International Journal of
Molecular Sciences

MDPI

Article

Evaluation of Low-Coverage Sequencing Strategies for Whole-Genome Imputation in Pacific Abalone *Haliotis discus hannai*

Chengxia Fei [1,2,†], Shoudu Zhang [2,3,†], Xiangrui Chen [1], Junyu Liu [4], Wenzhu Peng [4], Guofan Zhang [2,4], Weiwei You [4] and Fucun Wu [2,5,6,*]

1 School of Marine Sciences, Ningbo University, Ningbo 315211, China; ouhaifeichengxia@163.com (C.F.); xiangruichen@126.com (X.C.)
2 CAS and Shandong Province Key Laboratory of Experimental Marine Biology, Center for Ocean Mega-Science, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266000, China; shouduzhang@163.com (S.Z.); gfzhang@qdio.ac.cn (G.Z.)
3 Marine Science Research Institute of Shandong Province (National Oceangraphic Center, Qingdao), Qingdao 266104, China
4 State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen 361102, China; liujy@ysfri.ac.cn (J.L.); pengwenzhu2019@163.com (W.P.); wwyou@xmu.edu.cn (W.Y.)
5 Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Qingdao 266000, China
6 National and Local Joint Engineering Laboratory of Ecological Mariculture, Qingdao 266000, China
* Correspondence: wufucun@qdio.ac.cn; Tel.: +86-(532)-8289-8713
† These authors contribute equally to this work.

Abstract: Low-coverage whole-genome sequencing (lcWGS) followed by imputation is emerging as a cost-effective method for generating a substantial number of single nucleotide polymorphism (SNP) in aquatic species with highly heterozygous and complex genomes. This study represents the first systematic investigation into the application of low-coverage whole-genome sequencing (lcWGS) combined with imputation for genotyping in Pacific abalone (*Haliotis discus hannai*) without a reference panel. We utilized 1059 Pacific abalone individuals sequenced at an average depth of 7.86×, as well as 16 individuals sequenced at 20×, as sample materials. To assess the genotype imputation accuracy for lcWGS without a reference panel, we simulated data with varying sequencing depths (0.5–4×) and examined the effects of sample size, chromosome length, and minor allele frequency (MAF) using BaseVar and STITCH strategies. Results showed that STITCH achieved high accuracy when the sample size exceeded 400, with a genotype correlation (R^2) of 0.98 ± 0.002 and genotype concordance (GC) of 0.99 ± 0.001. Imputation accuracy plateaued when the sample size exceeded 400 and sequencing depth surpassed 1×. Chromosome length had minimal effects, with all three chromosomes achieving an accuracy of approximately 0.98. However, the accuracy for rare MAF (<0.05) was lower, falling below 0.99. A second imputation with Beagle significantly increased SNP detection by 3.9–8.3 folds for a sequencing depth of 0.5–4×, apparently without sacrificing accuracy. To our knowledge, this is the first study of lcWGS analysis conducted in abalone. The findings demonstrate that lcWGS with imputation can achieve high accuracy with moderate sample sizes ($n \geq 400$) in Pacific abalone, offering a cost-effective approach for genotyping in aquaculture species.

Keywords: *Haliotis discus hannai*; low-coverage whole-genome sequencing (lcWGS); genotype imputation; accuracy

Academic Editor: Hongyan Xu

Received: 14 March 2025
Revised: 22 April 2025
Accepted: 7 May 2025
Published: 11 May 2025

Citation: Fei, C.; Zhang, S.; Chen, X.; Liu, J.; Peng, W.; Zhang, G.; You, W.; Wu, F. Evaluation of Low-Coverage Sequencing Strategies for Whole-Genome Imputation in Pacific Abalone *Haliotis discus hannai*. *Int. J. Mol. Sci.* **2025**, *26*, 4598. https://doi.org/10.3390/ijms26104598

1. Introduction

The Pacific abalone (*Haliotis discus hannai*) is a commercially and ecologically significant species distributed along the coasts of China, Japan, and Korea. It plays a vital role in marine ecosystems by providing abundant food resources for humans and contributing significantly to the ecological balance of ocean environments [1]. In addition to its ecological importance, *H. discus hannai* is a highly valued aquaculture species in East Asia. In China, the annual production of farmed abalone and its hybrids exceeds 200,000 metric tons, making it a major contributor to the aquaculture industry [2]. The cultivation of this species has become a critical livelihood for coastal residents in China, particularly for small- and medium-scale farmers. Likewise, abalone farming in Korea has seen rapid development in recent years, positioning the country as the second-largest producer globally. Furthermore, *H. discus hannai* has been introduced into several European and American countries for aquaculture purposes, further underscoring its global economic significance [3].

In recent years, the rapid growth of the aquaculture industries, combined with advances in high-throughput sequencing technologies, has propelled Pacific abalone research into the genomic era [4]. The evolution of genotyping technologies from low-throughput gel electrophoresis methods to widely used high-throughput approaches has revolutionized the field. High-throughput sequencing technologies are particularly effective in addressing challenges such as high heterozygosity, repetitive sequences, and complex genomic structures [5]. These technologies enable the rapid and comprehensive acquisition of genomic variation data at both the individual and population levels [6], facilitating a deeper understanding of genetic diversity, gene function, and the mechanisms underlying complex traits in aquaculture species [7]. While high-coverage whole-genome sequencing (WGS) remains the gold standard for genotyping, its high cost limits large-scale applications in aquaculture. To overcome this challenge, several cost-effective genotyping technologies have been explored, including SNP chip genotyping [8,9] and reduced representation genome sequencing [10,11]. While SNP chips offer advantages such as lower cost, fast processing, and high data quality, they are inherently limited by their fixed content, which prevents comprehensive coverage of all genetic variants and trait-associated genomic regions [12]. This limitation can compromise the accuracy and completeness of downstream analyses [13].

Low-coverage whole-genome sequencing (lcWGS) has revolutionized genomic studies by offering a cost-effective genotyping method, heralding a new era in genomic studies. Typically performed at $\leq 1\times$ sequencing depth [14], this technological advancement enables large-scale genetic variant detection through computational imputation, where sparse sequencing data are statistically reconstructed using high-coverage reference genomes [15]. Notably, lcWGS outperforms SNP microarrays in capturing comprehensive variation profiles, particularly in non-coding and structural genomic regions [16]. Although low sequencing depths inherently increase genotype uncertainty, sequencing a large number of individuals can compensate for this by capturing comprehensive population-level genetic diversity [17]. Given the substantial number of missing data points inherent in low-coverage sequencing, genotype imputation plays a pivotal role in integrating population-level information for subsequent analyses [18,19]. Several genotype imputation methods have been proposed in the literature, including Beagle software [20], which uses a Monte Carlo Markov chain (MCMC) algorithm and is widely applied for imputation [21]. In contrast, the STITCH algorithm [8] is notable for its ability to infer population-level ancestral haplotypes from shared haplotype data across a large number of samples. This feature makes it particularly suitable for non-model or non-human species, such as aquatic organisms, where reliable reference panels are often unavailable. STITCH has demonstrated robust performance even at ultra-low sequencing depths [22]. For aquatic species, which

often lack large reference populations for functional gene mapping and genomic breeding, STITCH represents a promising approach. However, to date, there have been no published reports on its application in abalone species.

Low-coverage whole-genome sequencing (lcWGS) has proven to be a highly accurate and cost-effective approach for whole-genome SNP genotyping, genomic prediction, and genome-wide association analysis [22–24]. Significant progress has been made in its application to livestock and poultry [25,26], and lcWGS has also shown promising potential in aquaculture species such as yellow croaker [27], corals [28], Pacific oysters [29], scallops [30], and salmonid [31]. However, critical barriers persist in adapting lcWGS to marine shellfish like *Haliotis discus hannai*, a species characterized by extreme heterozygosity, fragmented genomes, and absent reference panels. In a study on the yellow croaker, a sequencing depth of just 0.5× across more than 500 individuals achieved an imputation accuracy comparable to that of 8× sequencing [27]. In addition, in another study where low-coverage whole-genome sequencing (lcWGS) was used for genotype imputation in rainbow trout, the concordance obtained was 99.1% [32]. Nonetheless, the application of lcWGS in aquaculture remains in its infancy [33], largely due to the high genetic diversity and recombination rates common in aquatic species. For example, a study on 360 Pacific oyster individuals using lcWGS at an average sequencing depth of 2.82× achieved a genotype accuracy of only 0.860 ± 0.055 [29]. Moreover, aside from the Pacific oyster study, the lack of large-scale reference panels in molluscan species such as Pacific abalone presents additional challenges for genetic inference using lcWGS. This limitation is particularly critical, given the species' pivotal role in global aquaculture, where rapid genetic improvement is urgently needed to address the environmental stressors and disease outbreaks threatening production sustainability. Currently, optimal genotyping strategies without reference panels have not been established for this species. Therefore, there is an urgent need to develop a systematic and efficient lcWGS-based workflow tailored to the genomic characteristics of aquatic species. Our study bridges this gap by establishing the first methodological framework tailored for high-heterozygosity aquatic species, offering a scalable solution to empower genomic-driven breeding programs in resource-limited aquaculture systems.

