Molecular Ecology Resources
July 2014, Volume 14, Issue 4, Pages 820-830
http://dx.doi.org/10.1111/1755-0998.12230

Development of SNP genotyping arrays in two shellfish species

S. Lapègue[1,4,*], E. Harrang[1], S. Heurtebise[1], E. Flahauw[1], C. Donnadieu[2], P. Gayral[3,4], M. Ballenghien[3], L. Genestout[5], L. Barbotte[5], R. Mahla[5], P. Haffray[6], C. Klopp[7]

[1] Ifremer, SG2M-LGPMM, Laboratoire de Génétique et Pathologie des Mollusques Marins, La Tremblade, France
[2] INRA, UMR444 Laboratoire de Génétique Cellulaire, Plateforme GeT-PlaGe Genotoul, Castanet-Tolosan, France
[3] Université Montpellier 2, CNRS UMR 5554, Institut des Sciences de l'Evolution de Montpellier, Montpellier, France
[4] Université François Rabelais, CNRS UMR 7261, Institut de Recherche sur la Biologie de l'Insecte, Faculté des Sciences et Techniques, Tours, France
[5] LABOGENA, Domaine de Vilvert, Jouy-en-Josas, France
[6] SYSAAF, Station LPGP/INRA, Rennes, France
[7] INRA, Sigenae, UR875 Biométrie et Intelligence Artificielle, Castanet-Tolosan, France

*: Corresponding author : Sylvie Lapègue, fax: (33) 5 46 76 26 11 ; email address : Sylvie.Lapegue@ifremer.fr

Abstract:

Use of SNPs has been favored due to their abundance in plant and animal genomes, accompanied by the falling cost and rising throughput capacity for detection and genotyping. Here, we present *in vitro* (obtained from targeted sequencing) and *in silico* discovery of SNPs, and the design of medium-throughput genotyping arrays for two oyster species, the Pacific oyster, *Crassostrea gigas*, and European flat oyster, *Ostrea edulis*. Two sets of 384 SNP markers were designed for two Illumina GoldenGate arrays and genotyped on more than 1000 samples for each species. In each case, oyster samples were obtained from wild and selected populations and from three-generation families segregating for traits of interest in aquaculture. The rate of successfully genotyped polymorphic SNPs was about 60% for each species. Effects of SNP origin and quality on genotyping success (Illumina functionality score) were analyzed and compared with other model and non-model species. Furthermore, a simulation was made based on a subset of the *C. gigas* SNP array with a minor allele frequency of 0.3 and typical crosses used in shellfish hatcheries. This simulation indicated that at least 150 markers were needed to perform an accurate parental assignment. Such panels might provide valuable tools to improve our understanding of the connectivity between wild (and selected) populations and could contribute to future selective breeding programs.

Keywords: SNP genotyping ; GoldenGate technology ; Oysters ; *Ostrea edulis*\ ; *Crassostrea gigas*

1. Introduction

Single Nucleotide Polymorphisms (SNPs) are the most common type of heritable variation at the molecular level (Zhu *et al.* 2003). Their abundance in animal and plant genomes has led to their increased use in biological studies, from ecology to biology of evolution, and makes them an important tool for population and quantitative genetics studies. They are widely used, even in non-model species, to analyze genetic diversity and characterize genetic population structure, but are also useful in high-resolution genetic mapping, fine mapping of Quantitative Trait Loci (QTLs), linkage disequilibrium-based association mapping, parentage analyses, and marker-assisted selection (Garvin *et al.* 2010; Stapley *et al.* 2010; Ekblom et Galindo, 2011). However, their detection and genotyping still remain a challenge in non-model species for which the whole genome is not yet sequenced and assembled (Seeb *et al.* 2011).

Even if a complete oyster genome sequence was released recently (Zhang *et al.* 2012), mollusks were not considered as model species although they are an important source of food for human consumption and include major fishery and aquaculture species all over the world. Indeed they account for 23.6% of total aquaculture production (FAO 2011). By 2011, global production of oyster species has reached 4.52 million tones, among which 105 300 tones for the Pacific oyster, *Crassostrea gigas* in Europe. Additionally in Europe, mussels and the

European flat oyster, *Ostrea edulis*, remain important aquaculture species, with culture based on natural recruitment from wild populations (mussels: 460 000 t; *O. edulis*: 2300 t; FAO, 2011). As a result of their economic value, molecular markers have been developed over several decades to improve characterization of these species and their populations and for use as tools in breeding programs. This work began with the use of allozymes and has since expanded to include mitochondrial sequences, AFLPs and microsatellites markers. Several hundred markers are now available in *C. gigas* allowing the development of the first genetic maps (Hubert and Hedgecock 2004, Sauvage *et al*. 2010). Although microsatellites remain the most commonly used markers for population genetic studies on non-model organisms (Guichoux *et al.* 2011) and for parentage assignment in aquaculture (Estoup *et al*. 1998; Haffray *et al.* 2012), they may be soon replaced by SNPs due to the recent increase of commercially available genome-wide SNP arrays and high-throughput customer-designed genotyping of targeted variants for biologically-focused research, but also the unreliability of microsatellites, mainly because of allelic dropout.

