RESEARCH ARTICLE

Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*)

Haigang Qi[1,2,3☯], **Kai Song**[1,3,4☯], **Chunyan Li**[1,2,3], **Wei Wang**[1,2,3], **Busu Li**[1,2,3], **Li Li**[1,3,4*], **Guofan Zhang**[1,2,3*]

1 Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China, **2** Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China, **3** National & Local Joint Engineering Laboratory of Ecological Mariculture, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China, **4** Laboratory for Marine Fisheries and Aquaculture, Qingdao National Laboratory for Marine Science and Technology, Qingdao, Shandong, China

☯ These authors contributed equally to this work.
* lili@qdio.ac.cn (LL); gfzhang@qdio.ac.cn (GZ)

Citation: Qi H, Song K, Li C, Wang W, Li B, Li L, et al. (2017) Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*). PLoS ONE 12(3): e0174007. https://doi.org/10.1371/journal.pone.0174007

Editor: Peng Xu, Xiamen University, CHINA

Received: December 11, 2016

Accepted: March 1, 2017

Published: March 22, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. All the raw data stored in CEL format have been deposited in the NCBI GEO database with an accession number of GSE94633.

Funding: This work was supported by the National Natural Science Foundation of China (31530079, 31572620), the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB06010101), the Technological Innovation Project Financially Supported by Qingdao National Laboratory for Marine Science and Technology (2015ASKJ02-03), the Earmarked Fund for Modern

Abstract

Single nucleotide polymorphisms (SNPs) are widely used in genetics and genomics research. The Pacific oyster (*Crassostrea gigas*) is an economically and ecologically important marine bivalve, and it possesses one of the highest levels of genomic DNA variation among animal species. Pacific oyster SNPs have been extensively investigated; however, the mechanisms by which these SNPs may be used in a high-throughput, transferable, and economical manner remain to be elucidated. Here, we constructed an oyster 190K SNP array using Affymetrix Axiom genotyping technology. We designed 190,420 SNPs on the chip; these SNPs were selected from 54 million SNPs identified through re-sequencing of 472 Pacific oysters collected in China, Japan, Korea, and Canada. Our genotyping results indicated that 133,984 (70.4%) SNPs were polymorphic and successfully converted on the chip. The SNPs were distributed evenly throughout the oyster genome, located in 3,595 scaffolds with a length of ~509.4 million; the average interval spacing was 4,210 bp. In addition, 111,158 SNPs were distributed in 21,050 coding genes, with an average of 5.3 SNPs per gene. In comparison with genotypes obtained through re-sequencing, ~69% of the converted SNPs had a concordance rate of >0.971; the mean concordance rate was 0.966. Evaluation based on genotypes of full-sib family individuals revealed that the average genotyping accuracy rate was 0.975. Carrying 133 K polymorphic SNPs, our oyster 190K SNP array is the first commercially available high-density SNP chip for mollusks, with the highest throughput. It represents a valuable tool for oyster genome-wide association studies, fine linkage mapping, and population genetics.

Agro-industry Technology Research System (CARS-48), and the Taishan Scholars Climbing Program of Shandong.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Oysters (phylum Mollusca, class Bivalvia) constitute an essential component of many aquatic ecosystems and are economically important in the fisheries and aquaculture industries [1]. Between 2010 and 2014, worldwide production of oysters increased from 4.5 million tonnes to 5.2 million tonnes (FAO 2014; http://www.fao.org), representing the highest global aquaculture production among marine animals. The Pacific cupped oyster *Crassostrea gigas* (Thunberg 1793) originated in northeastern Asia, and is natively distributed along the coasts of China, Japan, and Korea. It was introduced into Europe [2], Australia [3], and America [4] during the 20th century and has since established naturalized populations in most countries where it has been introduced for aquaculture purposes. However, owing to its ability to spread spontaneously and establish itself permanently in new habitats [5, 6], in many regions it is considered as an invasive species. In 2014, production of the Pacific oyster exceeded 625 kilotonnes, with an estimated value of 1.2 billion US dollars (FAO 2014). On the basis of its rapid growth rate, high disease resistance, and adaptability to different environments, the Pacific oyster is one of the most economically important bivalves worldwide.

Sustainable factory farming of oysters requires the development of high-quality oyster strains with rapid growth rates and high disease and stress resistance. Recent studies of the Pacific oyster industry have primarily focused on growth and yield [7–10], summer mortality [11–14], germplasm diversity [15–18], and viral infection [14, 19–21]. In the last decade, studies have mapped important traits [12, 22–24]and whole genome sequencing of *C. gigas* has been completed; this facilitates our understanding of the mechanisms of stress adaptation and shell formation in oysters [25]. Nevertheless, despite considerable progress in the oyster industry in recent decades, the Pacific oyster remains at an early stage of domestication, and the molecular mechanisms that modulate the commercially complex traits of this species and help it to survive in the variable marine environment remain unclear.

Single nucleotide polymorphisms (SNPs) are widespread nucleotide variations among individuals of a population, and they constitute the most abundant type of molecular marker in plant and animal genomes. Owing to their high abundance, co-dominant mode of inheritance, and ease of high-throughput detection, SNPs are widely used in biological research [26, 27]. The oyster possesses one of the highest levels of genomic polymorphism among animal species [25], and numbers of SNPs have been identified for various research purposes [28–30]. Nevertheless, oyster SNPs have not been extensively applied in high-resolution genetic research because of the lack of a high-throughput genotyping platform that can simultaneously type thousands of loci in multiple individuals. Such a platform is essential for fine mapping of important traits via extensive linkage or association analysis.

Since the release of the first commercial SNP array by Affymetrix (Santa Clara, CA) in 1996 [31], the use of microarrays and microarray technology has been a feasible choice for large-scale SNPs genotyping. A variety of SNP array platforms have been developed, of which the Affymetrix Custom Array, the Illumina BeadChip (Illumina, San Diego, CA), and the Sequenom MassArray (Sequenom, San Diego, CA) are most popular. These arrays differ in their principles for SNP detection, as well as in their requirements for marker numbers, cost, and sample size. In addition to the human SNP array, SNP arrays have been developed in many animal and plant species, including chicken [32], pig [33], cattle [34], horse [35], catfish [36], common carp [37], Atlantic salmon [38], rainbow trout [39], rice [40], soybean [41], maize [42], and strawberry [43]. In mollusks, a medium-throughput genotyping array involving 384 SNPs has been developed for the Pacific oyster [44]; however, to the best of our knowledge, a high-density oyster SNP array has not previously been available.

Owing to the increasing accessibility of next-generation sequencing (NGS) technologies, genotyping by sequencing (GBS) technologies—which usually detect SNPs through whole or reduced genome sequencing—have become a powerful genetic analysis tool [45]. GBS methods—especially those based on reduced genome sequencing—may be cost-effective for genome-wide SNP discovery or genotyping; however, the disadvantages of GBS arise because NGS data frequently suffer from high error rates derived from multiple factors, including base-calling and alignment errors. In general, for low-coverage sequencing, the larger the number of individuals, the higher the frequency of missing allele calls. For high-coverage sequencing, the increased cost—especially in the case of large genomes—cannot be ignored. When using whole genome sequencing for diploid species, a sequencing depth of more than 15–20 folds is essential for accurate SNP typing [46]. In addition, GBS is dependent on complicated library preparation ensured through rigorous quality control (QC) and intensive subsequent bioinformatics processing steps, including reads cleaning and filtering, reads mapping, brush-fire alignment adjustment, and SNP calling or genotyping; hence, GBS approaches are complex and time-consuming.

