https://doi.org/10.1093/nar/gkac424

Aquaculture Molecular Breeding Platform (AMBP): a comprehensive web server for genotype imputation and genetic analysis in aquaculture

Qifan Zeng[1,2,3,†], Baojun Zhao[1,†], Hao Wang[1,†], Mengqiu Wang[1], Mingxuan Teng[1], Jingjie Hu[1,2], Zhenmin Bao[1,2,3] and Yangfan Wang[1,3,*]

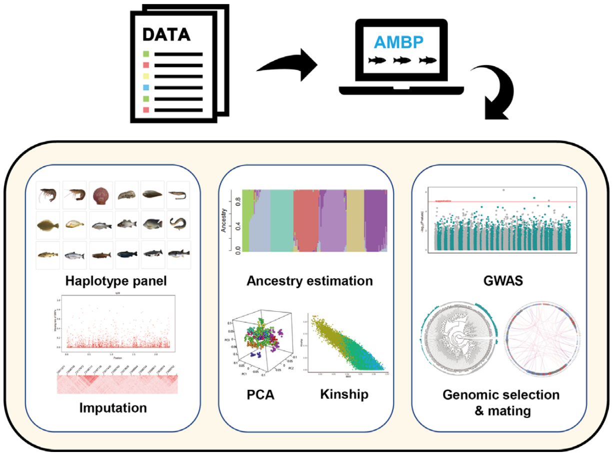
[1]MOE Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China, [2]Key Laboratory of Tropical Aquatic Germplasm of Hainan Province, Sanya Oceanog Inst, Ocean Univ China, Sanya 572000, Peoples R China and [3]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

Received March 21, 2022; Revised April 19, 2022; Editorial Decision April 29, 2022; Accepted May 09, 2022

ABSTRACT

It is of vital importance to understand the population structure, dissect the genetic bases of performance traits, and make proper strategies for selection in breeding programs. However, there is no single webserver covering the specific needs in aquaculture. We present Aquaculture Molecular Breeding Platform (AMBP), the first web server for genetic data analysis in aquatic species of farming interest. AMBP integrates the haplotype reference panels of 18 aquaculture species, which greatly improves the accuracy of genotype imputation. It also supports multiple tools to infer genetic structures, dissect the genetic architecture of performance traits, estimate breeding values, and predict optimum contribution. All the tools are coherently linked in a web-interface for users to generate interpretable results and evaluate statistical appropriateness. The webserver supports standard VCF and PLINK (PED, MAP) files, and implements automated pipelines for format transformation and visualization to simplify the process of analysis. As a demonstration, we applied the webserver to Pacific white shrimp and Atlantic salmon datasets. In summary, AMBP constitutes comprehensive resources and analytical tools for exploring genetic data and guiding practical breeding programs. AMBP is available at http://mgb.qnlm.ac.

GRAPHICAL ABSTRACT



INTRODUCTION

Aquaculture supplies over 16% of the seafood for human diets, contributing significantly to the Sustainable Development Goals of the United Nations (1). To match the ever-increasing food demands of the growing population, aquatic food production should increase about five-fold in the next three decades, with an annual growth rate larger than most sectors of the food industry (2). As in recent years, overexploitation of wild stocks is placing high pressure on aquatic ecosystems and causing irreversible impacts on environments (3). The capture fisheries are urged to be restricted within tolerable limits (4). Mariculture is evidently developing toward large-scale, intensive, and sustainable to afford dietary animal protein. Therefore, there is an urgent need for domestication and genetic improvement programs to increase efficiency and reduce the environmental impacts of aquaculture.

[*]To whom correspondence should be addressed. Tel: +86 0532 82031960; Email: yfwang@ouc.edu.cn
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Genomic tools could benefit the domestication and genetic improvement continuum in many aspects. It could be applied to characterize the population histories of wild resources and inform the construction of base populations for specialized breeding nuclei. Furthermore, genomic resources and analyses are essential to dissecting the genetic architecture of performance traits. Genomics holds great promise to increase the efficiency of selective breeding and management of farmed stocks. As a well-known example, the identification of a quantitative trait locus (QTL) for infectious pancreatic necrosis (IPN) resistance in Atlantic salmon greatly facilitated selective breeding, which eliminated the incidence of IPN in a few generations (5). Genomic selection (GS) makes use of genome-wide markers to capture quantitative trait loci (QTL) and can accurately predict breeding values (6). GS has revolutionized modern animal breeding leading to a more rapid genetic gain and a further reduction of generation interval than classic selection methods (7). Nevertheless, GS may increase inbreeding and produce a more rapid depletion of genetic variability of the selected traits in future generations. The loss of genetic diversity limits long-term gain for the trait under GS selection, and it also jeopardizes future breeding for other traits. In contrast, genomic mating (GM) provides a strategy that balances genetic gain and diversity (8). GM allows obtaining desirable genetic gain while constraining the rate of inbreeding in the progeny by restricting the relationships among selected parents. Therefore, GM appears to be a sustainable strategy for the genetic improvement in aquaculture breeding.

Despite the rapid advancement and reduced cost of genotyping and sequencing techniques, incorporation of genomic tools such as GS and GM typically requires genotyping of thousands of animals per generation, which is still prohibitively expensive and not practical to be routinely used. An alternative strategy is to genotype the target population with low-coverage sequencing or low-density SNP arrays and obtain genome-wide genotypes through imputation (9,10). This strategy is pronounced for aquaculture species, mostly without commercially available SNP array (11). Imputation is also very useful for meta-analysis of datasets from different genotyping platforms (12). It has been successfully applied in several aquaculture species, such as Atlantic salmon (13) and large yellow croaker (14). However, a high-quality reference panel is crucial for accurate genotype imputation. In contrast to crop and livestock, most aquaculture species have no high-quality reference panel, limiting the broad application of genotype imputation in aquaculture genetic studies. In addition, most current software for imputation, GWAS, GS, and other genetic analyses require a specific background knowledge of bioinformatics and quantitative genetics, making it challenging for general geneticists and biologists to perform the analysis. Therefore, it is essential to develop a user-friendly webserver to fill the gap.

To address this need, we developed the Aquaculture Molecular Breeding Platform (AMBP). By implementation of high-quality reference panels of 18 aquatic species of great economic value, and automated pipelines of genotype imputation, kinship deduction, population structure inference, GWAS, GS, and genomic mating (GM), we intend to develop the webserver into a portal for genetic data exploration and breeding strategy development of aquaculture species.