There are two main differences between microsatellites and SNPs. First, as already mentioned, SNPs are more numerous than microsatellites in the genomes of most species. Oysters are known for their high genetic variability (Hedgecock *et al.* 2005). In *Crassostrea gigas*, one SNP per 60 bp has been observed in coding regions and one every 40 bp in non-coding regions (Sauvage *et al.* 2007). In *C. virginica*, the density of SNPs was estimated to be one every 20 bp on average (Zhang *et al*. 2010). In the European flat oyster a recent study estimated an average density of one SNP every 76 bp in coding regions and one every 47 bp in non-coding regions (Harrang *et al.* 2013). Another important difference is that the mutation rate per generation differs drastically between the two types of markers, being several orders

of magnitude lower for microsatellites than for SNPs (Ellegren, 2000). As a consequence, SNPs are typically diallelic (Krawczak 1999), whereas microsatellites generally have high allelic richness, reaching up to 20 alleles per locus in European flat oyster (Taris *et al.* 2009; Lallias *et al.* 2010) and even 40 alleles per locus in Pacific oyster (Rohfritsch *et al.* 2013). Consequently, several SNPs are usually needed to attain the same resolution as a microsatellite. On the basis of simulations, one may need two to three SNPs to equal the power of a microsatellite in linkage studies (Seddon *et al.* 2005). The ratio can be about five SNPs to one microsatellite in parentage analyses (Glaubitz *et al.* 2003), four to twelve SNPs for one microsatellite in genetic structure studies (Rosenberg *et al*. 2003), or five to twenty in association studies (Ohashi and Tokunaga, 2003). In QTL detection and fine genome mapping, increasing the number of markers would clearly be of interest and their usefulness would depend on the polymorphism present in the initial biological material.

The abundance of SNPs in plant and animal genomes, together with falling costs of detection and genotyping, and increased throughput have favored the use of SNPs in a wider range of studies. A large number of SNP-genotyping technologies have been developed in recent years (Gupta *et al.* 2008) including medium- to high-throughput SNP-genotyping methods in non-model species (Garvin *et al.* 2010). Among these, the GoldenGate Genotyping technology from Illumina has been reported to be a reliable technology allowing the genotyping of large collections of samples (up to 480) for thousands of SNP markers (up to 3072), with high levels of SNP conversion rate (number of polymorphic SNPs divided by the total number of SNPs in the array) and reproducibility (Ganal *et al.* 2009). This assessment was corroborated in outbred species with high levels of nucleotide diversity (Grattapaglia *et al.* 2011). We therefore chose this approach for the two oyster species in this study, which are

undomesticated and very polymorphic at the nucleotide level. SNPs can be detected by amplicon resequencing targeting specific genes using large *in silico* sequence resources. In both Pacific oysters and European flat oysters large EST (Expressed Sequence Tags) databases already exist (Fleury *et al.* 2009; Morga *et al.* 2011, 2012; Cahais *et al.* 2012, Gayral *et al.* 2013), allowing for an *in silico* approach to be used, whereas resequencing of amplicons has already been performed to detect a few SNPs in previous studies on linkage (Sauvage *et al.* 2010), population structure, or evolution (Rohfritsch *et al.* 2013; Harrang *et al.* 2013). This paper describes the discovery and design of the first sets of 384 SNP markers, combining *in vitro* and *in silico* detection approaches, in both the Pacific and European flat oysters, using the Illumina BeadXpress genotyping system. As breeding programs are being developed for the Pacific oyster, we also included a simulation analysis for this species in order to investigate the potential use of a sub-panel of the array in parentage analysis.

Material and methods

Biological material

For the Pacific oyster, biological material came from 1084 individuals from four different three-generation outbred pedigrees used for linkage and QTL mapping of juvenile ('spat') resistant to summer mortality. Furthermore, we added 194 samples from five European wild and selected populations, for a total of 1278 individuals. For the European flat oyster, fewer individuals were used (1070 in total), but 434 of them were similarly from four different three-generation outbred pedigrees used for linkage and QTL mapping of bonamiosis resistance and 636 from 16 natural populations sampled all over Europe.