Further to the completion of our oyster genome project, we are currently conducting an oyster genome-wide association studies (GWAS) project using a re-sequencing approach to search for genes related to certain complex and important traits. The re-sequencing data generated from wild oysters will provide extensive resources for SNP mining and array design. The aim of the present study was to develop a high-density SNP genotyping array for the Pacific oyster (*C. gigas*) based on the Affymetrix Axiom platform, and to assess the potential of this array for future genetic and genomic research.

Materials and methods

Ethics statement

The Pacific oyster is a marine bivalve that is broadly distributed in large wild populations in coastal areas. The oysters used in this study were either directly collected from wild populations or cultured by the authors at a local farm. All experiments were performed according to local and national standard regulations. This study was approved by the Animal Care and Use Committee of Institute of Oceanology.

Sample collection and genome re-sequencing

In this study, we sampled 472 wild Pacific oysters—418 from 18 China coastal locations, 15 from Japan, 24 from Korea, and 15 from Canada. Genomic DNA was extracted from mantle tissues using a standard phenol-chloroform method. For each individual, 5 μg of DNA was sheared into fragments of 200–800 bp using the Covaris system (Life Technologies, Carlsbad, CA). DNA fragments were then treated according to the Illumina DNA sample preparation protocol. Fragments were end-repaired, A-tailed, ligated to paired-end adaptors, and PCR amplified with 300–500 bp inserts for library construction. Sequencing was performed to generate 100-bp paired-end reads with coverage of ≥20 folds on the HiSeq 2000 platform (Illumina) according to the manufacturer's standard protocols.

SNP identification

Filtered reads from all individuals were aligned to the oyster reference genome using Burrows-Wheeler Aligner (BWA) software [47] with the parameter "mem -M -t 10 -T 20." The aligned bam files were sorted and indexed with PICARD tools (https://broadinstitute.github.io/picard/). The reads from SNPs around the insertions and deletions (INDELs) in the bam files were then realigned using the Genome Analysis Toolkit (GATK) module RealignerTargetCreator and IndelRealigner [48]. To obtain high-quality variants, the GATK HaplotypeCaller module and Samtools [49] were used for variants calling of each sample, only concordance variants were selected, the SNPs were filtered with the parameter "QD < 2.0 & FS > 30.0 & MQ < 40.0 & DP < 6 & DP > 888 & ReadPosRankSum < −8.0 & BaseQRankSum < −8", and the INDELs were filtered with the parameter "QD < 2.0 & FS > 30.0 & ReadPosRankSum < −8.0". These variants were used to perform Base Quality Score Recalibration (BQSR), and reads were printed for population variants calling. Population variants calling was processed using the GATK HaplotypeCaller module with the parameter "-genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30".

SNP selection

Candidate SNPs were filtered in multiple steps using several criteria to eliminate possible false positive sites and distribute SNPs relatively evenly across the genome. For each oyster, if a SNP was of low and excessive read coverage (DP < 10 or DP > 100) and low genotype quality (GQ < 20), the SNP genotype call of the individual was considered to be missing or invalid. For a SNP, the missing rate was defined as the proportion of individuals with a missing SNP genotype in all individuals. The minor allele frequency (MAF) was calculated based on the allele information of individuals without a missing SNP genotype. In this step, SNPs within 20 bp around a predicted INDEL or with MAF < 0.05 or with a missing rate of > 0.1 were filtered out. The remaining SNPs, together with SNPs that were significantly related to important traits such as glycogen, amino acid, fatty acid, and heavy metals contents from our GWAS project (unpublished data), were submitted to Affymetrix for design score assessment using the Axiom myDesign GW bioinformatics pipeline. For each SNP, the sequence feather, binding energies, expected degree of nonspecific binding, possibility of hybridization to multiple genomic regions, and impacts originated from adjacent SNPs were taken into account; a p-convert value—which denotes the probability of the SNP converting to a reliable SNP assay on the Axiom array system—was assigned to each of the two probes flanking an SNP. SNPs with probes having high p-convert values were more likely to be convertible. In addition, a design proposal categorized as "recommended", "neutral", "not recommended" and "not possible" was assigned to each of the two probes. The SNP with at least a "recommended" or "neutral" proposal was retained and annotated with in-house custom Perl scripts. The GWAS SNPs and a small number of "large-effect" SNPs—which are predicted to cause stop or start codon loss, introduce a new stop codon, or cause a splice acceptor or donor variant—were initially selected and added into the final array SNP set. For the next selection, every SNP was assigned a priority level value, i.e., 4 for non-synonymous coding, 3 for synonymous coding, 2 for 2-kb upstream/downstream of a gene or in the intron region, and 1 for the intergenic region. The planned SNP capacity was ~200 K, and the assembled oyster genome size was 560 million; hence, the calculated average SNP interval was ~2,800 bp. A 3-kb sliding window was used to scan the oyster genome, and the SNP with the highest priority level value in the window was selected. Next, for genome scaffolds having an observed array SNP number lower than the expected SNP number, a 1-kb sliding window was used to fill in the remaining array probe space. The final list of SNPs was submitted to Affymetrix for production of the Axiom genotyping array.

Evaluation of the SNP array

To assess the performance of the oyster SNP array, we genotyped 96 Pacific oysters. Among these oysters, 44 were randomly selected from those re-sequenced for SNP discovery. Each oyster was deeply sequenced; hence, the genotype obtained using the oyster SNP array could be compared and cross-checked with that obtained through the GBS approach. We included 24 off-springs of a full-sib family designed for linkage mapping; both parents of this full-sib family were among the 44 randomly selected oysters mentioned above. In addition, 28 oysters were collected from a wild population in Qingdao, China. Genomic DNA was extracted from mantle tissues using a standard phenol-chloroform method. Genomic DNA samples were placed in a 96-well microtiter plate, and adjusted to a final concentration of ~50 ng/μL with a final volume of 10 μL. Genotyping with the oyster SNP chip was conducted using the Affymetrix GeneTitan Multi-Channel (MC) Instrument according to the standard operating instructions. The raw data stored in the form of CEL files were imported into Affymetrix Axiom Analysis Suite software version 1.1.1.66 for QC analysis and genotype calling, using the Axiom GT1 cluster algorithm with default parameters. The genotyping results were processed using SNPolisher software version 1.5.2. To assess the usability of the array in population genetics, all samples were subjected to principal component analysis (PCA) using Eigensoft software [50].

Results and discussion

Discovery and selection of high-confidence SNPs for array construction

For SNP discovery, we re-sequenced 472 wild oysters with coverage of >20 folds for each individual. After aligning the reads to oyster genome reference sequences through BWA, we identified 54 million high-quality SNPs using Samtools and GATK software; these SNPs may constitute the largest available oyster SNP dataset with broad representation of oyster resources. By applying several filters (see Materials and Methods), we reduced the initial SNP number by a factor of 20, to ~2.7 million putative SNPs. After *in silico* analysis for reproducibility prediction by Affymetrix, 886 K SNPs passed the evaluation; based on the distribution and function of each SNP, we submitted a final list of 190,420 SNPs to Affymetrix for production of an oyster genotyping array (Fig 1, S1 Dataset). We synthesized 192,789 probes for the 190,420 SNPs on the chip; 188,051 SNPs were tiled with one set of probes, and the remaining 2,369 SNPs were tiled with two sets of probes. For the 886 K candidate SNPs categorized by Affymetrix as "recommended" or "neutral", approximately 84% of the p-convert values of probes were >0.60. Probes with higher p-convert values are more likely to be converted; hence, 96% of the p-convert values of the 192,789 probes for the 190 K on-chip SNPs were >0.60 (Fig 2).

