Analysis of Population Genetic Diversity of *Cynoglossus Semilaevis* in Three Groups from Northern China Sea

ZHANG Bo¹, LIU KeFeng², ZHENG DeBin³, JIA Lei*

Tianjin Bohai Sea Fisheries Research Institute, Bohai Sea Fisheries Research Center of Chinese Academy of Fishery Sciences, Tianjin 300457, China
E-mail: jialei@163.com

Abstract. Taking the Cultured and Wild Populations of *Cynoglossus semilaevis* in Tianjin Coast and cultured Populations in Haiyang as study objects, there were 32746 SNP loci screened by 2b-RAD sequencing. The genetic relationship, genetic structure, genetic diversity in these three groups were also analyzed by SNP markers. The genetic variation within populations (Fis), the genetic variation between groups (Fst), and total variation (Fit) in each SNP locus in these three groups were estimated, which shown that Fst value was ranged from 0.0731~0.1635. The genetic differentiation between Tianjin culture population and Tianjin wild population is the smallest, and the genetic differentiation between Tianjin and Haiyang culture populations, Tianjin wild population and Haiyang culture population is great. The results of selective sweep showed that the polymorphism of wild population was highest in Tianjin, followed by Tianjin culture population, and lowest in Haiyang breeding population. The polymorphism information content (PIC) statistics of Haiyang, Tianjin wild and Tianjin cultured populations respectively in each SNP site shown that Tianjin wild was up to 0.220, followed by Haiyang cultured and Tianjin breeding, the lowest is 0.175. The differentiation between Tianjin wild and Tianjin culture populations is not significant. Principal component analysis was carried out by using software GCTA, and PCA showed that the contribution rates of the first two principal components PC1 and PC2 were 16.25% and 9.54%, respectively, and the cumulative contribution was 25.79%. Two dimensional clustering results show that a small part of the region in Tianjin wild population had a small intervals with that of Tianjin cultured population, but the two principal components generally can be clearly divided into two categories, and these two populations had a clear boundaries with haiyang cultured one. Based on the above SNP locus diversity, polymorphism information content, genetic diversity comparison, the genetic diversity of Tianjin wild population is higher than Haiyang farming groups, but not significant difference between wild and cultured populations in tianjin revealed that Tianjin populations may have a certain extent mixture, which means the possibility of population mixture between wild and culture individuals may exist.

1. Introduction
The molecular marker assisted breeding technology is the reseach focus and hotspot in recent years in the field of molecular breeding of fish. With the development of biotechnology, different molecular markers, such as RAPD, RFLP, SSR and SNP, have been used to assist molecular breeding, and the methods of developing new markers have been further developed. The development of complete genome sequencing has brought huge amounts of DNA marker, and the group diversity analysis, population structure analysis, genetic relationship and estimation of linkage disequilibrium degree can be explored by using these molecular markers, all these explorations can make the breeding ideas more clearly and the breeding goal easier to achieve ([11])

---

**Creative Commons Attribution 3.0 licence.** Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
2b-RAD method is a streamlined restriction site-associated DNA (RAD) genotyping method based on sequencing the uniform fragments produced by type II restriction endonucleases, which (for example, BsaXI and AluI) cleave genomic DNA upstream and downstream of the target site, producing tags of uniform length that are ideally suited for sequencing on existing next-generation platforms.[2] The method provides a cost-effective and flexible genotyping platform for aquaculture species lacking sufficient genomic resources. The simplicity of the 2b-RAD protocol makes it particularly suitable for high-throughput genotyping as required for linkage mapping and profiling genetic variation in natural populations. 2b-RAD can be used in population genetic diversity analysis, population evolution research, genome assisted assembly, genome selection breeding and other fields[3]. In this study, three populations of C. semilaevis have been screened and 32746 SNPs have been found, further more, diversity analysis and SNP polymorphism analysis have also been investigated.