DNA was extracted from the oyster gills. A Wizard® DNA Clean-up System (Promega, Madison, Wisconsin, USA) was used when the samples were collected in 70% alcohol, and a QiAamp DNA mini kit (Qiagen) when the samples where fresh or frozen. DNA quantification was performed with a NanoDrop spectrophotometer (NanoDrop Technologies, LLC, Wilmington, DEL, USA). It should be noted that half of the individuals from the *C. gigas* QTL families were dying when sampled, thus reducing the DNA quality.

SNP discovery and selection for array construction

For each oyster species, two sets of SNPs were considered. Those that were detected in sequences obtained by sequencing a subset of individuals for candidate genes will be referred to as *in vitro* SNPs. Those that were detected by aligning different sequences from contig databases will be referred to as *in silico* SNPs.

Pacific cupped oyster

For the Pacific oyster, *in vitro* sequencing investigated 103 loci from ESTs retrieved from the Genbank database (http://www.ncbi.nlm.nih.gov/) or from specific libraries that had been obtained to detect genes differentially regulated during summer mortality events (Fleury *et al.* 2009). Primers were designed using the Primer3 software package (Rozen and Skaletsky 2000). For a first set of ESTs (n = 61), 24 oysters belonging to a third generation of selection for summer mortality resistance were used in the SNP discovery phase (Sauvage 2008; Sauvage *et al.* 2007). A second set of ESTs (n = 42) was then added and 10 of the 24 oysters were used for sequencing, as described in Sauvage *et al.* (2007), together with a third set of

five SNPs from the 20 developed by Bai *et al.* (2009). Sequence alignment was performed with ClustalW via the BioEdit interface (Hall 1999) and DNAMAN version 4.1 (www.lynnon.com). The validity of each SNP was checked manually on the chromatograms and sequence alignments.

For the *in silico* SNPs, we investigated in 2009 the 6th assembly of the *Crassostrea gigas* EST database (http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas/index.html). The database contained results of the assembly of 55,851 public ESTs from dbEST and 417 Genbank mRNA sequences. The assembly, performed with TGICL (http://compbio.dfci.harvard.edu/tgi/software/; parameters -l 60 -p 96 -s 100000 -O '-p 75 -s 500'), produced an alignment file from which 1370 SNPs were extracted. We looked for SNPs that complied with the initial criteria: a minimum depth of seven sequences, with a minimum allele count of three, and the absence of any other SNP in the 60 bp segment flanking the analyzed SNP to the left or right. As these conditions appeared too stringent, and did not produce many SNPs, we relaxed the criteria to a minimum depth of five sequences with a minimum allele count of two, and allowed there to be a SNP within 120 bp of the SNP of interest, as long as there was only one and it was not close.

European flat oyster

For the European flat oyster, *in vitro* sequencing investigated 40 loci from two EST libraries (Morga *et al.* 2011, 2012). Primers were designed using Primer3 software package (Rozen and Skaletsky 2000). A total of 22 oysters, 16 from four different natural populations collected on the Atlantic and Mediterranean coasts and six belonging to the first generations of three selected families for resistance to bonamiosis were used to investigate

polymorphisms. The PCR and sequencing protocols used were the same as those given in Harrang *et al.* (2013). Sequence alignment was performed with ClustalW via the BioEdit interface (Hall 1999). The validity of each SNP was checked individually on nucleotide sequences and sequence alignments.

For the *in silico* SNPs, we investigated *O. edulis* transcriptome sequence data from eight individuals from the natural range (http://kimura.univ-montp2.fr/PopPhyl/index.php?section=data,http://datadryad.org/resource/doi:10.5061/dryad.5g32f94b/1, Cahais et *al.* 2012, Gayral *et al.* 2011). For the present study, the 454 and Illumina reads were assembled using a multi-kmer strategy (kmers: 37, 41, 45, 49, 53, 57 and 61, assembled with Velvet version 1.1.03). Contigs longer than 100 bp from every assembly were then meta-assembled with TGICL (http://compbio.dfci.harvard.edu/tgi/software/). The Illumina reads were remapped on the contigs using BWA (0.5.9-r16) and a compressed alignment file was produced using SAMtools view (version 0.1.11). The alignment file was then used to call the SNPs with SAMtools pileup and varFilter (version 0.1.11). In this database, we looked for SNPs that represented different contigs, with a depth ranging from 20 to 500 at the position and no other SNPs in the surrounding 120 bp. The SNP quality score was initially set at 20 but finally, due to the high number of SNPs available, we only used SNPs with the highest score of 227. Finally in both species, a list of SNPs was submitted to Illumina in order to accommodate 384 attempted SNPs. An output file gave us important performance values including the Illumina Functionality Score that indicates the expected success of each SNP designed in a given array.

