and YT datasets. BAM files were down-sampled to lcWGS levels of 0.5 × , 1 × , 2 × , 3 × , and 5 × using Picard. For GLIMPSE2 imputation, the optimal reference panel was constructed following the third strategy described above. To accurately estimate ancestral haplotypes for STITCH imputation without a reference panel, all 1007 samples were included in the process. Three chromosomes (chr1, chr8, and chr24) from lcWGS data were selected for genotype imputation, and the GC and R^2 values between the imputed and true genotypes were compared for both GLIMPSE2 and STITCH.

2.7. Application of imputed lcWGS data in genomic prediction for growth traits

Owing to the high heterozygosity and extensive SNP density within the *Lateolabrax maculatus* genome, the utility of imputed SNP markers for selective breeding programs, particularly concerning polygenic traits, requires further validation. Therefore, we conducted genomic prediction (GP) using imputed lcWGS data for TL trait for DY and TS datasets, and BW trait for YT datasets. Based on the imputation performance in the three datasets, imputed 3 × lcWGS data generated by GLIMPSE2, combined with an optimal reference panel, were selected for subsequent genomic prediction, and the results were compared with those obtained using hcWGS data. Genomic predictions were carried out using a 10-fold cross-validation approach with five replicates. Specifically, 10 % of the samples were randomly selected as the validation set, while the remaining 90 % served as the training set for GWAS analysis using GEMMA v0.98.1 (Zhou and Stephens, 2012). Different numbers of SNPs with the smallest genome-wide *p* values from GWAS were chosen to model the genotype and true phenotype data in the training set, which was then used to predict the phenotypes in the validation set. Predictive accuracy was calculated as the Pearson correlation coefficient between the true and predicted phenotypes. Support Vector Machine (SVM), a powerful machine-learning algorithm from the kernel-based family, was employed for genomic prediction due to its outstanding predictive performance in complex trait analysis (Wang et al., 2022).

3. Results

3.1. SNP identification and statistics

After high-throughput sequencing and filtering, 3.91, 19.05, 13.50 and 21.94 billion pairs of clean reads were generated for 100data, DY, TS and YT datasets, respectively, with average sequencing depths of 15.93 × , 10.00 × , 10.04 × and 10.48 × (Fig. 1 **A**). Following variant calling and quality control, a total of 5244,698 SNPs were identified as common across all datasets, which were subsequently used as imputation markers in this study. These SNPs spanned a total physical distance of 622.44 Mb, with an average density of SNP/118 bp, indicating a dense and uniform distribution throughout the genome. Among the 24 chromosomes, chromosome 22 harbored the highest SNP marker density of SNP/105 bp, while chromosome 6 possessed the lowest SNP marker density of SNP/137 bp (Fig. S1, Table S1).

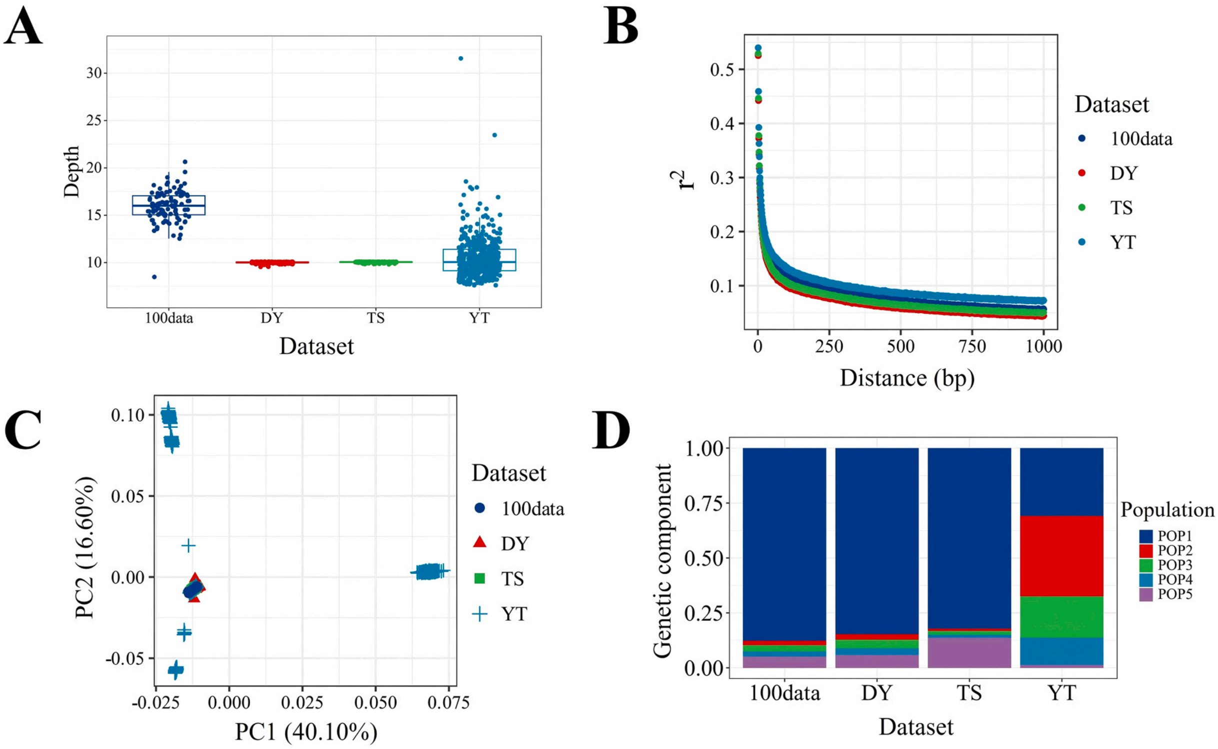

Fig. 1. (A) Box plot showing the sequencing depth for the four datasets. (B) LD decay plot of SNPs for the four datasets. (C) PCA plot of all individuals from the four datasets based on PC1 and PC2. (D) Bar plot showing the genetic components of the four datasets based on Admixture analysis. The numbers represent the corresponding populations.

