Development of a medium density combined-species SNP array for Pacific and European oysters (Crassostrea gigas and Ostrea edulis)


Published in:
G3: Genes, Genomes, Genetics

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright © 2017 Gutierrez et al. This is an open access article published under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen’s institutional repository that provides access to Queen’s research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen’s academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date:26. Oct. 2023
Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (Crassostrea gigas and Ostrea edulis)

Alejandro P. Gutierrez,* Frances Turner,† Karim Gharbi,† Richard Talbot,† Natalie R. Lowe,* Carolina Peñaloza,* Mark McCullough,‡ Paulo A. Prodöhl,‡ Tim P. Bean,§ and Ross D. Houston*,1

*The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, United Kingdom, †Edinburgh Genomics, Ashworth Laboratories, University of Edinburgh, EH9 3FL, United Kingdom, ‡Institute for Global Food Security, School of Biological Sciences, Queen’s University Belfast, BT7 1NN, United Kingdom, and §Centre for Environment Fisheries and Aquaculture Science, Cefas Weymouth Laboratory, Dorset DT4 8UB, United Kingdom

ABSTRACT SNP arrays are enabling tools for high-resolution studies of the genetic basis of complex traits in farmed and wild animals. Oysters are of critical importance in many regions from both an ecological and economic perspective, and oyster aquaculture forms a key component of global food security. The aim of our study was to design a combined-species, medium density SNP array for Pacific oyster (Crassostrea gigas) and European flat oyster (Ostrea edulis), and to test the performance of this array on farmed and wild populations from multiple locations, with a focus on European populations. SNP discovery was carried out by whole-genome sequencing (WGS) of pooled genomic DNA samples from eight C. gigas populations, and restriction site-associated DNA sequencing (RAD-Seq) of 11 geographically diverse O. edulis populations. Nearly 12 million candidate SNPs were discovered and filtered based on several criteria, including preference for SNPs segregating in multiple populations and SNPs with monomorphic flanking regions. An Affymetrix Axiom Custom Array was created and tested on a diverse set of samples (n = 219) showing ~27 K high quality SNPs for C. gigas populations, and restriction site-associated DNA sequencing (RAD-Seq) of 11 geographically diverse O. edulis populations. A high proportion of SNPs were segregating in each of the populations, and the array was used to detect population structure and levels of linkage disequilibrium (LD). Further testing of the array on three C. gigas nuclear families (n = 165) revealed that the array can be used to clearly distinguish between both families based on identity-by-state (IBS) clustering parental assignment software. This medium density, combined-species array will be publicly available through Affymetrix, and will be applied for genome-wide association and evolutionary genetic studies, and for genomic selection in oyster breeding programs.

KEYWORDS
Pacific oyster
flat oyster
single nucleotide polymorphism (SNP)
array
aquaculture

Oyster farming is one of the most important aquaculture activities worldwide, providing a socioeconomic contribution to many coastal communities. Among the numerous farmed oyster species, the Pacific oyster (Crassostrea gigas) is one of the most widely cultivated, with a global annual production estimated at 583 K tons in 2015 (FAO 2017). Starting in the 1960s, C. gigas was successfully introduced from Japan to all continents for cultivation (Troost 2010) due to its high acclimation ability, rapid growth, and high production, and as an alternative to replace the flat oyster farms affected by persistent disease outbreaks (Pernet et al. 2016). Accordingly, the European flat oyster (Ostrea edulis), an endemic species to Europe, has suffered a decrease in global production from 30 K tons in 1960 to 3 K tons produced in 2014. O. edulis is now a target for conservation efforts to help restore native populations (Lallis et al. 2010), and is also a niche aquaculture product, particularly in Europe and the USA.

In the past decade, there has been increasing interest from researchers and industry in the development of genomic resources for oysters, mainly because of the economic and ecological importance of both C. gigas and O. edulis. The genomic toolbox for C. gigas includes a moderate number of genetic markers, such as microsatellites (Li et al. 2003; Sekino et al. 2003) and SNPs (Fleury et al. 2009; Sauvage et al. 2007; Wang et al. 2015). Low density linkage maps have been developed, containing both microsatellites and SNPs (Hedgecock et al. 2015; Hubert and Hedgecock 2004). In addition, quantitative trait loci (QTL)
analyses have been carried out to identify genomic regions associated with desirable traits for aquaculture (Sauvage et al. 2010; Guo et al. 2012; Zhong et al. 2014). In addition, a reference genome sequence assembly is available for C. gigas (Zhang et al. 2012), although a number of putative assembly errors have been identified (Hedgecock et al. 2015). In contrast, genomic tools and resources are scarce for O. edulis, and only a limited number of markers, mostly microsatellites and amplified fragment length polymorphisms (AFLPs), have been utilized for the development of a linkage map (Lallias et al. 2007, 2009). Recently, the generation of genomic resources led to the development of a database containing genomic and transcriptome resources for O. edulis (Pardo et al. 2016; Vera et al. 2016).