In this study, we explored whether low-coverage whole-genome sequencing (lcWGS) combined with an imputation strategy could achieve high genotyping accuracy in *H. discus hannai* without relying on reference panels. A total of 1075 Pacific abalone individuals were sequenced at an average depth exceeding 7× and used as the source material. Low-coverage WGS datasets with depths ranging from 0.5× to 4× were generated by down-sampling reads from these high-coverage data. We evaluated the performance of the low-coverage whole-genome sequencing (lcWGS) strategy in Pacific abalone and investigated the factors affecting genotype imputation accuracy across various sequencing depths, sample sizes, chromosome lengths, and minimum allele frequencies, in the absence of a reference panel. The study aimed to explore the optimal strategies for cost-effective genotype imputation in highly heterozygous marine species. Our findings demonstrate that the BaseVar+STITCH method, followed by secondary imputation with Beagle, performs well and is a robust approach for genotype imputation in low-coverage sequencing data. This research provides valuable recommendations for future whole-genome SNP genotyping in Pacific abalone, with implications for genomic selection, functional gene studies, and other related areas of research in this species.

2. Results

2.1. SNP Genotyping

After paired-end sequencing and quality control (minor allele frequency, MAF > 0.05), a total of 1479,738 clean reads were generated from 1059 individuals with an average

sequencing depth of 7.86×, and 16 individuals with a depth of 20×. The genome quality assessment revealed the Core eukaryotic genes mapping approach (CEGMA) completeness of 85% and benchmarking universal single-copy orthologs (BUSCO) completeness of 93%, with a high small-fragment read pairing rate of 96.95%, and the majority of reads were properly paired. The population genetic structure analysis indicated an average heterozygosity between 0.1 and 0.25 for the 1059 individuals (Figure 1). By calculating Tajima Test (Tajima's D, positive bias), fixation index (Fst, 0–0.12), and nucleotide diversity (π, 0.001–0.003) for the 1059 individuals (Figure S1), the results showed a rich genetic diversity among these individuals but overall genetic homogeneity. The distribution of 1059 individuals' SNPs on eighteen chromosomes is shown in Figure 2A; the SNPs across the chromosomes are generally distributed in a relatively uniform manner with an average density of 1260.44 SNPs/Mb. The results show that the distribution of Group1 with 1059 individuals was wide and its internal genetic diversity was high, while the distribution of Group2 with 16 individuals was relatively concentrated, indicating that its genetic homogeneity was strong. However, there were similar genetic backgrounds between the two, which further indicated that there was a genetic exchange between the populations (Figure 2B), and the phylogenetic tree further described the kinship between the two as not extremely close but still moderate (Figure 2C).

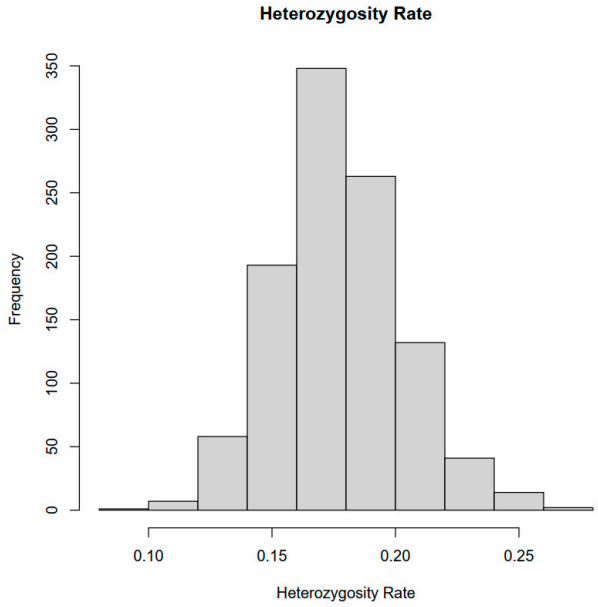

Figure 1. Genome-wide heterozygosity distribution of 1075 Pacific abalone (*Haliotis discus hannai*) samples based on whole-genome sequencing data after quality filtering (average depth 7.86×).

2.2. Effects of Sequencing Depth and Sample Size on Imputation Accuracy

In our study, the BaseVar+STITCH strategy was employed to evaluate the impact of different sequencing depths and sample size on imputation accuracy. The results indicate that, for varying sample sizes, an increase in sequencing depth from 0.5× to 1× resulted in a 0.1–0.2 improvement in accuracy as measured by the squared Pearson correlation coefficient of genotype dosage (R^2). With the augmentation of both sample size and sequencing depth, accuracy rose from 0.95 (sample size = 200 and sequencing depth = 0.5×) to 0.99 (sample size = 1075 and sequencing depth = 4×). When the sample size exceeded 400, both genotype imputation accuracy and genotype concordance (GC) content reached a plateau phase (Figure 3). Notably, at a sequencing depth of 0.5× and a sample size of 200, the STITCH imputation accuracy exceeded 0.95. As the sample size surpassed 400 and sequencing depth exceeded 1×, accuracy reached a plateau phase, consistently exceeding 0.98. At a

depth of 0.5×, as the sample size increased from 200 to 400, accuracy rose from 0.95 to 0.98. At a depth of 1×, as the sample size increased from 200 to 400, accuracy increased from 0.97 to 0.98. Concurrently, with the increase in sample size and sequencing depth, the GC content also continued to rise, all surpassing 0.97.

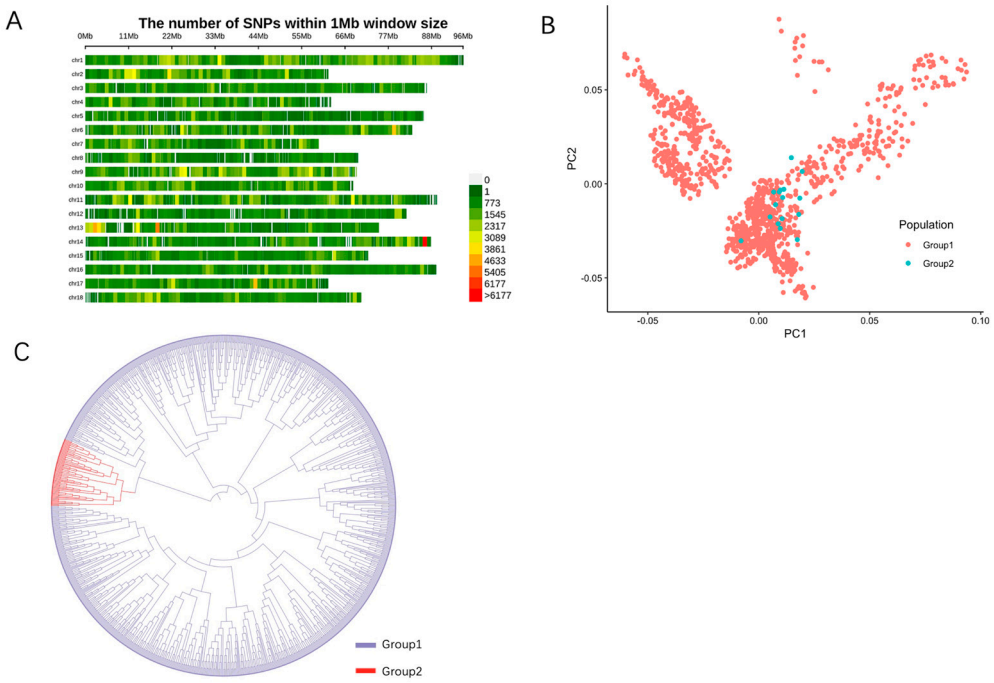

Figure 2. SNP distribution in 1 Mb windows across the genome and genetic structure analysis of the two groups in Pacific abalone. (**A**) Distribution of 1059 Pacific abalone samples' SNPs in 1 Mb windows across the genome. (**B**) Principal component analysis (PCA) of the first three principal components (Group1, 1059 samples with average coverage depth of 7.86×; Group2, 16 samples with average coverage depth of 20×). (**C**) Phylogenetic tree of samples based on genetic distance (Group1 and Group2 are the same as shown in (**B**)).

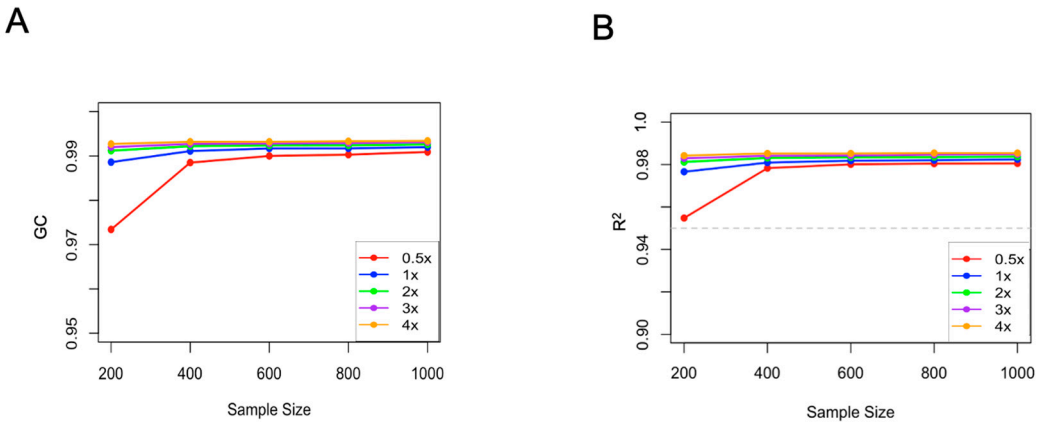

Figure 3. BaseVar+STITCH accuracy of imputation genotype and genotype concordance at different sequencing depths with different population sizes for chromosome 1 in *H. discus hannai*. (**A**) The imputation accuracy of different sequencing depths with different population sizes was measured by genotype concordance (GC). (**B**) The imputation accuracy of different sequencing depths with different population sizes was measured by the squared Pearson correlation coefficient of genotype dosage (R^2). The dashed line represents the threshold of 0.95.

