Development of a 38K Single Nucleotide Polymorphism Array and Application in Henomic Selection for Resistance Against Vibrio harveyi in Chinese Tongue sole, Cynoglossus Semilaevis

Sheng Lu
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Qian Zhou
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Yadong Chen
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Yang Liu
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Yangzhen Li
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Lei Wang
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Yingming Yang
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
Songlin Chen (✉ chensl@ysfri.ac.cn)
Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute https://orcid.org/0000-0002-7007-0640

Research

Keywords: Genomic selection, SNP array, Cynoglossus semilaevis, Vibrio harveyi, Disease resistance

DOI: https://doi.org/10.21203/rs.3.rs-48867/v1

Abstract

Background: In recent years, the disease outbreak caused by *Vibrio harveyi* upset the booming development of the Chinese tongue sole (*Cynoglossus semilaevis*) farming industry. Genomic selection (GS) is a powerful method to improve the traits of interest, which has been proved in livestock and some fishes. Besides, the single nucleotide polymorphism (SNP) array is an efficient genotyping platform that can be used for genetic studies. To improve *V. harveyi* resistance in *C. semilaevis*, we firstly constructed a reference group of 1,572 individuals and investigated accuracies of four genomic methods (genomic best linear unbiased prediction (GBLUP), weighted GBLUP, BayesB, and BayesC) at predicting the genomic estimated breeding value (GEBV) using five-fold cross-validation and SNPs varying from 0.5 k to 500 k. Then, an SNP array was developed using the Affymetrix Axiom technology, and its accuracy in genotyping was evaluated by comparing SNPs generated by the array and by the re-sequencing technology. Finally, we selected 44 candidates as the parents of 23 families of *C. semilaevis* to evaluate the feasibility of the SNP array for GS.

Results: all genomic methods outperformed the pedigree-based BLUP (ABLUP) when at least 50 k SNPs used for prediction, of which GBLUP resulted in better estimation than ABLUP when more than 1 k SNPs used. A 38 k SNP array, "Solechip No.1", was developed with an average of 10.5 kb inter-spacing between two adjacent SNPs. The SNPs generated by the array and by the re-sequencing reached an average consistency of 94.8 %, of which 79.3 % of loci had a more than 90 % of the consistency. The survival rates of these 23 offspring families had a correlation of 0.706 with the family GEBVs (mid-parental GEBVs), and the average survival rate of the top five families in GEBVs (79.1 %) is higher than the bottom five families (58.1 %).

Conclusion: GS is an efficient method to improve the *V. harveyi* resistance in *C. semilaevis*, and the SNP array "Solechip No.1" is a convenient and reliable tool for the Chinses tongue sole selective breeding practice.

Background

Chinese tongue sole (*Cynoglossus semilaevis*) is an economic marine flatfish which was cultured widely across the coastal area of China and known for its delicious taste, high nutrition, and easy domestication. During the past decades, the booming development of large-scale and industrialized culturing of *C. semilaevis* has provided high-quality fish foods for consumers and brought significant economic profits to the aquaculture practitioners. However, in recent years, the farming practice of *C. semilaevis* has encountered great challenges, such as frequent disease occurrence, reduction in fertility, and germplasm degeneration, which have resulted in huge economic losses. *Vibrio harveyi* (Gram-negative) is one of the most epidemic and high lethal pathogens in the *C. semilaevis* farming industry. Usually, the infected fishes appeared some typical symptoms, including dermal ulceration, eye lesions, and tail bleeding. Since many studies have proven that selective breeding is an efficient method to improve the traits of interest [1], disease-resistance breeding became an important concern for aquatic breeders.

In aquaculture, the artificial challenge test using a specific pathogeny is a common strategy to investigate individuals' ability of disease resistance. Since the disease resistance is the trait that could not be measured directly as well as the surviving fishes from the test were not suitable to be parents, selection based on the survival rate or estimation based on pedigree (ABLUP) is not precise enough for identifying individuals with great breeding potential. In 2001, Meuwissen et al. proposed a methodology that utilizes SNP markers to predict the genetic breeding values for increasing the genetic gains substantially, i.e. genomic selection (GS) [2]. After Schaeffer et al. exhibited the benefits of GS in dairy cattle breeding with simulated data [3], research and application of GS has been rapidly developed in livestock and poultry [4, 5]. Recent years, studies about genomic prediction have been reported in some farmed fishes, such as Atlantic salmon (*Salmo salar*) [6–9], rainbow trout (*Oncorhynchus mykiss*) [10–12], common carp (*Cyprinus carpio*) [13], large yellow croaker (*Larimichthys crocea*) [14], Japanese flounder (*Paralichthys olivaceus*) [15], and Nile tilapia (*Oreochromis niloticus*) [16], etc., and all genomic methods revealed more accurate than the ABLUP at selecting individuals with excellent breeding potential. However, studies on GS about resistance to *V. harveyi* in fish have not been reported yet.

With the rapid development and dramatic reduction of costs on next-generation sequencing technology, the genome sequencing, re-sequencing, and genome-wide marker discovery have become accessible to aquatic species. Except for the sequencing technology, the genotyping array also provides an efficient and convenient way to detect hundreds of thousands of single nucleotide polymorphism (SNP) simultaneously, which is a benefit for the genetic studies, such as population structure identification, genome-wide association study (GWAS), quantitative trait loci (QTL) mapping, and GS, etc. In aquaculture, the SNP array has been developed for several artificially cultured fishes, such as Atlantic salmon, rainbow trout, and catfish. For Atlantic salmon, a 50 k SNP array and two high-density (132 k and 200 k) arrays have been developed and applied to genetics studies [17, 18]. For catfish, a 250 k SNP array was developed and exhibited good application in population and genetic analysis [19]. For rainbow trout, a 57 k SNP array was also developed [20].

Based on the data of challenge test fitted in four models, Li et al. reported that *V. harveyi* resistance in *C. semilaevis* belongs to a low heritability trait (ranging from 0.11 to 0.28), and there are moderate positive genetic correlations (0.27 to 0.51) between the resistance and growth performance (body weight and body length) [21]. Benefit from the availability of the whole-genome sequence of *C. semilaevis* [22], studies on the genetics of complex traits and high-efficient SNP array could be developed in Chinese tongue sole. Zhou et al. identified several genes, such as *plekha7*, *nucb2*, and *fgfr2*, etc., that might be potentially related to resistance against *V. harveyi* in Chinese tongue sole using GWAS with 505 re-sequenced individuals [23]. However, no GS study about *C. semilaevis* has been reported, especially for the *V. harveyi* resistance. Besides, no SNP array is available for the disease-resistance selective breeding in the Chinese tongue sole yet. The main purposes of this study are to 1) investigate the accuracies of four genomic methods on predicting the genomic estimated breeding value (GEBV); 2) develop an SNP array for *C. semilaevis* breeding practice; 3) evaluate the selection efficiency of GS using the SNP array.