Genotyping performance of the SNP array

We performed a series of QC steps to ensure the accuracy of the outcomes during genotype calling. The signals of all 96 samples possessed a dish QC (DQC) value of 0.92–1.0, which was much greater than the default Axiom DQC cutoff value of 0.82. However, one wild oyster sample did not pass the call rate (CR) QC because its CR was 0.954, which was slightly lower than the default cutoff value of 0.970; hence, the overall rate of passing samples was 0.99 (95/96). As predicted, the passing rates of the 44 re-sequencing samples and the 24 full-sib family samples were 1.0, and there was no significant difference between the passing rate of the 28 wild samples (27/28 = 0.96) and that of the 44 re-sequencing samples (Fisher's Exact Test, $P = 0.39$). All the raw data stored in CEL format have been deposited in the NCBI GEO database with an accession number of GSE94633.

The genotypes of the oyster SNPs were classified into the following six categories according to their clustering performance (S1 Fig, S2 Dataset): (i) "PolyHighResolution"—three clusters were formed with good resolution; (ii) "NoMinorHom"—two clusters were observed with no samples of one homozygous genotype; (iii) "MonoHighResolution"—a single cluster of a

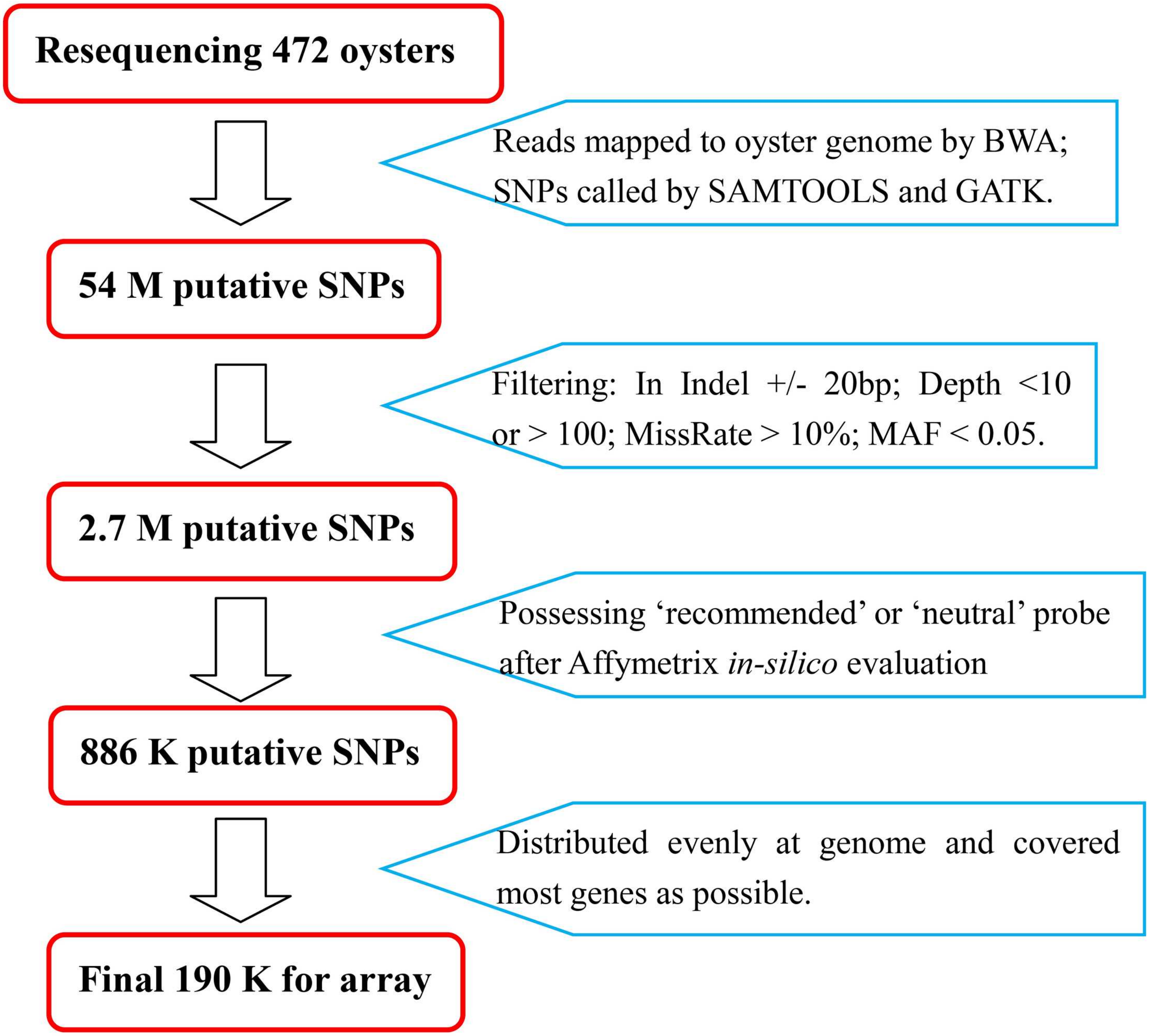
Resequencing 472 oysters
Reads mapped to oyster genome by BWA; SNPs called by SAMTOOLS and GATK.
54 M putative SNPs
Filtering: In Indel +/- 20bp; Depth <10 or > 100; MissRate > 10%; MAF < 0.05.
2.7 M putative SNPs
Possessing 'recommended' or 'neutral' probe after Affymetrix *in-silico* evaluation
886 K putative SNPs
Distributed evenly at genome and covered most genes as possible.
Final 190 K for array

Fig 1. Flow diagram for the SNP selection steps with major criteria.

https://doi.org/10.1371/journal.pone.0174007.g001

homozygous genotype was formed; (iv) "OTV," off-target variants—three good clusters were formed, with a single additional off-target cluster caused by variants in the SNP flanking region; (v) "CallRateBelowThreshold"—the SNP call rate was below the threshold (0.970), but other cluster properties were above the threshold; and (vi) "Other"—the SNPs were not grouped into any of the previous categories. We found that 101,193 SNPs (representing 53.1% of the total on-chip SNPs) were polymorphic, with three observed SNP genotypes (Table 1). Together with the 32,791 "NoMinorHom" SNPs, 133,984 (~70%) of the SNPs were validated as polymorphic and successfully converted on the chip.