SNPs have become the marker of choice in genetics research due to their high abundance, codominant mode of inheritance, ease of high-throughput discovery, and low cost of genotyping per locus. Next-generation sequencing technologies enable efficient identification of many thousands of SNPs in a single experiment using either WGS or reduced representation approaches such as RAD-Seq (Baird et al. 2008; Davey et al. 2011). While the medium density SNP arrays typically generated by direct genotyping-by-sequencing approaches have been widely applied in aquaculture species (Robledo et al. 2017), SNP arrays can offer a higher density genotyping platform that is simpler to use. SNP arrays have been developed for most terrestrial livestock species such as cattle, pig, and chicken (Matukumalli et al. 2009; Roses et al. 2009; Kranis et al. 2013), and also for farmed finfish species such as Atlantic salmon, rainbow trout, catfish, and carp among others (Houston et al. 2014; Yánez et al. 2016; Palti et al. 2015; Liu et al. 2014; Xu et al. 2014). These arrays have formed the basis of genome-wide association studies for traits of economic importance such as resistance to pathogens (Geng et al. 2015; Correa et al. 2015; Tsai et al. 2016) and the application of genomic selection in aquaculture breeding (Odegård and Meuwissen 2012; Tsai et al. 2015, 2016; Vallejo et al. 2016).

For oyster species, low density SNP arrays for C. gigas and O. edulis have been developed, with 384 markers per species (Lapègue et al. 2014), and these have been applied for parentage assignment. In addition, a C. gigas-specific high density array was recently developed, which contains ~134 K SNP markers shown to be polymorphic across populations sampled from China, Japan, Korea, and Canada (Qi et al. 2017). However, a medium density, combined-species platform is a worthy addition to the genomic toolbox for oysters because: (i) the performance of the higher density (133 K) array in farmed C. gigas populations from other global regions (e.g., Europe) is not known, (ii) medium density arrays are adequate for many genetics and breeding studies at substantially lower cost than high density arrays, and (iii) there is not yet a medium or high density genotyping platform for O. edulis. The major aim of the current study was to design and test a medium density, combined-species SNP array for two key oyster species, C. gigas and O. edulis, and to test the performance of the array on hatchery and wild populations from multiple locations, as well as nuclear families from pair-crosses.

**MATERIALS AND METHODS**

Sample collection and sequencing

The DNA sequencing protocols for SNP discovery were tailored to the status of genomic tools available for the two species. Since C. gigas has a reference genome sequence (Zhang et al. 2012), a whole-genome resequencing approach was taken with reads subsequently aligned to the reference assembly as described below. There was no reference sequence available for O. edulis, so a RAD-Seq approach was taken since this is suitable for de novo assembly and discovery of SNPs within RAD loci (Baird et al. 2008).

Samples from eight C. gigas populations from different geographical locations (primarily from hatcheries in the UK and France) were obtained, each comprising 13-47 individuals (Table 1). These included a population of 16 samples from lines of oysters that had been selected for resistance to Oyster Herpes Virus by Ifremer (France). Genomic DNA from all individuals was extracted via the CTAB (cetyl trimethylammonium bromide) protocol described by Richards et al. (2013). Briefly, oyster tissue was incubated at 56°C in lysis solution (3% CTAB, 100 mM Tris-HCl, pH 7.5, 25 mM EDTA, and 2 mM NaCl) with 0.2 mg/ml proteinase K and 5 μl of RNase (10 mg/ml). After lysis, a chloroform extraction was performed twice and three volumes of CTAB dilution solution were added (1% CTAB, 50 mM Tris-HCl, pH 7.5, and 10 mM EDTA, pH 8). The pellet was then washed in 0.4 M NaCl in TE, resuspended in 1.42 M NaCl in TE, and finally precipitated overnight in 1 ml ethanol (99%) at -4°C. Within each population, DNA samples were then pooled in equimolar concentrations, and these pools were prepared for WGS using the TruSeq Nano DNA Library Prep kit (Illumina, San Diego). Libraries were sequenced across five lanes of Illumina Hisequation 2500 to produce 125 bp paired end reads.

Samples from 11 O. edulis wild populations from diverse geographical locations were obtained (Table 1). Each population sample comprised 13–15 individuals, and genomic DNA had previously been extracted from these samples using a phenol-chloroform method. Equimolar pools of genomic DNA were generated for each population and the pooled genomic DNA was digested using the endonuclease PstI. Standard RAD libraries were constructed in three replicates following the standard protocol described by Baird et al. (2008). Equimolar amounts of all libraries were combined and sequenced on a single Illumina Hisequation 2500 lane to produce 125 bp paired end reads.