MATERIALS AND METHODS

AMBP was hosted on a dedicated rack server at the Platform for High-Performance Computing and Systematic Simulation at Qingdao National Laboratory for Marine Science and Technology. It allows users to explore genotype data in three major sections (Figure 1). The impute section was designed for the phasing and imputation of low-density SNP array or low-coverage sequencing datasets. High-density haplotype reference panels of 18 aquaculture species were pre-constructed and implemented in the pipeline. Users can browse the SNP information by chromosome coordinate and retrieve the reference panels of each species. In the population characterizing section, users could get indications of the genetic structure by ancestry estimation and principal component analysis (PCA). Pairwise relationships could be inferred by kinship coefficients and identical-by-descent (IBD) segments. The genetic breeding section includes models for GWAS, genomic prediction, and mating allocation. Users could overview the genetic architecture of performance traits and make optimum selections with constrained inbreeding. Furthermore, simulation analysis could help develop breeding strategies depending on short-term genetic gain and long-term potential.

Cookie statement

The AMBP service stores session information in cookies to provide information described in the Privacy Statement. Personal information is not directly stored in cookies. The username and password will be encrypted for transmission. Users cannot access the session information and need to restart a new session after 30 minutes of inactivity.

Imputation

The WGS data of 18 aquaculture species were collected from the NCBI SRA database to construct reference panels (Supplementary Table S1). Low-quality reads were trimmed using the trimmomatic v0.39. The clean reads were aligned to the current standard reference genome using Burrows-Wheeler Aligner (BWA) (15). The duplicated reads were marked and removed using MarkDuplicates of Picard Tools. Variant calling was carried out using HaplotypeCaller and GenotypeGVCFs algorithms in GATK (16). The raw SNPs were filtered using vcffilter with the parameters GQ $<$ 20, QD $<$ 10, FS $>$ 10, MQ $<$ 40, ReadPosRankSum $<$ -8, SOR $>$ 4, and MQRankSum $<$ -12.5. The SNPs with MAF $<$ 0.01 or located within 5bp flanking regions of other variants were excluded with bcftools (17). Finally, the biallelic SNPs were used to construct the reference panels of 18 species with the read-aware phasing method implemented in SHAPEIT2 (18). The polymorphic spectrum and potential effect of each SNP in the reference panels can be browsed by the ‘SNP Search’ function. Reference panels can also be retrieved on the ‘Download’ page for local analysis.

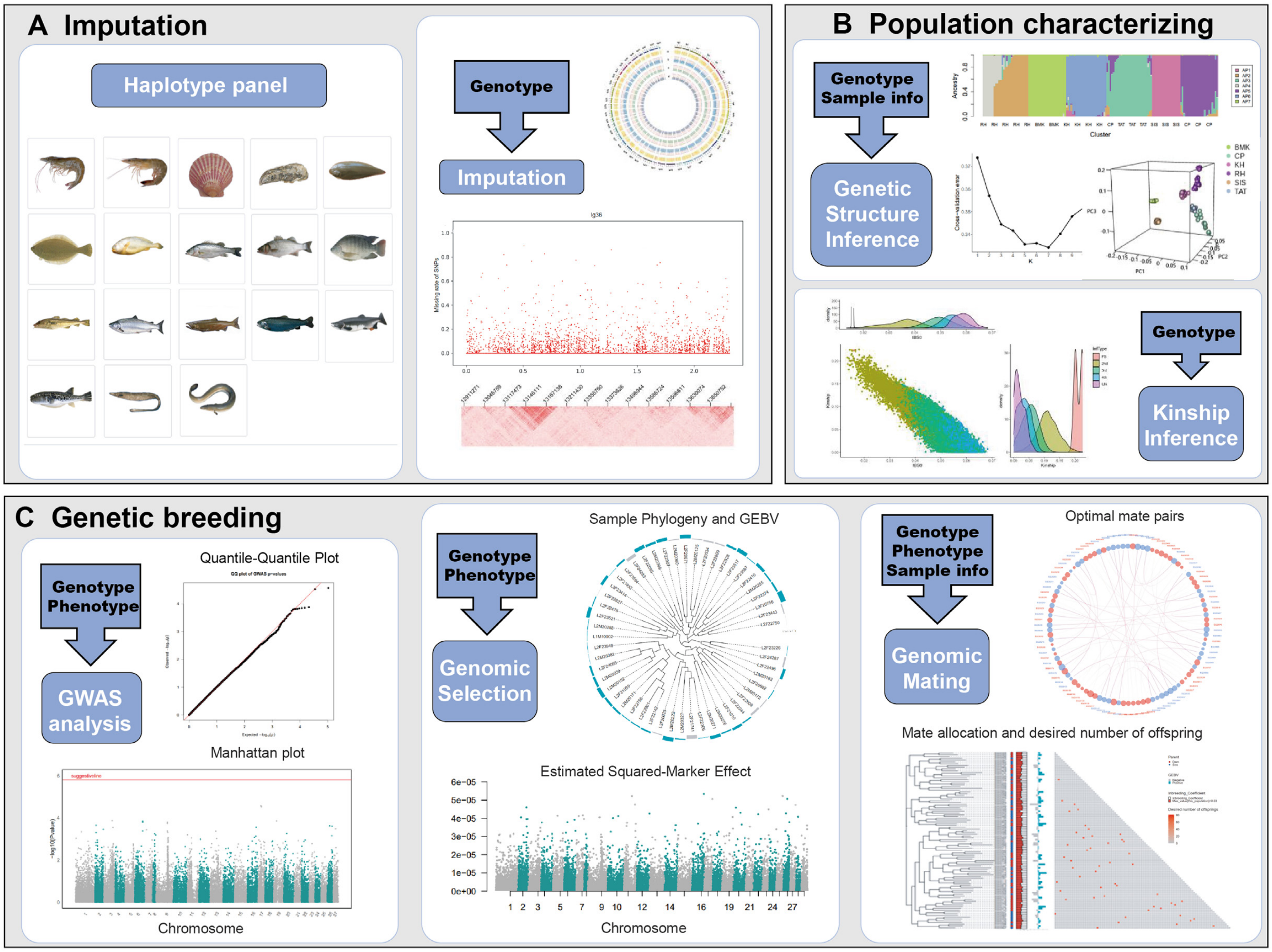


Figure 1. Functions of AMBP can be classified into three major categories. (**A**) High-density haplotype reference panels of 18 aquaculture species were implemented in the pipeline for users to impute genotype data with low density markers. The quality of imputation and LD patterns could be visualized online for a better interpretation. (**B**) Users could characterize the population structure and pairwise kinship using genotype data. Genetic structure inferred by ancestral components and PCA can be used to validate sample information. (**C**) Users could dissect the genetic landscape of performance traits from genotype and phenotype data by GWA analysis. GEBVs and mating allocations could be predicted by GS and GM. The effects of GS and GM could be further compared by simulating multiple generations.