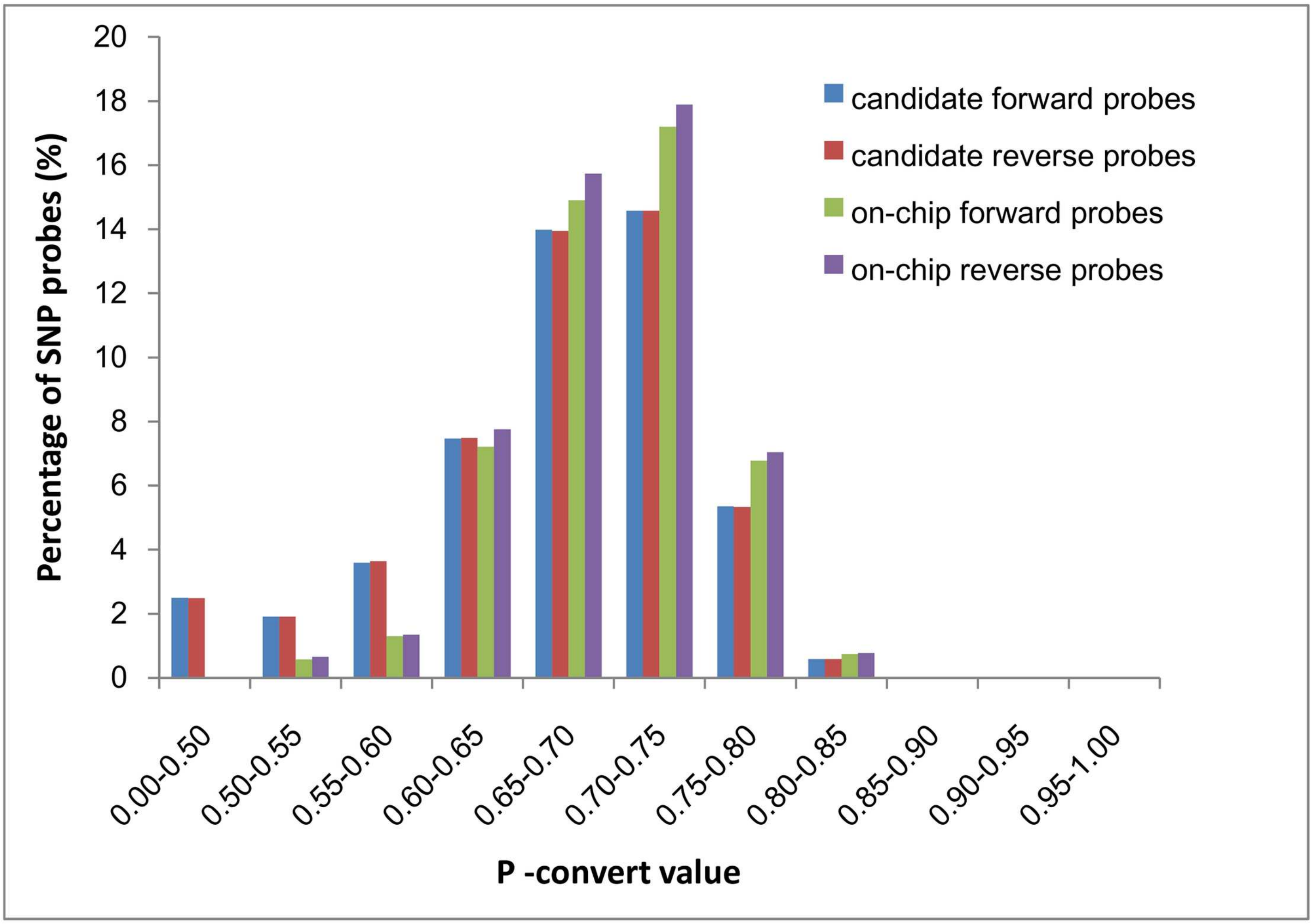

Fig 2. Distribution of the p-convert values for candidate and on-chip probes.

https://doi.org/10.1371/journal.pone.0174007.g002

Statistical analysis of the converted SNPs

Transition SNPs comprised 73% of the polymorphic SNPs, including 49,108 A/G and 48,442 C/T; transversion SNPs involved 14,863 A/C, 15,047 G/T, 4,437 A/T, and 2,087 C/G (Table 2). The average conversion rate was 0.70, and the conversion rate of each type of SNP was between 0.66 and 0.79. According to the positions of the SNPs in the genome or their effects on the predicted

Table 1. The counts of the SNPs clustering categories.

SNP Category	Probe No.	Percent	SNP No.	Percent
PolyHighResolution	101,722	52.8	101,193	53.1
NoMinorHom	33,033	17.1	32,791	17.2
MonoHighResolution	4,177	2.2	4,145	2.2
OTV	3,472	1.8	3,409	1.8
CallRateBelowThreshold	17,336	9.0	17,015	8.9
Other	33,049	17.1	31,867	16.7
Total	192,789	100.0	190,420	100.0

https://doi.org/10.1371/journal.pone.0174007.t001

Table 2. The counts of the SNPs types.

SNP type	On-chip	Percent	Converted	Percent	Conversion Rate
A/G	68,655	36.1	49,108	36.7	0.72
C/T	67,800	35.6	48,442	36.2	0.71
G/T	22,345	11.7	15,047	11.2	0.67
A/C	22,241	11.7	14,863	11.1	0.67
A/T	6,753	3.5	4,437	3.3	0.66
C/G	2,626	1.4	2,087	1.6	0.79
Total	190,420	100	133,984	100	0.70

https://doi.org/10.1371/journal.pone.0174007.t002

protein sequences, ~44.4% (59,485) of the converted SNPs were in the coding region—these SNPs included 39,579 synonymous SNPs and 19,906 non-synonymous SNPs (S3 Dataset). The SNPs in the intron region, 2-kb upstream/downstream of a gene, and in the intergenic region accounted for 22.5%, 16.1%, and 17.0% of the converted SNPs, respectively (Table 3). Thus, 111,158 (83%) SNPs were in or near a predicted coding gene of the oyster genome and they may facilitate future gene-related analysis of oysters. The on-chip SNPs were located in 4,315 genome scaffolds that spanned a total length of 539.7 million; the average interval spacing of the SNPs was 2,960 bp. The converted SNPs were located in 3,595 scaffolds with a length of ~509.4 million; the average interval spacing was 4,210 bp. To assess the evenness of the SNPs, we calculated the distribution of the converted SNPs. We found that 12,458 (9.6%) SNPs had a small inter-SNP spacing (<200 bp) and 10,487 (8.0%) SNPs had a small SNP spacing (200–500 bp) (Fig 3). In addition, 25,756 (19.8%), 36,547 (28%), 15,959 (12.2%), 8,701 (6.7%), and 5,246 (4.0%) SNPs had a medium marker spacing of 500–1000 bp, 1000–2000 bp, 2000–3000 bp, 3000–4000 bp, and 4000–5000 bp, respectively. Cumulatively, 37.4% SNPs had a marker spacing of <1000 bp, 51.0% SNPs had a marker spacing of 1001–5000 bp, and 11.6% SNPs had a marker spacing of >5000 bp. Moreover, the 111,158 SNPs were distributed in 21,050 coding genes, with an average of 5.3 SNPs per gene. Of the genes, 2,750 (13.1%) had only one SNP converted on the chip,

Table 3. Summary of the SNPs according to their positions or functions.