3.2. LD and population structure analysis

The analysis of linkage disequilibrium (LD) revealed a rapid decline in the squared correlation coefficient (r^2) between loci as the distance between SNP pairs increased. At a distance of 250 bp, the r^2 values were 0.0932, 0.0768, 0.0826 and 0.1045 for 100data, DY, TS and YT datasets, respectively (Fig. 1B). Among these, YT dataset exhibited a relatively higher level of LD. The PCA results suggested that the YT dataset comprised complex genetic groups, while individuals from the other three datasets were more genetically homogeneous and clustered closely with a subset of YT individuals (Fig. 1 C). Admixture analysis identified the optimal population stratification number as 5 for all individuals (Fig. S2). Based on *q* values of the genetic components, 1, 1, 2, and 4 populations were assigned to the 100data, DY, TS, and YT datasets, respectively, further highlighting the higher genetic diversity in the YT dataset (Fig. 1D). Additionally, analysis of genetic relatedness revealed no detectable relatedness within the DY dataset, while weak genetic relatedness was observed in the TS dataset. Notably, significantly stronger genetic relatedness was identified within the YT dataset (Fig. S3).

3.3. Computational efficiency comparison of genotype imputation pipelines

To assess computational efficiency of different genotype imputation pipelines, we performed chromosome 1 (267,335 SNPs) imputation (Table S1) benchmarking at 3 × sequencing depth using Intel H3C R4900 G5 clusters with 60-thread parallelization, while constraining GATK to its default 4-thread implementation. The actual CPU hours consumed for different pipelines revealed significant disparities (Table S2). The BEAGLE pipeline consumed 79.0 CPU-hours, comprising 47.2 CPU-hours for GATK variant calling (11.8 h × 4 threads) and 31.8 CPU-hours for imputation (0.53 h × 60 threads). STITCH required 252.0 CPU-hours (4.2 h × 60 threads) despite direct BAM imputation. In contrast, GLIMPSE2 achieved optimal efficiency at 10.8 CPU-hours (0.18 h × 60 threads) through direct BAM imputation. This represents an 86.3 % reduction for GLIMPSE2 versus BEAGLE and a 95.7 % reduction versus STITCH, establishing GLIMPSE2 as the most computationally efficient pipeline.

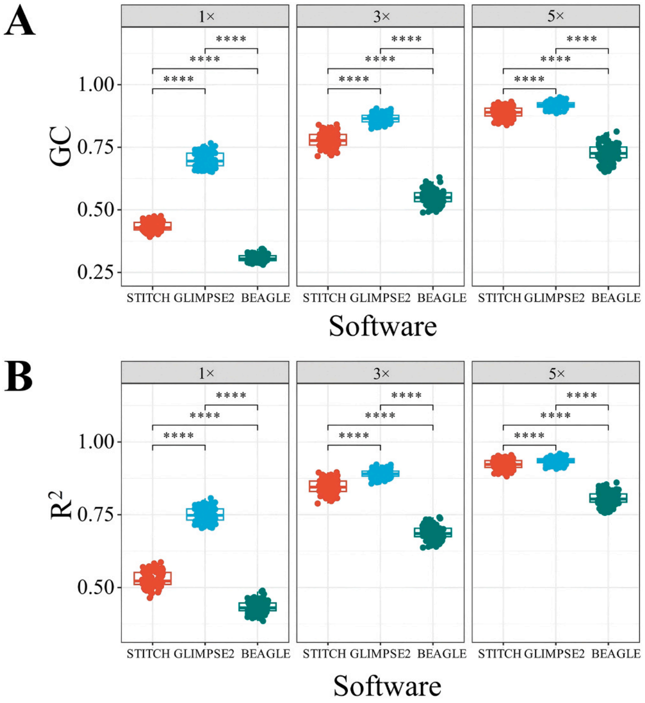

Fig. 2. The comparison of genotype imputation accuracy bewteen different imputation methods and sequencing depths for 100data dataset. (A) Estimated genotype concordance (GC) between imputed genotypes and true genotypes. (B) Estimated squared Pearson correlation coefficient (R^2) for genotype dosage between imputed and true genotypes.

3.4. Accuracy evaluation of different genotype imputation pipelines for 100data dataset

To preliminary determine the optimal genotype imputation pipeline, 100data dataset, characterized by a simple population structure (Fig. 1 C and D), was selected as test data for three imputation pipelines. One sample with a low sequencing depth (8.48 ×) was excluded due to deviation from the average (Fig. 1 A), and the imputation accuracy of three methods across various sequencing depths was assessed using GC and R^2 metrics (Fig. 2). Our results revealed significant differences ($P < 0.0001$) in accuracy among the three methods at various sequencing depths, although imputation accuracy generally improved with increasing sequencing depth (Fig. 2). Of which, GLIMPSE2 demonstrated the highest imputation accuracy, BEAGLE the lowest, and STITCH exhibited intermediate imputation accuracy. Notably, as sequencing depth increased, the accuracy gap between STITCH and GLIMPSE2 gradually narrowed. For example, at a sequencing depth of 5 × , GC values for STITCH, GLIMPSE2, and BEAGLE were 0.889, 0.919 and 0.727 (Fig. 2A, Table S3), respectively, and the corresponding R^2 values were 0.923, 0.935 and 0.806 (Fig. 2B, Table S3). Therefore, GLIMPSE2 was deemed a superior reference panel-based imputation method compared to BEAGLE. However, due to the relatively small sample size used for STITCH imputation, its performance warrants further evaluation.

3.5. Effect of reference and target data on GLIMPSE2 imputation accuracy

To further refine the optimal imputation pipeline, the effect of reference data on GLIMPSE2 imputation accuracy was evaluated for YT datasets. At a relatively high sequencing depth of 5 × , imputation accuracy across the three reference panels was comparable, with GC values of 0.935, 0.929 and 0.911, and R^2 values of 0.946, 0.942 and 0.926 for ALL, YT and “DY+TS” reference panels, respectively (Fig. 3, Table S4). However, as sequencing depth decreased, particularly at 0.5 × and 1 × , the imputation accuracy of “DY+TS” reference panel dropped significantly compared to the YT and ALL panels. At 0.5 × , the GC and R^2 values of “DY+TS” reference panel were 0.639 and 0.677, respectively (Fig. 3A and B, Table S4). These results indicate that when the reference panel lacks individuals of target population, the imputation accuracy is considerably reduced. When a subset of YT individuals was selected to construct reference panel (YT), the imputation accuracy significantly improved, with GC and R^2 values of 0.830 and 0.854, respectively, at 0.5 × (Fig. 3A and B, Table S4). Moreover, increasing population size within the reference panel (ALL) only slightly enhanced the imputation accuracy, with GC and R^2 values of 0.862 and 0.884, respectively, at 0.5 × (Fig. 3A and B, Table S4). Based on the optimal reference panel construction strategy (ALL), the imputation accuracies for DY and TS datasets were significantly lower than those of YT datasets at lower sequencing depths, especially at 0.5 × , the GC values were 0.601, 0.771 and 0.862 for DY, TS and YT datasets, respectively, with corresponding R^2 values of 0.645, 0.805 and 0.884 (Fig. 3C and D, Table S5).