Methods

Fish materials and trait definition

From 2014 to 2018, we established tens of families of *C. semilaevis* each year for the disease resistance and growth performance breeding. After mating, each family was reared separately and provided culture conditions as identical as possible before tagging. Details about family establishing was followed the describes of Chen et al. [24].

When fishes grew up to approximately 10 cm, 80 to 100 juveniles were selected randomly from each family for challenging with *V. harveyi* via intraperitoneal injection. The concentration of *V. harveyi* for formal challenge test was determined through a pre-experiment. In 2014, the concentration for formal test was 2.5×10^7 CFU/mL [21]. Followed the methods described by Chen et al. and Li et al. [21, 24], the concentration of *V. harveyi* in 2016 and 2018 was 8.3×10^5 CFU/mL and 8.0×10^4 CFU/mL, respectively. Each individual received a dose of 0.1 mL per 10 g body weight. After injection, all fishes from the same family were moved to a 0.7 to 1.0 m^3 separate tank and reared in filtered running water. Every 8 h observed the status of subjects, and the test terminated when no more dead juveniles appeared. The dead fish was moved from the tank at each observation and recorded relevant details, such as family information, body weight and body length. Besides, the tail fin was also sampled and preserved in absolute ethanol.

In this study, survival trait was defined as a binary trait, i.e. scoring 0 when the subject died in the test, scoring 1 for the alive. Based on the results of challenge test, individuals from the families established in 2014, 2016, and 2018 were selected as the reference group for GS study.

Whole genome re-sequencing and SNP calling

Genomic DNA of these fishes was isolated and purified from the tail fin using the TIANamp Marine Animals DNA Kit (Tiangen Biotech, Beijing, China). Illumina pair-ended (PE) libraries were constructed for each fish with an insert size of approximately 300 bp according to the manufacturer's protocol (Illumina). The PE libraries were then sequenced using Illumina HiSeq 2000 platform, generating 2*100 bp paired reads. The raw sequencing reads were subjected to quality control steps using QC-Chain and the tag sequences, low quality bases/reads, and ambiguous nucleotides were removed [25]. Then the clean reads were aligned to the reference genome of *C. semilaevis* (NCBI accession No. AGRG00000000.1) using BWA with default settings, generating the alignment results in BAM format. Prior to SNP identification, the BAM files were processed with Picard tools (version 1.119) to decrease false positive SNPs from duplicated genome regions. The SNP calling was then performed by SAMtools [26], with multiple filtering standards, including the mapping quality ≥ 20, the base quality ≥ 30 and the SNP quality score ≥ 20.

Quality control and SNP selection

To obtain reliable variants for the SNP array and GS study, the SNP selection was performed with several quality control (QC) steps, which decreased the false positives and ensured an even distribution of the SNPs across the genome. Firstly, raw SNPs were filtered based on the minor allele frequency (MAF), missing rate, and the Hardy-Weinberg equilibrium (HWE). SNPs with MAF less than 0.01, missing rate more than 0.1, and deviation from HWE (p = 0.001) were discarded using PLINK (version 1.90b6.6) [27]. For the GS analysis, ten subsets of genotypes with the number of SNP varying from 0.5 k to 500 k were extracted from these filtered SNPs.

Secondly, the preliminary selected SNPs were further filtered for the SNP array if (1) they were located in repetitive sequences; (2) there were other SNPs occurring in each side of 35 bp flanking sequences; (3) Fewer than two copies of each allele were detected in the sequence reads; (4) GC content of their 35 bp flanking sequences were more than 70% or less than 30%. In addition, the pairwise linkage disequilibrium was estimated to avoid the SNPs that fell within the same haplotype were simultaneously selected.

Finally, the remainder SNPs were submitted to Affymetrix® *in-silico* analytical pipeline to assess the potential of probe design. The p-converted value represents the probability of conversion to a reliable SNP assay. Based on that, potential probes were designed for each SNP in both the forward and reverse direction, and were categorized as "recommended", "neutral", "not recommended" or "not possible". Only probes classified into "recommended" or "neutral" were selected. The position and spacing of the selected SNPs were assessed to ensure an entire coverage in 20 autosomes. In addition, to filter the background signal from the assay, totally 2000 dish quality control (DQC) probes were designed and anchored on the SNP array as negative controls. The effects of SNPs selected for the SNP array were predicted by the software SnpEff (version 4.3t) [28]. The MAF distribution of all SNPs tilling on the SNP array was plotted using R-ggplot2 [29].

Statistical methods

Four genomic methods, including genomic best linear unbiased prediction (GBLUP), weighted GBLUP (wGBLUP), BayesB and BayesC, were used to predict the GEBVs of *V. harveyi* resistance in *C. semilaevis*. The threshold model (probit) used for these GS methods was defined as:

$$\varvec{y} = \varvec{X}\varvec{b} + \varvec{W}\varvec{g} + \varvec{e}$$

where \varvec$y$ is the vector of phenotypes, the element of 0 and 1 represent the fish died and survived from the challenge test, respectively; vector \varvec$b$ contains fixed effects, including intercept, locations and years of families establishing; vector \varvec$g$ contains random additive genomic effects distributed as $N(0,$ \varvec$Gs_g^2)$, of which \varvec$G$ is the genomic relationship matrix that was constructed using

\raisebox$1ex$\$(\varvec$M -$ \varvec$P$)\varvec$D$(\varvec$M -$ \varvec$P)'$\$ $\Big/$ \raisebox$-1ex$\$$2 ? p_k(1-p_k)$\$

[30]. In brief, the element of matrix \varvec$M$ (M_{jk}) represents the genotype of fish *j* at locus *k*, and the genotypes AA/Aa/aa were recoded as 0/1/2; the element of matrix \varvec$P$ is defined as $2p_k$, and p_k represents the observed second allele frequency at locus *k*. In addition, formula ?\varvec$G + (1 - ?)$\varvec$A_{22}$ was used for adjusting matrix \varvec$G$ to avoid the potential of singularity, of which ω equals 0.95, subscript 2 expresses the genotyped individual, and matrix \varvec$A$ is the relationship matrix based of pedigree (four generations). Matrix \varvec$X$ and \varvec$Z$ are incidence matrices that relate phenotypes to fixed effects and individuals, respectively. The difference between GBLUP and wGBLUP is whether matrix \varvec$G$ was weighted through matrix \varvec$D$. For GBLUP, matrix \varvec$D$ is the identity matrix. For wGBLUP, matrix \varvec$G$ was weighted through an iterative approach which was described by Wang et al. [31]. In brief, the weight of each SNP (d_k) was estimated from the SNP effect and frequency of the minor allele of SNP. At the first iteration, each SNP had same weights which was 1, i.e. matrix \varvec$D$ is the identity matrix. The SNP effects (\varvec$\hat{u}$) were calculated using

$2 ? p_k(1-p_k)$\varvec$D$(\varvec$M -$ \varvec$P)'$\varvec$G^{\text{\varvec} *}$\varvec$\hat{g}$, of which \varvec$\hat{g}$ is the GEBVs from GBLUP. Then, element of matrix \varvec$D$ (d_k) for weighting matrix \varvec$G$ in next iterations was estimated as described by Zhang et al.: \varvec$\hat{u}_k^{\,2} 2 ? p_k(1-p_k)$ [32]. In this study, results of wGBLUP were derived from the second iteration, and GEBVs were re-estimated in each iteration. Genetic variances and GEBVs of GBLUP and wGBLUP were estimated by R-ASReml 3 [33].