Users can upload one or more VCF files and carry out online imputation with corresponding reference panels by Beagle (version 5.1) and Glimpse (version 1.1.1) (19,20). The parameters for imputation and quality control can be adjusted accordingly to accommodate the datasets. Once the analysis is done, the global imputation rate and linkage disequilibrium (LD) patterns will be visually illustrated for each chromosome (Figure 1A). Users will get imputed genotypes in vcf.gz format. A PED and MAP file will also be generated to fit subsequent analysis.

Population characterizing

Understanding the genetic structure and population history of the breeds of interest plays an essential role in correcting of population stratification (21) and predicting of heterosis (22). It is also of vital importance in the practical management of farmed stocks (22). One commonly used method is through low-dimensional projections by PCA. An alternative method is to estimate genomic breed compositions via a likelihood-based admixture model (23). We implemented both algorithms in the 'Genetic Structure' function (24,25). Users can select the appropriate one for samples clustering. As samples of close relatives may reduce the explanatory power of ancestry inference in population characterizing. We incorporated a robust inference model via estimating kinship coefficients and IBD segments (26).

The input files of 'Genetic Structure' and 'Kingship' functions should be in the PLINK standard text format for pedigree and genotypes. For admixture and PCA analysis, the presumed cluster of samples is expected to be provided in a sample information file. In case of absence, the estimations of ancestry fractions are used for sample clustering. Users can set a *K* value slightly larger than the presumed number of ancestries. The optimal *K* is determined by cross-validation, which typically has the lowest level of error rate. After analysis, individual genomic compositions, population structures, and kinship distributions are graphically exhibited (Figure 1B). Detailed results, including CV error and pairwise relationships, can be retrieved from the 'Download' page.

Genetic breeding

We implemented the genome-wide efficient mixed-effects model in the 'GWAS' function, which is characterized by efficient computation of exact values of standard test statistics in linear mixed models (27). The algorithm fits a standard linear mixed model for association tests with the statistical significance examined by the Wald test. It calculates the centered relatedness matrix based on the genotypes and does not require prior knowledge about the genetic structure. Users can check the significance of genome-wide markers and deviation of the observed *P*-values from the null hypothesis by Manhattan and quantile-quantile plot, respectively (Figure 1C). The estimated effects and *P*-values for each SNP could be retrieved from the results.

The 'Genomic Selection' function integrates RR-GBLUP (6), Bayes Lasso (28), and sparse neural networks (SNN) (29). As the accuracy of each model largely depended on multiple factors, such as the trait heritability, the reference population size, and marker density, we implemented a ten-fold CV in the pipeline to help compare and select the appropriate model for each study. Details regarding the models for prediction and accuracy evaluation were provided in the supplementary files (Supplementary Method). The allele substitution effects of genome-wide markers and genomic estimated breeding values (GEBVs) for each candidate can be graphically illustrated after the analysis (Figure 1C). Inbreeding coefficients and sample phylogeny are also provided to users on the 'Download' page as additional files.

Although GS provides a revolutionary tool for genetic improvement, it may result in rapid depletion of genetic variability of the selected traits in future generations. In contrast, GM represents an alternative approach to maximizing genetic gain with a constrained inbreeding rate (8). The pre-calculated GEBVs could be stored in a sample information file and imported into the 'Genomic Mating' pipeline. Mating allocations and desired numbers of offspring would be indicated graphically to users. By leveraging the marker effects estimated from the reference population, one could compare the genetic improvements made by GS and GM across multiple simulated generations. This is particularly useful for making decisions on short-term genetic gain or long-term potential.

RESULTS AND DISCUSSION

Haplotype reference panel

We collected and analyzed the WGS data of 2,294 samples from 18 aquaculture species, including two crustaceans, two mollusks, and fourteen finfish (Table 1). A total of 144 million SNPs have been identified and included in the haplotype reference panels. The detailed information regarding each species, including the background introduction, genome size, number of chromosomes, and size of samples for SNP mining is provided on the page of 'Reference Panel'. To capture the polymorphism comprehensively, we exhaustively searched and collected the publicly available WGS data for each species. The accession numbers of datasets and their corresponding studies/projects are summarized at the bottom of the webpage. For species with datasets from multiple sources, the genome-wide distribution of SNPs, sample phylogeny, and genetic structure are analyzed and illustrated, respectively (Supplementary Figures S1-S15).

To evaluate the performance of imputation with the reference panels, the accuracy of imputed genotypes was assessed by a five-fold CV for the haplotype panels that were constructed with over 100 samples (Table 2). In each test round, samples of each species were divided into five groups. Four of the five groups were used for reference panel construction and the remaining one was down-sampled to ~1x coverage and imputed by Beagle and Glimpse, respectively. The accuracy was evaluated by concordance rate (CR) and squared correlation (R^2) of dosages between the imputed and true genotypes. Despite that Beagle is faster and requires relatively fewer resources, imputation with Glimpse is overall more accurate (Table 2). The average CRs of Glimpse exceed 0.924 for all the test species, indicating a good and stable performance for practical analysis.

A case study of characterizing the Pacific white shrimp population

Pacific white shrimp (*Litopenaeus vannamei*) is native to the eastern Pacific Ocean and has been introduced to a wide range of areas since the late 1970s. It has now become one of the top aquatic species of commercial importance around the world. The annual global yield of *L. vannamei* exceeded 4.4 million tonnes with a value of over 26.7 billion USD, accounting for 80% of the total cultured shrimp production (1). To explore the genetic signatures of domestication and artificial selection, we recently collected and sequenced *L. vannamei* broodstock from two artificially selective breeds: Renhai No. 1 (RH), Kehai No. 1 (KH); and four market-leading companies: Benchmark Genetics (BMK), Charoen Pokphand (CP), Shrimp Improvement Systems (SIS), and Top Aquaculture Technology (TA). The genotype data of 3.8 million loci from 150 samples were prepared in PED format; the chromosomal positions of markers were stored in MAP format; other meta-information, such as the source of samples, were deposited in TSV format. We imported these files into the AMBP for population characterizing by the 'Genetic Structure' and 'Kinship' functions.