2.3. Effects of Chromosome Length on Imputation Accuracy

We selected three chromosomes, Chr1, Chr13, and Chr7, of varying lengths, for imputation using 1075 individuals with sequencing depths of 1×, prompting an investigation into the effects of chromosome length on accuracy. The results showed that the imputation accuracy for all three chromosomes exceeded 0.98, with consistency in GC values surpassing 0.97. The accuracy of chromosome 1 was the lowest accuracy, with an R^2 of 0.982 and GC of 0.992, and that of chromosome 13 was the highest, with an R^2 of 0.988 and GC of 0.965 (Table 1). The squared Pearson correlation coefficient of genotype dosage (R^2) and GC did not differ much among the three chromosomes, with Chr 7 having the highest R^2 and Chr 1 having the highest GC (Figure 4).

Table 1. The squared Pearson correlation coefficient of genotype dosage (R^2), genotype concordance (GC), chromosome length, and SNP density revealed by lcWGS approach among different chromosomes in *H. discus hanna*.

Chr	R^2	GC	Chromosome Length	SNP Density (bp/SNP)
1	0.982	0.992	96,096,873	96.983
7	0.985	0.966	59,286,736	131.189
13	0.988	0.965	74,588,556	102.116

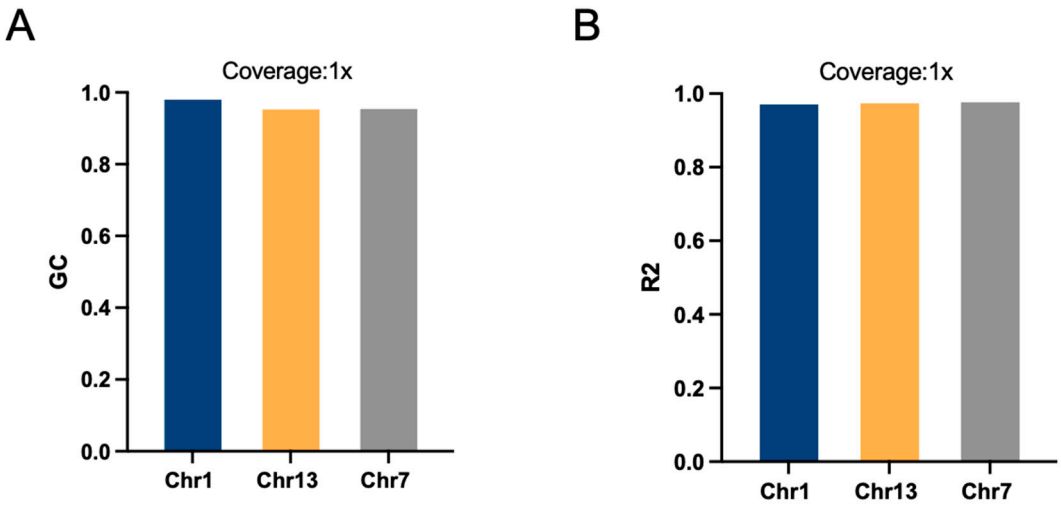

Figure 4. BaseVar+STITCH results of imputation genotype at different chromosome length in *H. discus hannai*. (**A**) The genotype concordance of imputation genotype (GC); (**B**) The squared Pearson correlation coefficient of genotype dosage (R^2); sample size = 1075, sequencing depth = 1×.

2.4. Effects of Minor Allele Frequency on Imputation Accuracy

We used a dataset with sample size of 1075 and a sequencing depth of 1× to evaluate the effect of minor allele frequency (MAF) on accuracy of different methods. The SNPs were divided into eight bins according to their MAF as follows: [0–0.005], [0.005–0.01], [0.01–0.05], [0.05–0.1], [0.1–0.2], [0.2–0.3], [0.3–0.4], and [0.4–0.5]. Figure 5 shows that the effect of MAF was obvious for rare variants with MAF < 0.05; both the genotypic accuracy and the genotypic concordance were greatly affected by MAF; and the accuracy increased rapidly with the increase in MAF. However, MAF > 0.05 had minimal impact on the accuracy of genotype imputation, while genotypic concordance decreased slightly with the increase in MAF.

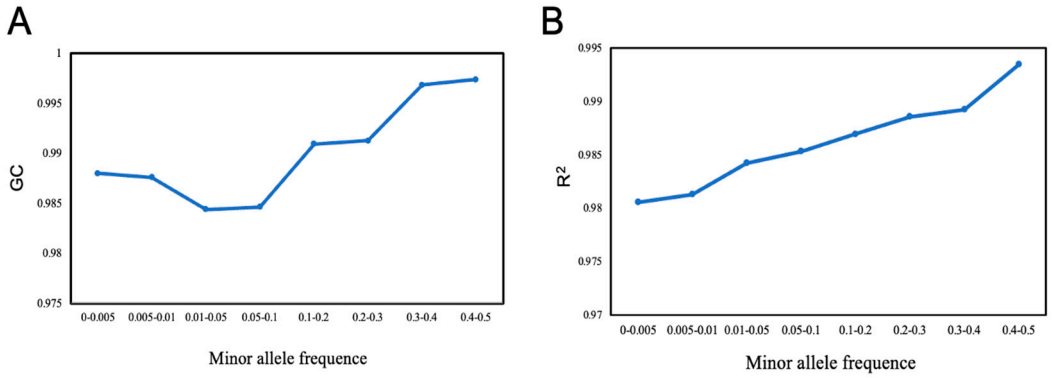

Figure 5. BaseVar+STITCH results of imputation genotype at minor allele frequency (MAF) for chromosome 1 in *H. discus hannai*. (**A**) The genotype concordance of imputation genotype (GC); (**B**) The squared Pearson correlation coefficient of genotype dosage (R^2); sample size = 1075, sequencing depth = 1×; The SNPs were divided into 8 bins according to their MAF as follows: [0–0.005], [0.005–0.01], [0.01–0.05], [0.05–0.1], [0.1–0.2], [0.2–0.3], [0.3–0.4], and [0.4–0.5].

2.5. Evaluation of STITCH and Beagle Strategy

Due to the persistent high rate of missing data post-STITCH imputation, leading to a reduced count of shared SNPs across individuals, we tried to use the BEAGLE software for a second imputation to the results of STITCH. We assessed the performance of STITCH and BEAGLE with a sample size = 1075 individuals with different sequencing depths. We kept those imputing SNPs in the dataset that were missing rate in ≤20% of the individuals after STITCH imputation and subsequently filled in these missing genotypes using Beagle. With the escalation of sequencing depth and sample size, imputation accuracy exhibited a corresponding increase (Figure 6). As expected, the BEAGLE imputation yielded a substantial rise in the count of SNPs. Table 2 shows the number of SNPs common to all individuals on chromosome 1 after Beagle estimation and their respective estimation accuracies. For missing rate ≤ 20%, compared with the STITCH results, the numbers of SNPs increased from 397.4% (for sequencing depth of 4×) to 646.5% (for sequencing depth of 0.5×), while the imputation accuracies were only slightly reduced. These results proved that the integration of STITCH and Beagle was an effective strategy to imputation and still maintained its high accuracy.

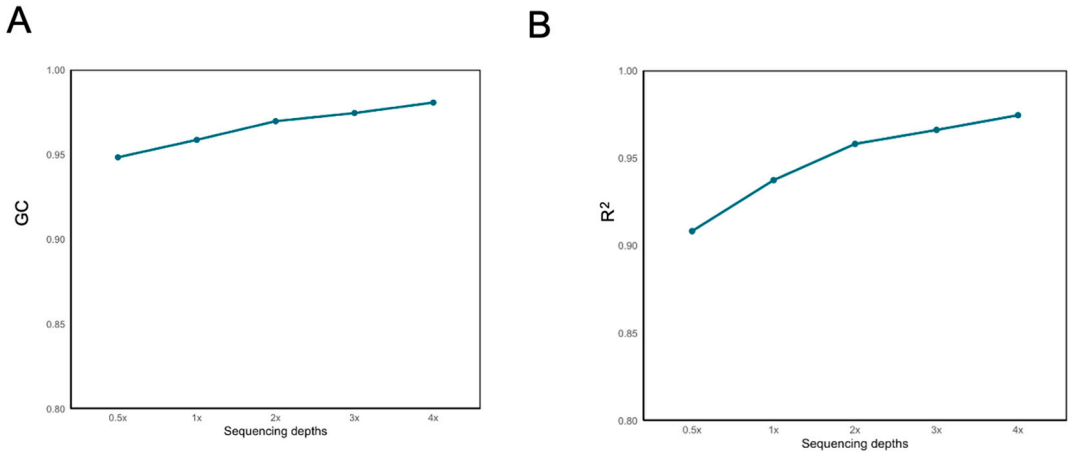

Figure 6. STITCH+BEAGLE strategy results of imputation genotype at different sequencing depths for chromosome 1 in *H. discus hannai*. (**A**) The genotype concordance of imputation genotype (GC); (**B**) The squared Pearson correlation coefficient of genotype dosage (R^2); sample size = 1075, sequencing depth = 1×.