SNP genotyping array

Illumina GoldenGate technology (Illumina Inc., San Diego, CA, USA) was used for the genotyping reactions, which were performed in accordance with the manufacturer's protocol (Lin *et al.* 2009, Fan *et al.* 2006). Ten and four DNA samples were duplicated across the different plates for the Pacific and flat oyster arrays, respectively, in order to assess the reproducibility of the genotyping assay. In total, 1278 cupped oyster and 1070 flat oyster samples were genotyped. A negative control (water) was also added to each of the 15 plates of each assay. Data generated from the BeadXpressTM reader were analyzed with GenomeStudio for automated genotype clustering and calling. A quality score was produced for each genotype. GenCall and GenTrain scores measure the reliability of SNP detection based on the distribution of genotypic classes (AA, AB and BB). A GenCall score cutoff of 0.25 was applied to define valid genotypes for each SNP and only those which attained a minimum GenTrain score of 0.25 were kept. These scores are the same stringent thresholds as those applied in previous studies on other species (human: Fan *et al.* 2003; trout: Sanchez *et al.* 2009; pine: Lepoittevin *et al.* 2010).

The call rate was then calculated for each oyster sample as the number of SNPs for which a genotype was obtained, divided by the total number of SNPs. If the call rate of a sample was below the call rate of the negative control, the sample was discarded from the analysis. Furthermore, samples were discarded if they showed a GC50 (rate of genotyping success for at least 50% of the SNPs for a single individual) too different (±0.01) from the mean of all the samples.

Clusters were visually inspected to ensure high data quality (Fig. 1). SNPs that did not show clear patterns of cluster separation were excluded from further analysis. Exclusion usually corresponded to situations of cluster compression (*i.e.* when the normalized theta values of

the homozygous clusters were not in the [0, 0.1] or [0.9, 1] ranges). Indeed, the compression of the BB homozygous cluster towards the AA cluster could result from a paralogous gene matching the A allele, which would increase the signal for the A dye for both BB and AB genotypes.

The minor allele frequencies of each SNP (the frequency of the allele in minority: MAF) were calculated in both arrays.

Parental assignment power

In the Pacific oyster, for which several breeding programs are being developed, we performed an analysis to develop an efficient and accurate SNP panel for parental assignment. We used AccurAssign® software (Barbotte *et al.* 2012), which uses the probabilities of Mendelian transmission of the markers and maximum likelihood for *in silico* validation. Briefly, the genotypes of parents and offspring of a given cross are simulated on the basis of the allelic frequencies of the markers in a set of populations (we used the set provided by the four population samples of Pacific oyster). This method allows genotyping errors to be taken into account. Then, the likelihood of a pair of parents for a given offspring is estimated. Four situations can arise: (1) the pair of parents is correct, (2) incorrect, (3) one parent cannot be assigned, or (4) neither of the parents can be assigned. In this study we compared six panels of 50, 75, 100, 150, 175 or 200 SNPs. The cross used to generate the simulated offspring involved 14 sires and 16 dams (each sire was crossed with 3 dams), for 3 generations. Two hundred offspring were generated in total. The genotypes of all parents and 200 offspring were simulated, and a genotyping error rate of 1% and missing genotype rate of 5% were applied. We then compared the likelihood of each pair of parents, considering the difference

between the first and second most likely pair of parents, and the four possible situations described above.

Results

Construction of the SNP arrays

For the Pacific oyster, 321 *in vitro* SNPs were detected in the first dataset of 61 sequenced fragments, and 380 in the second dataset of 42 sequenced fragments. Among those 701 SNPs, 72 were selected (39 and 33 from the two datasets, respectively) because they had high functionality scores and no neighboring polymorphisms. However, as we wanted to be sure that some genes of interest were represented in the SNP dataset, for several ESTs we kept two SNPs. Therefore, our 72 selected *in vitro* SNPs were obtained from 65 different ESTs. Adding the five SNPs from Bai *et al.* (2009), this gave a total of 77 *in vitro* SNPs to be included in the array, representing 70 different gene fragments. A total of 307 *in silico* SNPs completed the array, which corresponded to 148 different contigs. The Illumina Functionality score (0 to 1), calculated by Illumina for each SNP, ranged from 0.406 to 1 (Fig. 2).

For the flat oyster, a total of 420 *in vitro* SNPs were detected in the dataset of 40 sequenced fragments. Among them, the indels (n = 34) were discarded. Moreover, 347 SNPs were also discarded because of neighboring polymorphisms or low functionality scores. However, as we wanted some genes of interest to be represented in the SNP dataset, we kept some (n = 13) that had neighboring polymorphisms. To favor genotyping, those polymorphic nucleotides were treated as degenerated nucleotides. In total, 52 *in vitro* SNPs were included in the array, representing 35 different gene fragments. In contrast, a total of 1305 *in silico* SNPs fitted the

proposed criteria. Therefore, we chose the 332 *in silico* SNPs with the highest Illumina Functionality score (>0.935) to complete the array (Fig. 2).