As these samples were collected from six groups, we set the maximum number of *K* as 10 for the analysis. The cross-validation revealed the lowest error rate at $K = 7$, indicating that these samples may be derived from seven ancestral populations (Figure 2A, Supplementary Figure S16). The estimated structure was consistent with their recorded sources, exhibiting a near 1 to 1 correspondence. The RH breed was developed by crossbreeding two selective breeds from Miami and Oahu, which may explain the inferred within-breed stratification. We further checked the clustering in case of no records of sampling sources. In this scenario, the inferred ancestry fractions of each sample were used for clustering, which generated identical results with the optimal *K* value of seven (Figure 2B, Supplementary Figure S17). As a com-

Table 1. Data summary of the haplotype reference panels in AMBP

Species	Genomic Assembly	No. of Samples	No. of SNPs
Litopenaeus vannamei	ASM378908v1	180	3,926,527
Fenneropenaeus chinensis	ASM1920278v1	43	11,595,609
Argopecten irradians	QAU_Airr_1.1	40	11,325,844
Crassostrea gigas	cgigas_uk_roslin_v1	220	3,873,608
Cynoglossus semilaevis	Cse_v1.0	53	1,207,365
Paralichthys olivaceus	Flounder_ref_guided_V1.0	120	6,367,725
Larimichthys crocea	L_crocea_2.0	253	9,998,395
Lateolabrax maculatus	ASM402354v1	99	7,036,107
Dicentrarchus labrax	dlabrax2021	76	6,531,380
Oreochromis niloticus	O_niloticus_UMD_NMBU	166	5,164,107
Gadus morhua	gadMor3.0	220	5,936,877
Salmo salar	ICSASG_v2	281	5,873,967
Oncorhynchus kisutch	Okis_V2	60	7,332,974
Oncorhynchus mykiss	USDA_OmykA_1.1	179	17,012,766
Oncorhynchus gorbuscha	Ogor_1.0	62	6,544,976
Takifugu rubripes	TakRub1.2	61	1,888,317
Anguilla japonica	Ajp_01	84	20,550,847
Anguilla anguilla	fAngAng1	97	11,934,548

Table 2. The imputation accuracy using reference panels in AMBP

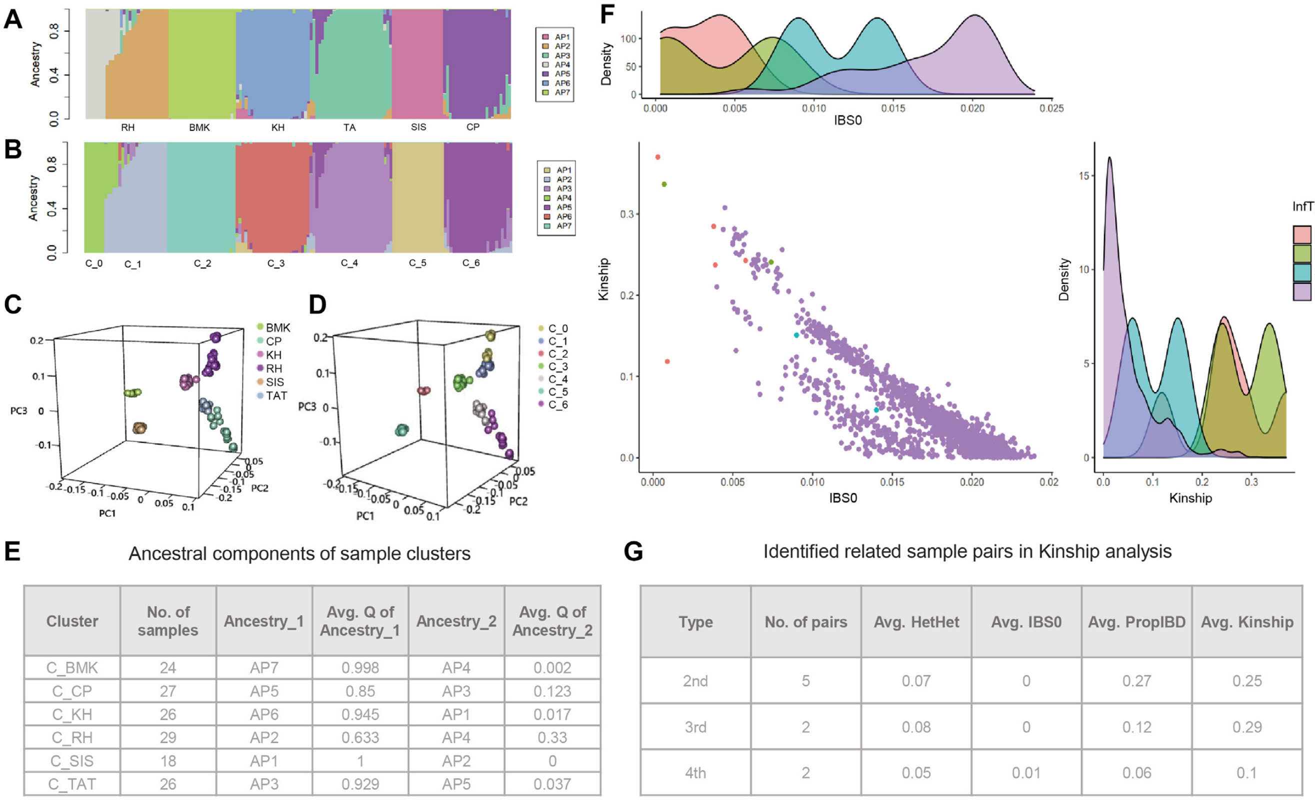
Species	Glimpse		Beagle	
	CR (mean ± SD)	R2 (mean ± SD)	CR (mean ± SD)	R2 (mean ± SD)
Litopenaeus vannamei	0.984 ± 0.008	0.938 ± 0.033	0.927 ± 0.014	0.937 ± 0.033
Crassostrea gigas	0.925 ± 0.016	0.789 ± 0.045	0.881 ± 0.005	0.732 ± 0.014
Paralichthys olivaceus	0.958 ± 0.028	0.905 ± 0.063	0.856 ± 0.029	0.741 ± 0.067
Larimichthys crocea	0.956 ± 0.030	0.854 ± 0.103	0.894 ± 0.013	0.714 ± 0.038
Oreochromis niloticus	0.924 ± 0.050	0.712 ± 0.124	0.875 ± 0.058	0.607 ± 0.099
Gadus morhua	0.959 ± 0.018	0.855 ± 0.064	0.907 ± 0.006	0.738 ± 0.022
Salmo salar	0.949 ± 0.022	0.882 ± 0.068	0.856 ± 0.015	0.738 ± 0.092
Oncorhynchus mykiss	0.966 ± 0.019	0.919 ± 0.053	0.889 ± 0.044	0.795 ± 0.084