3.6. Comparison of imputation accuracy between GLIMPSE2 and STITCH for three datasets

To compare the imputation accuracy between GLIMPSE2 and STITCH for adequate sample size, we conducted genotype imputation for all 1007 samples including three datasets. As shown in Fig. 4, the

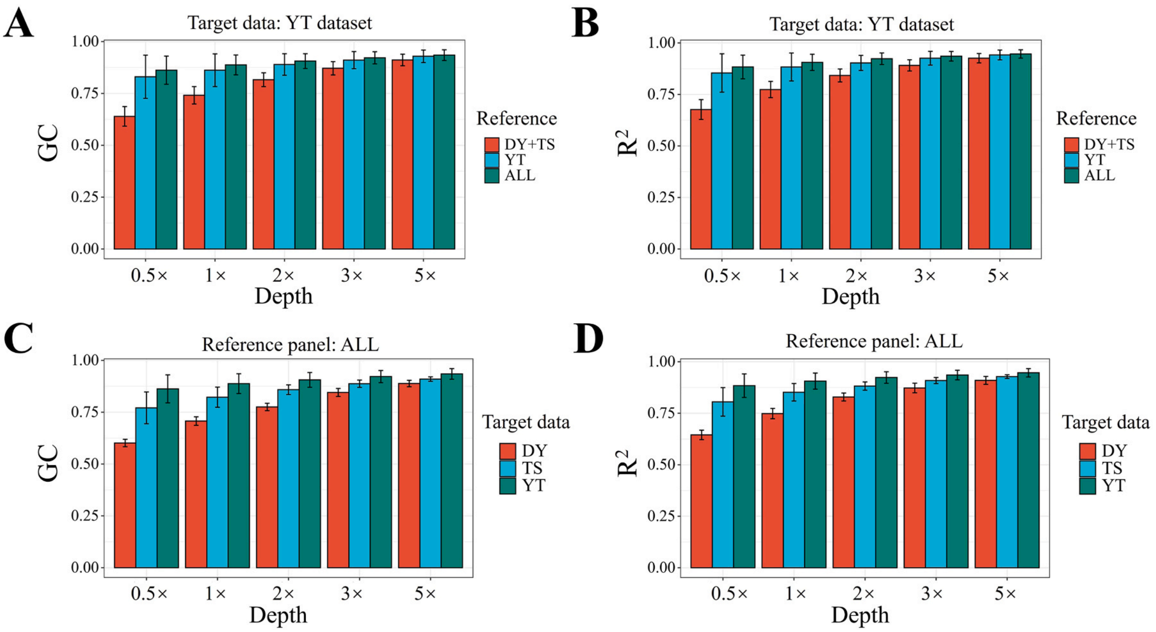

Fig. 3. The effects of reference and target data on GLIMPSE2 imputation accuracy. The effects of different reference panels on imputation accuracy for YT dataset based on (A) GC and (B) R^2. The effects of different target datasets on imputation accuracy using the optimal reference panel construction strategy based on **(C)** GC and **(D)** R^2.

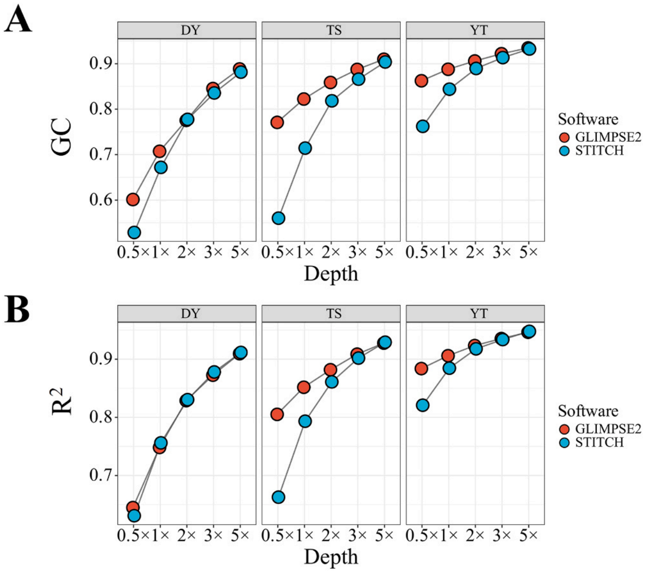

Fig. 4. The comparison of (A) GC and (B) R^2 between GLIMPSE2 and STITCH across different sequencing depths for DY, TS and YT datasets. GLIMPSE2 imputation was performed using the optimal reference panel construction strategy, and STITCH imputation included all 1007 samples.

imputation accuracy of both GLIMPSE2 and STITCH generally improved with increasing sequencing depth. At depths of 2 × , 3 × and 5 × , there was no significant difference between two methods in terms of GC and R^2 values for three datasets, except for higher GC values with GLIMPSE2 in the TS dataset at 2 × depth. However, at lower sequencing depths of 0.5 × and 1 × , GLIMPSE2 generally exhibited superior imputation accuracy compared to STITCH, particularly for TS and YT datasets (Fig. 4, Table S6). Notably, for DY dataset, which has a simple population structure, imputation accuracy including GC and R^2 remained consistently lower (< 0.8) at 0.5 × and 1 × depths for both GLIMPSE2 and STITCH (Fig. 4, Table S6).