Finally, for both species, the SNPs were mainly chosen in the highest]0.9–1] class of Illumina Functionality scores (Fig. 2), especially for the *in silico* SNPs of *O. edulis* (100%). However, some SNPs with lower scores were also selected (sometimes those with scores even lower than the 0.6 recommended by Illumina) because: (1) for *in vitro* SNPs, we wanted to try to keep as many SNPs as possible that were located in genes related to traits of interest, (2) for *C. gigas* we did not have many SNPs, even after we relaxed the selection criteria. The list and characteristics of the SNPs in the two arrays are given in Tables S1 and S2.

Reproducibility and success rate

According to the call rate and GC50 estimated for the samples, 24 *C. gigas* and 29 *O. edulis* samples were excluded from further analyses, which correspond to a loss of respectively 2.2% and 2.7 % of the samples. For the 10 replicated samples in the Pacific oyster and the 4 in the flat oyster, the same genotype was observed over the replicates within or between plates, giving a reproducibility rate of 100%.

Based on the GenTrain score and visual inspection of the cluster separation (Fig. 1), 232 SNPs were considered as polymorphic for *C. gigas*, while 38 were monomorphic and 114 failed. For *O. edulis*, 234 SNPs were polymorphic, 67 were monomorphic and 83 failed. The main difference between the two arrays, when considering the two types of SNPs (*in silico* and *in vitro*), comes from the higher percentage of *in vitro* SNPs that failed compared with *in silico* SNPs for *O. edulis* (Fig. S1). Hence, for *O. edulis*, *in silico* SNPs showed significantly higher genotyping success compared to *in vitro* SNPs (+30% with χ^2-test; p-value = 1.9e-06).

For *C. gigas*, we observed independence between the category of the SNPs (failed, polymorphic or monomorphic) and their origin (*in silico, in vitro*) with a χ^2-test p-value of 0.05951.

The GenTrain score distributions of the polymorphic SNPs were very similar for the two species: 4% and 10% below 0.5, 50% and 50% between 0.5 and 0.8, and 47% and 39% above 0.8 for *C. gigas* and *O. edulis*, respectively.

We measured the overall genotyping success rate by dividing the number of SNPs that were successfully genotyped by the total number of SNPs in the assay. This rate included both monomorphic and polymorphic SNPs. Then we measured the conversion rate, which only corresponds to the number of polymorphic SNPs, divided by the total number of SNPs in the assay, as defined by Fan *et al.* (2003). For *C. gigas*, the overall success rate was 70.3% and the conversion rate was 60.4%. These rates were very similar to those measured for *O. edulis* where there was an overall success rate of 78.4% and a conversion rate of 60.9%.

Interestingly, this success was not strictly correlated with the Illumina functionality score (Fig. 3), but showed a significant increase after the score of 0.7, between 65 and 80% for *C. gigas* and 60 and 81% for *O. edulis* (p-values of 0.0027 and 0.0001, respectively).

Finally, MAF indicated a bias toward lower allele frequencies for *C. gigas* (30% of SNPs, Fig.4), whereas in *O. edulis* MAF distribution was more homogeneous between classes. In *C. gigas*, we focused on a subset of 203 SNPs successfully genotyped on 156 samples from four wild European and selected populations in order to assess the assignment power of the SNP panel. Estimations of mean MAF in these four populations were all very close to 0.3 (Fig. S2).

Power of SNP panels for parental assignment

Based on the estimation of MAF in a subset of 203 SNPs and four wild and selected populations of *C. gigas*, simulations were performed in order to test the power of assignment of several SNP panels of different sizes (Fig. 5). With 150 markers, no pairs of parents were incorrect and the group of offspring assigned to both parents was distinguished from false positive results linked to one or two parents missing. Therefore, at least 150 markers from this panel with a mean MAF of 0.3 are needed to perform a powerful parental assignment.

Discussion

In non-model species, a classical approach to discovering SNPs relies on sequencing random genomic DNA fragments (Primmer *et al.* 2002) or targeting conserved regions of orthologous sequences from closely related species (Aitken *et al.* 2004). In oysters, this second strategy was made possible by the availability of candidate genes for interesting aspects of the physiology, reproduction and immunology of the species (Huvet *et al.* 2004; Renault *et al.* 2011). Although this approach yielded a relatively low number of markers for a large effort, these markers have made an important contribution, allowing clearer characterization of the genetic structure of populations (Rohfritsch *et al.* 2013), development of genetic maps and detection of QTLs of traits of interest in aquaculture (Sauvage *et al.* 2010). Moreover, SNPs derived from ESTs are located in functional genes, which can help us to establish potential links between functional genetic variation and these traits.