2. Material and Methods

2.1. Animals
Here we picked 10 healthy C. semilaevis weighing 25.2 ± 3.4 g were cultured in the Huanghai fishery company in haiyang city in Shandong Province, defined as ‘haiyang culture group’ and 10 healthy C. semilaevis weighing 22.1 ± 2.4 g were cultured in the XINGSHENG fishery company in Tianjin, defined as ‘Tianjin culture group’, also, 10 healthy C. semilaevis weighing 21.1 ± 5.7 g were captured from Bohai sea, which are considered as ‘Tianjin wild group’. All the fish were kept in three 70 L tanks at 22 ℃ in aerated seawater. Before the experiment, three fish were euthanized and sampled for the presence of DNA extraction. The animal experiments were conducted in accordance with the “Regulations for the Administration of Affairs Concerning Experimental Animals” from the State Science and Technology Commission of Tianjin. The study was approved by the ethics committee of Bohai Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences[4].

2.2. Extraction of genomic DNA
For DNA extraction, the larvae or fins fixed in anhydrous ethanol were placed in 1.5 mL centrifuge tubes. The minced tissues were homogenized in 400 μL TNEs lysis buffer (10 mM Tris- HCl, pH 7.5; 400 μM NaCl; 100 mM EDTA; 0.6% SDS, and freshly added 10 μL proteinase K, 10 mg/mL), then the homogenate was lysed at 55 ℃. After that, 140 μL saturated NaCl solution was added, and the mixture was centrifuged at 12000 rpm for 30 min at 4 ℃. The DNA was precipitated by two volumes of ethanol which had been refrigerated at -20 ℃. Then the DNA pellets were washed twice with 70% ethanol, dried and dissolved in 50 μL TE buffer (10 mM Tris- HCl, pH 8.0, 10 mM EDTA).

2.3. 2b-RAD library preparation
The concentrations of all DNA samples were adjusted to 50 ng/μL. The DNA (200 ng) was digested in 10 μL reaction volume of NEB Buffer 4 with 1 U BsaXI at 37 ℃ for 2 h. An additional DNA was digested simultaneously to detect the digestion efficiency by 1% agarose gel electrophoresis. The primary DNA band disappeared and became disperse, indicating a successful digestion. Then the ligation reaction was completed in the same tube as the digestion, combining the remaining digested DNA with 2 μL of the adapter1, 2 μL of the adapter2, 2 μL ligase buffer and 400 U of the T4 ligase. BsaXI could not be inactivated by heating, so the digested productions was recommended performing the ligations in 4 ℃ for 1 hour, then hold on ice.

2.4. Genotyping and quality analysis
30 samples were sequenced on the HiSeq X-Ten platform by Paired-end sequencing. In the procedure, the standard protocol features adaptors with fully degenerate cohesive ends (5′-NNN-3′). Eliminate the sequences containing no BsaXI recognition sites, excluding low quality sequences, keeping the high quality original reads whose ratio was above 75% on average, and the label depth was between 14-25, the average depth was 20, all these showed the Library of good quality. After genotypes procedure, totally 32,746 SNPs yielded by SOAPsnp were confirmed. All of the markers information was showed in Additional file the results in the original reads filter is in Table 1:

| O.  | N | Original reads | High quality reads | Percentage | Unique | Depth |
| --- | --- | --- | --- | --- | --- | --- |
| HY-1 | 6,396,349 | 5,054,386 | 79.02% | 143,337 | 20 |
| HY-2 | 6,396,349 | 5,587,054 | 87.35% | 145,862 | 21 |
| HY-3 | 6,396,349 | 5,798,571 | 90.65% | 145,740 | 22 |
3.1 Results and Analysis

According to the 32746 SNP loci obtained, the genotyping data have been analysed of each SNP at the individuals in 3 groups to get the value of Fis. Fst. Fit by Genepop software (version 4.2.2) ([5]). θπ and Tajima D have been calculated by GD software (version: 0.8) to evaluate the nucleotide diversity according to the genotyping data and mismatch analysis.