SNP identification and filtering

C. gigas WGS reads were aligned to the C. gigas genome (GCA_000297895.1) using BWA-mem (v0.7.10) (Li and Durbin 2009) with the -M flag. Potential duplicated reads originating from PCR were then removed using Picard Tools (v1.69) MarkDuplicates and SAMTools (v1.2) (Li et al. 2009). Local realignment around indels was performed using the GATK (v3.4.0) (McKenna et al. 2010) and alignments with a quality phred score > 20 were retained. SNP calling was performed using PoPoolation2 (Koller et al. 2011), filtering to discard bases with a call quality phred score < 30.

O. edulis RAD-Seq reads were trimmed with Cutadapt (v1.7.1) (Martin 2011). Data from each of the three replicates described above were combined. Read 1 reads were clustered using ustats (v1.30) with the parameters “-m 2 -M 5 -H,” followed by cstacks (Catchen et al. 2013) with the parameter “-n 2,” to create consensus sequences for each locus. RAD loci absent from ≥8 of the 11 pooled samples were
discarded. Read 1 trimmed reads from each of the samples were then aligned to the set of RAD consensus sequences using BWA (v0.7.9a) (Li and Durbin 2009) (step 1). Reads mapping to each separate consensus sequence were then identified, and the corresponding read 2 sequences extracted from the trimmed data. These read 2 sequences for each locus were then assembled using IBDA-UD (Peng et al. 2012) (step 2). The read 1 consensus sequences and the associated assembled read 2 sequences for each locus were merged using flash (v1.2.2) (Magoč and Salzberg 2011). For SNP discovery, the trimmed sequences corresponding to each locus were then mapped to the merged consensus sequence using bwa (v0.7.6). Duplicate reads were marked using Picard tools (v1.115) and realignments around indels performed using GATK indel realigner (v3.4.0) (McKenna et al. 2010).

SNPs were identified and genotyped using PoPoolation2 and SAMtools (v1.3) pileup. Reads with a mapping quality phred score < 20 and bases with a call quality phred score < 20 were discarded.

### SNP selection for Axiom array design

A list of candidate SNPs from both species (containing 1,691,005 and 117,235 priority SNPs from C. gigas and O. edulis, respectively) was provided to Affymetrix as 71-mer nucleotide sequences from the forward strand, with the alleles at the target SNP highlighted at position 36. A “p-convert” value (representing the probability of a given SNP converting to a reliable SNP assay on the Axiom array system) was computed by Affymetrix for each submitted SNP sequence. Probes are assessed for each SNP in both the forward and reverse direction, in return each strand is designated as “recommended,” “neutral,” or “not recommended” based on p-convert values.

The list of recommended markers (1,316,870 SNPs for C. gigas and O. edulis combined) was much greater than the total capacity of the Axiom MyDesign custom array. Therefore, additional filtering steps were carried out. For C. gigas, starting from the 1,216,467 Affymetrix-recommended SNPs, those with evidence for a 20 bp flanking mono-morphic region covered by at least 36 reads from each pooled sample were retained (n = 186,948). For O. edulis, the Affymetrix-recommended SNPs (n = 100,403) were filtered so that each RAD locus contained a maximum of one SNP. When a RAD locus had multiple recommended SNPs, only the best SNP (based on the p-convert scores) was included (resulting in 59,976 candidate SNPs). Subsequently, to filter the SNPs to the required number for the array, SNPs for both species were selected according to the following additional filtering criteria: (i) highest p-convert values, (ii) even distribution across the reference genome (with at least 1000 bp distance between pairs of SNPs for C. gigas), and (iii) preference for those with a positive hit (minimum e-value $10^{-14}$) against the BLASTx NCBI NR database or against the C. gigas genome (for O. edulis). In addition, most A/T and C/G SNP transversions were discarded since these require double the space on the Affymetrix Axiom array platform. Additionally, 463 SNPs identified and validated by Hedgecock et al. (2015) passed the SNP filtering and scoring process and were included in the final array design.

### SNP array validation

A plate of 384 individual genomic DNA samples (274 C. gigas and 110 O. edulis) was sent to Edinburgh Genomics (Edinburgh, UK) for genotyping using the array. Of these 384 samples, 219 were used for testing and validating the array’s performance and quantifying the number of segregating SNPs in the various sampled populations. These included 109 C. gigas samples of individuals of unknown relatedness from eight populations (the same eight populations used for SNP discovery, plus an additional set of 28 broodstock oysters from Guernsey Sea Farms (Guernsey, UK). The validation samples also included 110 O. edulis samples corresponding to the 11 population samples used for SNP discovery (Table 1), with $n = 10$ from each population. The remaining 165 samples were offspring of three nuclear families derived from parents from Guernsey Sea Farms, reared at the Centre for Environment, Fisheries and Aquaculture Science (Cefas, UK). These were analyzed separately to test parentage assignment, genetic structure, and within-family linkage LD levels (see below).