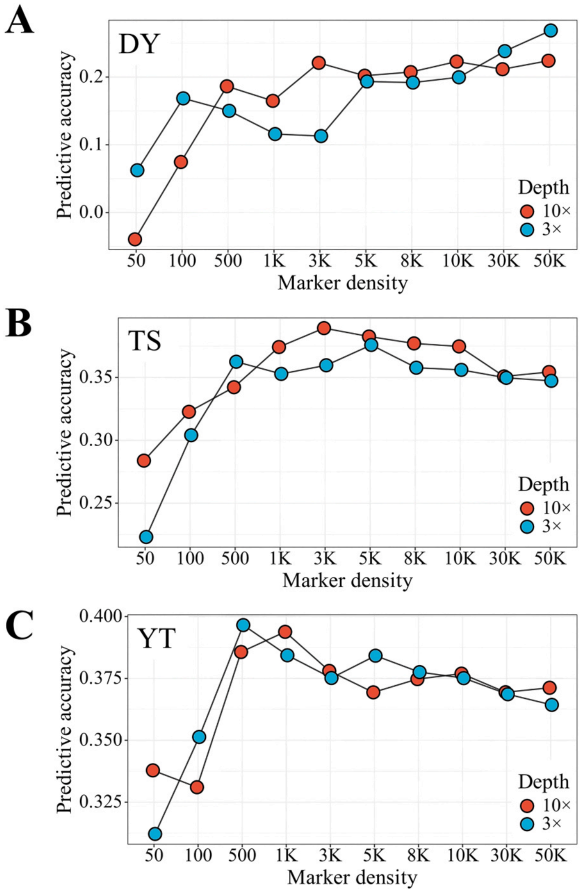

Fig. 5. The comparison of predictive accuracies for growth traits using lcWGS (3 ×) and hcWGS (10 ×) data for (A) DY, (B) TS and (C) YT datasets. The SVM model was used for genomic prediction.

3.7. Genomic prediction using lcWGS data after GLIMPSE2 imputation

Given that imputation accuracy at 3 × depth achieved relatively high levels (GC > 0.85 and $R^2 > 0.87$) across all three datasets (Fig. 4, Table S6), we selected the 3 × lcWGS data after GLIMPSE2 imputation, along with hcWGS data, to perform genomic prediction and further evaluate its application potential. For DY dataset (Fig. 5A), lcWGS and hcWGS data alternated in predictive accuracy across marker numbers ranging from 50 to 3000. Notably, the predictive accuracy of hcWGS data reached plateau (0.221) at 3000 SNPs, whereas at least 10,000 SNPs were required to achieve predictive accuracy above 0.2 (Table S7). For TS dataset (Fig. 5B), both lcWGS and hcWGS data exhibited a similar predictive trend. Predictive accuracy increased gradually with the number of markers, reaching the predictive plateau (0.389 and 0.376) at 3000 and 5000 SNPs for hcWGS and lcWGS data, respectively, and decreased slightly with further increases in marker number (Table S7). For YT dataset (Fig. 5C), despite fluctuations in predictive accuracy, both lcWGS and hcWGS data generally followed an upward trend followed by a decline. The predictive peak was 0.397 for lcWGS data at 500 SNPs and 0.394 for hcWGS data at 1000 SNPs (Table S7). Overall, as the number of SNPs increased, there was no significant discrepancy in predictive accuracy between lcWGS and hcWGS data. The required SNP number largely depended on imputation accuracy, with 5000, 100, and 100 SNPs needed for DY, TS and YT datasets, respectively (Fig. 5, Table S7).

4. Discussion

The rapid advancement of sequencing and biostatistics technologies has facilitated the widespread use of genomic resources to dissect genetic mechanisms and enhance genetic gains for economically important traits in aquaculture species (Houston et al., 2020; Yáñez et al., 2020). However, the high cost of genotyping remains a barrier, limiting the number of sequenced individuals and hindering the full utilization of genomic resources. The development of low-cost genotyping strategies and genotype imputation offers a promising solution to address cost challenges and increase the number of genotyped individuals (Zhang et al., 2023). Among these, lcWGS stands out for its extensive coverage of genetic variation across the genome and cost-effectiveness (Davies et al., 2021; Lou et al., 2021) and has been successfully applied in genotype imputation research in human (Gilly et al., 2016), cattle (Zhang et al., 2023) and large yellow croaker (Zhang et al., 2021). Therefore, the development of an appropriate and efficient imputation pipeline leveraging lcWGS data is essential for spotted sea bass that demand urgent genetic improvement (Zhang et al., 2023).

In this study, we evaluated three widely used imputation software - STITCH, GLIMPSE2, and BEAGLE - to determine the optimal imputation pipeline. These tools employ diverse imputation strategies, varying dependence on reference panels and their use of BAM or VCF files for imputation. To minimize computational redundancy and enhance imputation efficiency, the 100data dataset, characterized by a small sample size, high sequencing depth (15.93 ×), and simple genetic structure (Fig. 1 C and D), was selected as the most suitable dataset for the preliminary screening of imputation pipelines. Of which, GLIMPSE2 and BEAGLE, both reliant on same reference panel, demonstrated the highest and lowest imputation accuracy across varying sequencing depths, respectively (Fig. 2). The relatively poor performance of BEAGLE than GLIMPSE2 or STITCH has also been observed in previous imputation studies involving both aquaculture and livestock species (Teng et al., 2022; Wang et al., 2025; Yang et al., 2024; Yang et al., 2021), suggesting that BEAGLE imputation is unable to accurately estimate missing genotypes despite having a high quality reference panel. Furthermore, we quantified computational efficiency through genotype imputation benchmarking for chr1 at 3 × sequencing depth. Tests were conducted on Intel H3C R4900 G5 clusters with 60-thread parallelization, measuring actual CPU-hour consumption. The BEAGLE pipeline necessitates intermediate VCF generation via GATK variant calling, which is constrained by a default 4-thread implementation (McKenna et al., 2010). This architectural limitation resulted in high resource demands totaling 79.0 CPU-hours, -comprising 47.2 CPU-hours for GATK (11.8 h × 4 threads) and 31.8 CPU-hoursfor BEAGLE (0.53 h × 60 threads). In contrast, GLIMPSE2's direct BAM processing achieved superior efficiency at 10.8 CPU-hours (0.18 h × 60 threads), representing an 86.3 % reduction relative to BEAGLE. Therefore, this combination of computational efficiency and accuracy establishes GLIMPSE2 as the optimal imputation approach for spotted sea bass when reference panels are available.