parison, the PCA analysis revealed six clusters for all the samples and failed to detect the segregation in RH (Figure 2C, D, and E). Therefore, combining the results of PCA and admixture could provide us with a deeper understanding of the genetic structures. As a step further, we would like to know if the inferred population structure is biased by sampling. The 'Kinship' function was used to check the presence of close relatives and separate them from the unrelated pairs. As shown in the result, only nine pairs of samples shared a kinship higher than the 4th-degree and exhibited distinguished distributions of IBS0 and kingship coefficient compared with the vast majority (Figure 2E and G, Supplementary Figure S18). Thus, the revealed population differentiation can help users understand the process of shrimp domestication and selection.

A case study of genomic breeding using low-density marker panels in Atlantic Salmon

Atlantic salmon (*Salmo salar*) is naturally distributed in the temperate and subarctic regions of the North Atlantic Ocean. It is widely known for its importance in aquaculture and has been widely studied as a model organism for salmonid species. Atlantic salmon breeding programs are the most advanced of all aquaculture species. Since the first salmon high-density SNP arrays demonstrated their utility in accurately predicting breeding values, genomic information has been routinely incorporated in practical breeding (30). As there are typically many thousands of fish to test, controlling the price for genotyping is usually a key point in a breeding program. A strategy is to genotype a few selected founders at high density, while the rest are genotyped at low density, followed by imputation to acquire high-density genotypes (31). The strategy is promising but can only be applied to fish with close relationships. To overcome this limit, we constructed a high-density haplotype reference panel of *S. salar* and implemented it in the AMBP. As a demonstration imputation and genetic breeding analysis in AMBP, we retrieved a *S. salar* dataset of 524 individuals that were genotyped by a high-density panel with 78K SNPs (31).

About 10K SNPs were randomly sampled from the total 78K to represent a low-density panel. Genotypes for the 10K SNPs of all the samples were uploaded for 'Online Imputation'. The imputation was performed by Beagle with the default parameters. We set GP = 0.5 and Missing Rate = 0.2, to filter less accurate or non-informative SNPs. Finally, we kept a total of 118,039 SNPs for the subsequent analysis. Figures for the global LD pattern and missing rate were also generated to facilitate quality control (Figure 3A, Supplementary Figure S19). The 'GWAS' function was used to determine which individual SNP was associated with the trait of 'weight'. The GWA analysis revealed no significant SNPs according to the genome-wide threshold of suggestive, indicating a polygenic genetic architecture (Figure 3B, Supplementary Figure S20). One SNP on chromosome 17 has the smallest *P*-value and an estimated effect size of 0.049 (Figure 3C).



E Ancestral components of sample clusters

Cluster	No. of samples	Ancestry_1	Avg. Q of Ancestry_1	Ancestry_2	Avg. Q of Ancestry_2
C_BMK	24	AP7	0.998	AP4	0.002
C_CP	27	AP5	0.85	AP3	0.123
C_KH	26	AP6	0.945	AP1	0.017
C_RH	29	AP2	0.633	AP4	0.33
C_SIS	18	AP1	1	AP2	0
C_TAT	26	AP3	0.929	AP5	0.037

G Identified related sample pairs in Kinship analysis

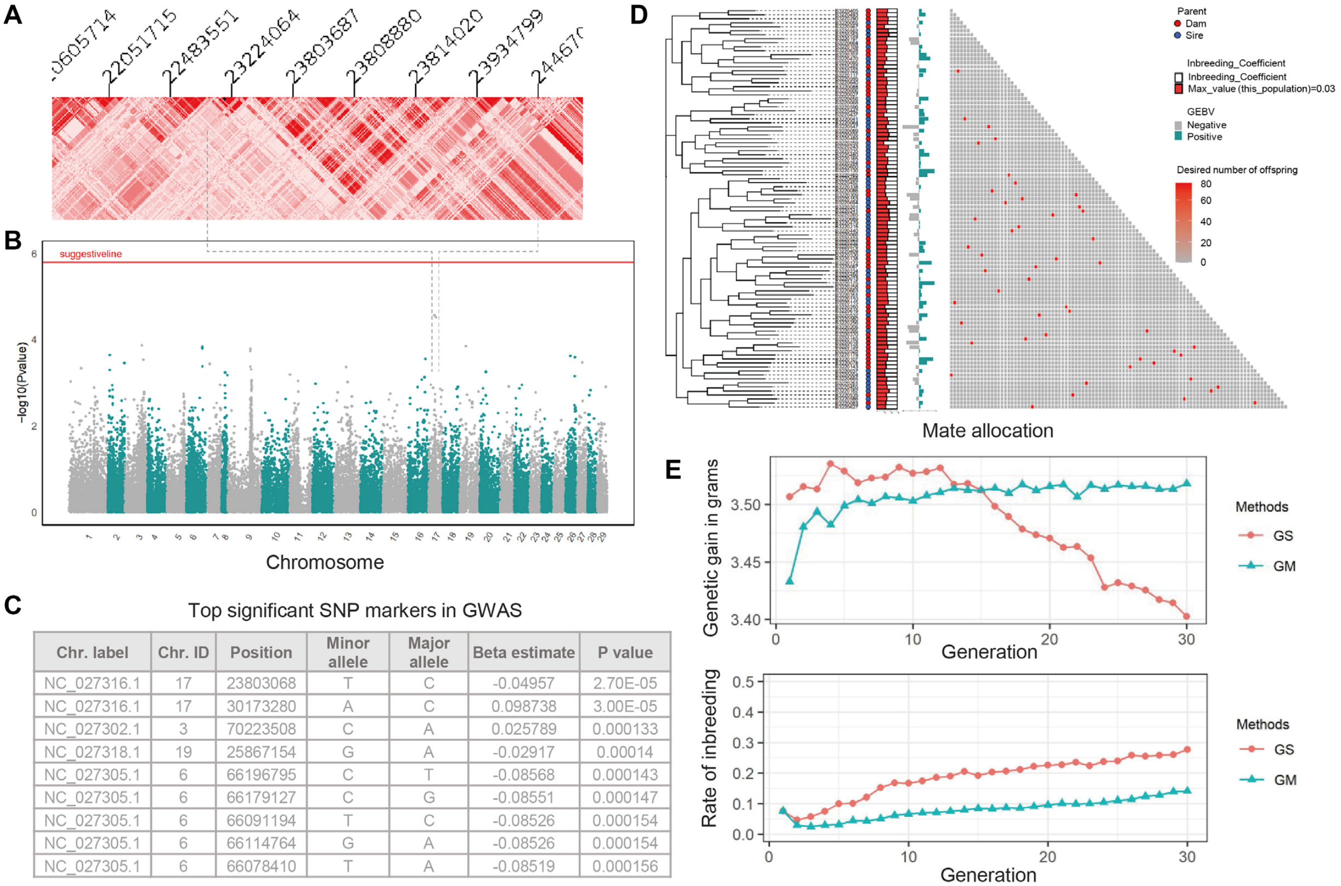
Type	No. of pairs	Avg. HetHet	Avg. IBS0	Avg. PropIBD	Avg. Kinship
2nd	5	0.07	0	0.27	0.25
3rd	2	0.08	0	0.12	0.29
4th	2	0.05	0.01	0.06	0.1