Raw data containing the results of the intensity calculations (CEL files) was imported into the Axiom Analysis Suite (v2.0.035, Affymetrix) for quality control analysis and genotype calling. Samples with a dish quality control (DQC) value > 0.82 and QC call rate > 0.97 threshold (following the “Best Practices Workflow” recommended by Affymetrix), were considered to have passed the quality control assessment. The quality control analysis classifies the SNPs into categories according to their clustering performance with respect to various Axiom-generated quality control criteria: (i) “polymorphic high resolution” where the SNP passes all QC, (ii) “monomorphic high resolution” where the SNP passes all QC except the presence of a minor allele in two or more samples, (iii) “call rate below threshold” where genotype call rate is < 97%, (iv) “no minor homozygote” where the SNP passes all QC but only two clusters are observed, (v) “off-target variant” where atypical cluster properties arise from variants in the SNP flanking region, and (vi) “other” where the SNP does not fall into any of the previous categories. For further analyses, only SNPs from categories (i) and (iv) were included and classified as “good quality,” as they are most likely to be reliable and informative SNPs.

---

**Table 1** Detail of populations sampled for sequencing and SNP discovery

<table>
<thead>
<tr>
<th>Population</th>
<th>Location (Lat, Long)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. gigas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guernsey, England</td>
<td>49.497, -2.502</td>
<td>47</td>
</tr>
<tr>
<td>Maldon, England</td>
<td>51.724, 0.710</td>
<td>15</td>
</tr>
<tr>
<td>Sea Salter, England</td>
<td>51.378, 1.212</td>
<td>13</td>
</tr>
<tr>
<td>Ifremer, France</td>
<td>n/a</td>
<td>16</td>
</tr>
<tr>
<td>Hatchery 1 (Marinove), France</td>
<td>46.987, -2.238</td>
<td>29</td>
</tr>
<tr>
<td>Hatchery 2 (SATMAR), France</td>
<td>46.948, -2.052</td>
<td>26</td>
</tr>
<tr>
<td>Hatchery 3 (France Naissain), France</td>
<td>47.514, -2.666</td>
<td>29</td>
</tr>
<tr>
<td>Hatchery 4 (Novostrea), France</td>
<td>46.954, -2.044</td>
<td>28</td>
</tr>
<tr>
<td>O. edulis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatia</td>
<td>42.855, 17.688</td>
<td>14</td>
</tr>
<tr>
<td>Lough Foyle, Ireland</td>
<td>55.130, -7.087</td>
<td>15</td>
</tr>
<tr>
<td>Lake Grevelingen, The Netherlands</td>
<td>51.709, 4.017</td>
<td>15</td>
</tr>
<tr>
<td>Larne, Northern Ireland</td>
<td>54.817, -6.751</td>
<td>14</td>
</tr>
<tr>
<td>Mersea, England</td>
<td>51.776, 0.9646</td>
<td>15</td>
</tr>
<tr>
<td>Baie de Quiberon, France</td>
<td>47.548, -2.996</td>
<td>15</td>
</tr>
<tr>
<td>Rossmore (Cork), Ireland</td>
<td>51.883, -8.247</td>
<td>15</td>
</tr>
<tr>
<td>Sveio, Norway</td>
<td>59.519, 5.227</td>
<td>15</td>
</tr>
<tr>
<td>Swansea Bay, England</td>
<td>51.604, -3.981</td>
<td>15</td>
</tr>
<tr>
<td>Tralee, Ireland</td>
<td>52.316, -10.028</td>
<td>13</td>
</tr>
<tr>
<td>Damariscotta, Maine</td>
<td>44.028, -69.534</td>
<td>14</td>
</tr>
</tbody>
</table>

Lat, latitude; Long, longitude.
RESULTS AND DISCUSSION

Evaluation of the SNP array in C. gigas and O. edulis

The oyster array was evaluated in C. gigas by analyzing the “validation populations” of 109 samples corresponding to eight distinct populations from France and the UK (Table 2). All but one sample passed the DQC and genotype call rate ≥97% threshold. The classification of SNPs according to their quality showed that 68.2% (n = 27,697) had probes classified as good quality (either “Poly High Resolution” or “No Minor Hom”), which is similar to the percentage of informative markers obtained by the recently published C. gigas 134 K array (Qi et al. 2017). The MAF of these good quality SNPs (MAF > 0) in the combined 108 samples varied between 0.005 and 0.5 with a median of 0.18 (Table 2). From the 110 O. edulis samples genotyped (Table 3), two samples failed the DQC and genotype call rate ≥97% threshold, resulting in genotypes for 108 samples. A total of 74.6% of SNPs (n = 11,151) were classified as good quality as described above. The MAF of these good quality SNPs (combining all the 108 samples and SNPs with a MAF > 0) also varied between 0.005 and 0.5 with a median of 0.21 (Table 3).