In two Bayesian methods, the GEBV of individual *j* (\varvec$\hat{g}_j$) was expressed as $?_{k=1}^{m} M_{jk} u_k$, where *m* is the number of SNP; M_{jk} is the genotypes of individual *j* at locus *k*, u_k is the SNP genetic effects. For BayesB, π is the proportion of SNP that had genetic effects. We tested π equals 0.01, 0.02, 0.03, 0.04, and 0.05 via five-fold cross validation to determine the value of π (results not shown here). In this study, π = 0.02 was selected for the final GS analysis. The genetic effects of SNPs in BayesB and BayesC were estimated using R-BGLR through the MCMC Gibbs sampling with 30,000 iterations and the first 10,000 runs being as burn-in (Pérez and de los Campos, 2014).

Cross validation

In this study, the predictive accuracies of four genomic methods with the number of SNP varying from 0.5 k to 500 k on predicting GEBVs were investigated by five-fold cross validation. In brief, the full data set was randomly divided into five non-overlapping subsets. In a single independent estimation, one subset was used as the validation set that the phenotypes of this subset were treated as missing, the rest of the subsets were used for model training. The above steps repeat five times, i.e. each non-overlapping subset was used as the validation set once. The predictive accuracy was defined as $Acc =$ \raisebox$1ex$\$$r_{(GEBV,y)}$\$/\raisebox$-1ex$\$$h$\$ [34], where $r_{(GEBV,y)}$ is the Pearson's correlation between GEBVs and phenotypes of validation set; $h$ is the square root of the heritability of *V. harveyi* resistance in *C. semilaevis*. In this study, heritability used here equals 0.16, which was reported by Li et al. [21]. The final accuracy is the average accuracy of five validation sets.

Evaluation of the SNP array performance

The DNA hybridization, staining and array scanning were completed using GeneTitan Multi-Channel Instrument (Thermo Fisher, USA). The Axiom Analysis Suite software was used to control the quality of samples with parameters: DQC ≥ 0.82; call rate ≥ 97%; percent of passing samples ≥ 95%; and average call rate for passing samples ≥ 98.5%. In addition, the default settings of SNP QC criteria were used to filter the SNPs. After the computer automatically analysis, SNPs could be divided into six classifications, including "NoMinorHom" (SNPs formed two good clusters but none of the minor homozygous genotypes were detected), "MonoHighResolution" (SNPs were detected as the monomorphic SNPs with good cluster resolution), "PolyHighResolution" (SNPs were detected as the polymorphic SNPs with three high-resolution clusters), "Other" (SNPs were identified as the low-quality genotypes), "CallRateBelowThreshold" (SNPs that were below the threshold of SNP call rate), and " Off Target Variant" (OTV, SNPs that few hybridization signals were captured). Therefore, SNPs with "PolyHighResolution", "MonoHighResolution", and "NoMinorHom" classification were recommended for downstream analysis.

To further evaluate the performance of the SNP array, the genotyping accuracy was also evaluated through comparing the SNPs generated by the SNP array and by the re-sequencing technology. For that purpose, 24 individuals of Chinese tongue sole were selected randomly from the reference group.

Application of the SNP array in breeding

Based on the reference group described above and the results of comparison of GS methods, 44 candidates were selected as the parents of 23 Chinese tongue sole families and were genotyped by the SNP array. The GEBVs of these individuals were estimated through the highest accuracy genomic method which was investigated previously. The family GEBV was expressed as the mid-parental GEBV. The Pearson's correlation between the family GEBVs and survival rates of the corresponding families was estimated to evaluate the selection efficiency of GS.

Results

Challenge test

The description of challenge test of families established in 2014 was reported by Li et al. [21]. In 2016, we challenged 3,714 fishes from 23 families with *V. harveyi* through intraperitoneal injection. The mortality is ranging from 10.3% to 100.0% with the mean, median, and standard error (SE) is 67.6%, 64.5%, and 23.7%, respectively. In 2018, the range of mortality of 23 families established at TS is from 11.6% to 59.7%, of which the average, median, and SE is 35.6%, 34.8%, and 13.8%. For families established at LZ, we selected 57 of 75 families for the test, the range of mortality is from 5.2% to 100.0% with the mean, median, and SE is 76.9%, 86.3%, and 24.7%.

Genomic selection with re-sequencing data

Based on the results of the challenge test conducted in 2014, 2016, and 2018, we selected 1,572 Chinese tongue sole (863, 182, and 527 fishes from 2014, 2016, and 2018) as the reference group and genotyped them through the whole-genome re-sequencing with an average sequencing depth of 6.2×. We identified 23.57 M putative SNPs after the calling procedure. Then, the SNPs with missing rate ≥ 0.1, MAF ≤ 0.01, severely departed from HWE (p = 0.001) were removed. These filtering resulted in a rapid decline of the SNPs number from 23.57 M to 2.08 M. Finally, we extracted ten subsets of genotypes containing 0.5 k, 1 k, 5 k, 8 k, 10 k, 30 k, 50 k, 100 k, 300 k, and 500 k markers from these filtered 2.08 M SNPs for downstream analysis.

The accuracies of GBLUP, wGBLUP, BayesB, and BayesC in estimating GEBVs of *V. harveyi* resistance in *C. semilaevis* were evaluated using the five-fold cross validation strategy and the ten SNP subsets. As showed in **Fig. 1**, the dashed line in grey represented the predictive accuracy of ABLUP (0.599), and solid line in black with rhombi, squares, triangles, and circles showed the accuracy variation of GBLUP, wGBLUP, BayesB, and BayesC, respectively. At first, the accuracy of the four genomic methods had a growing tendency with the increasing number of SNPs used for prediction, and then basically maintained steady or declined slightly after reached the highest accuracy. For GBLUP, more than 1 k SNPs could result in better estimation than ABLUP. For BayesB and BayesC, 8 k or more SNPs were essential for good prediction. For wGBLUP, at least 50 k SNPs were needed. The accuracy of GBLUP increased sharply from 0.468 to 0.681 when SNPs added from 0.5 k to 5 k and increased gently to 0.705 with the number of markers added to 30 k. When 50 k SNPs were used for prediction, the accuracy of GBLUP reached the peak (0.724) and then remained steady when the number of SNPs continually increased to 500 k. BayesB and BayesC showed similar variation in accuracy. Their accuracies both had huge increases (BayesB: 0.345 to 0.639; BayesC: 0.360 to 0.619) when the density of markers increased from 0.5 k to 8 k, and declined a little (BayesB: 0.610; BayesC: 0.608) in the 10 k SNPs subset. Then, their accuracies rose gradually and had the best prediction (BayesB: 0.659; BayesC: 0.662) in the 50 k SNPs subset. Finally, the accuracies of BayesB and BayesC decreased slightly when more than 50 k SNPs used for prediction. The accuracy changes in wGBLUP were different from the other three methods. At first, the accuracy increased rapidly from 0.220 to 0.534 with the density of SNP added from 0.5 k to 5 k. Secondly, the accuracy fluctuated between 0.498 to 0.545 when SNPs varying from 8 k to 30 k and then increased sharply to 0.688 with 50 k SNPs used. Thirdly, the accuracy of wGBLUP had the maximum (0.746) in 300 k SNPs subset, and the accuracy using 100 k and 500 k SNPs were 0.690 and 0.712, respectively.