Figure 2. AMBP deciphered the genetic structure of shrimp populations. (**A**) Sample clustering with known sampling sources. AP: ancestral population. (**B**) Sample clustering based on genomic ancestral compositions in the absence of sampling information. PCA analysis with (**C**) and without (**D**) sampling information. (**E**) Ancestral components for each cluster, Ancestry_1: the first largest ancestral population in the cluster, Avg. Q of Ancestry_1: the average fraction of Ancestry_1 in the cluster, Ancestry_2: the second largest ancestral population in the cluster, Avg. Q of Ancestry_2: the average fraction of Ancestry_2 in the cluster. (**F**) Distribution of the pairwise kinship coefficient and IBS0. IBS0: proportion of genotypes with zero IBS. (**G**) Close relatives in the population. Avg. HetHet: the average Proportion of SNPs with double heterozygotes, Avg.IBS0: the average proportion of genotypes with zero IBS, Avg. PropIBD: the average Proportion of genomes shared identical-by-descent, Avg.Kinship: the average kinship coefficient.

The imputed genotypes were then used for genomic prediction via the function of 'Genomic Selection'. The accuracies of RR-GBLUP, Bayes Lasso, and Sparse Neural Network (SNN) revealed by ten-fold CV were about 50.83%, 37.85%, and 50.97%, respectively (Table 3). As a comparison, the prediction accuracies based on genotypes prior to imputation were only 44.68%, 35.46% and 38.46% for the three models. Our SNN model exhibited the largest improvement with the increase of marker density, suggesting its superior in dealing with complex traits that require large datasets. As the heritability of weight was estimated to be 0.61. The results also agreed with the finding of Bellot et al. that the predictive accuracy of linear models depended highly on heritability (32). RR-GBLUP always outperforms Bayes models in scenarios when the heritability of a complex trait is relatively high (>0.5). The 'Genomic Selection' also allows users to upload a candidate population for GEBV prediction. The predicted GEBV, sample phylogeny, and estimations of marker effects (for linear models) will also be graphically exhibited (Supplementary Figure S21).

Although GS is a revolutionary tool for genetic improvement, it can increase inbreeding and may produce an increased depletion of genetic variability in future generations. On the other hand, GM represents an approach to maximizing genetic gain with controlled inbreeding. In contrast to the GS, the selective intensity depended on the size of retained offspring rather than the parents. It could predict the mate allocation for selected candidates as well as their optimum sizes of offspring (Figure 3D, Supplementary Figure S22).

To compare the long-term effects of GS and GM, we conducted a simulation analysis based on the salmon data. Briefly, SNP effects were first estimated using RR-GBLUP. The GEBV was taken to be the sum of estimated SNP effects. Progeny genotypes were generated by stochastic simulation. Recombination events were modeled as the number of crossovers per Morgan, following a Poisson distribution with the mean equaling one. A mutation rate of 1e-6 per nucleotide was assumed for all chromosomes. The top 1,000 SNPs with the largest effects were selected as QTLs and the inbreeding coefficients were computed as runs of homozygosity (ROH) (24). A total of 400 offspring were retained in each generation for both the GS and GM. For GS, the offspring were randomly generated from the top 90 males and 90 females with the highest GEBV. The simulation was performed for 30 consecutive generations. As shown in Figure 3E, despite that the genetic gain of GS was higher for the first 15 generations, it kept decreasing since the 12th generation and eventually became lower than GM after the 15th



Top significant SNP markers in GWAS

Chr. label	Chr. ID	Position	Minor allele	Major allele	Beta estimate	P value
NC_027316.1	17	23803068	T	C	-0.04957	2.70E-05
NC_027316.1	17	30173280	A	C	0.098738	3.00E-05
NC_027302.1	3	70223508	C	A	0.025789	0.000133
NC_027318.1	19	25867154	G	A	-0.02917	0.00014
NC_027305.1	6	66196795	C	T	-0.08568	0.000143
NC_027305.1	6	66179127	C	G	-0.08551	0.000147
NC_027305.1	6	66091194	T	C	-0.08526	0.000154
NC_027305.1	6	66114764	G	A	-0.08526	0.000154
NC_027305.1	6	66078410	T	A	-0.08519	0.000156

Figure 3. AMBP dissect the genetic mechanism of performance trait and guide salmon breeding. (**A**) Global LD pattern generated by 'Online imputation', numbers indicated the SNP coordinate on the chromosomes. (**B**) Manhattan plot of the marker statistical significance across the whole genome. (**C**) Top significant SNPs revealed by GWA analysis. (**D**) Matting allocation and optimum numbers of offspring for each pair. (**E**) Simulated effects of GS and GM in 30 consecutive generations.