The quality of the reference panel critically influences imputation accuracy (Charon et al., 2021; Li et al., 2023), with population genetic diversity and sample size serving as key determinants in panel construction (Fernandes Garcia et al., 2022). To systematically evaluate these factors for GLIMPSE2 imputation, we selected the YT dataset, characterized by diverse population structures (Fig. 1 C and D), ensuring population diversity alignment between reference and target data. Notably, for GLIMPSE2 imputation with "DY+TS" reference panel, the sample size of target and reference data was 493 and 514, respectively. While for imputation using YT panel, we employed a five-fold cross-validation approach with five imputations, with each imputation involving a target sample size of 99 and a reference size of 394, both significantly smaller than those of the "DY+TS" panel. Crucially, at lower sequencing depths (0.5 ×, 1 × and 2 ×), the YT panel demonstrated significantly higher accuracy than the larger "DY+TS" panel (Fig. 3A and B, Table S4), despite the latter having a larger reference panel. This superior imputation performance of YT panel is mainly due to its diverse population structure congruence with target data. These results indicate that optimizing reference panels for population genetic diversity rather than sheer sample size maximizes imputation accuracy, aligning with previous genotype imputation studies in cattle and tilapia (Fernandes Garcia et al., 2022; Zhang et al., 2023). This principle is reinforced by diminishing returns in accuracy gains: the ALL panel showed only marginal improvements over YT panel (3.9 % for GC and 3.3 % for R^2 at 0.5 ×) (Fig. 3A and 3B, Table S4), confirming that panel size is not the primary accuracy determinant when population genetic diversity is matched (Yang et al., 2024; Zhao et al., 2021). Nevertheless, maximizing the size of reference panels remains crucial to achieve the highest possible accuracy, especially when sequenced animals exhibit limited genetic diversity. It is also noteworthy that, despite employing an identical construction strategy of reference panel (ALL), the imputation accuracies for DY and TS datasets remain significantly lower than those of YT datasets at lower sequencing depths (Fig. 3C and D, Table S5). This superior imputation performance of YT datasets mainly stems from a relatively higher level of LD and genetic relatedness, and diverse population structure. Furthermore, although the DY dataset had a larger sample size than TS dataset, its imputation accuracy was notably lower than that due to the lower LD, weaker genetic relatedness, and simple population structure (Fig. 1B and D, Fig. S3). These results about the impact of reference and target data on imputation accuracy collectively emphasize that population structure, genetic relatedness and LD level between the haplotype reference data and the lcWGS data to be imputed are important factors affecting imputation performance. Therefore, when performing genotype imputation using lcWGS data, prioritizing these factors in both the reference and target data is essential, and increasing the sample size as much as possible will further enhance accuracy.

STITCH, a leading imputation software that operates without a reference panel, has effectively addressed the challenge of accurate genotype imputation in many non-model species that lack high-quality genotype reference panels (Davies et al., 2016). Its outstanding performance has been demonstrated in humans, mice, pigs, and cattle (Davies et al., 2016; Nicod et al., 2016; Teng et al., 2022; Yang et al., 2021). For aquatic species, lcWGS data imputed using STITCH achieved predictive accuracy comparable to WGS genotype data in both real and simulated datasets in large yellow croaker, highlighting its potential in genomic selection (Zhang et al., 2021). Similarly, STITCH combined with lcWGS has proven to be a high-throughput and cost-effective genotyping method in Pacific oyster (Yang et al., 2024). In our study, we first evaluated the impact of founders or ancestral haplotypes (K) on STITCH imputation accuracy using the 100data dataset. While the highest imputation accuracy was observed at K = 30 (Fig. S4), this required longer computation times. Therefore, K = 25 was selected as the optimal parameter for further imputation. Although the imputation performance of STITCH was slightly lower than that of GLIMPSE2 for 100data dataset, this result aligns with expectations, as STITCH relies solely on sequencing reads in BAM format to estimate optimal ancestral haplotypes. Clearly, the 100data dataset was insufficient to generate effective haplotype information for accurate imputation. A similar pattern was seen in Pacific oyster imputation, where accuracy stabilized after the sample size reached 300 at sequencing depths of 1 × and 2 × (Yang et al., 2024). To maximize STITCH's performance, 1007 samples were included for imputation process, and the results were compared with GLIMPSE2 using ALL panel across three datasets. Despite this, GLIMPSE2 consistently outperformed STITCH in three datasets (Fig. 4). Consequently, GLIMPSE2 using ALL panel was identified as the optimal imputation pipeline for spotted sea bass up to now. Notably, both STITCH and GLIMPSE2 exhibited lower imputation accuracy for DY and TS datasets compared to YT dataset (Fig. 4), reinforcing the idea that the genetic diversity of reference and target data is a crucial factor affecting imputation accuracy, regardless of whether a reference panel is used.

Genomic selection has proven to be an effective method for accelerating breeding progress and reducing the costs associated with breeding programs (Georges et al., 2019). The application potential of GS based on hcWGS data for growth traits has been demonstrated in our previous study (Zhang et al., 2024). To further reduce genotyping cost in genomic selection for spotted sea bass, the impact of lcWGS data on genomic prediction was investigated in this study. Genomic predictive performance of lcWGS data showed a high correlation to their imputation accuracy within three datasets. A significant difference in predictive accuracy between hcWGS and lcWGS data was observed only in DY dataset, with a relatively lower R^2 value (0.873). However, the comparable predictive accuracy between hcWGS and lcWGS data was observed once the number of markers exceeded 5000 (Fig. 5, Table S7). Furthermore, DY dataset has previously shown poor predictive performance in GP using SNP and InDel markers, likely due to its simple genetic structure and low genetic relatedness (Zhang et al., 2023; Zhang et al., 2024), indicating that it may not be suitable for GS. Our study primarily focused on assessing the potential of lcWGS data in GP, although investigating factors affecting predictive accuracy warrants further exploration. The predictive accuracy between imputed lcWGS data and WGS genotype data was remarkably consistent for TS and YT datasets when marker numbers exceeded 50, with R^2 values of 0.909 and 0.935, respectively (Fig. 5B and C, Table S6). Although slightly lower accuracy was observed for lcWGS compared to hcWGS at 50 markers for two datasets, both datasets could reach the predictive plateau with a similar number of markers (Fig. 5B and C, Table S7). This suggests that lcWGS, combined with genotype imputation, can effectively capture genetic variation and achieve comparable prediction performance to hcWGS data (Song et al., 2024; Zhang et al., 2021; Zhang et al., 2022). However, due to the relatively low genetic relatedness observed in the three datasets (Fig. S3), there remains considerable potential for improving imputation accuracy, particularly at lower sequencing depths. Future efforts will focus on expanding both the reference and target data with additional family representatives to further optimize the genotype imputation pipeline.