Figure 1 The accuracy curves of GBLUP with SNPs varying from 0.5 k to 500 k.

SNP identification and characterization of the SNP array “Solechip No.1”

The dataset used for SNP selection of SNP array was derived from the reference group of 2014 and their parents. Using the next-generation sequencing technology, the average sequencing depth of 863 reference individuals and their 196 parents was 3.3 × and 7.5×, respectively. After SNP calling and quality control, 1.07 M were remained for selection.

Based on the criteria of Affymetrix *in-silico* probe design, the flanking sequence of SNPs and GC content, and the potential effects of the SNPs predicted by SNP annotation, etc., 38,295 SNPs were included in the SNP array with an average p-convert value of 0.687. Since 2,000 DQC probes were serving as negative controls, the total number of SNP probes on the array was 40,295. This 38 k SNP array was named as “Solechip No.1”. These 38,295 SNPs distributed throughout the whole genome with an average of 10.5 kb inter-spacing between two adjacent SNPs (Fig. 2). A total of 2,242 SNPs had an inter-spacing less than 4 kb and 1,202 SNPs had an interval that is more than 10 kb. The greatest number of SNPs (3,934) had an inter-spacing ranging from 5 kb to 5.5 kb, and the least number of SNPs (451) were belonging to an inter-spacing of 9.5 kb to 10 kb. 52.3% of SNPs (20,029) had an interval ≤ 7 kb of adjacent SNPs, and cumulative 96.9% of SNPs (37,073) had an SNP-spacing of ≤ 10 kb. The MAF distribution of all SNPs tilling on the array showed in **Fig. 3**. These SNPs had an average MAF of 0.055 and a median of 0.039. Most SNPs had a MAF less than 0.1, of which 22,874 SNPs with MAF between 0.01 to 0.05 and 9,440 SNPs between 0.05 to 0.1. Moreover, according to the results of predicted SNP effects, most of these 38,295 SNPs belong to intergenic (31.44%) and intronic (30.51%) variants, and 19.12% and 12.30% of SNPs were identified as the upstream and downstream of the gene, respectively (Table 1).

Table 1
Summary of the SNP effects on the customized 38 k SNP array “Solechip No.1”

SNP classification	Quantity	Percentage (%)
Intergenic region	12,041	31.44
Intronic variant	11,682	30.51
Missense variant	712	1.86
Stop gained	37	0.10
Stop lost	26	0.07
Stop retained variant	18	0.05
Synonymous variant	1,418	3.70
Splice acceptor variant	4	0.01
Splice donor variant	6	0.02
Splice region variant	317	0.83
Upstream gene variant	7,323	19.12
Downstream gene variant	4,709	12.30
Failure	2	0.01
Total	38,295	-

Figure 2 Distribution of inter-spacing of adjacent SNPs on the SNP array “Solechip No.1”.

Figure 3 Minor allele frequency distribution of SNPs tilling on the SNP array “Solechip No.1”.

Evaluation of the SNP array “Solechip No.1”

To test the genotyping quality of the SNP array “Solechip No.1”, 24 individuals of Chinese tongue sole were selected randomly from the reference group for genotyping again using the SNP array. After automatic analysis with the Axiom Analysis Suite software, all subjects passed the quality control processing, and 28,016 SNPs (73.2%) were recommended for downstream analysis. Among these loci, 63.2% of SNPs (including 6,668 (23.8%) “PloyHighResolution” and 11,043 (39.4%) “NoMinorHom”) were classified into the polymorphic and 36.8% (including 10,305 “MonoHighResolution”) were categorized as the monomorphic. Besides, to evaluate the genotyping accuracy, we compared the SNP of each locus generated by the array and by the re-sequencing technology. Among the recommended 28,016 SNPs, the average consistency of SNPs reached 94.8%, of which 53.5% of these SNPs had a consistency of 100% (Fig. 4). Cumulatively, a total of 79.3% of loci had a more than 90% of the consistency; 91.4% of loci had a more than 85% of the consistency. A proportion of 8.5% of loci had a consistency between 0 to 85%, and 0.2% of loci are completely inconsistency.

Figure 4 Consistency of genotypes yielded by the SNP array “Solechip No.1” and by the re-sequencing technology.

Genomic selection using the SNP array “Solechip No.1”

We selected 44 fishes as the candidates and were genotyped through the SNP array “Solechip No.1”. Then, these candidates were used as the parents of 23 families of *C. semilaevis*, and their GEBVs were estimated by the GBLUP.

The GEBVs of these 23 families were expressed as the mid-parental GEBVs. The Pearson correlation between the family GEBVs and survival rate of the corresponding family is 0.706, which is belonging to a strong positive correlation. The average survival rate of the top five families in GEBVs is 79.1%, which is higher than the bottom five families (58.1%).

Discussion

The artificial challenge test is a commonly used method to identify subjects' ability of disease resistance. Based on that, the evaluation of genetic parameters and the selective breeding schemes have been achieved in some farmed fishes [1]. From the results of the test conducted in 2016 and 2018, we observed abundant genetic variation of *V. harveyi* resistance among these Chinese tongue sole populations, which implied that the *V. harveyi* resistance of these populations could be improved by an appropriate strategy of selective breeding.

The accuracies of four genomic methods at estimating GEBVs of *V. harveyi* resistance in *C. semilaevis* were investigated using the five-fold cross validation strategy and the number of SNP varying from 500 to 500 k. The accuracy of GBLUP reached the peak in the 50 k SNPs subset and became steady when more than 50 k SNPs were used for prediction. This might imply that it is sufficient for GBLUP to construct a precise genetic relationship matrix using 50 k SNPs. Different from the GBLUP, BayesB and BayesC relied on the magnitude of linkage disequilibrium (LD) between genetic markers and QTLs [35]. Though BayesB and BayesC had the highest accuracy using 50 k SNPs, a few decreases were observed when more than 50 k SNPs were used for prediction. Theoretically, more SNPs meant more LD information could be captured, likewise implied more SNPs would be LD phase with each other. Since BayesB and BayesC assumed that only a fraction of SNPs had genetic effects, an excess of markers used for prediction might affect the sampling results of MCMC. This might be the reason why slight decreases in accuracy were observed when more than 50 k SNPs were used for estimation in BayesB and BayesC. Besides, it is interesting to note that wGBLUP showed fierce changes in accuracy than the other three GS methods. This phenomenon might be caused by the nature of estimation of weights of SNPs. Since the SNP effect is a regressed value instead of the true value, the SNP effects derived from the GEBVs were suboptimal [31]. Therefore, it is worth to study the substitution of the regressed SNP effects by the effects estimated from the Bayesian methods.