### 3.1.1 The value of Fst among three groups

| Groups | HY     | TJCL   |
|--------|--------|--------|
| TJCL   | 0.1635 |        |
| TJYS   | 0.1778 | 0.0731 |

For the species which have reference genomes, according to the genotyping strategy of RAD-typing, the markers are classified as follows: Construction of reference sequences: extract the tags includes BsaXI restriction site label as a reference sequence from the Cynoglossus semilaevis reference genomes. 

*Cynoglossus semilaevis* 

Cynoglossus semilaevis reference genome reference: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000523025.1_Cse_v1.0/GCF_000523025.1_Cse_v1.0_genomic.fna.gz

For Individual SNP genotyping: Mapping the high quality reads to the reference sequence of the individual by using SOAP software (the parameter is: -M 4 - V 2 - R 0). Site genotyping was performed using the maximum likelihood ML. In order to ensure the accuracy and preciseness of typing of SNP loci, the following conditions were filtered: in individuals, the tag depth was below 500; Leave only 3 SNP tagged sites within the all tags; the tags can be typed in 80% of the individuals. The minimum allele frequency MAF ≥ 0.05; There were a total of 32746 genotyping have been typed.

| HY     | 6,396,349 | 5,545,037 | 86.69% | 146,059 | 21 |
|--------|-----------|-----------|--------|---------|----|
| HY-5   | 6,396,349 | 4,045,274 | 63.24% | 144,101 | 16 |
| HY-6   | 7,065,887 | 5,445,752 | 77.07% | 144,571 | 21 |
| HY-7   | 7,065,887 | 5,972,173 | 84.52% | 144,777 | 23 |
| HY-8   | 7,065,887 | 6,462,181 | 91.46% | 145,127 | 25 |
| HY-9   | 7,065,887 | 6,444,063 | 91.20% | 143,720 | 25 |
| HY-10  | 7,065,887 | 4,748,605 | 67.20% | 143,837 | 19 |
| TJcl-1 | 6,646,800 | 4,688,066 | 70.53% | 145,915 | 18 |
| TJcl-2 | 6,646,800 | 5,804,242 | 87.32% | 146,395 | 21 |
| TJcl-3 | 6,646,800 | 5,728,174 | 86.18% | 145,486 | 21 |
| TJcl-4 | 6,646,800 | 5,573,184 | 83.85% | 146,024 | 21 |
| TJcl-5 | 6,646,800 | 4,244,575 | 63.86% | 145,211 | 16 |
| TJcl-6 | 7,027,055 | 5,482,653 | 78.02% | 145,998 | 21 |
| TJcl-7 | 7,027,055 | 6,021,305 | 85.69% | 146,142 | 22 |
| TJcl-8 | 7,027,055 | 6,082,922 | 86.56% | 145,434 | 23 |
| TJcl-9 | 7,027,055 | 5,766,252 | 82.06% | 146,083 | 22 |
| TJcl-10| 7,027,055 | 4,655,306 | 66.25% | 146,354 | 18 |
| TJYS-1 | 5,866,332 | 4,209,487 | 71.76% | 143,403 | 16 |
| TJYS-2 | 5,866,332 | 5,149,297 | 87.78% | 143,053 | 20 |
| TJYS-3 | 5,866,332 | 5,136,678 | 87.60% | 142,469 | 20 |
| TJYS-4 | 5,866,332 | 4,993,461 | 85.12% | 143,425 | 19 |
| TJYS-5 | 5,866,332 | 3,645,928 | 62.15% | 142,983 | 14 |
| TJYS-6 | 7,404,738 | 4,473,240 | 60.41% | 145,430 | 16 |
| TJYS-7 | 7,404,738 | 5,977,990 | 80.73% | 143,673 | 22 |
| TJYS-8 | 7,404,738 | 5,375,319 | 72.59% | 143,216 | 20 |
| TJYS-9 | 7,404,738 | 6,021,680 | 81.32% | 143,343 | 22 |
| TJYS-  | 7,404,738 | 4,184,500 | 56.51% | 143,167 | 15 |
The statistical data from table 2 showed that the value of Fst among three groups ranged from 0.0731-0.1635. The genetic differentiation between Tianjin cultured group and Tianjin wild group was the smallest, and the genetic differentiation between Tianjin cultured group and Haiyang group, Tianjin wild group and Haiyang group were great. As we all know the coefficient of genetic differentiation (Fst) was less than 0.05, indicating that the variation was mainly from within the group and no differentiation among the group. The value of Fst was greater than 0.05, indicating that genetic variation occurred mostly among groups, but there was low differentiation among the groups.