Table 2. Number of SNPs and Genotype imputation accuracy with the squared Pearson correlation coefficient of genotype dosage(R^2) under different sequencing depths for STITCH and STITCH + Beagle (sample size = 1075, chromosome 1, Missing rate $\leq$ 0.2).

Sequencing Depth	STITCH		STITCH+BEAGLE	
	No. of SNPs	R^2	No. of SNPs	R^2
0.5×	114,056	0.9805	851,527	0.9082
1×	147,449	0.9824	1,369,351	0.9373
2×	210,381	0.9837	1,369,982	0.9580
3×	247,669	0.9848	1,369,982	0.9660
4×	275,461	0.9854	1,370,182	0.9744

3. Discussion

In aquaculture species, the prevailing genotyping strategy for genomic studies has been based on high-coverage sequencing data, typically analyzed using the Genome Analysis Toolkit (GATK) approach. Recently, the introduction of low-coverage whole-genome sequencing (lcWGS) followed by genotype imputation has emerged as a promising, cost-effective alternative for obtaining genome-wide genotypic data in aquaculture [34,35]. The aim of this study was to evaluate the impact of the lcWGS genotyping approach on both the accuracy of SNP imputation and the number of SNPs identified using different strategies in the Pacific abalone *Haliotis discus hannai*. To our knowledge, this is the first study to assess such an approach in Pacific abalone or any other abalone species. Our findings suggest that lcWGS genotyping, coupled with imputation, provides a highly effective method for detecting genome-wide SNP variants in Pacific abalone, offering a robust alternative to traditional high-coverage sequencing for genetic studies in aquaculture. There are various methods for SNP detection, including liquid phase SNP chips [36,37], whole-genome sequencing (WGS) [38], and reduced representation sequencing (RADseq) [39]. However, WGS is prohibitively expensive for large-scale population studies, while reduced representation sequencing and SNP chips have limitations that prevent full coverage of all loci across the genome, making it extremely difficult to identify true quantitative trait loci [40]. This limitation substantially impedes the elucidation of the genetic architecture underlying critical traits and diminishes the reliability of marker-assisted selection in molecular breeding programs.

Missing genotype imputation is a critical component of low-coverage whole-genome sequencing (lcWGS) and plays a pivotal role in its application to aquaculture species [34,41,42]. The effectiveness of the lcWGS genotyping approach is influenced by several factors, including the imputation method, sequencing depth, the number of sequenced individuals (sample size), and the availability and size of a reference panel [43]. In this study, we assessed the imputation performance of lcWGS data in Pacific abalone with respect to these key factors. Specifically, 1075 Pacific abalone individuals, each with sequencing depths exceeding 7×, were used as down-sampling materials for lower coverage depths, while a validation set of 16 individuals with a sequencing depth of 20× was employed to evaluate imputation accuracy. Our findings indicate that high imputation accuracy (>98%) can be achieved in the Pacific abalone population using the BaseVar+STITCH strategy, even in the absence of a reference panel. This represents a significant advancement over the previous aquaculture studies; in *Crassostrea gigas*, lcWGS at a depth of 2.8× achieved only 0.86 accuracy with similar sample sizes [29]. This is particularly relevant, as aquaculture species typically lack large haplotype reference panels and pre-existing variant catalogs [27]. These results are aligned with the previous reports showing that STITCH does not require reference panels for imputation, making it a more suitable approach for aquaculture species [27–29]. Moreover, we demonstrate that sequencing depth and

sample size significantly affect imputation accuracy. Specifically, a positive correlation was observed between the number of individuals, sequencing depth, and the accuracy of genotype imputation. Notably, an imputation accuracy with and R^2 exceeding 0.95 was achieved with a sequencing depth of 0.5× and a sample size of 200 individuals. At a sequencing depth of 2×, all sample sizes yielded imputation accuracies surpassing 0.98. However, accuracy tended to plateau beyond a certain threshold. The plateau effect observed at n = 400 mirrors the thresholds reported in poultry genomics [25]. These findings are consistent with prior studies in other species, such as Mexican Holstein cattle [44,45], underscoring the critical influence of sequencing depth and sample size on imputation accuracy. Based on our results, we recommend that an imputation accuracy exceeding 98% can be achieved with a sample size of approximately 400 individuals in the Pacific abalone population. Furthermore, sequencing depths between 0.5× and 1× are suggested as optimal for large populations (0.5× for 600 samples and 1× for 400 samples) in the lcWGS approach for Pacific abalone. Above all, we recommend sequencing depths of 1× as optimal for large populations. When the sample size exceeded 400 samples, lcWGS performed very well. This study also highlights the comparable or superior imputation accuracy of the lcWGS approach in aquaculture species, as summarized in Table 3. The impacts of low-coverage depth and sample size on imputation accuracy can be explained by the fact that STITCH uses reads from all BAM files to reconstruct founder haplotypes and perform imputation. As the total number of sequenced individuals and the depth of each sample increase, more reads are available for haplotype reconstruction, leading to a more accurate imputation of the missing whole-genome genotypes [27,46].

Table 3. Previous publications on aquaculture animals, with species, sequencing depth, sample size, imputation accuracies, and number of SNPs.

Species	Sequencing Depth	Sample Size	Imputation Accuracy	SNPs Number	Publication
Crassostrea gigas	2.8×	≥300	0.860	11,000,000	[29]
Larimichthys crocea	0.5×	536	0.795	5,949,426	[27]
Acipenser gueldenstaedtii	2×	≥300	0.882	>5,514,392	[24]
Acropora millepora	1.5×	193	0.94	Unknow	[28]
Chlamys farreri	0.5×	174	0.91	3,968,417	[30]
Haliotis discus hannai	1×	≥400	0.98	147,449	the present study

The STITCH algorithm, introduced by Davies et al. [8], stands out for its ability to infer genotypes accurately even without a reference panel. It is reported that STITCH had a drawback in that a significant proportion of SNPs remained unimputed even after applying the STITCH imputation method, resulting in the low count of SNPs shared across all individuals as compared to the other methods [25]. In the present study, we conducted a comparative analysis between the STITCH and the STITCH+Beagle imputation strategies. Following a subsequent imputation with Beagle, there was a substantial increase in the number of SNPs, whereas the imputation accuracies were only slightly reduced. This is in accordance with the reports on livestock of Holstein cattle and donkey [25,26]. The present study indicated that STITCH and Beagle were an effective strategy to make up for STITCH's limitation of reduced SNP yield, while maintaining its advantage of high accuracy [25]. The observed increase in SNP numbers can be attributed to the complementary nature of these imputation methods. STITCH, a reference-free tool, infers missing genotypes from low-coverage sequencing data using population-wide linkage disequilibrium (LD) patterns, but it may struggle with rare variants due to the lack of a reference panel [8]. Beagle, a haplotype-based imputation tool, improves genotype inference by leveraging both LD

information and, when available, external reference panels. Its robust algorithm enables accurate imputation across diverse sample types, including low-coverage sequencing data, admixed populations, and pedigrees, demonstrating universal applicability regardless of population structure or sequencing depth [47]. By refining STITCH imputed genotypes with Beagle, additional SNPs can be identified, particularly low-frequency alleles that STITCH alone may miss. Beagle's superior haplotype reconstruction further enhances SNP detection, explaining the observed SNP increase [25,48]. Using imputation-based sequencing data, the performance of genomic wide association study or evolutionary studies—such as genomic introgression analysis—appears to be strongly influenced by SNP density in genome-wide sequencing [49,50]. For Pacific abalone with no reference panel available, STITCH followed by Beagle would be an optimal strategy to increase the number of SNPs discovered without reducing accuracy.