Table 3. Prediction accuracy for dataset prior to and after imputation

Fold	Prediction accuracy (%) of original genotypes			Prediction accuracy (%) after imputation			Improvement (%)		
	RR-GBLUP	Bayes Lasso	SNN	RR-GBLUP	Bayes Lasso	SNN	RR-GBLUP	Bayes Lasso	SNN
1	56.80	48.18	28.69	54.81	52.03	57.17	-1.98	3.85	28.48
2	17.19	23.28	10.21	15.47	21.78	15.81	-1.72	-1.49	5.59
3	61.21	44.77	34.39	68.23	30.72	70.11	7.02	-3.72	35.72
4	54.13	44.77	61.30	62.67	37.32	61.78	8.53	-7.45	0.47
5	22.77	9.97	30.38	39.52	19.36	43.05	16.74	9.38	12.67
6	56.01	52.07	52.35	68.68	58.43	68.48	12.67	6.36	16.12
7	40.19	30.19	45.59	48.83	42.16	48.26	8.64	11.97	2.66
8	63.76	53.62	51.03	65.05	63.61	65.08	1.28	9.97	14.05
9	52.01	46.92	43.66	58.95	31.95	54.05	6.95	-14.96	10.38
10	22.70	11.18	26.98	26.09	21.10	25.86	3.39	9.92	-1.11
Avg ACC	44.68	35.46	38.46	50.83	37.85	50.97	6.15	2.38	12.51
Sd ACC	17.62	16.33	15.12	18.37	15.89	18.17	6.03	8.96	11.96

generation. In the scenario of GM, the rate of genetic gain kept increasing and maintained at a relatively stable level after the first 5 generations. Furthermore, GM had a constrained increasing rate of inbreeding coefficient that surpassed 0.1 after the 20th generation. As a comparison, the inbreeding coefficient of GS exceeded 0.1 at the 5th generation, which explained the drastically reduced genetic gain. Therefore, GM exhibited superior potential in genetic improvement, especially for long-term breeding programs.

Comparisons with existing webservers of similar functions

We compared AMBP with the existing webservers of similar functions according to the practicability for imputation, population characterizing, and genetic breeding (Table 4). The functional innovations of AMBP are summarized in the following three aspects: (1) AMBP uniquely provides high-density haplotype reference panels for 18 aquaculture species and supports multiple tools to infer genetic struc-

Table 4. Functional innovations of AMBP compared with other webservers

Functions	AMBP	Michigan Imputation Server (33)	Animal-ImputeDB (34)	Plant-ImputeDB (35)	StructuRly (36)	SNiPlay (37)	CASSAVABASE (38)	easyGWAS (39)
Imputation								
Reference panels for aquaculture species	✓							
Imputation	✓	✓	✓	✓				
Illustration of LD pattern	✓							
Quality control	✓							
Population characterizing								
Ancestry estimation	✓				✓	✓	✓	
PCA analysis	✓				✓	✓		
Kinship inference	✓					✓	✓	
Genetic breeding								
GWAS analysis	✓	✓				✓	✓	✓
Genomic prediction	✓						✓	
Neural network model	✓							
Cross validation	✓							
Genomic mating	✓							
Simulation analysis	✓							
Total numbers	**13**	**2**	**1**	**1**	**2**	**4**	**4**	**1**

tures, dissect the genetic architecture of performance traits, estimate breeding values, and predict optimum contribution. (2) In the breeding programs, it is crucial to balance two conflicting objectives: the selection for genetic gain and the maintenance of genetic diversity. Finding the optimal strategy has always been a challenge for breeders. AMBP provides GS to acquire maximum genetic improvement in short generations, and also incorporates GM as an alternative strategy for retaining long-term potential. Users can evaluate their performance through cross-validation and simulation modules, which enables the comparison of statistical appropriateness under different circumstances. (3) Furthermore, AMBP joints each tool coherently and generates paper-ready figures for a better interpretation and exploration of data. As an outlook to the future, we will keep improving the service by including more resources and new functionalities on aquaculture genetics.

DATA AVAILABILITY

AMBP pipeline is freely available as both the webserver and standalone versions. The webserver can be accessed via the following link: http://mgb.qnlm.ac. The standalone version can be downloaded from Docker hub with the following address: https://hub.docker.com/r/ouc2021mgb/ambp. The datasets used for constructing the haplotype reference panels were listed in Supplementary Table S1. The shrimp and salmon data for demonstrations were saved as example datasets at AMBP and deposited at FigShare with the fowling link: https://figshare.com/articles/dataset/AMBP_case_study/19390652.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank graduate students Mingyang Zhao, Chao Qin, Qianqian Zhao, Jingyu Song, Lu Lin, Lingyun Tang, Hongyu Lv, and Shengtao Gao for their assistance in data processing.

FUNDING

This work was supported by Project of Sanya Yazhouwan Science and Technology City Management Foundation [SKJC-KJ-2019KY01], National Key Research and Development Program of China [2021YFD1200805], National Natural Science Foundation of China [32002388 and 32072976], Key R&D Project of Hainan Province [ZDYF2021XDNY133], Key R&D Project of Shandong Province [2021ZLGX03], Major Science and Technology Program of Hainan Province [ZDKJ2021017]. Science Foundation of Ministry of Education of China [2021ZL09]

Conflict of interest statement. We declare that all authors have no conflicts of interest in this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

REFERENCES

1. Food and Agricultural Organization (2020) In: *The state of world fisheries and aquaculture*.
2. Costello,C., Cao,L., Gelcich,S., Cisneros-Mata,M.A., Free,C.M., Froehlich,H.E., Golden,C.D., Ishimura,G., Maier,J., Macadam-Somer,I. *et al.* (2020) The future of food from the sea. *Nature*., **588**, 95–100.
3. Longo,S.B., Clark,B., York,R. and Jorgenson,A.K. (2019) Aquaculture and the displacement of fisheries captures. *Conserv. Biol.*, **33**, 832–841.
4. Froehlich,H.E., Runge,C.A., Gentry,R.R., Gaines,S.D. and Halpern,B.S. (2018) Comparative terrestrial feed and land use of an aquaculture-dominant world. *P Natl Acad Sci USA.*,**115**, 5295–5300.
5. Norris,A. (2017) Application of genomics in salmon aquaculture breeding programs by ashie norris: who knows where the genomic revolution will lead us?*Mar Genomics*., **36**, 13–15.
6. Meuwissen,T.H., Hayes,B.J. and Goddard,M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*., **157**, 1819–1829.
7. Wiggans,G.R., Cole,J.B., Hubbard,S.M. and Sonstegard,T.S. (2017) Genomic selection in dairy cattle: the USDA experience. *Annu Rev Anim Biosci*., **5**, 309–327.