SNP region/function	On-chip	Converted
Coding region		
synonymous	46,285	39,579
missense	24,215	19,604
stop_gained	245	186
start_lost	57	41
stop_lost	45	30
stop_retained	45	40
initiator_codon	6	5
Intron region		
intron	43,304	27,790
splice_donor	83	61
splice_acceptor	63	45
splice_region	2,745	2,259
2Kbp-up/down-stream of a gene		
up	17,630	11,880
down	15,133	9,638
Intergenic region	40,564	22,826

https://doi.org/10.1371/journal.pone.0174007.t003

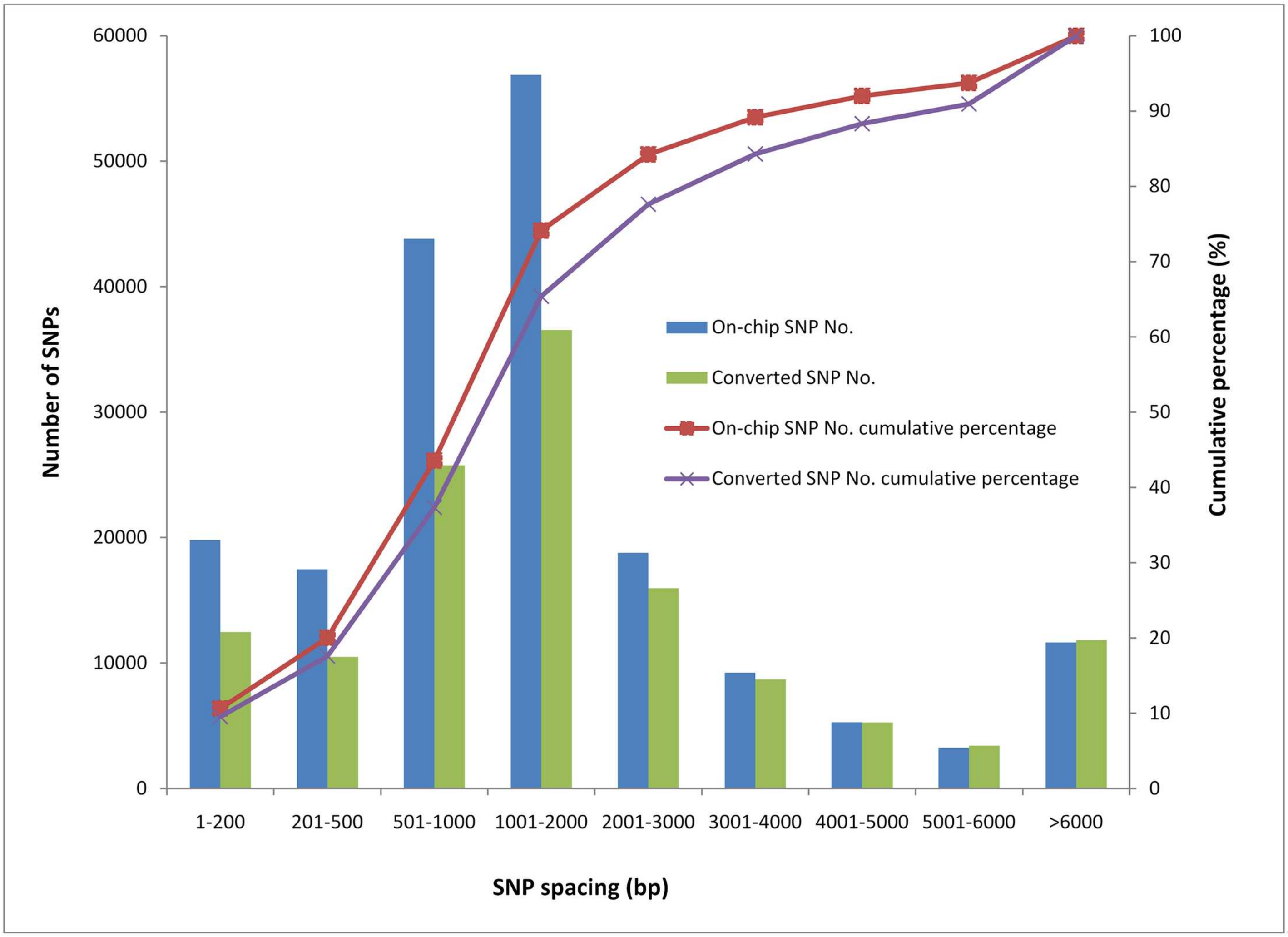

Fig 3. Distribution of the interval spacing of the SNPs on the array.

https://doi.org/10.1371/journal.pone.0174007.g003

11,460 (54.4%) had 2–5 SNPs converted on the chip, 4,561 (21.7%) had 6–10 SNPs converted on the chip, and 2,279 (10.8%) had more >10 SNPs converted on the chip.

Concordance of the SNP genotypes between array and re-sequencing

To further evaluate the quality of the converted SNPs, we compared the genotypes obtained from the SNP array with those obtained through high-depth re-sequencing. We found that 60,805 SNPs (representing 45.4% of the total on-chip SNPs) had a concordance rate of 1.000, 31,181 SNPs (representing 23.3% of the total on-chip SNPs) had a concordance rate of 0.971–0.977, and 14,777 SNPs (representing 11.0% of the total on-chip SNPs) had a concordance rate of 0.950–0.955 (Fig 4). Cumulatively, 68.7% of the converted SNPs had a concordance rate of >0.971, and the mean concordance rates for the converted SNPs and all the on-chip SNPs were 0.966 and 0.927, respectively. The non-concordance rate (~0.034) implied the presence of errors either in the array or in the GBS-derived genotypes; however, the data showed that the oyster SNP array and the GBS technology were both appropriate for oyster SNP genotyping and provided high-quality genotyping outcomes with an average accuracy rate of >0.96.