In this study, we also explored the impact of different chromosome lengths on imputation accuracy. The findings show that the length of the chromosome did not significantly impact the imputation genotype accuracy. This conclusion serves as a guide when discussing the results of imputing low-depth sequencing data on different chromosome lengths using the BaseVar+STITCH strategy. The result was consistent with the previous investigations [51]. In addition, since rare variants are observed only a few times within a population, their imputation is more challenging than that of common variants. This increases the difficulty of establishing haplotype templates while also reducing the available set of matching haplotype references [52]. We delved into the influence of minimum allele frequency (MAF) on imputation accuracy, uncovering a notable decrease in accuracy for MAF < 0.05. This observation parallels previous findings in maize [53] and in sheep [54], highlighting the significance of MAF in imputation accuracy assessments. This finding has important implications for genome-wide association studies (GWAS) and genomic selection (GS), particularly in species like *H. discus hannai*, where rare variants may play crucial roles in key traits such as growth and disease resistance. Due to their low frequency in the population, traditional genotype imputation methods may struggle to accurately infer the true genotypes of these rare variants, potentially impacting subsequent genetic analyses [55]. Compared to the imputation strategy of STITCH, an advantage of another strategy such as GLIMPSE is that it is very robust to MAF. GLIMPSE has been reported to perform quite well (accuracy > 0.9) even for SNPs with an MAF lower than 0.001 [25]. Currently, there are no systematic reports on strategies to improve rare variant imputation using population-specific reference panels. Some studies suggest that selecting loci with a higher MAF can enhance overall imputation accuracy [56]. Alternatively, selecting SNP sites for imputation using an evenly spaced approach can help reduce the genotyping errors caused by low-frequency MAF variants [57].

In aquatic species with highly heterozygous and complex genomes, a coverage of more than 10× should be suitable for obtaining sufficient SNPs with high accuracy [58]. In this study, to assess the imputation accuracies from lcWGS data, we utilized genotyped data from 16 individuals with a sequencing depth of 20× as a validation set. It is important to note that the 1059 Pacific abalone individuals—with a sequencing depth of 7.86×—and 16 individuals—with a sequencing depth of 20×—used for imputation accuracy assessment were able to accurately detect whole-genome SNP variations. From the genomic relationship analysis between the group of 1059 animals and the group of 16 animals in this study, there exists a similar genetic structure but still moderate genetic differentiation between the two populations. This is in accordance with studies where genetic differentiations occurred during a multi-generation selection program in aquaculture [59]. Previous studies in WGS and lcWGS have shown that higher imputation accuracies can be achieved in populations with a close genetic relationship [29]. For instance, in Pacific abalone, improved imputation

accuracy is observed in populations with similar genetic structures when using high-depth coverage sequencing data [60]. However, further investigation is needed to determine whether the lcWGS approach is equally effective for multiple populations within the Pacific abalone species. A larger reference panel that provides more comprehensive information about linkage disequilibrium (LD) patterns, as well as stronger LD among SNPs across the genome, can enhance imputation accuracy [27,61]. Such characteristics could contribute to higher imputation accuracy, even across genetically diverse populations.

The findings of this study differ slightly from those reported for other molluscan species, such as the Pacific oyster [29]. In a previous study, genotype imputation using lcWGS data in 300 Pacific oyster individuals at a sequencing depth of 2× achieved an imputation accuracy of GC = 0.951 and R^2 = 0.890 [29], lower than the imputation accuracy of GC = 0.97 and R^2 = 0.95 obtained in our study with 200 individuals (sequencing depth of 0.5×). This discrepancy highlights the impact of genomic background on imputation genotype accuracy. It is well documented that the Pacific oyster *Crassostrea gigas* has one of the most complex genomes among molluscan species, characterized by high heterozygosity and genomic variability [62]. As shown in Figure 1, although the Pacific abalone genome is also highly heterozygous and complex, its heterozygosity is approximately 0.15, which is significantly lower than that of the Pacific oyster. Moreover, compared to the relatively low sequencing depth used for validation in the oyster study, our study employed high-depth sequencing (20×) to assess the additional factors influencing imputation accuracy. Our results revealed a strong association between minor allele frequency (MAF) and imputation accuracy. Specifically, imputation accuracy for rare variants was lower, likely due to the significant role that MAF plays in the genetics of complex traits with larger genetic effects [63]. Therefore, enhancing imputation accuracy for rare variants remains an important avenue for future research, as suggested by the present study. Overall, the use of a 20× sequencing depth in this study for accuracy assessment strengthened the reliability of our findings. Additionally, the relatively large sample size enhanced the statistical power of our analyses, improved the imputation performance, facilitated rare variant detection, and provided a more comprehensive understanding of the genetic diversity and population structure in *H. discus hannai*.

This study demonstrates the promising potential of low-coverage whole-genome sequencing (lcWGS) for genomic analysis in Pacific abalone, particularly in the context of cost-effective genomic selection. We evaluated the imputation performance of lcWGS data in relation to sequencing depth and the number of individuals sequenced. Our findings provide valuable insights for both scientific research and applied breeding programs, offering a cost-efficient genotyping strategy that can be integrated into selective breeding pipelines. By optimizing sequencing depth and sample size, hatcheries and breeding programs can enhance genetic evaluations while minimizing costs, ultimately improving production efficiency and sustainability in aquaculture. However, to enhance the reliability and broader applicability of these findings. Future research should focus on developing high-quality reference panels to improve imputation accuracy, particularly for rare variants. Exploring advanced imputation algorithms and hybrid methods could further enhance performance. Increasing sample sizes and optimizing sequencing depth across diverse populations would improve statistical power and cost-effectiveness. Comparative studies on genomic background effects and integrating long-read sequencing or multi-omics approaches could refine imputation strategies. Additionally, haplotype-based methods and functional annotations may enhance rare variant detection. These improvements will strengthen the application of low-coverage whole-genome sequencing in Pacific abalone breeding and genomic studies.

4. Materials and Methods

4.1. Sample Materials and Whole-Genome Sequencing

A total of 1075 Pacific abalone individuals were randomly selected as samples from a selective breeding program in Fuda Abalone Farm (Jinjiang, China) for whole-genome sequencing, as described by Liu et al. [36]. To ensure population representativeness, individuals were selected from 114 families, including 86 paternal half-sibling families and 16 maternal half-sibling families, with 57 fathers and 106 mothers. In this study, normality tests were conducted using the phenotyping data of the animal samples that were chosen for sequencing. The whole-genome sequencing (WGS) data of 1059 abalone were obtained, with an average sequencing depth of 7.86× (at Novogene Corporation (Beijing, China), using the Illumina NovaSeq 6000 platform (150-bp paired-end; Illumina, Sacramento, CA, USA). In addition, in order to eliminate the potential impact of inaccurate genotyping due to relatively low measurement depth (7.86×), this study used an additional 16 individuals with an average resequencing depth of 20× as validation samples.

4.2. Sequencing Analysis and Variant Calling

A total of 1075 individuals were sequenced using the Illumina sequencing platform, with an average depth of 7.86×. The raw data were filtered using fastp (v0.23.4, Open-gene, Shenzhen, China) to remove low-quality or adapter sequences [64]. The clean reads were then aligned to the reference genome of abalone (GenBank: GCA_044707095.1) (1.4 Gb) [36] using the mem algorithm of BWA (Burrows–Wheeler Aligner, v0.7.17, Boston, MA, USA) [65], and the resulting BAM files were sorted using SAMtools (v1.18, Boston, MA, USA). The sorted and aligned BAM files were processed with the MarkDuplicates algorithm from GATK (v4.2.2, Broad Institute, Cambridge, MA, USA) to flag duplicates [66], with default parameters. Variants were filtered with the GATK VariantFiltration parameter "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < 12.5 || ReadPos-RankSum < −8.0". Following hard filtering with GATK, the SNPs were further filtered using VCFtools (v0.1.16, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK) [67] for subsequent analyses.

To evaluate the genomic relationship between the 16 samples with a sequencing depth of 20× and 1059 samples with an average sequencing depth of 7.86×, imputation of the missing genotypes in the whole-genome sequencing data was performed using Beagle (v5.1, University of Washington, Seattle, WA, USA) [68]. Variants with a minor allele frequency (MAF) lower than 0.05 and a deviation from the Hardy–Weinberg equilibrium (HWE) (p value < 10^{-7}) were excluded using the PLINK software (v 1.90, Broad Institute, Cambridge, MA, USA) [69]. Furthermore, due to the high level of LD in the genome, most SNPs were redundant. LD pruning was performed using PLINK (v 1.9) [69] to remove variants with high LD (R^2 > 0.9). After pruning, 1,642,965 SNPs were retained in the whole-genome sequencing dataset. Principal component analysis (PCA) was performed on the genomic relationship matrix using the GCTA software (v1.25.3, Westlake University, Hangzhou, China) [70]. This resulted in a matrix of eigenvectors in descending order that represented principal components (PCs), where PC1 had the largest eigenvalue. The overall structure of genetic variation was visualized using a scatterplot of the top few PCs. Furthermore, we used MEGA (v 11.0.13, Temple University, Philadelphia, PA, USA) [71] to construct a phylogenetic tree to illustrate the genetic relationships.

4.3. Genotype Imputation

We used two methods to impute low-depth sequencing data. The Beagle software (v 5.1) [68] was employed as the first method, which leveraged the linkage disequilibrium information between genetic markers to impute the genotypes for missing loci, thereby

enhancing the accuracy of the genotype data in the samples. The second method used was the STITCH software (v1.7.0, University of Oxford, Oxford, Oxfordshire, UK) (Sequencing to Imputation Through Constructing Haplotypes, which performed genotype imputation based on a reference haplotype library, eliminating the need for additional reference panels [8]. In this study, the following two strategies were primarily adopted for imputation: BaseVar+STITCH for the initial imputation and Beagle for the secondary imputation. BaseVar (v0.8.0, Beijing Genomics Institute, Shenzhen, Guangdong, China) was primarily employed to infer allele frequencies and identify polymorphic sites, followed by imputation using STITCH. Subsequently, the SNPs with a missing rate of 0.2 were selected for the secondary imputation using Beagle.