In the 50 k SNPs subset, the accuracy of GBLUP outperformed the other three GS methods with an accuracy of 0.724 and followed by the wGBLUP (0.688). BayesB and BayesC yielded a close estimation with an accuracy of 0.659 and 0.662, respectively. Some studies on GS about the trait of disease resistance in fish reported that Bayesian methods had a close or higher accuracy than GBLUP [10, 15, 36, 37]. However, Vallejo et al. reported that single-step GBLUP (ssGBLUP) had a better prediction than weighted ssGBLUP, BayesB, and BayesC at improving resistance to bacterial cold water disease in rainbow trout [38], which was similar to the situation showed here. According to the results of GWAS conducted by Zhou et al. [23], we could conclude that resistance against *V. harveyi* in Chinses tongue sole belonged to a polygenic genetic architecture. Therefore, the method based on an assumption of equal variance for all SNPs, such as GBLUP and ssGBLUP, might be more suitable than that method based on different variance for a fraction of SNPs, such as Bayesian methods and weighted single-step method for estimation of GEBV of resistance to *V. harveyi* in *C. semilaevis*. However, GBLUP or Bayesian methods only utilized the data of genotyped individuals. Therefore, the performance of the ssGBLUP which integrated phenotypes, pedigree, and genotypes of all breeding populations is worth investigating in the follow-up study [39, 40].

The re-sequencing technology and SNP array are both efficient tools to obtain abundant SNPs. However, the SNP array has advantages at convenience and timesaving, which might have a priority in breeding practice. In this study, we developed a 38 k SNP array for the disease resistance selective breeding of Chinese tongue sole, which is names as “Solechip No.1”. After several steps of selection, 38,295 SNPs were selected for the SNP array from 1.07 M high-quality SNPs with an average of 10.5 kb inter-spacing between two adjacent SNPs and an average MAF of 0.055. A high proportion of SNPs with MAF between 0.01 to 0.05 were observed. These rare variants (MAF less than 0.05) might exist in some families, which would be beneficial to the GWAS and GS analysis. If the SNPs obtained by the re-sequencing were considered to precise enough, it is worth comparing the SNPs generated by the array and by the re-sequencing to evaluate the genotyping accuracy of the “Solechip No.1”. We randomly selected 24 re-sequencing individuals and genotyped them again using the SNP array “Solechip No.1”. All subjects passed the QC procedures and yielded 28,016 SNPs for subsequent analysis. Among these markers, 53.5% of SNPs had a complete consistency with the SNPs obtained by the re-sequencing and 0.2% of SNPs were fully different. We inferred that all these inconsistent SNPs might be caused by the false positive, rare bases among some families, and genotype imputation of the reference group. From these results, we could conclude that the SNP array “Solechip No.1” possessed high accuracy at genotyping, which could generate high-quality SNP to meet the needs of genetic analysis, such as GWAS and GS.

GS refers to the use of genomic prediction in selection. Here, we selected 44 candidates as the parents of 23 families of *C. semilaevis* and genotyped them using the SNP array “Solechip No.1”. Their GEBVs were estimated by GBLUP, and the family GEBV was expressed as the mid-parental GEBV. The Pearson’s correlation between the family GEBVs and survival rate of these 23 offspring families is 0.706, and the average survival rate of the top five families in GEBVs (79.1%) is higher than the bottom five families (58.1%). It suggested that families with high GEBV are more likely to possess low mortality after infection, which is accordant with the conclusion of the high efficiency in the selection of GS in Japanese flounder (*Paralichthys olivaceus*) disease-resistant breeding reported by Liu et al. [15].

Conclusion

We successfully constructed a reference group of *C. semilaevis* containing 1572 individuals for *V. harveyi* resistance breeding and investigated the accuracies of GBLUP, wGBLUP, BayesB, and BayesC at estimating the GEBVs. Among these genomic methods, GBLUP had a better estimation than ABLUP when more than 1 k SNPs used for prediction. Therefore, it could be an optimal method to improve the *V. harveyi* resistance in *C. semilaevis*. Besides, we developed the 38 k SNP array “Solechip No.1” for the Chinese tongue sole breeding practice. This array could generate high-quality SNPs for genetic analysis and had a high accuracy at genotyping. Last, we selected 44 candidates of *C. semilaevis* as the parents of 23 families. The GEBVs of these candidates were estimated using GBLUP and SNPs generated by the SNP array “Solechip No.1”. The Pearson’s correlation between the family GEBVs and the survival rates of these 23 offspring families was 0.706, which indicated that selection based on high GEBV is practicable. Our study demonstrated that GS is a feasible and efficient method to improve the *V. harveyi* resistance in *C. semilaevis*, and the SNP array “Solechip No.1” is a convenient and reliable tool for the Chinses tongue sole breeding practice.

Abbreviations

ABLUP
best linear unbiased prediction using pedigree
CallRateBelowThreshold
SNPs that were below the threshold of SNP call rate in the SNP array analysis
DQC
dish quality control, probes used as the negative control in the SNP array
GBLUP
genomic best linear unbiased prediction

GEBV

genomic estimated breeding value

GS

genomic selection

GWAS

genome-wide association study

HWE

Hardy-Weinberg equilibrium

LD

linkage disequilibrium

MAF

minor allele frequency

MonoHighResolution

SNPs that were detected as the monomorphic SNPs with good cluster resolution in the SNP array analysis

NoMinorHom

SNPs formed two good clusters but none of the minor homozygous genotypes were detected in the SNP array analysis

OTV

SNPs that few hybridization signals were captured in the SNP array analysis

PolyHighResolution

SNPs were detected as the polymorphic SNPs with three high-resolution clusters in the SNP array analysis

QC

quality control

QTL

quantitative trait loci

SNP

single nucleotide polymorphism

ssGBLUP

single-step genomic best linear unbiased prediction

wGBLUP

weighted genomic best linear prediction

Declarations

Ethics approval

The challenge test carried out in this study was in accordance with the recommendations of the Care and Use of Laboratory Animals of the Chinese Academy of Fishery Sciences. The protocol of the test was approved by the Animal Care and Use Committee of the Chinese Academy of Fishery Sciences.

Consent for publication

Not applicable

Competing interests

All authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by National Nature Science Foundation (31530078), Central Public-interest Scientific Institution Basal Research Fund, CAFS (2020TD20), China Agriculture Research System (CARS-47-G03), AoShan Talents Cultivation Program Supported by Qingdao National Laboratory for Marine Science and Technology (No.2017ASTCP-OS15).