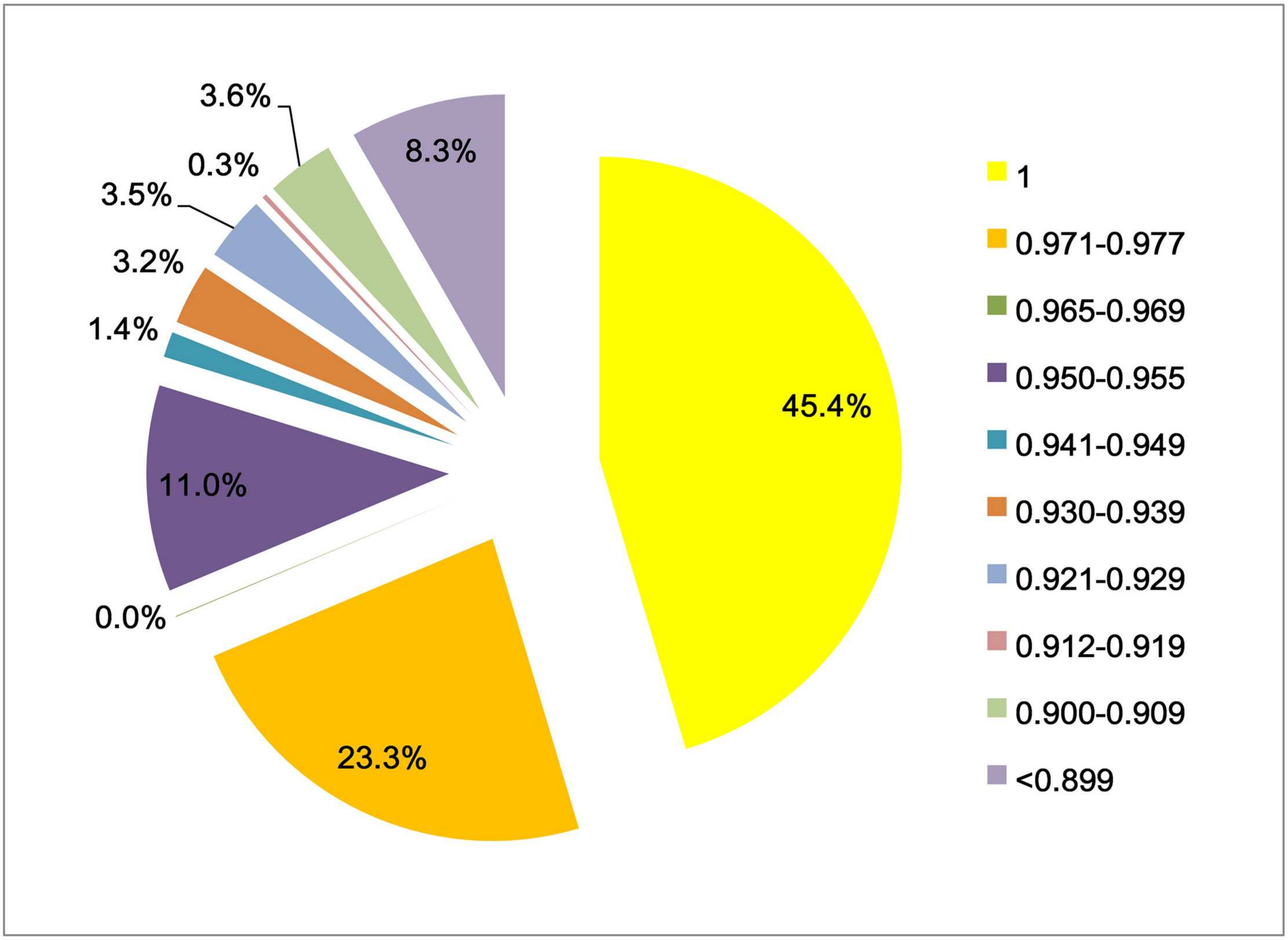

Fig 4. Distribution of the concordance rate of the SNPs on the array.

https://doi.org/10.1371/journal.pone.0174007.g004

Evaluation of the genotyping accuracy according to family data

We further evaluated the genotyping accuracy using data obtained from the two parents and 24 off-springs of a full-sib family. The numbers of genotype combinations of the two parents that could be used in linkage analysis for types "AA × AB", "AB × AA" and "AB × AB" were 20,716, 20,814, and 10,271, respectively (Table 4). No errors were detected in the markers of

Table 4. Estimation of genotyping accuracy by family data.

Parents genotypes	All no.	Error no.	Error no. Percentage	Call rate	Concordance rate	Error rate
AA × AA	73,246	2,040	2.8	0.996	0.971	0.002
AA × AB	20,716	2,199	10.6	0.995	0.964	0.025
AA × BB	8,842	3,342	37.8	0.994	0.943	0.235
AB × AA	20,814	2,276	10.9	0.995	0.962	0.026
AB × AB	10,271	0	0.0	0.995	0.962	0.000
	133,889	9,857	7.4	0.995	0.966	0.025

https://doi.org/10.1371/journal.pone.0174007.t004

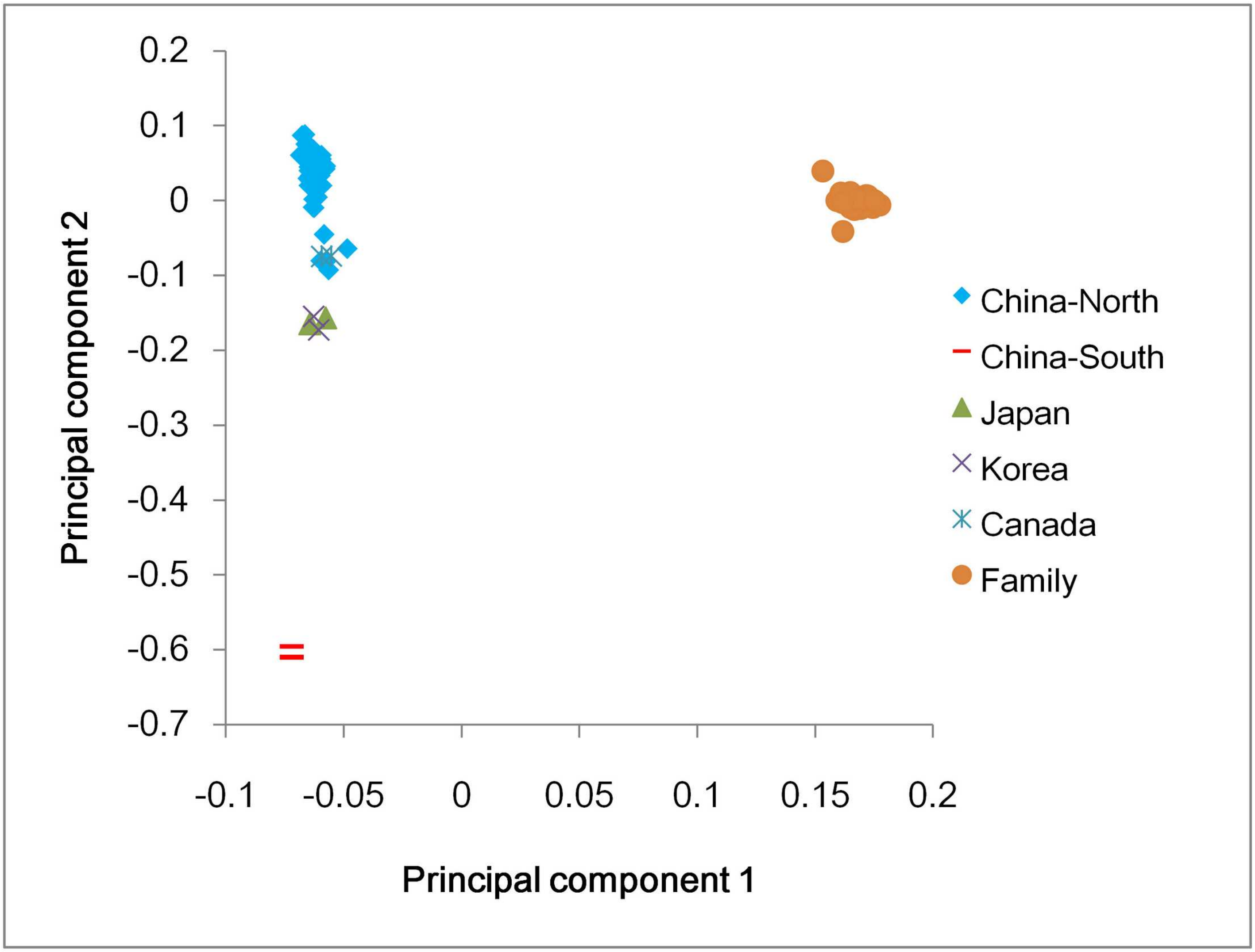

Fig 5. Principal component analysis of all samples. The first principal component (PC1) was assigned to X axis, and the second principal component (PC2) was assigned to Y axis. “China-North”, “China-South”, “Japan”, “Korea”, and “Canada” represented the Pacific oysters collected in northern China, southern China, Japan, Korea, and Canada, respectively. “Family” represented the parents and 24 off-springs of a full-sib family. The parents of the full-sib family were also collected in northern China.

https://doi.org/10.1371/journal.pone.0174007.g005

type “AB × AB” and the error rate for types “AA × AB” and “AB × AA” was ~0.025. In the case of type “AA × BB”, all off-spring genotypes were expected to be “AB” heterozygosity; however, some homozygous individuals were detected and thus a very high genotype error rate was observed in markers of type “AA × BB”. We detected 9,857 SNPs with unexpected genotypes; these SNPs accounted for 7.4% of the 133,889 SNPs that could be typed in both parents. For each individual, the number of unexpected genotypes ranged from to 3,228 to 3,445, and the average accuracy rate was 0.975. The average CR and concordance rate of the 9,857 SNPs were lower than the mean levels, indicating that these SNPs could not be well typed and that their genotyping results must be treated with caution in subsequent analysis. The oyster genome is highly polymorphic; hence, we deduced that these SNPs may be artificial and that the genotyping errors may result from the unspecific binding of probes.