3.2. Tajima's D test
Tajima's D is a statistical test to distinguish between a DNA sequence evolving randomly ("neutral") and one evolving under a non-random process. A randomly evolving DNA sequence contains mutations with no effect on the fitness and survival of an organism. The randomly evolving mutations are called "neutral", while mutations under selection are "non-neutral". The purpose of Tajima's test is to identify sequences which do not fit the neutral theory modeateequilibriumbetween mutation and genetic drift.

Table 3. \( \theta \pi \) and Tajima D among three groups

|        | HY    | TJCL  | TJYS  |
|--------|-------|-------|-------|
| \( \theta \pi \) | 0.118343 | 0.185802 | 0.206763 |
| Tajima D | 1.457026 | 0.448594 | 0.59124 |

In the standard neutral evolution model, the theoretical value of Tajima's D is zero, which means the randomly evolving mutations happened. If the actual Tajima's D value deviates significantly from zero, the actual allele frequency is found to be biased against the expectation of the neutral evolution model, which means mutations under selection are "non-neutral". In Table 3, the Tajima D in HY group was highest, which shown that the group deviated from the neutral pattern and may be affected by population expansion and natural selection. \( \theta \pi \) is an important index to measure the population polymorphism. Table 3 showed that there was the highest degree of polymorphism among the wild population in Tianjin, followed by the Tianjin culture group and the lowest in Haiyang group.

3.3. SNP polymorphism analysis
According to the 32746 SNP loci were obtained, respectively, TJCL, TJYS and HY three groups of each SNP locus were analysed by using PowerMarker software (version 3.25), the polymorphism information content (PIC), heterozygosity observed (Ho) and heterozygosity expected (He) were shown in table 4.

Table 4. SNP polymorphism analysis

| Group | He       | Ho       | PIC    |
|-------|----------|----------|--------|
| HY-   | range    | 0~0.5    | 0~1    | 0~0.375 |
|       | average  | 0.2212   | 0.3115 | 0.1752  |
| TJCL  | range    | 0~0.5    | 0~1    | 0~0.375 |
|       | average  | 0.2531   | 0.2877 | 0.2066  |
| TJYS  | range    | 0~0.5    | 0~1    | 0~0.375 |
|       | average  | 0.2473   | 0.2805 | 0.220   |

The polymorphism information content of HY, TJCL and TJYS respectively in each SNP site (PIC) were be statistically analysed, TJYS up to 0.220, followed by HY and TJCL, the lowest is 0.175, the differences between TJCL and TJSY were not significant.

3.4. Principle Component Analysis, PCA
Principal component analysis was carried out using software GCTA (version, 1.25.0). According to the first two principal components, each individual is depicted as a scatter diagram. In general, the individual can be clustered by the first two principal components. The result is shown below.(Fig1.)
should conservation population significant fact, In populations from wild diversity genome tags sequencing 2b 4. Discussion were obtained 3.5 boundaries components clustering 16.25% PCA ([Dark blue dots indicate TJYS group;Light blue dots indicate TJCL group; Pink dots indicate HY group]) PCA results showed that the contribution rates of PC1 and PC2 in the first two principal components were 16.25% and 9.54%, respectively, and the cumulative contribution was 25.79%(Fig1). Two dimensional clustering results shown that a small part of TJYS had a smaller intervals with TJCL, but the two principal components generally can be clearly divided into two categories, and the two groups in Tianjin had clear boundaries with the Haiyang cultured group

3.5. Construction of phylogenetic tree based on individual shared tag sequences Select the SNPs shared among 30 individuals, and if the corresponding loci are missing, use "-" instead. The obtained sequence were used to construct NJ (Neighbor-joining) tree by treebest (Version: 1.9.2), and the NJ (Neighbor-joining) tree was tested by bootstrap method (repeated 1000 times)(Fig2); the percentage of bootstrap were represented as the value at internal nodes in the graph, as the object name is at leaf node.