Authors' contributions

SC initiated, managed, and conceived this study. SL, QZ, and SC wrote this article. SL, QZ, YC, and YL analyzed data. LW, YL, and YY performed the challenge test and provided biological samples. All authors read and approved the final manuscript.

References

1. Ødegård J, Baranski M, Gjerde B, Gjedrem T. Methodology for genetic evaluation of disease resistance in aquaculture species: challenges and future prospects. Aquac Res. 2011;42:103 – 14.

2. Meuwissen THE., Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

3. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123:218 – 23.

4. Hayes BJ, Lewin HA, Goddard ME. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends Genet. 2013;29:206 – 14.

5. Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP, et al. Implementation of genomic selection in the poultry industry. Anim Front. 2016;6:23–31.

6. Ødegård J, Moen T, Santi N, Korsvoll SA, Kjøglum S, Meuwissen THE. Genomic prediction in an admixed population of Atlantic salmon (*Salmo salar*). Front Genet. 2014;5:402.

7. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, et al. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics. 2015;16:969.

8. Tsai HY, Hamilton A, Tinch AE, Guy DR, Bron JE, Taggart JB, et al. Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet Sel Evol. 2016;48:47.

9. Yoshida GM, Carvalheiro R, Lhorente JP, Correa K, Figueroa R, Houston RD, et al. Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. Aquaculture. 2018;491:147 – 54.

10. Vallejo RL, Leeds TD, Gao G, Parsons JE, Martin KE, Evenhuis JP, et al. Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genet Sel Evol. 2017;49:17.

11. Vallejo RL, Silva RMO, Evenhuis JP, Gao G, Liu S, Parsons JE, et al. Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: evidence that long-range LD is a major contributing factor. J Anim Breed Genet. 2018;135:263 – 74.

12. Yoshida GM, Carvalheiro R, Rodríguez FH, Lhorente JP, Yáñez JM. Single-step genomic evaluation improves accuracy of breeding value predictions for resistance to infectious pancreatic necrosis virus in rainbow trout. Genomics. 2018;111:127 – 32.

13. Palaiokostas C, Kocour M, Prchal M, Houston RD. Accuracy of genomic evaluations of juvenile growth rate in common carp (*Cyprinus carpio*) using genotyping by sequencing. Front Genet. 2018;9:82.

14. Dong L, Xiao S, Wang Q, Wang Z. Comparative analysis of the GBLUP, emBayesB, and GWAS algorithms to predict genetic values in large yellow croaker (*Larimichthys crocea*). BMC Genomics. 2016;17:460.

15. Liu Y, Lu S, Liu F, Shao C, Zhou Q, Wang N, et al. Genomic selection using BayesCπ and GBLUP for resistance against *Edwardsiella tarda* in Japanese flounder (*Paralichthys olivaceus*). Mar Biotechnol. 2018;20,559 – 65.

16. Lu S, Zhu J, Du X, Sun S, Meng L, Liu S, et al. Genomic selection for resistance to *Streptococcus agalactiae* in GIFT strain of *Oreochromis niloticus* by GBLUP, wGBLUP, and BayesCπ. Aquaculture. 2020;523:735212.

17. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, et al. Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). BMC Genomics. 2014;15:90.

18. Correa K, Lhorente JP, López ME, Bassini L, Naswa S, Deeb N, et al. Genome-wide association analysis reveals loci associated with resistance against *Piscirickettsia salmonis* in two Atlantic salmon (*Salmo salar* L.) chromosomes. BMC Genomics. 2015;16:854.

19. Liu S, Sun L, Li Y, Sun F, Jiang Y, Zhang Y, et al. Development of the catfish 250K SNP array for genome-wide association studies. BMC Res Notes. 2014;7:135.

20. Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR, et al. The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. Mol Ecol Resour. 2014;15:662 – 72.

21. Li Y, Wang L, Yang Y, Li X, Dai H, Chen S. Genetic analysis of disease resistance to *Vibrio harveyi* by challenge test in Chinese tongue sole (*Cynoglossus semilaevis*). Aquaculture. 2019;503:430-5.

22. Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, et al. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat Genet. 2014;46:253 – 60.

23. Zhou Q, Su Z, Li Y, Liu Y, Wang L, Lu S, et al. Genome-wide association mapping and gene expression analyses reveal genetic mechanisms of disease resistance variations in *Cynoglossus semilaevis*. Front Genet. 2019;10:1167.

24. Chen S, Du M, Yang J, Hu Q, Xu Y, Zhai J, et al. Development and characterization for growth rate and disease resistance of families in half-smooth tongue sole (*Cynoglossus semilaevis*). J Fish China. 2010;34:1789-94.

25. Zhou Q, Su X, Wang A, Xu J, Ning K. QC-Chain: fast and holistic quality control method for next-generation sequencing data. PLoS ONE. 2013;8:e60234.

26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078-9.

27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559 – 75.

28. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6:80–92.

29. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer. 2016.

30. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.

31. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res. 2012;94:73–83.

32. Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE. 2010;5:e12648.

33. Gilmour AR, Gogel BJ, Cullis BR, Thompson R, Butler D. ASReml user guide release 3.0. Hemel Hempstead: VSN International Ltd. 2009.

34. Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of genomic selection in mice. Genetics. 2008;180:611-8.

35. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389-97.

36. Bangera R, Correa K, Lhorente JP, Figueroa R, Yáñez JM. Genomic predictions can accelerate selection for resistance against *Piscirickettsia salmonis* in Atlantic salmon (*Salmo salar*). BMC Genomics. 2017;18:121.

37. Palaiokostas C, Cariou S, Bestin A, Bruant JS, Haffray P, Morin T, et al. Genome-wide association and genomic prediction of resistance to viral nervous necrosis in European sea bass (*Dicentrarchus labrax*) using RAD sequencing. Genet Sel Evol. 2018;50:30.

38. Vallejo RL, Leeds TD, Fragomeni BO, Gao G, Hernandez AG, Misztal I, et al. Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. Front Genet. 2016;7:96.

39. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci. 2009;92:4648-55.

40. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42:2.