Population structure analysis

The overall genome similarity of any two samples and the population stratification has frequently been estimated in population genetics and GWAS analyses [51]. Based on the genotypes of the 133,984 converted SNPs, the genetic differentiation was directly observed and the samples were divided into three groups according to the first and second principle component (Fig 5). Individuals in the same family clustered together and were always distinct from others. Pacific oysters collected in southern China, which were identified as subspecies of *C. gigas* [52], clearly differed from Pacific oysters collected in northern China, Japan, Korea, and Canada. The largest group consisted of Pacific oysters collected in northern China, Canada, Korea, and Japan; moreover, the samples collected in northern China could be further distinguished from those collected in Korea and Japan (Fig 5, S5 Dataset). The sample size in our present study was not large enough for comprehensive research; nevertheless, our results were in concordance with our expectations and may, to some extent, provide an insight into the population genetics of Pacific oysters.

Conclusions

In this study, we have established the largest commercially available Pacific oyster SNP array, composed of 190 K SNPs. Genotyping of 96 oysters revealed that ~133 K SNPs (~70%) were polymorphic and successfully converted on the chip. The SNPs were distributed evenly throughout the oyster genome, located in 3,595 scaffolds; the average interval spacing was 4,210 bp. In addition, 111,158 SNPs were distributed in 21,050 coding genes, with an average of 5.3 SNPs per gene. Comparison of the array-derived genotypes with those obtained through re-sequencing revealed that the mean concordance rate was 0.966. Moreover, evaluation based on the genotypes of two parents and 24 off-springs of a full-sib family revealed that the average accuracy rate was 0.975. Our results indicate that the oyster SNP array constitutes an alternative platform for genome-wide SNP genotyping and represents a valuable tool for research into the genetics of *C. gigas*.

Supporting information

S1 Dataset. The sequences of all the on-chip SNPs.
(GZ)

S2 Dataset. The clustering category for each SNP probe.
(GZ)

S3 Dataset. The function annotation of all the on-chip SNPs.
(GZ)

S4 Dataset. The SNP genotypes for the 96 samples.
(GZ)

S5 Dataset. The PCA data of the 96 samples.
(GZ)

S1 Fig. Examples of the six SNP cluster categories. SNPs could be classified into six categories according to the probeset clustering properties: (i) 'PolyHighResolution'; (ii) 'NoMinorHom'; (iii) 'MonoHighResolution'; (iv) 'OTV'; (v) 'CallRateBelowThreshold'; and (vi) 'Other'.
(TIF)

Acknowledgments

We thank the staff at Compass Biotechnology (Beijing, China) for their help with microarray data analysis. We also thank the high performance computing center (HPCC) of the Institute of Oceanology for providing assistance with bioinformatics software installation.

Author Contributions

Conceptualization: HQ LL GZ.

Data curation: HQ KS.

Formal analysis: HQ KS.

Funding acquisition: GZ LL.

Investigation: HQ CL WW BL.

Methodology: HQ KS.

Project administration: HQ LL GZ.

Resources: HQ CL WW BL LL.

Software: HQ KS.

Supervision: HQ LL GZ.

Validation: HQ KS CL WW BL.

Visualization: HQ KS.

Writing – original draft: HQ.

Writing – review & editing: HQ.

References

1. Quilang J, Wang S, Li P, Abernathy J, Peatman E, Wang Y, et al. Generation and analysis of ESTs from the eastern oyster, Crassostrea virginica Gmelin and identification of microsatellite and SNP markers. Bmc Genomics. 2007; 8:157. https://doi.org/10.1186/1471-2164-8-157 PMID: 17559679
2. Grizel H, Heral M. Introduction into France of the Japanese Oyster (Crassostrea-Gigas). J Conseil. 1991; 47(3):399–403.
3. Ward RD, English LJ, McGoldrick DJ, Maguire GB, Nell JA, Thompson PA. Genetic improvement of the Pacific oyster Crassostrea gigas (Thunberg) in Australia. Aquac Res. 2000; 31(1):35–44.
4. Lavoie RE. Oyster culture in North America: history, present and future. The 1st International Oyster Symposium Proceedings, Oyster Research Institute News. 2005;No.17
5. Schmidt A, Wehrmann A, Dittmann S. Population dynamics of the invasive Pacific oyster Crassostrea gigas during the early stages of an outbreak in the Wadden Sea (Germany). Helgoland Mar Res. 2008; 62(4):367–76.
6. Wrange AL, Valero J, Harkestad LS, Strand O, Lindegarth S, Christensen HT, et al. Massive settlements of the Pacific oyster, Crassostrea gigas, in Scandinavia. Biol Invasions. 2010; 12(6):1453–8.
7. Alunno-Bruscia M, Bourles Y, Maurer D, Robert S, Mazurie J, Gangnery A, et al. A single bio-energetics growth and reproduction model for the oyster *Crassostrea gigas* in six Atlantic ecosystems. J Sea Res. 2011; 66(4):340–8.
8. Olivier D, Heinecken L, Jackson S. Mussel and oyster culture in Saldanha Bay, South Africa: potential for sustainable growth, development and employment creation. Food Secur. 2013; 5(2):251–67.
9. Devos A, Dallas LJ, Voiseux C, Lecomte-Pradines C, Jha AN, Fievet B. Assessment of growth, genotoxic responses and expression of stress related genes in the Pacific oyster Crassostrea gigas following chronic exposure to ionizing radiation. Mar Pollut Bull. 2015; 95(2):688–98. https://doi.org/10.1016/j.marpolbul.2015.03.039 PMID: 25843441

10. Kochmann J, Crowe TP. Effects of native macroalgae and predators on survival, condition and growth of non-indigenous Pacific oysters (Crassostrea gigas). J Exp Mar Biol Ecol. 2014; 451:122–9.

11. Fleury E, Huvet A. Microarray Analysis Highlights Immune Response of Pacific Oysters as a Determinant of Resistance to Summer Mortality. Mar Biotechnol. 2012; 14(2):203–17. https://doi.org/10.1007/s10126-011-9403-6 PMID: 21845383

12. Sauvage C, Boudry P, de Koning DJ, Haley CS, Heurtebise S, Lapegue S. QTL for resistance to summer mortality and OsHV-1 load in the Pacific oyster (Crassostrea gigas). Anim Genet. 2010; 41(4):390–9. https://doi.org/10.1111/j.1365-2052.2009.02018.x PMID: 20096029

13. Schmitt P, Santini A, Vergnes A, Degremont L, de Lorgeril J. Sequence polymorphism and expression variability of Crassostrea gigas immune related genes discriminate two oyster lines contrasted in term of resistance to summer mortalities. PLoS One. 2013; 8(9):e75900. https://doi.org/10.1371/journal.pone.0075900 PMID: 24086661

14. Mortensen S, Strand A, Bodvin T, Alfjorden A, Skar CK, Jelmert A, et al. Summer mortalities and detection of ostreid herpesvirus microvariant in Pacific oyster Crassostrea gigas in Sweden and Norway. Dis Aquat Organ. 2016; 117(3):171–6. https://doi.org/10.3354/dao02944 PMID: 26758650

15. Keightley J, von der Heyden S, Jackson S. Introduced Pacific oysters Crassostrea gigas in South Africa: demographic change, genetic diversity and body condition. Afr J Mar Sci. 2015; 37(1):89–98.