5. Conclusion

This study systematically evaluated different genotype imputation pipelines, differing in their reliance on reference panels and the use of BAM or VCF files. Due to lower accuracy and excessive computational demands, BEAGLE was not considered for further analyses. We also explored the effects of reference and target data on GLIMPSE2 imputation, finding that population genetic diversity outweighed sample size in constructing reference panels. And population structure, genetic relatedness and LD level between reference and target data are important factors affecting imputation accuracy. In addition, we developed the first publicly available reference panel, comprising 1107 spotted sea bass samples. The imputation accuracy of STITCH and GLIMPSE2 for adequate sample size were compared for three datasets, and GLIMPSE2 imputation using ALL panel emerged as the most effective imputation pipeline for spotted sea bass. Finally, we demonstrated that lcWGS data combined with GLIMPSE2 imputation provides genomic prediction results comparable to those obtained with hcWGS data. These insights contribute to advancing large-scale genotyping efforts for spotted sea bass and can serve as a reference for genotype imputation in other aquaculture species.

CRediT authorship contribution statement

Cong Liu: Visualization, Software. **Lingyu Wang:** Visualization, Software. **Yani Dong:** Visualization, Software. **Donglei Sun:** Resources, Methodology. **Chong Zhang:** Writing – original draft, Software, Methodology, Conceptualization. **Shaosen Yang:** Funding acquisition, Conceptualization. **Yun Li:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Yonghang Zhang:** Visualization, Software. **Pengyu Li:** Visualization, Software. **Xin Qi:** Resources, Conceptualization. **Haishen Wen:** Resources, Funding acquisition, Conceptualization. **Kaiqiang Zhang:** Resources, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the support of the High-Performance Biological Supercomputing Center at the Ocean University of China for this research. This research was funded by the Fundamental Research Funds for the Central Universities [grant number: 202461043], Research on Breeding Technology of Candidate Species for Guangdong Modern Marine Ranching [grant number: 2024-MRB-00–001], and China Agriculture Research System (CARS for Marine Fish Culture Industry) [grant number: CARS-47].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aqrep.2025.103088.

Data availability

Data will be made available on request.

References

Alexander, D., Novembre, J., Lange, K., 2009. Fast Model-based estimation of ancestry in unrelated individuals. Genome Res. 19 (9), 1655–1664. https://doi.org/10.1101/gr.094052.109.

Alicia, R.M., Elizabeth, G.A., Sinéad, B.C., Anne, S., Rocky, E.S., Tamrat, A., Zukiswa, Z., Neuro, G.A.P.P.C., 2021. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. Am. J. Hum. Genet. 108, 656–668. https://doi.org/10.1016/j.ajhg.2021.03.012.

Arthur, G., Karoline, K., Lorraine, S., Daniel, S., Rachel, M., Giorgio, E.M.M., Konstantinos, H., Aliki-Eleni, F., Graham, R., Jeremy, S., Petr, D., Britt, K., Martin, O. P., Xiangyu, G., Heather, E., William, J.A., Tao, J., Adam, B., Nicole, S., Emmanouil, T., Maria, K., George, D., Eleftheria, Z., 2019. Very low depth whole genome sequencing in complex trait association studies. Bioinformatics 35 (15), 2555–2561. https://doi.org/10.1093/bioinformatics/bty1032.

Browning, B.L., Browning, S.R., 2009. 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. https://doi.org/10.1016/j.ajhg.2009.01.005.

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A One-Penny imputed genome from Next-Generation reference panels. Am. J. Hum. Genet. 103, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

Charon, C., Allodji, R., Meyer, V., Deleuze, J.-F., 2021. Impact of pre- and post-variant filtration strategies on imputation. Sci. Rep. 11, 6214. https://doi.org/10.1038/s41598-021-85333-z.

Chen, B., Zhou, Z., Shi, Y., Gong, J., Li, C., Zhou, T., Li, Y., Zhang, D., Xu, P., 2023. Genome-wide evolutionary signatures of climate adaptation in spotted sea bass inhabiting different latitudinal regions. Evolut. Appl. 16, 1029–1043. https://doi.org/10.1111/eva.13551.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Davies, R.W., Flint, J., Myers, S., Mott, R., 2016. Rapid genotype imputation from sequence without reference panels. Nat. Genet. 48, 965–969. https://doi.org/10.1038/ng.3594.

Davies, R.W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C.M., Chan, Y.F., Myers, S., 2021. Rapid genotype imputation from sequence with reference panels. Nat. Genet. 53, 1104–1111. https://doi.org/10.1038/s41588-021-00877-0.

Fang, C., Zhong, H., Lin, Y., Chen, B., Han, M., Ren, H., Lu, H., Luber, J., Xia, M., Li, W., Stein, S., Xu, X., Zhang, W., Drmanac, R., Wang, J., Yang, H., Hammarström, L., Kostic, A., Kristiansen, K., Li, J., 2018. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. GigaScience 7 (3), 1–8. https://doi.org/10.1093/gigascience/gix133.

Fernandes Garcia, B., Yoshida, G., Carvalheiro, R., Yáñez, J., 2022. Accuracy of genotype imputation to whole genome sequencing level using different populations of Nile tilapia. Aquaculture 551, 737947. https://doi.org/10.1016/j.aquaculture.2022.737947.