8. Akdemir,D. and Sanchez,J.I. (2016) Efficient breeding by genomic mating. *Front Genet.*, **7**, 210.
9. Li,Y., Willer,C., Sanna,S. and Abecasis,G. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.
10. Zan,Y., Payen,T., Lillie,M., Honaker,C.F., Siegel,P.B. and Carlborg,O. (2019) Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: a cost-efficient approach. *Genet. Sel. Evol.*, **51**, 44.
11. Alex Buerkle,C. and Gompert,Z. (2013) Population genomics based on low coverage sequencing: how low should we go?*Mol. Ecol.*, **22**, 3028–3035.
12. Johnson,E.O., Hancock,D.B., Levy,J.L., Gaddis,N.C., Saccone,N.L., Bierut,L.J. and Page,G.P. (2013) Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum. Genet.*, **132**, 509–522.
13. Yoshida,G.M., Carvalheiro,R., Lhorente,J.P., Correa,K., Figueroa,R., Houston,R.D. and Yáñez,J.M. (2018). Accuracy of genotype imputation and genomic predictions in a two-generation farmed atlantic salmon population using high-density and low-density SNP panels. *Aquaculture*, **491**, 147–154.
14. Zhang,W., Li,W., Liu,G., Gu,L., Ye,K., Zhang,Y., Li,W., Jiang,D., Wang,Z. and Fang,M. (2021) Evaluation for the effect of low-coverage sequencing on genomic selection in large yellow croaker. *Aquaculture*, **534**, 736323.
15. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997, 26 May 2013, preprint: not peer reviewed.
16. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
17. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience.*, **10**, giab008.
18. Delaneau,O., Zagury,J.F. and Marchini,J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods.*, **10**, 5–6.
19. Browning,B.L., Zhou,Y. and Browning,S.R. (2018) A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, **103**, 338–348.
20. Rubinacci,S., Ribeiro,D.M., Hofmeister,R.J. and Delaneau,O. (2021) Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.*, **53**, 120–126.
21. Mebratie,W., Reyer,H., Wimmers,K., Bovenhuis,H. and Jensen,J. (2019) Genome wide association study of body weight and feed efficiency traits in a commercial broiler chicken population, a re-visitation. *Sci. Rep.*, **9**, 922.
22. Akanno,E.C., Abo-Ismail,M.K., Chen,L., Crowley,J.J., Wang,Z., Li,C., Basarab,J.A., MacNeil,M.D. and Plastow,G.S. (2018) Modeling heterotic effects in beef cattle using genome-wide SNP-marker genotypes. *J. Anim. Sci.*, **96**, 830–845.
23. He,J., Guo,Y., Xu,J., Li,H., Fuller,A., Tait,R.G. Jr, Wu,X.L. and Bauck,S. (2018) Comparing SNP panels and statistical methods for estimating genomic breed composition of individual animals in ten cattle breeds. *BMC Genet.*, **19**, 56.
24. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
25. Alexander,D.H., Novembre,J. and Lange,K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
26. Manichaikul,A., Mychaleckyj,J.C., Rich,S.S., Daly,K., Sale,M. and Chen,W.M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics.*, **26**, 2867–2873.
27. Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.
28. de Los Campos,G., Hickey,J.M., Pong-Wong,R., Daetwyler,H.D. and Calus,M.P. (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics.*, **193**, 327–345.
29. Wang,Y., Mi,X., Rosa,G.J.M., Chen,Z., Lin,P., Wang,S. and Bao,Z. (2018) Technical note: an r package for fitting sparse neural networks with application in animal breeding. *J. Anim. Sci.*, **96**, 2016–2026.
30. Houston,R.D., Taggart,J.B., Cezard,T., Bekaert,M., Lowe,N.R., Downing,A., Talbot,R., Bishop,S.C., Archibald,A.L., Bron,J.E. *et al.* (2014) Development and validation of a high density SNP genotyping array for atlantic salmon (Salmo salar). *BMC Genomics.*, **15**, 90.
31. Tsai,H.Y., Matika,O., Edwards,S.M., Antolin-Sanchez,R., Hamilton,A., Guy,D.R., Tinch,A.E., Gharbi,K., Stear,M.J., Taggart,J.B. *et al.* (2017) Genotype imputation to improve the cost-efficiency of genomic selection in farmed atlantic salmon. *G3.*, **7**, 1377–1383.
32. Bellot,P., de Los Campos,G. and Perez-Enciso,M. (2018) Can deep learning improve genomic prediction of complex human traits?*Genetics.*, **210**, 809–819.
33. Das,S., Forer,L., Schonherr,S., Sidore,C., Locke,A.E., Kwong,A., Vrieze,S.I., Chew,E.Y., Levy,S., McGue,M. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
34. Yang,W., Yang,Y., Zhao,C., Yang,K., Wang,D., Yang,J., Niu,X. and Gong,J. (2020) Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic. Acids. Res.*, **48**, D659–D667.
35. Gao,Y., Yang,Z., Yang,W., Yang,Y., Gong,J., Yang,Q.Y. and Niu,X. (2021) Plant-ImputeDB: an integrated multiple plant reference panel database for genotype imputation. *Nucleic. Acids. Res.*, **49**, D1480–D1488.
36. Criscuolo,N.G. and Angelini,C. (2020) StructuRly: a novel shiny app to produce comprehensive, detailed and interactive plots for population genetic analysis. *PLoS One.*, **15**, e0229330.
37. Dereeper,A., Homa,F., Andres,G., Sempere,G., Sarah,G., Hueber,Y., Dufayard,J.F. and Ruiz,M. (2015) SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. *Nucleic. Acids. Res.*, **43**, W295–W300.
38. Fernandez-Pozo,N., Menda,N., Edwards,J.D., Saha,S., Tecle,I.Y., Strickler,S.R., Bombarely,A., Fisher-York,T., Pujar,A., Foerster,H. *et al.* (2015) The sol genomics network (SGN)–from genotype to phenotype to breeding. *Nucleic. Acids. Res.*, **43**, D1036–D1041.
39. Grimm,D.G., Roqueiro,D., Salome,P.A., Kleeberger,S., Greshake,B., Zhu,W., Liu,C., Lippert,C., Stegle,O., Scholkopf,B. *et al.* (2017) easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell.*, **29**, 5–19.