16. An HS, Lee JW, Kim WJ, Lim HJ, Kim EM, Byun SG, et al. Comparative genetic diversity of wild and hatchery-produced Pacific oyster (Crassostrea gigas) populations in Korea using multiplex PCR assays with nine polymorphic microsatellite markers. Genes Genom. 2013; 35(6):805–15.

17. Rohfritsch A, Bierne N, Boudry P, Heurtebise S, Cornette F, Lapegue S. Population genomics shed light on the demographic and adaptive histories of European invasion in the Pacific oyster, Crassostrea gigas. Evol Appl. 2013; 6(7):1064–78. https://doi.org/10.1111/eva.12086 PMID: 24187588

18. Meistertzheim AL, Arnaud-Haond S, Boudry P, Thebault MT. Genetic structure of wild European populations of the invasive Pacific oyster Crassostrea gigas due to aquaculture practices. Mar Biol. 2013; 160(2):453–63.

19. Dundon WG, Arzul I, Omnes E, Robert M, Magnabosco C, Zambon M, et al. Detection of Type 1 Ostreid Herpes variant (OsHV-1 mu var) with no associated mortality in French-origin Pacific cupped oyster Crassostrea gigas farmed in Italy. Aquaculture. 2011; 314(1–4):49–52.

20. Roque A, Carrasco N, Andree KB, Lacuesta B, Elandaloussi L, Gairin I, et al. First report of OsHV-1 microvar in Pacific oyster (Crassostrea gigas) cultured in Spain. Aquaculture. 2012; 324:303–6.

21. Degremont L, Guyader T, Tourbiez D, Pepin JF. Is horizontal transmission of the Ostreid herpesvirus OsHV-1 in Crassostrea gigas affected by unselected or selected survival status in adults to juveniles? Aquaculture. 2013; 408:51–7.

22. Hubert S, Hedgecock D. Linkage maps of microsatellite DNA markers for the Pacific oyster Crassostrea gigas. Genetics. 2004; 168(1):351–62. https://doi.org/10.1534/genetics.104.027342 PMID: 15454548

23. Li L, Guo X. AFLP-based genetic linkage maps of the pacific oyster Crassostrea gigas Thunberg. Mar Biotechnol (NY). 2004; 6(1):26–36.

24. Plough LV, Hedgecock D. Quantitative trait locus analysis of stage-specific inbreeding depression in the Pacific oyster Crassostrea gigas. Genetics. 2011; 189(4):1473–86. https://doi.org/10.1534/genetics.111.131854 PMID: 21940682

25. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012; 490(7418):49–54. https://doi.org/10.1038/nature11413 PMID: 22992520

26. Brookes AJ. The essence of SNPs. Gene. 1999; 234(2):177–86. PMID: 10395891

27. Park YJ, Lee JK, Kim NS. Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. Molecules. 2009; 14:4546–69. https://doi.org/10.3390/molecules14114546 PMID: 19924085

28. Wang J, Qi H, Li L, Que H, Wang D, Zhang G. Discovery and validation of genic single nucleotide polymorphisms in the Pacific oyster Crassostrea gigas. Mol Ecol Resour. 2015; 15(1):123–35. https://doi.org/10.1111/1755-0998.12278 PMID: 24823694

29. Zhong XX, Li Q, Yu H, Kong LF. Development and validation of single-nucleotide polymorphism markers in the Pacific oyster, Crassostrea gigas, using high-resolution melting analysis. J World Aquacult Soc. 2013; 44(3):455–65.

30. Sauvage C, Bierne N, Lapegue S, Boudry P. Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster Crassostrea gigas. Gene. 2007; 406(1–2):13–22. https://doi.org/10.1016/j.gene.2007.05.011 PMID: 17616269

31. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 1998; 280 (5366):1077–82. PMID: 9582121

32. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. Bmc Genomics. 2013; 14:59. https://doi.org/10.1186/1471-2164-14-59 PMID: 23356797

33. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009; 4(8)

34. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 2009; 4(4):e5350. https://doi.org/10.1371/journal.pone.0005350 PMID: 19390634

35. McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, et al. A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. Plos Genetics. 2012; 8(1):e1002451. https://doi.org/10.1371/journal.pgen.1002451 PMID: 22253606

36. 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. https://doi.org/10.1186/1756-0500-7-135 PMID: 24618043

37. Xu J, Zhao Z, Zhang X, Zheng X, Li J, Jiang Y, et al. Development and evaluation of the first high-throughput SNP array for common carp (Cyprinus carpio). Bmc Genomics. 2014; 15:307. https://doi.org/10.1186/1471-2164-15-307 PMID: 24762296

38. Houston RD, Taggart JB, Cezard 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. https://doi.org/10.1186/1471-2164-15-90 PMID: 24524230

39. 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. 2015; 15(3):662–72. https://doi.org/10.1111/1755-0998.12337 PMID: 25294387

40. Singh N, Jayaswal PK, Panda K, Mandal P, Kumar V, Singh B, et al. Single-copy gene based 50 K SNP chip for genetic studies and molecular breeding in rice. Sci Rep. 2015; 5:11600. https://doi.org/10.1038/srep11600 PMID: 26111882

41. Lee YG, Jeong N, Kim JH, Lee K, Kim KH, Pirani A, et al. Development, validation and genetic analysis of a large soybean SNP genotyping array. Plant J. 2015; 81(4):625–36. https://doi.org/10.1111/tpj.12755 PMID: 25641104

42. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. Bmc Genomics. 2014; 15:823. https://doi.org/10.1186/1471-2164-15-823 PMID: 25266061

43. Bassil NV, Davis TM, Zhang H, Ficklin S, Mittmann M, Webster T, et al. Development and preliminary evaluation of a 90 K Axiom(R) SNP array for the allo-octoploid cultivated strawberry Fragaria x ananassa. Bmc Genomics. 2015; 16:155. https://doi.org/10.1186/s12864-015-1310-1 PMID: 25886969

44. Lapegue S, Harrang E, Heurtebise S, Flahauw E, Donnadieu C, Gayral P, et al. Development of SNP-genotyping arrays in two shellfish species. Mol Ecol Resour. 2014; 14(4):820–30. https://doi.org/10.1111/1755-0998.12230 PMID: 24447767

45. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011; 6(5):e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248

46. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011; 12(6):443–51. https://doi.org/10.1038/nrg2986 PMID: 21587300

47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

48. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

50. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–9. https://doi.org/10.1038/ng1847 PMID: 16862161

51. Bouaziz M, Ambroise C, Guedj M. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. PLoS One. 2011; 6(12):e28845. https://doi.org/10.1371/journal.pone.0028845 PMID: 22216125

52. Wang H, Qian L, Liu X, Zhang G, Guo X. Classification of a common cupped oyster from southern China. J Shellfish Res. 2010; 29(4):857–66.