4.3.1. Evaluation of the Impact of Different Sequencing Depths and Sample Sizes on Imputation Accuracy

To detect whether and how the accuracy of genotype imputation is influenced by sequencing depth and sample size in Pacific abalone, we used the BaseVar+STITCH strategy to further investigate the relationships. The BaseVar software, developed by BGI, was applied to extract variant position information and infer allele frequencies [72]. In this study, we randomly down-sampled paired-reads from the 1075 sequenced individuals to produce different sequencing datasets with depths of 0.5×, 1×, 2×, 3×, and 4×, respectively, using the DownsampleSam module of the Picard software (v2.9.0, Broad Institute, Cambridge, MA, USA) [73]. Sequencing depths (0.5×, 1×, 2×, 3×, 4×) were chosen based on cost–benefit thresholds observed in previous aquatic studies [29]. Additionally, we sampled 200, 400, 600, and 800 individuals that were randomly selected from the 1075 sequenced samples to generate samples of diminished sizes. The sample sizes were selected to span from the minimum effective population size to the saturation points identified in previous genomic studies [25]. STITCH was then employed to impute the data on chromosome 1, enabling the assessment of imputation accuracy.

4.3.2. Evaluation of the Impact of Chromosome Length on Imputation Accuracy

To evaluate the influence of chromosome length on genotype imputation in Pacific abalone, considering the differences in imputation stability observed across chromosomes [74] and the computational time limitations [75], we selected a sample size of 1075 and three chromosomes—Chr 1, Chr 13, Chr 7—representing the long, medium, and short chromosomes, respectively, to assess the imputation accuracy at a sequencing depth of 1×. The variant position information of Chr 1, Chr 7, and Chr 13 was extracted from all the de-weighted bam files using the BaseVar software (v0.8.0). The three chromosomes were then imputed and assessed for accuracy.

4.3.3. Evaluation of the Effect of Allele Frequency on Imputation Accuracy

The minor allele frequency (MAF) is indicative of the frequency of a specific allele within a population, impacting the genotypes considered and imputed during genotype imputation. Lower MAF values may result in the neglect or erroneous imputation of rare genotypes, while higher MAF values could enhance imputation accuracy but also increase computational time. To assess the impact of MAF on imputation accuracy, the MAF was divided into eight intervals using the vcftools (v0.1.16) software—[0–0.005], [0.005–0.01], [0.01–0.05], [0.05–0.1], [0.1–0.2], [0.2–0.3], [0.3–0.4], and [0.4–0.5]—and the accuracy was compared at a sequencing depth of 1×.

4.4. Evaluation of Imputation Accuracy

To assess the accuracy of genotype imputation in the low-coverage whole-genome sequencing (lcWGS) of Pacific abalone, a validation set comprising 16 individuals with

an average sequencing depth exceeding 20× was randomly selected. In this study, we considered the use of 16 validation individuals to be sufficient. For example, in the study by Yang et al., 36 individuals with a sequencing depth of 12× were used for validation, while another study used a subset of 18 individuals sequenced at a depth of 30× as validation samples [12,29]. High sequencing depth provides highly accurate genotype information, and the selected individuals share a similar genetic background with the 1059 imputed individuals, making them a reliable "true" genotype reference. The increased sequencing depth significantly reduces sequencing errors and random noise in genotype calls, ensuring the reliability of the validation results. The following two evaluation metrics were employed: genotype concordance (GC) and the squared Pearson correlation coefficient of genotype dosage (R^2). Genotype concordance (GC) evaluates the consistency between imputed genotypes and those determined through high-depth sequencing, serving as a measure to compare the agreement between the two [76]. The squared Pearson correlation coefficient of genotype dosage (R^2) is used to quantify the accuracy of imputed genotypes as compared to genotypes obtained from high-depth sequencing. It reflects the squared correlation between the posterior expectation of imputed allele dosages and the true genotypes derived from the high-depth sequencing data; the high-depth data serve as the reference, and low-depth data serve as the test set [54]. Genotyping was performed on the sequencing data of these individuals. Both metrics were calculated for each SNP, and the mean value across all SNPs was determined. This systematic evaluation, utilizing GC and R^2 metrics, established a robust framework for assessing the accuracy of genotype imputation within the lcWGS context, ensuring a comprehensive and rigorous evaluation of the imputation quality against high-depth sequencing data.

5. Conclusions

This study provides a comprehensive evaluation of low-coverage whole-genome sequencing (lcWGS) as a cost-effective and reliable genotyping method for Pacific abalone. The following key conclusions can be drawn from our research: (1) Our results suggest that an imputation accuracy above 98% can be achieved with a sample size of approximately 400 individuals, making lcWGS a robust method for genomic studies in aquaculture species. (2) A sequencing depth of 1× is recommended as the optimal balance between cost and accuracy for large-scale population studies. When the sample size exceeds 400 individuals, lcWGS exhibits outstanding performance in variant detection efficiency. (3) For Pacific abalone lacking a reference panel, the combination of STITCH followed by Beagle represents an optimal strategy to increase the number of SNPs discovered, which is crucial for improving genetic selection in Pacific abalone. (4) These findings establish low-coverage whole-genome sequencing (lcWGS) as a viable and cost-effective genotyping method for Pacific abalone, providing a solid foundation for future analyses in this species and potentially other aquaculture species with complex genomes. Collectively, this study provides valuable insights into the implementation of high-throughput genotyping technologies, which can accelerate the genetic analysis of economic traits in aquaculture species.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms26104598/s1.

Author Contributions: C.F. and S.Z.: Data Curation, Formal Analysis, Writing—Original Draft; J.L. and W.P.: Data Curation; X.C., W.Y. and G.Z.: Review and Discussion; F.W.: Conceptualization, Methodology, Formal Analysis, Writing—Original Draft, Writing—Review and Editing, Supervision, Project Administration, Funding Acquisition. All authors contributed to the article and approved the submitted version. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (31972790 to F.W.), the Key Research and Development Program of Shandong (2023CXGC010410 and 2022LZGC015, 2024LZGCQY003 and ZFJH202309 to F.W.), Fujian-CAS (Chinese Academy of Sciences) STS Program (2024T3049 to F.W.), and the Dalian Science and Technology Plan Project (2023YF19SN034 to F.W.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: This study was supported by the Oceanographic Data Center, Institute of Oceanology, Chinese Academy of Sciences. The authors thank Ning Chao and Teng Jun at Shandong Agriculture University for their valuable discussions.