Within-population segregation of SNPs: The segregation of the SNPs was evaluated within each of the eight genotyped C. gigas population samples. From the 27,697 high quality SNPs defined across all population samples, the majority of SNPs (MAF > 0) were segregating within each of the populations (Figure S1), with an average of 22,486 SNPs segregating within each population, ranging from 20,141 (Hatchery 2) to 26,549 (Guernsey) (Table 2). Among the UK populations (sampled from Guernsey, Maldon, and Sea Salter), 19,613 SNPs were shared, while Guernsey had the highest number of exclusive SNPs (n = 2373) (Figure S2). This is likely to be due to the fact that the Guernsey population was the most highly represented within the sequenced populations used for SNP discovery (Table 1) and the validation samples selected for array design and production.

Table 2 Descriptive population genetic estimates for the sampled C. gigas populations included in the validation of the array

<table>
<thead>
<tr>
<th>Sample N</th>
<th># SNPs</th>
<th>Average MAF</th>
<th>Ho</th>
<th>He</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK (combined)</td>
<td>56</td>
<td>27,313</td>
<td>0.186</td>
<td>0.294</td>
</tr>
<tr>
<td>GSF + parents</td>
<td>38</td>
<td>26,549</td>
<td>0.19</td>
<td>0.308</td>
</tr>
<tr>
<td>Maldon</td>
<td>9</td>
<td>22,079</td>
<td>0.216</td>
<td>0.308</td>
</tr>
<tr>
<td>Sea Salter</td>
<td>9</td>
<td>22,821</td>
<td>0.214</td>
<td>0.317</td>
</tr>
<tr>
<td>Average within UK populations</td>
<td></td>
<td>23,816</td>
<td>0.207</td>
<td>0.311</td>
</tr>
<tr>
<td>France (combined)</td>
<td>52</td>
<td>26,891</td>
<td>0.182</td>
<td>0.240</td>
</tr>
<tr>
<td>Ifremer</td>
<td>13</td>
<td>23,010</td>
<td>0.203</td>
<td>0.312</td>
</tr>
<tr>
<td>Hatchery 1</td>
<td>10</td>
<td>21,479</td>
<td>0.217</td>
<td>0.321</td>
</tr>
<tr>
<td>Hatchery 2</td>
<td>10</td>
<td>20,141</td>
<td>0.221</td>
<td>0.322</td>
</tr>
<tr>
<td>Hatchery 3</td>
<td>10</td>
<td>21,730</td>
<td>0.215</td>
<td>0.302</td>
</tr>
<tr>
<td>Hatchery 4</td>
<td>9</td>
<td>22,052</td>
<td>0.214</td>
<td>0.317</td>
</tr>
<tr>
<td>Average within French populations</td>
<td></td>
<td>21,682</td>
<td>0.214</td>
<td>0.315</td>
</tr>
<tr>
<td>All populations (combined)</td>
<td>108</td>
<td>27,697</td>
<td>0.182</td>
<td>0.268</td>
</tr>
</tbody>
</table>

MAF, minor allele frequency; #, number; SNPs, single nucleotide polymorphisms; Ho, level of genetic variability in terms of observed heterozygosity; He, level of genetic variability in terms of expected heterozygosity; GSF, Guernsey Sea Farm.

Values obtained by the analysis of the combined dataset, not the average of the individual populations.

Values represent the within-population average.

Descriptive statistics and family assignment

Calculations of minor allele frequencies (MAF), levels of heterozygosity, discriminant analysis of principal components (DAPC), LD, and IBS followed by multi-dimensional scaling (MDS) were carried out using Plink (Purcell et al. 2007), adegenet 1.3-1 package in R (Jombart and Ahmed 2011), and Genepop (Rousset 2008). Family assignment for the C. gigas families was performed using Cervus 3.07 (Kalinowski et al. 2007). Cervus assigns offspring to their parent pairs based on the pairwise likelihood comparison approach generating locus-by-locus likelihood scores for each candidate parent for each offspring, and assigns parentage to a candidate parent with the highest LOD score.

Data availability

The Illumina sequencing data for the pooled C. gigas and O. edulis samples have been deposited into the European nucleotide archive under accession number PRJEB20253 (http://www.ebi.ac.uk/ena/data/view/PRJEB20253). The details of the SNP markers on the array are given in (Supplemental Material, File S1). O. edulis markers with significant alignment to the C. gigas genome (e-value 1E−20) are given in File S2.

RESULTS AND DISCUSSION

Sequencing and SNP selection

To discover and prioritize SNPs for inclusion on the combined-species oyster SNP array, species-specific DNA sequencing, SNP discovery, and filtering strategies were followed. For C. gigas, WGS data aligned to the oyster genome identified 12.4 million putative SNPs across all populations. The 1,216,467 putative SNPs that passed the Affymetrix evaluation were subsequently filtered using the criteria described above to 40,625 putative SNPs that were submitted for the final Axiom MyDesign array. For O. edulis, 588,266 putative SNPs were identified, of which 100,403 putative SNPs were recommended at least for one strand by Affymetrix. Further filtering based on the criteria described above reduced the set to 19,215 putative SNPs that were submitted for array design and production.