Among the variety of technologies available for SNP genotyping, our main concern was to use one that offered flexibility in terms of multiplexing, as we did not know how many SNPs could be retrieved after the *in silico* analyses. Illumina's GoldenGate arrays allowed 48-384

loci to be genotyped in a single reaction with the BeadXpress platform, and 96-1536 with BeadArray platform. When comparing the two oyster species, we can see that database availability had an important influence on the ease of obtaining an array. Nevertheless, we had to be less stringent for *C. gigas* SNP discovery in order to retrieve enough usable SNPs. This can be observed with the Illumina functionality score (Fig. 2) where *C. gigas* had a lower proportion of *in silico C. gigas* SNPs between 0.9-1 (60%) compared with the 100% of *in silico* SNPs with a score of >0.9 in *O. edulis*. However, when comparing the genotyping success between the species, the main difference between the two arrays is the higher percentage of *in vitro* SNPs that failed compared to *in silico* SNPs for *O. edulis* (with a 30% difference, Fig. S1) leading to a poor conversion rate. For *C. gigas*, there was less variation between *in silico* and *in vitro* SNPs for the failed category (a 10% difference), although Illumina functionality scores were very similar between *in vitro* SNPs for *C. gigas* and *O. edulis* (Fig. 2). Contrary to our observations, a different conversion rate pattern between *in silico* and *in vitro* SNPs would generally be expected to favor *in vitro* SNPs (e.g. Lepoittevin *et al.* 2010). Genotyping failures could arise from low quality of SNP flanking sequences, or from the presence of an exon-intron junction near the SNP of interest (Wang *et al.* 2008). However, sequencing errors could be observed both in *in vitro* and *in silico* SNPs because this step is performed in both cases, leading to false-positive SNPs (pseudo-SNPs). In our case, the difference between the two arrays could have arisen from the fact that some of the *in vitro O. edulis* SNPs were developed on oysters (from the wild or from selected families) that were eventually not included in the Illumina genotyping array. Hence the families they belonged to where not useful in our QTL analysis. Some of the *in vitro* SNPs proved to be very rare and were not successfully genotyped in the panel of 1070 genotyped flat oysters. This genotyping failure could be partially explained by the high proportion of low Illumina functionality score

for failed *in vitro O. edulis* SNPs (48% below 0.7; 72% below 0.9). Thus, the genotyping success of SNPs in *O. edulis* could be particularly sensitive to the quality of the flanking sequences.

It should be noted that the *C. gigas* database is regularly updated and that the last assembly performed was the eighth, done in March 2011 (http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas/index.html). Although the database we explored in 2009 was based on results of the sixth assembly of 55,851 public ESTs from dbEST and 417 Genbank mRNA sequences, this same database now contains results of the 8th assembly of 1,013,570 public ESTs from dbEST and Genbank mRNA sequences. Therefore, we can expect a far higher number of potential SNPs to be included in future genotyping arrays and the recently whole sequenced genome (Zhang *et al*. 2012) to be useful in such developments.

Without considering the SNP origin criteria, both arrays were very similar in terms of GenTrain score distributions for polymorphic SNPs, indicating a good rate of genotyping success. When the GenTrain scores were completed by a visual inspection of the clusters (Fig. 1), the overall success rates were 70.3% and 78.4% for *C. gigas* and *O. edulis*, respectively, and the conversion rates were 60.4% and 60.9%. The oyster rates fall within the range of values seen in other non-model species, such as 66.1% to 71% in Eucalyptus (Grattapaglia *et al.* 2011), or 66.9% and 68.75% in the maritime pine (Lepoittevin *et al.* 2010; Chancerel *et al.* 2011), but are higher than the 40.6% obtained in catfish (Wang *et al.* 2008). The success rates are, however, very different from those success rates obtained in model species, which often exceed 80% as they can draw on more genomic databases or even on whole sequenced genomes (Hyten *et al.* 2008, Rostoks *et al.* 2006, Deulvot *et al.* 2010, Yan *et al.* 2010,

Loridon *et al.* 2013). However, the conversion rate can drastically vary depending on the screening population used to detect SNPs. As a consequence, the conversion rate can be as low as 12.5-19.5% or 12.3-18.4% in F2 mapping families in maritime pine (Chancerel *et al.* 2011) or sugar pine (Jermstad *et al.* 2011) for SNPs that were not a priori screened for their polymorphism in the mapping pedigrees. On the contrary, the conversion rate reached 86% in a 384 array that was designed on the polymorphisms of the parents of the map in summer squashes (Esteras *et al.* 2012). In non-model species particularly, genotyping failures may be due to (1) low quality of SNP flanking sequences, (2) the presence of an exon-intron junction near the SNP of interest, or (3) the assembly and/or amplification of paralogous sequences, leading to false-positive SNP and patterns as observed in Figure 1d. In this last case, the shifting of the two homozygous clusters together has been observed, leading to a "cluster compression" (Hyten *et al.* 2008 Yan *et al.* 2010). Such SNPs patterns are discarded even when they show a clear cluster separation.