Fernandes Júnior, G.A., Carvalheiro, R., de Oliveira, H.N., Sargolzaei, M., Costilla, R., Ventura, R.V., Fonseca, L.F.S., Neves, H.H.R., Hayes, B.J., de Albuquerque, L.G., 2021. Imputation accuracy to whole-genome sequence in nellore cattle. Genet. Sel. Evol. 53, 27. https://doi.org/10.1186/s12711-021-00622-5.

Georges, M., Charlier, C., Hayes, B., 2019. Harnessing genomic information for livestock improvement. Nat. Rev. Genet. 20, 135–156. https://doi.org/10.1038/s41576-018-0082-2.

Gilly, A., Ritchie, G.R., Southam, L., Farmaki, A.-E., Tsafantakis, E., Dedoussis, G., Zeggini, E., 2016. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. Hum. Mol. Genet. 25, 2360–2365. https://doi.org/10.1093/hmg/ddw088.

Gong, J.A.-O., Zhao, J., Ke, Q., Li, B., Zhou, Z.A.-O., Wang, J., Zhou, T., Zheng, W., Xu, P. A.-O., 2021. First genomic prediction and genome-wide association for complex growth-related traits in rock bream (*oplegnathus fasciatus*). Evolut. Appl. 15 (4), 523–536. https://doi.org/10.1111/eva.13218.

Hayes, B., Bowman, P., Daetwyler, H.D., Kijas, J., Werf, J., 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43, 72–80. https://doi.org/10.1111/j.1365-2052.2011.02208.x.

Hayward, J.J., White, M.E., Boyle, M., Shannon, L.M., Casal, M.L., Castelhano, M.G., Center, S.A., Meyers-Wallen, V.N., Simpson, K.W., Sutter, N.B., Todhunter, R.J., Boyko, A.R., 2019. Imputation of canine genotype array data using 365 whole-genome sequences improves power of genome-wide association studies. PLOS Genet. 15, e1008003. https://doi.org/10.1371/journal.pgen.1008003.

Hickey, J.M., Crossa, J., Babu, R., de los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654–663. https://doi.org/10.2135/cropsci2011.07.0358.

Hofmeister, R.J., Ribeiro, D.M., Rubinacci, S., Delaneau, O., 2023. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK biobank. Nat. Genet. 55, 1243–1249. https://doi.org/10.1038/s41588-023-01415-w.

Höglund, J., Rafati, N., Rask-Andersen, M., Enroth, S., Karlsson, T., Ek, W.E., Johansson, Å., 2019. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. Sci. Rep. 9, 16844. https://doi.org/10.1038/s41598-019-53111-7.

Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M., Davie, A., Robledo, D., 2020. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 21, 389–409. https://doi.org/10.1038/s41576-020-0227-y.

Huang, Y., Hickey, J.M., Cleveland, M.A., Maltecca, C., 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44, 25. https://doi.org/10.1186/1297-9686-44-25.

Iheshiulor, O.O.M., Woolliams, J.A., Yu, X., Wellmann, R., Meuwissen, T.H.E., 2016. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. Genet. Sel. Evol. 48, 15. https://doi.org/10.1186/s12711-016-0193-1.

Ji, J., Yan, G., Chen, D., Xiao, S., Gao, J., Zhang, Z., 2019. An association study using imputed whole-genome sequence data identifies novel significant loci for growth-related traits in a duroc × erhualian F2 population. J. Anim. Breed. Genet. 136, 217–228. https://doi.org/10.1111/jbg.12389.

Lachance, J., Tishkoff, S.A., 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays 35, 780–786. https://doi.org/10.1002/bies.201300014.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing, S., 2009. The sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, J.H., Liu, A., Buerkle, C.A., Palmer, W., Belbin, G.M., Ahangari, M., Gibson, M.J.S., Flagel, L., 2023. The effects of reference panel perturbations on the accuracy of genotype imputation. bioRxiv. https://doi.org/10.1101/2023.08.10.552684.

Lou, R.N., Jacobs, A., Wilder, A.P., Therkildsen, N.O., 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol. Ecol. 30, 5966–5993. https://doi.org/10.1111/mec.16077.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M., 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110.

Nicod, J., Davies, R.W., Cai, N., Hassett, C., Goodstadt, L., Cosgrove, C., Yee, B.K., Lionikaite, V., McIntyre, R.E., Remme, C.A., Lodder, E.M., Gregory, J.S., Hough, T., Joynson, R., Phelps, H., Nell, B., Rowe, C., Wood, J., Walling, A., Bopp, N., Bhomra, A., Hernandez-Pliego, P., Callebert, J., Aspden, R.M., Talbot, N.P., Robbins, P.A., Harrison, M., Fray, M., Launay, J.-M., Pinto, Y.M., Blizard, D.A., Bezzina, C.R., Adams, D.J., Franken, P., Weaver, T., Wells, S., Brown, S.D.M., Potter, P.K., Klenerman, P., Lionikas, A., Mott, R., Flint, J., 2016. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat. Genet. 48, 912–918. https://doi.org/10.1038/ng.3595.

Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., Sullivan, P.F., Bergen, S., Moran, J.L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S.M., Haas, D.W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P.I.W., Reich, D., Price, A.L., 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. 44, 631–635. https://doi.org/10.1038/ng.2283.

Pedersen, B.S., Quinlan, A.R., 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

Peng, W., Xu, J., Zhang, Y., Feng, J., Dong, C., Jiang, L., Feng, J., Chen, B., Gong, Y., Chen, L., Xu, P., 2016. An ultra-high density linkage map and QTL mapping for sex and growth-related traits of common carp (*cyprinus carpio*). Sci. Rep. 6, 26693. https://doi.org/10.1038/srep26693.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for Whole-Genome association and Population-Based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795.

Rubinacci, S., Hofmeister, R.J., Sousa da Mota, B., Delaneau, O., 2023. Imputation of low-coverage sequencing data from 150,119 UK biobank genomes. Nat. Genet. 55, 1088–1090. https://doi.org/10.1038/s41588-023-01438-3.

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genom. 15, 478. https://doi.org/10.1186/1471-2164-15-478.