Figures

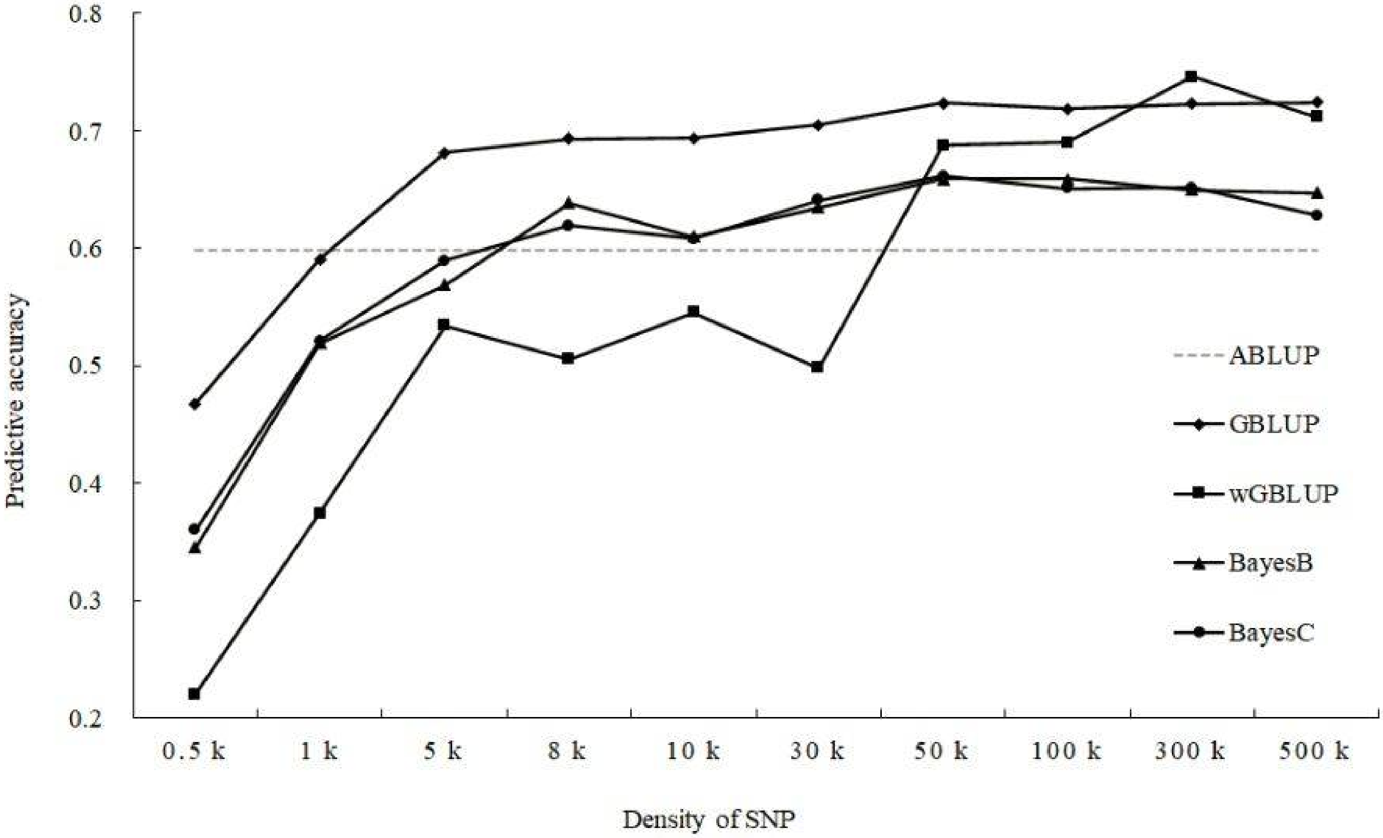

Figure 1

The accuracy curves of GBLUP with SNPs varying from 0.5 k to 500k.

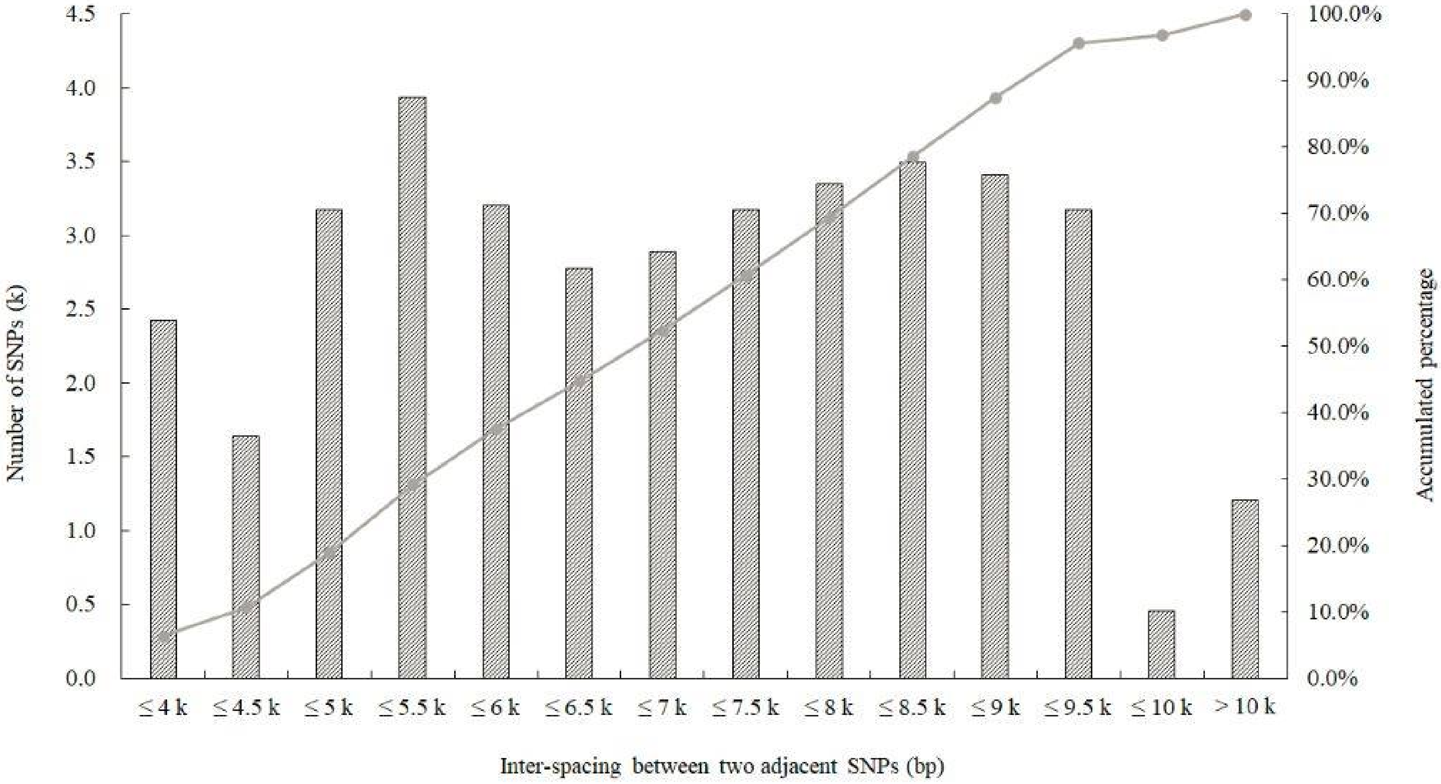

Figure 2

Distribution of inter-spacing of adjacent SNPs on the SNP array “Solechip No.1”.