GoldenGate technology is a robust technology for genotyping, even in non-model species. Although the functionality score calculated by the Illumina Assay Design Tool gives a first indication of the likelihood for a SNP to be successfully genotyped, even SNPs with low scores can still be genotyped. This is particularly important when SNPs of interest need to be included in an array, for example if they represent loci of interest (outliers in population genetics or QTLs to be included in a genotyping panel for use in selection). In the two oyster arrays, respectively 30 and 50% of SNPs with Illumina functionality scores below 0.7 could be genotyped. The genotyping success increased greatly, to between 60 and 80%, as this score increased above the threshold. The relationship between Illumina score and genotyping success is in agreement with a number of other studies performed in non-model species

(Lepoittevin *et al.* 2010; Pavy *et al.* 2008) and supports Illumina's recommendations of using SNPs with functionality scores above 0.6. However, in the oyster arrays, the threshold was 0.7 for a clear improvement in genotyping success. Both arrays in the present study were very powerful as they showed excellent reproducibility and only a few samples had to be excluded. It should be noted that although half of the individuals from the *C. gigas* QTL families were dying when sampled, thus reducing the DNA quality, there was no notable reduction in the genotyping results.

When considering the informativeness of arrays, the MAF parameter can reveal potential biases in interpreting genotyping data. In the oyster arrays, we observed a particularly high number of *C. gigas* SNPs showing a low MAF (31% of the SNPs had a MAF below 10%). This could be a consequence of the oyster panel used to detect *in vitro* SNPs. Hence, *in vitro* SNPs were detected on a small number of animals that were parents of families used in QTL studies and originated from Marennes-Oléron bay. This could induce a bias on the MAF of those SNPs. However the proportion of *in vitro* SNPs with a MAF below 0.1 among all SNPs was on the same order as the proportion of *in vitro* SNPs in the polymorphic SNP panel (respectively 20 and 17%). Consequently, the high number of SNPs with a low MAF is mainly based on *in silico* SNPs. This result in *C. gigas* could be explained by the difference in stringency applied for the SNP discovery between the two *C. gigas* and *O. edulis* arrays: far enough SNPs even with the highest quality score of 227 in *O.edulis* but relaxed criteria of minimum depth and minimum allele count in *C. gigas* to get enough SNPs. The contrast might also arise from differences in the nature of the databases in which the *in silico* SNPs were detected or differences in the relative rarity of SNPs situated in genes of interest compared with those located elsewhere in the genome.

Although next-generation genotyping promises to become a powerful tool for population genomics and genomic selection, low to medium arrays will continue to be sufficient for many applications, especially in non-model species (Seeb *et al.* 2011). The power and flexibility of the present array for European flat oyster will be used in several population studies at the European and bay scales in order to characterize the genetic diversity and structure of flat oyster populations, and to potentially identify discontinuities in the distribution of allele frequencies and signatures of selection. The *C. gigas* panel will be tested for characterization of the very closely related sub species *C. gigas* and *C. angulata* (Huvet *et al.* 2004) in order to confirm admixture observed in nature and characterize a set of diagnostic markers for the development of tools to be applied for conservation and management of both species (e.g. Pritchard *et al.* 2012; Burgarella *et al.* 2009). This panel will also be used to characterize French Pacific oyster resources used to start selection programs. Furthermore, it might allow a better characterization of the QTLs associated with survival during summer (Sauvage *et al.* 2010). Even without marker-assisted selection, medium-throughput genotyping technologies can be useful in oyster breeding programs that are in progress, to provide tools to improve stock management or genotyping of mixed families in the hatchery. With such strategies in mind, we performed a simulation analysis to make an initial estimate of the number of markers necessary to provide a powerful tool for parental assignment in the Pacific oyster. For the four wild oyster populations, we observed mean MAF levels close to 0.3 for a subset of our array, which is also the value observed after the genotyping of families produced in shellfish hatcheries (Haffray *et al.* 2013). We therefore based our simulations on this value and observed that at least 150 markers from this panel with a mean MAF of 0.3 were needed to perform a powerful parental assignment. This result was the same order of magnitude as that estimated for French bovine parental assignment (Barbotte *et al.* 2012),

while considering crosses relevant for shellfish hatcheries (P. Haffray, pers. com.). This panel could be useful in parental assignment in breeding programs but also in restocking programs that are being considered for the Pacific oyster. Finally these first SNP panels will likely be improved, especially that for *C. gigas*, which will benefit from the availability of new transcriptomic assemblies and the recently sequenced genome (Zhang *et al.* 2012).