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Li, J.; He, Q.; Sun, H.; Liu, X. Acclimation-Dependent Expression of Heat Shock Protein 70 in Pacific Abalone (Haliotis Discus Hannai Ino) and Its Acute Response to Thermal Exposure. *Chin. J. Oceanol. Limnol.* **2012**, *30*, 146–151. [CrossRef]
2. Gao, X.; Zhang, M.; Luo, X.; You, W.; Ke, C. Transitions, Challenges and Trends in China's Abalone Culture Industry. *Rev. Aquac.* **2023**, *15*, 1274–1293. [CrossRef]
3. FAO. *The State of World Fisheries and Aquaculture 2020*; FAO: Rome, Italy, 2020; ISBN 978-92-5-132692-3.
4. Ellegren, H. Genome Sequencing and Population Genomics in Non-Model Organisms. *Trends Ecol. Evol.* **2014**, *29*, 51–63. [CrossRef] [PubMed]
5. Hon, T.; Mars, K.; Young, G.; Tsai, Y.C.; Karalius, J.W.; Landolin, J.M.; Maurer, N.; Kudrna, D.; Hardigan, M.A.; Steiner, C.C.; et al. Highly Accurate Long-Read HiFi Sequencing Data for Five Complex Genomes. *Sci. Data* **2020**, *7*, 399. [CrossRef]
6. Mardis, E.R. DNA Sequencing Technologies: 2006–2016. *Nat. Protoc.* **2017**, *12*, 213–218. [CrossRef]
7. Song, H.; Dong, T.; Yan, X.; Wang, W.; Tian, Z.; Sun, A.; Dong, Y.; Zhu, H.; Hu, H. Genomic Selection and Its Research Progress in Aquaculture Breeding. *Rev. Aquac.* **2023**, *15*, 274–291. [CrossRef]
8. Davies, R.W.; Flint, J.; Myers, S.; Mott, R. Rapid Genotype Imputation from Sequence without Reference Panels. *Nat. Genet.* **2016**, *48*, 965–969. [CrossRef]
9. Robledo, D.; Palaiokostas, C.; Bargelloni, L.; Martínez, P.; Houston, R. Applications of Genotyping by Sequencing in Aquaculture Breeding and Genetics. *Rev. Aquac.* **2018**, *10*, 670–682. [CrossRef]
10. Baird, N.A.; Etter, P.D.; Atwood, T.S.; Currey, M.C.; Shiver, A.L.; Lewis, Z.A.; Selker, E.U.; Cresko, W.A.; Johnson, E.A. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **2008**, *3*, e3376. [CrossRef]
11. Peterson, B.K.; Weber, J.N.; Kay, E.H.; Fisher, H.S.; Hoekstra, H.E. Hoekstra He Double Digest RADseq: An Inexpensive Method for de Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* **2012**, *7*, e37135. [CrossRef]
12. Yang, R.; Guo, X.; Zhu, D.; Tan, C.; Bian, C.; Ren, J.; Huang, Z.; Zhao, Y.; Cai, G.; Liu, D.; et al. Accelerated Deciphering of the Genetic Architecture of Agricultural Economic Traits in Pigs Using a Low-Coverage Whole-Genome Sequencing Strategy. *GigaScience* **2021**, *10*, giab048. [CrossRef] [PubMed]
13. VanRaden, P.M.; Null, D.J.; Sargolzaei, M.; Wiggans, G.R.; Tooker, M.E.; Cole, J.B.; Sonstegard, T.S.; Connor, E.E.; Winters, M.; van Kaam, J.B.; et al. Genomic Imputation and Evaluation Using High-Density Holstein Genotypes. *J. Dairy Sci.* **2013**, *96*, 668–678. [CrossRef]
14. Li, Y.; Sidore, C.; Kang, H.M.; Boehnke, M.; Abecasis, G.R. Low-Coverage Sequencing: Implications for Design of Complex Trait Association Studies. *Genome Res.* **2011**, *21*, 940–951. [CrossRef] [PubMed]
15. Zan, Y.; Payen, T.; Lillie, M.; Honaker, C.F.; Siegel, P.B.; Carlborg, Ö. Genotyping by Low-Coverage Whole-Genome Sequencing in Intercross Pedigrees from Outbred Founders: A Cost-Efficient Approach. *Genet. Sel. Evol.* **2019**, *51*, 44. [CrossRef] [PubMed]
16. Deng, T.; Zhang, P.; Garrick, D.; Gao, H.; Wang, L.; Zhao, F. Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. *Front. Genet.* **2021**, *12*, 704118. [CrossRef]
17. Nielsen, R.; Paul, J.S.; Albrechtsen, A.; Song, Y.S. Genotype and SNP Calling from Next-Generation Sequencing Data. *Nat. Rev. Genet.* **2011**, *12*, 443–451. [CrossRef]
18. Fragoso, C.A.; Heffelfinger, C.; Zhao, H.; Dellaporta, S.L. Imputing Genotypes in Biallelic Populations from Low-Coverage Sequence Data. *Genetics* **2016**, *202*, 487–495. [CrossRef]

19. Ros-Freixedes, R.; Gonen, S.; Gorjanc, G.; Hickey, J.M. A Method for Allocating Low-Coverage Sequencing Resources by Targeting Haplotypes Rather than Individuals. *Genet. Sel. Evol.* **2017**, *49*, 78. [CrossRef]
20. Browning, S.R.; Browning, B.L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **2007**, *81*, 1084–1097. [CrossRef]
21. Pook, T.; Mayer, M.; Geibel, J.; Weigend, S.; Cavero, D.; Schoen, C.C.; Simianer, H. Improving Imputation Quality in BEAGLE for Crop and Livestock Data. *G3* **2020**, *10*, 177–188. [CrossRef]
22. Nicod, J.; Davies, R.W.; Cai, N.; Hassett, C.; Goodstadt, L.; Cosgrove, C.; Yee, B.K.; Lionikaite, V.; McIntyre, R.E.; Remme, C.A.; et al. Genome-Wide Association of Multiple Complex Traits in Outbred Mice by Ultra-Low-Coverage Sequencing. *Nat. Genet.* **2016**, *48*, 912–918. [CrossRef] [PubMed]
23. Gilly, A.; Ritchie, G.R.; Southam, L.; Farmaki, A.E.; Tsafantakis, E.; Dedoussis, G.; Zeggini, E. Very Low-Depth Sequencing in a Founder Population Identifies a Cardioprotective APOC3 Signal Missed by Genome-Wide Imputation. *Hum. Mol. Genet.* **2016**, *25*, 2360–2365. [CrossRef] [PubMed]
24. Song, H.; Dong, T.; Wang, W.; Jiang, B.; Yan, X.; Geng, C.; Bai, S.; Xu, S.; Hu, H. Cost-Effective Genomic Prediction of Critical. Economic Traits in Sturgeons through Low-Coverage Sequencing. *Genomics* **2024**, *116*, 110874. [CrossRef] [PubMed]
25. Teng, J.; Zhao, C.; Wang, D.; Chen, Z.; Tang, H.; Li, J.; Mei, C.; Yang, Z.; Ning, C.; Zhang, Q. Assessment of the Performance of Different Imputation Methods for Low-Coverage Sequencing in Holstein Cattle. *J. Dairy Sci.* **2022**, *105*, 3355–3366. [CrossRef]
26. Zhao, C.; Teng, J.; Zhang, X.; Wang, D.; Zhang, X.; Li, S.; Jiang, X.; Li, H.; Ning, C.; Zhang, Q. Towards a Cost-Effective Implementation of Genomic Prediction Based on Low Coverage Whole Genome Sequencing in Dezhou Donkey. *Front. Genet.* **2021**, *12*, 728764. [CrossRef]
27. Zhang, W.; Li, W.; Liu, G.; Gu, L.; Ye, K.; Zhang, Y.; Li, W.; Jiang, D.; Wang, Z.; Fang, M. Evaluation for the Effect of Low-Coverage Sequencing on Genomic Selection in Large Yellow Croaker. *Aquaculture* **2021**, *534*, 736323. [CrossRef]
28. Fuller, Z.L.; Mocellin, V.J.L.; Morris, L.A.; Cantin, N.; Shepherd, J.; Sarre, L.; Peng, J.; Liao, Y.; Pickrell, J.; Andolfatto, P.; et al. Population Genetics of the Coral Acropora Millepora: Toward Genomic Prediction of Bleaching. *Science* **2020**, *369*, eaba4674. [CrossRef]
29. 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* **2024**, *591*, 741134. [CrossRef]
30. Sui, M.; Liu, Z.; Huang, X.; Yang, Z.; Yu, H.; Cui, C.; Hu, Y.; Wang, X.; Shen, X.; Mu, Q.; et al. Development and Evaluation of a Haplotype Reference Panel of Zhikong Scallop (Chlamys Farreri) for Genotype Imputation. *Aquaculture* **2024**, *582*, 740497. [CrossRef]
31. Beemelmanns, A.; Bouchard, R.; Michaelides, S.; Normandeau, E.; Jeon, H.; Chamlian, B.; Babin, C.; Hénault, P.; Perrot, O.; Harris, L.N.; et al. Development of SNP Panels from Low-coverage Whole Genome Sequencing (lcWGS) to Support Indigenous Fisheries for Three Salmonid Species in Northern Canada. *Mol. Ecol. Resour.* **2025**, *25*, e14040. [CrossRef]
32. Liu, S.; Martin, K.E.; Snelling, W.M.; Long, R.; Leeds, T.D.; Vallejo, R.L.; Wiens, G.D.; Palti, Y. Accurate Genotype Imputation from Low-Coverage Whole-Genome Sequencing Data of Rainbow Trout. *G3 Genes | Genomes | Genetics* **2024**, *14*, jkae168. [CrossRef] [PubMed]
33. Zeng, Q.; Zhao, B.; Wang, H.; Wang, M.; Teng, M.; Hu, J.; Bao, Z.; Wang, Y. Aquaculture Molecular Breeding Platform (AMBP): A Comprehensive Web Server for Genotype Imputation and Genetic Analysis in Aquaculture. *Nucleic Acids Res.* **2022**, *50*, W66–W74. [CrossRef]
34. 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.; et al. Harnessing Genomics to Fast-Track Genetic Improvement in Aquaculture. *Nat. Rev. Genet.* **2020**, *21*, 389–409. [CrossRef] [PubMed]
35. Pasaniuc, B.; Rohland, N.; McLaren, P.J.; Garimella, K.; Zaitlen, N.; Li, H.; Gupta, N.; Neale, B.M.; Daly, M.J.; Sklar, P.; et al. Extremely Low-Coverage Sequencing and Imputation Increases Power for Genome-Wide Association Studies. *Nat. Genet.* **2012**, *44*, 631–635. [CrossRef]
36. Liu, J.; Peng, W.; Yu, F.; Lin, W.; Shen, Y.; Yu, W.; Gong, S.; Huang, H.; You, W.; Luo, X.; et al. Development and Validation of a 40-K Multiple-SNP Array for Pacific Abalone (Haliotis Discus Hannai). *Aquaculture* **2022**, *558*, 738393. [CrossRef]
37. Lin, W.; Xiao, Q.; Yu, F.; Han, Z.; Liu, J. Development of a Low-Density SNP Genotyping Panel by a Novel Technology mGPS and Its Application in Germplasm Identification of Abalone. *Aquaculture* **2023**, *565*, 739089. [CrossRef]
38. Kijas, J.; Hamilton, M.; Botwright, N.; King, H.; McPherson, L.; Krsinich, A.; McWilliam, S. Genome Sequencing of Blacklip and Greenlip Abalone for Development and Validation of a SNP Based Genotyping Tool. *Front. Genet.* **2018**, *9*, 687. [CrossRef]
39. Dimond, J.L.; Bouma, J.V.; Lafarga-De la Cruz, F.; Supernault, K.J.; White, T.; Witting, D.A. Endangered Pinto/Northern Abalone (Haliotis Kamtschatkana) Are Panmictic across Their 3700 Km Range along the Pacific Coast of North America. *Evol. Appl.* **2024**, *17*, e70040. [CrossRef]