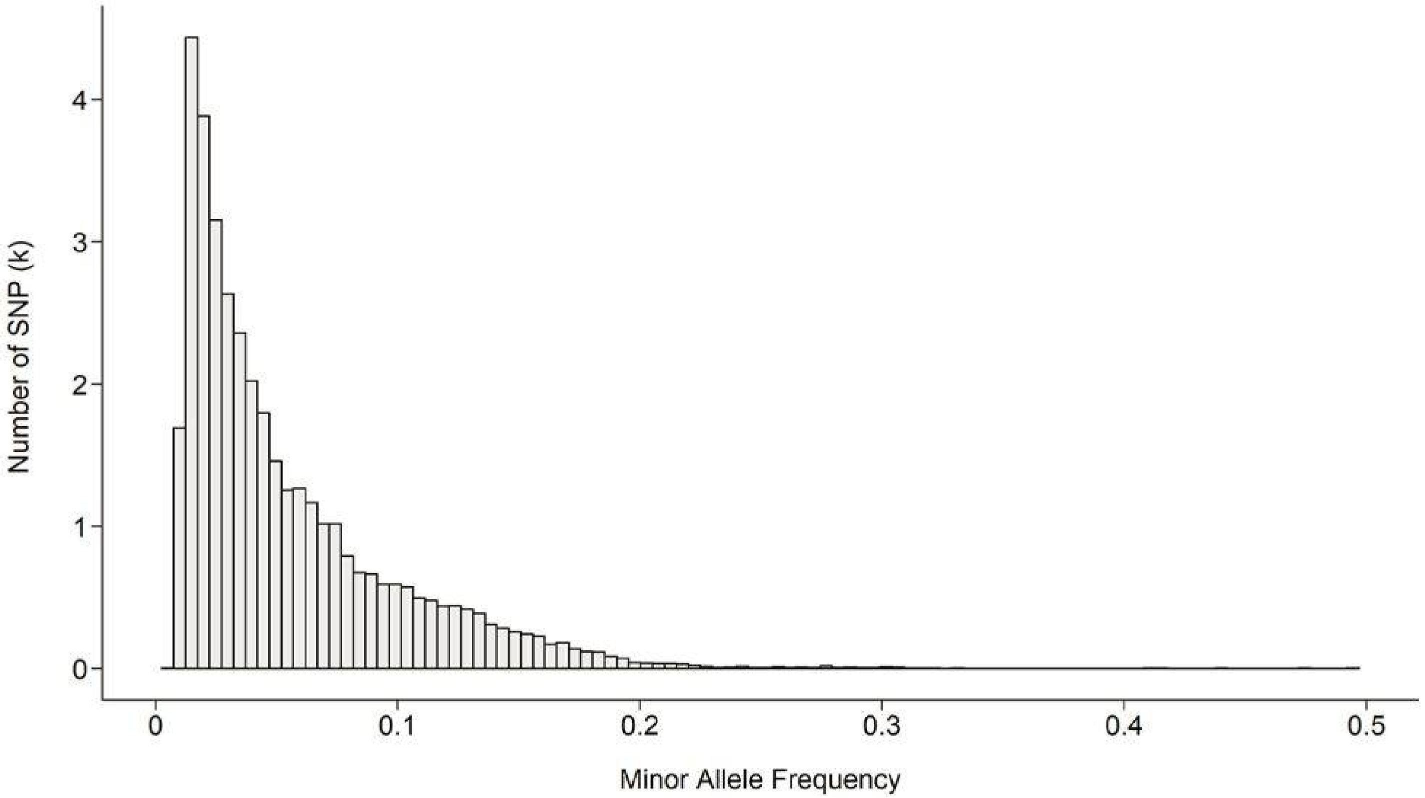

Figure 3

Minor allele frequency distribution of SNPs tilling on the SNP array “Solechip No.1”.

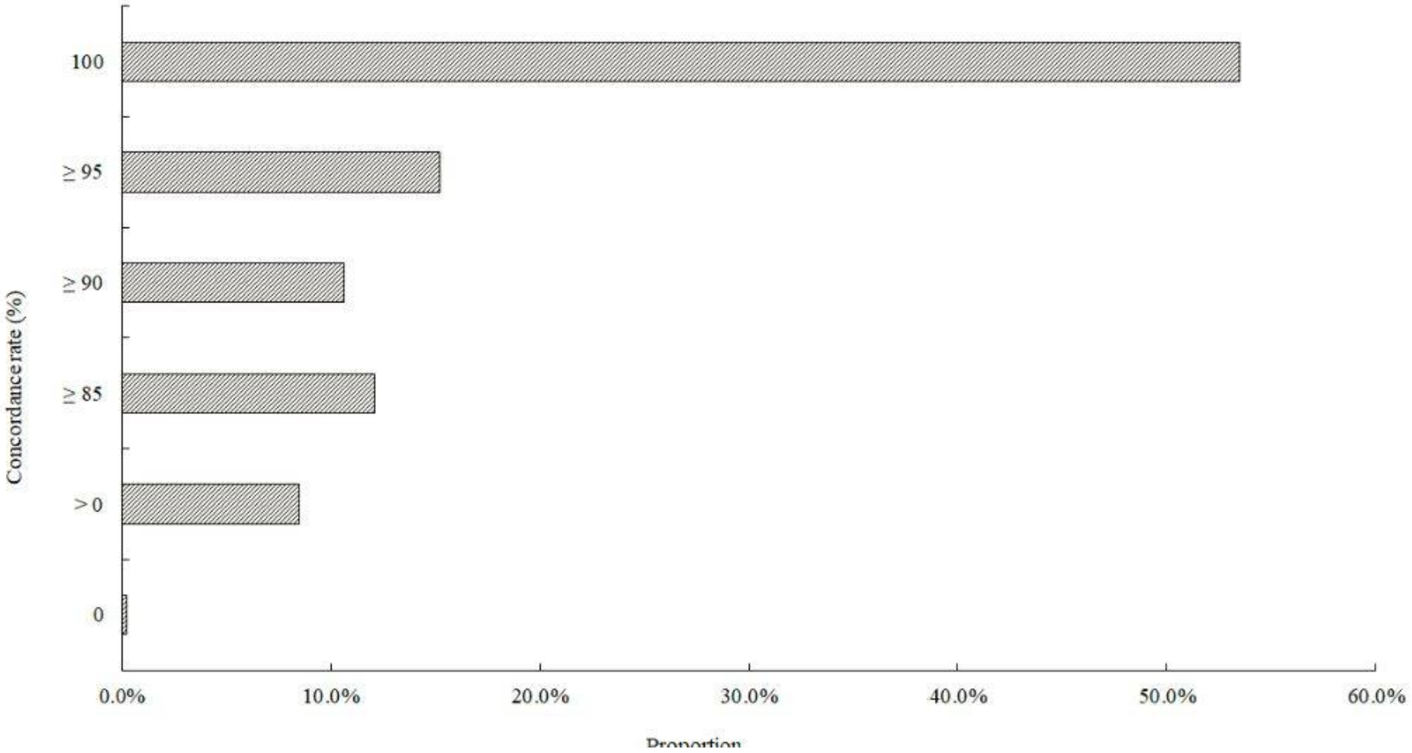

Figure 4

Consistency of genotypes yielded by the SNP array “Solechip No.1” and by the re-sequencing technology.