The final array contained 40,625 putative SNPs from C. gigas and 14,950 putative SNPs from O. edulis to give a total of 55,575 putative SNPs assayed by a total of 111,360 probes. There were a greater number of C. gigas SNPs placed on the array than O. edulis due to the anticipated greater future use of the array for genome-wide association studies and genomic prediction for economically important traits in breeding programs in this species. This includes an ongoing project to study host resistance to Oyster Herpes Virus based on genotyping samples collected from a large challenge experiment on oysters derived from Guernsey Sea Farm stocks. Nonetheless, it is anticipated that the ~15 K putative O. edulis SNPs will be widely applied for population and conservation genetics in future studies of this species.
Table 3 Descriptive population genetic estimates for the sampled O. edulis populations included in the validation of the array

<table>
<thead>
<tr>
<th>Sample</th>
<th>N</th>
<th># SNPs</th>
<th>Average MAF</th>
<th>Ho</th>
<th>He</th>
</tr>
</thead>
<tbody>
<tr>
<td>Croatia</td>
<td>9</td>
<td>8,474</td>
<td>0.234</td>
<td>0.323</td>
<td>0.320</td>
</tr>
<tr>
<td>Foyle_IRL</td>
<td>10</td>
<td>10,013</td>
<td>0.224</td>
<td>0.319</td>
<td>0.311</td>
</tr>
<tr>
<td>Grevelingen_NLD</td>
<td>10</td>
<td>9,946</td>
<td>0.224</td>
<td>0.319</td>
<td>0.310</td>
</tr>
<tr>
<td>Lame_NIRL</td>
<td>10</td>
<td>8,927</td>
<td>0.231</td>
<td>0.354</td>
<td>0.316</td>
</tr>
<tr>
<td>Mersea_UK</td>
<td>10</td>
<td>9,980</td>
<td>0.224</td>
<td>0.318</td>
<td>0.310</td>
</tr>
<tr>
<td>Quiberon_FR</td>
<td>10</td>
<td>9,973</td>
<td>0.226</td>
<td>0.315</td>
<td>0.312</td>
</tr>
<tr>
<td>Rossmore_IRL</td>
<td>10</td>
<td>9,846</td>
<td>0.228</td>
<td>0.327</td>
<td>0.314</td>
</tr>
<tr>
<td>Swansea_UK</td>
<td>9</td>
<td>9,696</td>
<td>0.224</td>
<td>0.319</td>
<td>0.311</td>
</tr>
<tr>
<td>Tralee_IRL</td>
<td>10</td>
<td>9,980</td>
<td>0.219</td>
<td>0.317</td>
<td>0.306</td>
</tr>
<tr>
<td>Maine_USA</td>
<td>10</td>
<td>9,614</td>
<td>0.221</td>
<td>0.317</td>
<td>0.305</td>
</tr>
<tr>
<td>Average within population</td>
<td>9597</td>
<td>0.225</td>
<td>0.323</td>
<td>0.312</td>
<td></td>
</tr>
<tr>
<td>All populations (combined)</td>
<td>108</td>
<td>11,151</td>
<td>0.210</td>
<td>0.292</td>
<td>0.311</td>
</tr>
</tbody>
</table>

MAF, minor allele frequency; #, number; SNPs, single nucleotide polymorphisms; Ho, level of genetic variability in terms of observed heterozygosity; He, level of genetic variability in terms of expected heterozygosity.

*a*Values represent the within-population average.

*b*Values were obtained by the analysis of the combined dataset, not the average of the individual populations.

(Table 2), giving a greater chance of detecting rare minor alleles. Among all the five French populations, 13,855 SNPs were shared, with few SNPs segregating exclusively in particular populations (Figure S3). Finally, 11,997 common SNPs were segregating in all the eight populations from both France and the UK (Figure S4). The average MAF (for markers showing a MAF > 0) was 0.207 across all UK populations, while it was 0.214 across all French populations. Analysis of the distribution of MAF values for polymorphic SNPs (MAF > 0) showed that the highest numbers of SNPs are located within a MAF value range between 0.01 and 0.2 in all populations, decreasing in frequency when the MAF approaches 0.5 (Figure S5). A similar situation was observed by Lapegue *et al.* (2014), who found a high proportion of low MAF SNPs within *C. gigas* populations. Based on an additional test of the array on a small number of Australian *C. gigas* samples (data not shown), the number of segregating SNPs was similar, indicating that the array is likely to perform comparably for geographically diverse populations.