4. Discussion
2b-RAD method is a streamlined restriction site-associated DNA (RAD) genotyping method based on sequencing the uniform fragments produced by type IIB restriction endonucleases, 33-36bp enzyme digesting tags were produced after enrichment for high-throughput sequencing reaction downstream. Through genome-wide high-throughput SNP screening and genotyping analysis, SNPs can be found for the genetic diversity analysis of population ([6]) The principal component analysis found that the genetic differentiation between Tianjin cultured population and Tianjin wild populations were small, genetic differentiation of Tianjin wild population and Haiyang cultured population was great. The genetic diversity of half smooth tongue sole in Tianjin population is higher degree of polymorphism than that of the population of in Haiyang,The population from Tianjin wild and Tianjin cultured had not a significant difference, revealed Tianjin wild and cultured populations may be mixed to a certain extent. In this study, the same geographical conditions, differences between wild and cultured populations of genetic diversity and differentiation in three groups of Cynoglossussemilaevis were examined by SNP genotyping([7]). In fact, this difference is not significant, because the implementation of all the restocking activities, make breeding groups may part after releasing into the wild population, resulting in two in the morning was statistically significant difference. Here are some suggestions of protection of germplasm and utilization of the wild population in C.semilaevis: 1) As wild resources of C.semilaevi were reduced sharply, protection and conservation of the spawning field should be paid more attention,2) Cynoglossussemilaevis wild populations deviated from the Hardy-Weinberg equilibrium,Therefore, the artificial restocking and releasing in the sea should pay attention to its genetic structure, to avoid man-made accelerated half smooth tongue natural germplasm resources ;3) The genetic diversity of populations should be analysed before the artificial restocking

![Figure 1. Scatter plot of PCA](image-url)

(Dark blue dots indicate TJYS group;Light blue dots indicate TJCL group; Pink dots indicate HY group)
and releasing in the sea, and the population genetic structure can not be too simple, causing disturbance to the natural population structure.

Figure 2. Neighbor-joining tree of 30 individuals in three different groups of *Cynoglossus semilaevis*  
(Q1——Q10 indicate HY1-HY10; T1——T10 indicate TJCL1-TJCL10; TW1-TW10 indicate TJYS1-TJYS10)

5. References

[1] V I, Delphine E M, Albrecht L, Claude et al. 2010, Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers[J]. Theoretical and Applied Genetics. 120 (7): 1289-1299.

[2] X Fu J, Dou J, Mao et al. 2013, Radtyping: An integrated package for accurate de novo codominant and dominant RAD genotyping in mapping populations. PlosOne, 8(11): E79960.

[3] J T, Gerdes C F, Behr J G, Coors et al. 1994. Compilation of North American maize breeding germplasm[M]. Madison, WI: CSSA,

[4] L J, Lu A J, Ma X A, Wang et al. 2013, Polymorphisms analysis of five populations of Takifugurubipes with microsatellite.[J]. Progress in Fishery Science, 34(4): 27-34.

[5] S K, Mao A J, Ma F H, Ding et al. 2016, Analysis of genetic structures of four wild geographic populations of mullet *Liza haematocheila* by using microsatellite marker technique.[J]. Progress in Fishery Science, 37(2): 67-75.

[6] S, Wang E, Meyer JK, Mckay MV, Matz2012, 2B-RAD: A simple and flexible method for genome-wide genotyping. Nat Methods. 9: 808-810. 10.1038/NMETH.2023.

[7] W J, Zhang Y H, Yu Y F, Shen et al. 2003, Advances on genetic analysis of microsatellite DNA in protozoology. [J]. Acta Hydrobiologica Sinica, 27(2): 185-190.