Song, H., Dong, T., Wang, W., Jiang, B., Yan, X., Geng, C., Bai, S., Xu, S., Hu, H., 2024. Cost-effective genomic prediction of critical economic traits in sturgeons through low-coverage sequencing. Genomics 116, 110874. https://doi.org/10.1016/j.ygeno.2024.110874.

Teng, J., Zhao, C., Wang, D., Chen, Z., Tang, H., Li, J., Mei, C., Yang, Z., Ning, C., Zhang, Q., 2022. Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. J. Dairy Sci. 105, 3355–3366. https://doi.org/10.3168/jds.2021-21360.

Tsai, H.-Y., Matika, O., Edwards, S.M., Antolín-Sánchez, R., Hamilton, A., Guy, D.R., Tinch, A.E., Gharbi, K., Stear, M.J., Taggart, J.B., Bron, J.E., Hickey, J.M., Houston, R.D., 2017. Genotype imputation to improve the Cost-Efficiency of genomic selection in farmed atlantic salmon. G3 Genes|Genomes|Genet. 7, 1377–1383. https://doi.org/10.1534/g3.117.040717.

Tsairidou, S., Hamilton, A., Robledo, D., Bron, J.E., Houston, R.D., 2020. Optimizing Low-Cost genotyping and imputation strategies for genomic selection in atlantic salmon. G3 Genes|Genomes|Genet. 10, 581–590. https://doi.org/10.1534/g3.119.400800.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J., 2017. 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. 101, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Wang, D., Xie, K., Wang, Y., Hu, J., Li, W., Yang, A., Zhang, Q., Ning, C., Fan, X., 2022. Cost-effectively dissecting the genetic architecture of complex wool traits in rabbits by low-coverage sequencing. Genet. Sel. Evol. 54, 75. https://doi.org/10.1186/s12711-022-00766-y.

Wang, K., Yang, B., Li, Q., Liu, S.A.-O.X., 2022. Systematic evaluation of genomic prediction algorithms for genomic prediction and breeding of aquatic animals, 13 (12), 2247. https://doi.org/10.3390/genes13122247.

Wang, Y., Yao, R., Zhao, L., Zhang, Q., Li, M., Kong, X., Liu, P., Huang, S., Hu, C., Bao, Z., Hu, X., 2025. Optimizing strategy for Whole-Genome genotype imputation in scallops. Aquaculture 595, 741492. https://doi.org/10.1016/j.aquaculture.2024.741492.

Yáñez, J.M., Joshi, R., Yoshida, G.M., 2020. Genomics to accelerate genetic improvement in tilapia. Anim. Genet. 51, 658–674. https://doi.org/10.1111/age.12989.

Yang, B., Li, Y., Liu, S., 2024. High-throughput and cost-effective genotyping by low-coverage whole genome sequencing with genotype imputation in pacific oyster, *crassostrea gigas*. Aquaculture 591, 741134. https://doi.org/10.1016/j.aquaculture.2024.741134.

Yang, R., Guo, X., Zhu, D., Tan, C., Bian, C., Ren, J., Huang, Z., Zhao, Y., Cai, G., Liu, D., Wu, Z., Wang, Y., Li, N., Hu, X., 2021. Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy. GigaScience 10, giab048. https://doi.org/10.1093/gigascience/giab048.

Yoshida, G.M., Yáñez, J.M., 2021. Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in Nile tilapia. BMC Genom. 22, 57. https://doi.org/10.1186/s12864-020-07341-z.

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., Yang, T.-L., 2019. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 35, 1786–1788. https://doi.org/10.1093/bioinformatics/bty875.

Zhang, C., Wen, H., Zhang, Y., Zhang, K., Qi, X., Li, Y., 2023. First genome-wide association study and genomic prediction for growth traits in spotted sea bass (*lateolabrax maculatus*) using whole-genome resequencing. Aquaculture 566, 739194. https://doi.org/10.1016/j.aquaculture.2022.739194.

Zhang, C., Zhang, Y., Liu, C., Wang, L., Dong, Y., Sun, D., Wen, H., Zhang, K., Qi, X., Li, Y., 2024. Genome-wide association study and genomic prediction for growth traits in spotted sea bass (*lateolabrax maculatus*) using insertion and deletion markers. Anim. Res. One Health 1–17. https://doi.org/10.1002/aro2.87.

Zhang, W., Li, W., Liu, G., Gu, L., Ye, K., Zhang, Y., Li, W., Jiang, D., Wang, Z., Fang, M., 2021. Evaluation for the effect of low-coverage sequencing on genomic selection in large yellow croaker. Aquaculture 534, 736323. https://doi.org/10.1016/j.aquaculture.2020.736323.

Zhang, Z., Ma, P., Zhang, Z., Wang, Z., Wang, Q., Pan, Y., 2022. The construction of a haplotype reference panel using extremely low coverage whole genome sequences and its application in genome-wide association studies and genomic prediction in duroc pigs. Genomics 114, 340–350. https://doi.org/10.1016/j.ygeno.2021.12.016.

Zhang, Z., Wang, A., Hu, H., Wang, L., Gong, M., Yang, Q., Liu, A., Li, R., Zhang, H., Zhang, Q., Shah, A.M., Wang, X., Wang, Y., Liu, Q., Gao, L., Zhang, Z., Wang, C., Ma, Y., Cai, Y., Jiang, Y., 2023. The efficient phasing and imputation pipeline of low-coverage whole genome sequencing data using a high-quality and publicly available reference panel in cattle. Anim. Res. One Health 1, 4–16. https://doi.org/10.1002/aro2.8.

Zhao, C., Teng, J., Zhang, X., Wang, D., Zhang, X., Li, S., Jiang, X., Li, H., Ning, C., Zhang, Q., 2021. Towards a Cost-Effective implementation of genomic prediction based on low coverage whole genome sequencing in dezhou donkey. Front. Genet. 12. https://doi.org/10.3389/fgene.2021.728764.

Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824. https://doi.org/10.1038/ng.2310.

Zhou, Z., Han, K., Wu, Y., Bai, H., Ke, Q., Pu, F., Wang, Y., Xu, P., 2019. Genome-Wide association study of growth and Body-Shape-Related traits in large yellow croaker (*larimichthys crocea*) using ddRAD sequencing. Mar. Biotechnol. 21, 655–670. https://doi.org/10.1007/s10126-019-09910-0.