From the 11,151 high quality SNPs segregating in the *O. edulis* populations, the average number of SNPs segregating (MAF > 0) in each population was 9597. The samples from Croatia showed the lowest number of segregating SNPs (*n* = 8474), while those from Foyle (Ireland) showed the highest (*n* = 10,013) (see Figure S6 and Table 3). A total of 4912 SNPs were shared between all (11) populations, with no particular population showing a high number of unique segregating SNPs within *C. gigas* populations. Based on an additional test of the array on a small number of Australian *C. gigas* samples (data not shown), the number of segregating SNPs was similar, indicating that the array is likely to perform comparably for geographically diverse populations.

Figure 1 Identity-by-state clustering of selected *O. edulis* populations. Neth, The Netherlands; N.Ire, Northern Ireland.
SNPs. The average MAF value across the populations was 0.225, with Croatia showing the highest value of 0.234. Analysis of the distribution of MAF values for polymorphic SNPs (MAF > 0) showed that most populations have a large number of SNPs within a MAF value range between 0.05 and 0.2, with the exception of Croatia and Swansea, which show a greater number of SNPs with a MAF higher than 0.1 (Figure S7).

The levels of genetic variability in terms of observed (Ho) and expected (He) heterozygosity (according to HWE) showed that most populations (C. gigas and O. edulis) had higher observed levels of heterozygosity than expected. Overall, no strong evidence of heterozygous deficiency was detected, in contrast to some previous studies that have described heterozygous deficiency in oysters and bivalves in general, albeit typically using a much lower number of microsatellites, SNPs, and allozymes (Appleyard and Ward 2006; English et al. 2000; Li et al. 2003; Sekino et al. 2003; Lapègue et al. 2014; Yu and Li 2007; Sobolewska and Beaumont 2005; Vercaemer et al. 2006). This discrepancy may be due to the fact that genome-wide SNP markers were used in the current study at a density not previously tested. In a larger-scale SNP assay-based evaluation of the bivalve mollusc Chlamys farreri, no evidence for heterozygote deficiency was detected (Jiao et al. 2014). It is also possible that the strict filtering process led to SNPs on the array being enriched for stable genomic regions with lower levels of variation, while genomic regions with higher variability (and potentially more prone to null alleles) might have been discarded.

Assessing population structure using IBS: The overall genetic similarity of any two samples can be evaluated by calculating average measures of IBS of the marker loci, which was then summarized using MDS to give indications of population (sub)structure (IBS clustering was also confirmed by DAPC analysis, data not shown). There was some
evidence of *C. gigas* samples clustering according to their hatchery origin, and French hatchery populations tended to cluster separately to UK hatchery populations (Figure S8). The *O. edulis* samples were typically from “wild” stocks from more diverse geographical locations than for the *C. gigas* samples (Figure S9 and Figure S10). Accordingly, certain populations did show evidence of genetic differentiation, notably Croatia, Larne (Northern Ireland), and Sveio (Norway), which are geographical outgroups (Figure 1 and Figure S10). Our results show evidence of a strong genetic similarity between Maine, Sveio (Norway), and Grevelingen (Netherlands) populations. Similarly, the origin of the Maine population has been linked to the Netherlands’ (Loosanoff 1955; Vercaemer et al. 2006), Netherlands populations have been linked to Denmark’s (Vera et al. 2016), and the genetic similarity between the Maine, Norway, Denmark, and Netherlands samples has also been observed using microsatellite markers (M. McCullough, personal communication). A lack of population structure according to geographical original was observed in the other *O. edulis* population samples tested, for example the majority of samples from the coast of the UK and Ireland (Figure S9). This is consistent with existing evidence that suggests that marine organisms with larval stages (such as bivalves) often show low genetic differentiation (Li et al. 2015; Shabtay et al. 2014; Rohfritsch et al. 2013; Giantisis et al. 2014), with temporal factors rather than geographical factors often playing the major role in population structure. It is also possible that historical stock translocations might have also played an important role in the lack of genetic structure and admixture of the *O. edulis* populations (Bromley et al. 2016).

**Evaluation of the SNP array in pair crosses of *C. gigas***: Three pair crosses between Guernsey Sea Farm parents were created, reared separately, and genotyped using the SNP array. Two of these nuclear families were half-siblings sharing a dam (F29 and F30). A total of 165 samples (161 offspring and their five parents) were genotyped. These families were analyzed separately from the population samples used to validate the array described above. In part, this was due to the difficulty in obtaining high quality genomic DNA from the juvenile oysters. From the 165 samples, 139 passed the DQC and genotype call rate ≥97% threshold, resulting in a total of 25,629 SNPs that were classified as good quality in these families. The vast majority of SNPs showed stable Mendelian inheritance in all samples, although there was an average of 395 SNPs (~2% of total informative SNPs) with evidence of a Mendelian error per individual.