40. Yang, B.; Zhai, S.Y.; Zhang, F.Q.; Wang, H.B.; Ren, L.T.; Li, Y.J.; Li, Q.; Liu, S.K. Genome-Wide Association Study toward Efficient Selection Breeding of Resistance to Vibrio Alginolyticus in Pacific Oyster, Crassostrea Gigas. *Aquaculture* **2022**, *548*, 737592. [CrossRef]
41. Calus, M.; Bouwman, A.C.; Hickey, J.M.; Veerkamp, R.F.; Mulder, H.A. Evaluation of Measures of Correctness of Genotype Imputation in the Context of Genomic Prediction: A Review of Livestock Applications. *Animal* **2014**, *8*, 1743–1753. [CrossRef]
42. Rutkoski, J.E.; Poland, J.; Jannink, J.L.; Sorrells, M.E. Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. *G3* **2013**, *3*, 427–439. [CrossRef] [PubMed]
43. Das, S.; Abecasis, G.R.; Browning, B.L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genom. Hum. Genet.* **2018**, *19*, 73–96. [CrossRef]
44. Aliloo, H.; Mrode, R.; Okeyo, A.M.; Ni, G.; Goddard, M.E.; Gibson, J.P. The Feasibility of Using Low-Density Marker Panels for Genotype Imputation and Genomic Prediction of Crossbred Dairy Cattle of East Africa. *J. Dairy Sci.* **2018**, *101*, 9108–9127. [CrossRef] [PubMed]
45. García-Ruiz, A.; Ruiz-Lopez, F.J.; Wiggans, G.R.; Van Tassell, C.P.; Montaldo, H.H. Effect of Reference Population Size and Available Ancestor Genotypes on Imputation of Mexican Holstein Genotypes. *J. Dairy Sci.* **2015**, *98*, 3478–3484. [CrossRef] [PubMed]
46. Butty, A.M.; Sargolzaei, M.; Miglior, F.; Stothard, P.; Schenkel, F.S.; Gredler-Grandl, B.; Baes, C.F. Optimizing Selection of the Reference Population for Genotype Imputation From Array to Sequence Variants. *Front. Genet.* **2019**, *10*, 510. [CrossRef]
47. VanRaden, P.M.; Sun, C.; O'Connell, J.R. Fast Imputation Using Medium or Low-Coverage Sequence Data. *BMC Genet.* **2015**, *16*, 82. [CrossRef]
48. Hui, R.; D'Atanasio, E.; Cassidy, L.M.; Scheib, C.L.; Kivisild, T. Evaluating Genotype Imputation Pipeline for Ultra-Low Coverage Ancient Genomes. *Sci. Rep.* **2020**, *10*, 18542. [CrossRef]
49. Delomas, T.A.; Hollenbeck, C.M.; Matt, J.L.; Thompson, N.F. Evaluating Cost-Effective Genotyping Strategies for Genomic Selection in Oysters. *Aquaculture* **2023**, *562*, 738844. [CrossRef]
50. Kriaridou, C.; Tsairidou, S.; Fraslin, C.; Gorjanc, G.; Looseley, M.E.; Johnston, I.A.; Houston, R.D.; Robledo, D. Evaluation of Low-Density SNP Panels and Imputation for Cost-Effective Genomic Selection in Four Aquaculture Species. *Front. Genet.* **2023**, *14*, 1194266. [CrossRef]
51. Heidaritabar, M.; Calus, M.P.L.; Vereijken, A.; Groenen, M.A.M.; Bastiaansen, J.W.M. Accuracy of Imputation Using the Most Common Sires as Reference Population in Layer Chickens. *BMC Genet.* **2015**, *16*, 101. [CrossRef]
52. Júnior, G.A.F.; 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. Imputation Accuracy to Whole-Genome Sequence in Nellore Cattle. *Genet. Sel. Evol.* **2021**, *53*, 27. [CrossRef]
53. Hayes, B.J.; Bowman, P.J.; Daetwyler, H.D.; Kijas, J.W.; van der Werf, J.H.J. Accuracy of Genotype Imputation in Sheep Breeds. *Anim. Genet.* **2012**, *43*, 72–80. [CrossRef] [PubMed]
54. Hickey, J.M.; Crossa, J.; Babu, R.; de los Campos, G. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Sci.* **2012**, *52*, 654–663. [CrossRef]
55. Yu, W.; Yan, S.; Zhang, S.; Ni, J.; Bin, L.; Pei, Y.; Zhang, L. Efficient Identification of Trait-associated Loss-of-function Variants in the UK Biobank Cohort by Exome-sequencing Based Genotype Imputation. *Genet. Epidemiol.* **2023**, *47*, 121–134. [CrossRef] [PubMed]
56. Boichard, D.; Chung, H.; Dassonneville, R.; David, X.; Eggen, A.; Fritz, S.; Gietzen, K.J.; Hayes, B.J.; Lawley, C.T.; Sonstegard, T.S.; et al. Design of a Bovine Low-Density SNP Array Optimized for Imputation. *PLoS ONE* **2012**, *7*, e34130. [CrossRef]
57. Yuan, M.; Fang, H.; Zhang, H. Correcting for Differential Genotyping Error in Genetic Association Analysis. *J. Hum. Genet.* **2013**, *58*, 657–666. [CrossRef]
58. Song, K.; Li, L.; Zhang, G. Coverage Recommendation for Genotyping Analysis of Highly Heterologous Species Using Next-Generation Sequencing Technology. *Sci. Rep.* **2016**, *6*, 35736. [CrossRef]
59. Diyie, R.L.; Agyarkwa, S.K.; Armah, E.; Amonoo, N.A.; Owusu-Frimpong, I.; Osei-Atweneboana, M.Y. Genetic Variations among Different Generations and Cultured Populations of Nile Tilapia (Oreochromis Niloticus) in Ghana: Application of Microsatellite Markers. *Aquaculture* **2021**, *544*, 737070. [CrossRef]
60. Sun, X.; Fei, C.; Mi, C.; Li, M.; Zhang, G.; Wu, F. Genetic Diversity and Population Structure of Pacific Abalone (Haliotis Discus Hannai) Using SNP Genotyping Data. *Aquaculture* **2024**, *593*, 741335. [CrossRef]
61. Zhang, C.; Dong, S.S.; Xu, J.Y.; He, W.M.; Yang, T.L. PopLDdecay: A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files. *Bioinformatics* **2019**, *35*, 1786–1788. [CrossRef]
62. Plough, L.V. Genetic Load in Marine Animals: A Review. *Curr. Zool.* **2016**, *62*, 567–579. [CrossRef] [PubMed]
63. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the Missing Heritability of Complex Diseases. *Nature* **2009**, *461*, 747–753. [CrossRef]

64. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics* **2018**, *34*, 884–890. [CrossRef]
65. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
66. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
67. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The Variant Call Format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef] [PubMed]
68. Browning, B.; Browning, S. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* **2009**, *84*, 210–223. [CrossRef]
69. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaScience* **2015**, *4*, s13742-015-0047–0048. [CrossRef]
70. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A Tool for Genome-Wide Complex Trait Analysis. *Am. J. Hum. Genet.* **2011**, *88*, 76–82. [CrossRef]
71. Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022–3027. [CrossRef]
72. Liu, S.; Huang, S.; Chen, F.; Zhao, L.; Yuan, Y.; Francis, S.S.; Fang, L.; Li, Z.; Lin, L.; Liu, R.; et al. Genomic Analyses from Non-Invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* **2018**, *175*, 347–359. [CrossRef] [PubMed]
73. Broad Institute. *Picard Toolkit*; Broad Institute of MIT and Harvard: Cambridge, MA, USA, 2019.
74. Korkuć, P.; Arends, D.; Brockmann, G.A. Finding the Optimal Imputation Strategy for Small Cattle Populations. *Front. Genet.* **2019**, *10*, 52. [CrossRef] [PubMed]
75. Jattawa, D.; Elzo, M.A.; Koonawootrittriron, S.; Suwanasopee, T. Imputation Accuracy from Low to Moderate Density Single Nucleotide Polymorphism Chips in a Thai Multibreed Dairy Cattle Population. *Asian-Australas J. Anim. Sci.* **2016**, *29*, 464–470. [CrossRef] [PubMed]
76. Lin, P.; Hartz, S.M.; Zhang, Z.; Saccone, S.F.; Wang, J.; Tischfield, J.A.; Edenberg, H.J.; Kramer, J.R.; M Goate, A.; Bierut, L.J.; et al. A New Statistic to Evaluate Imputation Reliability. *PLoS ONE* **2010**, *5*, e9697. [CrossRef]