Since the offspring from each nuclear family were physically tracked throughout the experiment, such that their family structure was known a priori, the utility of the SNP array to differentiate between families was assessed using IBS clustering with MDS scaling. The MDS scaling plot based on IBS clustering clearly shows a clear separate cluster for each of the families, as shown in Figure 2. Interestingly, the clustering and separation of the three nuclear families was more obvious than for the population samples, even for populations from very distant geographical locations. Four individuals were distant to any of the family clusters, which may suggest incorrect pedigree assignment according to the physical animal tracking. Family assignment successfully assigned all the individuals to their correspondent parents using 3000 randomly chosen SNPs, and confirmed that the four aforementioned individuals were not members of any of these three families. Microsatellites and SNP panels for parentage assignment have been described previously for oysters (Wang et al. 2010; Li et al. 2010; Lapègue et al. 2014; Jin et al. 2014). However, the successful parentage assignment in these physically tracked nuclear families, and the clear IBS-based differentiation of these families, bodes well for the utility of this SNP array for high resolution genetic mapping studies and selective breeding programs for oysters.

**Distribution of SNPs in the pacific oyster genome**: To assess the distribution of SNPs in the *C. gigas* genome (Zhang et al. 2012), SNPs were annotated according to the publicly available Ensembl oyster genome assembly (NCBI accession number: GCA_000237895.1). The
The oyster genome contains 7658 scaffolds (N50 = 401,585), 30,459 contigs (N50 = 31,239), and a total of ~558 Mb of assembled sequence. All 27,697 SNPs are mapped to the oyster genome according to BLAST alignment using their flanking region(s), with at least one SNP on 2007 of the scaffolds, which in total covered 501 Mb (89.6% of the total assembled genome sequence). The number of SNPs per scaffold was positively associated with scaffold length (Figure 3), with approximately one-fifth of the scaffolds containing only one SNP. Additionally, harnessing the publicly-available oyster genome annotation (GCA_000297895.133), the SNPs on the array were grouped into putative positional and functional categories using SNPeff (Cingolani et al. 2012). A total of 14.6, 13.1, 18.7, 17.6, and 2.8% of the SNPs were located in intergenic, intron, downstream, upstream, and exon regions, respectively. The remaining SNPs (33%) were identified as transcript, splice site donor, splice site acceptor, and splice site region.

The extent of LD between SNP pairs was assessed relative to their physical distance for the C. gigas populations. Pairwise $r^2$ was calculated using polymorphic SNPs with MAF $\geq 0.05$ as shown in Table 2. The mean $r^2$ was calculated for every kilobase and covering up to 500 kb, according to the physical distance on the oyster genome assembly, as shown in Figure 4. In general, low levels of LD with slow decay with increasing physical distance were observed. The Guernsey and Ifremer populations had lower levels of LD than the other populations. Although these LD levels are low compared to other aquaculture species, such as carp or tilapia (Hong Xia et al. 2015; Xu et al. 2014), they are in accordance with recent reports describing low levels and short extents of LD in wild C. gigas populations (Zhong et al. 2017). Moreover, differences in LD levels between populations can be related to the divergence of these populations and the number of generations for which they have been bred in isolation, as observed in cattle (de Roos et al. 2008).

There was a higher extent and slower decay of LD in the three nuclear families, and LD levels were substantially higher than those observed in the (presumably unrelated) validation populations, as would be expected (Figure 4 and Figure 5). A lower effective population size ($N_e$) brings higher levels of kinship between individuals and therefore a higher extent of LD (Sved 1971; Falconer and Mackay 1996).

**Conclusions**

This article describes the development and analysis of a high density SNP array for two oyster species. A very large database of SNP markers was developed for both C. gigas and O. edulis, using WGA and RAD-Seq, respectively. Following extensive filtering, SNP assays for these two oyster species were combined on the array with 40,625 high quality SNPs for C. gigas and 14,950 for O. edulis. Testing of the array on genomic DNA samples from diverse locations revealed that the array contains a high number of SNPs that are shared between populations, and that the array can be applied to detect population and family structure. This oyster SNP array will be publicly available and will facilitate the study of important economic and ecological traits for these two oyster species, with possible applications for genomic selection, QTL mapping, evolutionary genetics, and conservation programs.

**ACKNOWLEDGMENTS**

The authors acknowledge Pierrick Haffray, Anastasia Bestin, Florian Enez, and Anne-Sophie Tyran from the Syndicat des Selectionneurs Avicoles and Aquacoles, France for coordinating the provision of the French hatchery samples used for single nucleotide polymorphism (SNP) discovery and genotyping; Lucie Buockellyoen from Marinove, Emilie Vetois from Satmar, Adeline Lange from France Naissain, and Fiz Dacosta from Novostrea for agreeing to provide samples; Lionel Degremont from Ifremer, France for coordinating the provision of the Ifremer selection line samples; and Dennis Hedgecock from the University of Southern California for providing details of validated SNPs to include on the array. P.A.P. and M.M.C. were supported by funding from the European Union (EU) INTERREG IVA Program.


Communicating editor: B. J. Andrews