Acknowledgements

The authors wish to thank the organizers and partners of the projects that supported this research: GigADN, funded by FranceAgrimer (contract n°SIVAL NL: 2010-1021), with the active participation of Gildas Gautier (SFC breeding company, Noirmoutier France) and Marc Vandeputte (INRA); GAMETOGENES (ANR-08-GENM-041), funded by a grant from the Agence Nationale de la Recherche 'Génomique animale' Programme (ANR-08-GENM-041); SEAFARE (Sustainable and Environmentally Friendly Aquaculture for the Atlantic Region of Europe – n°2009-1/123), funded by the Atlantic Area Transnational Program (2007–2013); and Hi-Flo, funded by the Agence Nationale de la Recherche (ANR-08-blanc-0334-01).
This work was also supported by a European Research Council grant (ERC PopPhyl 232971) coordinated by Nicolas Galtier. The authors are grateful to all members of staff of the Ifremer hatchery in La Tremblade and of the nursery in Bouin for producing and rearing the oysters. The authors wish to acknowledge the support from the Genomic platform of Genopole Toulouse Midi Pyrénées, on which the genotyping work was performed. The authors also wish to thank Bureau de Traduction de l'Université de Brest (Helen McCombie) for English editing.

References

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals : a targeted-gene approach. *Molecular Ecology*, **13**, 1423-1431.

Bai J, Li Q, Kong LF *et al.* (2009) Characterization of 20 single nucleotide polymorphism markers in the Pacific oyster (*Crassostrea gigas*). *Animal Genetics*, **40**, 1004-1005.

Barbotte L, Boichard D, Genestout L, Haffray P, Chapuis H, Chantry-Darmon C, Boscher M. Y. (2012). A new parental assignment tool for complex mating and large number of animals. In *Abstract Book 33rd Conference of the International Society for Animal Genetics* (p. 100), Cairns, Australie.

Burgarella C, Lorenzo Z, Jabbour-Zahab R *et al.* (2009) detection of Hybrids in nature : application to oaks (*Quercus suber* and *Q. ilex*). *Heredity*, **102**, 442-452.

Cahais V, Gayral G, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N (2012) Reference-free transcriptome assembly in non-model animals next-generation sequencing data. *Molecular Ecology Resources*, **12**, 834-845.

Chancerel E, Lepoittevin C, Le Provost G et al. (2011) Development and implementation of a higly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. *BMC Genomics*, **12**, 368.

Deulvot C, Charrel H, Marty A, Jacquin F, Donnadieu C, Lejeune-Hénaut I, Burstin J, Aubert G (2010) Higly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. *BMC Genomics*, **11**, 468.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1-15.

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, **49**, 746-756.

Estoup A, Gharbi K, SanCristobal M, Chevalet C, Haffray P, Guyomard R (1998) Parentage assignment using microsatellites in turbot (*Scophtalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. Canadian Journal of Fisheries and Aquatic Sciences, **55**, 715-725.

Esteras C, Gomez P, Monforte AJ *et al.* (2012) High-throughput SNP genotyping in *Cucurbita pepo* for map construction and quantitative trait loci mapping. *BMC Genomics*, **13**, 80.

Fan JB, Oliphant A, Shen R *et al.* (2003) Highly parallel genotyping. *Cold Spring Harbor ymposia on Quantitative Biology*, **68**, 69-78.

Fan JB, Gunderson KL, Bibikova M *et al.* (2006) Illumina universal bead arrays. *Methods in Enzymology*, **410**, 57-73.

Fleury E, Huvet A, Lelong C, *et al.* (2009) Generation and analysis of a 29,745 unique Expressed Sequence Tags from the Pacific oyster (*Crassostrea gigas*) assembled into a publicly accessible database: the GigasDatabase. *BMC Genomics*, **10**, 341.

Ganal MW, Altmann T, Roder MS (2009) SNP identification in crop plants. *Current Opinion in Plant Biology*, **12**, 211-217.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915-934.

Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in non-model animals. *Molecular Ecology Resources*, **11**, 650-661.

Glaubitz JC, Rhodes OE, DeWoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039-1047.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology resources*, **11**, 591-611.

Grattapaglia D, Silva-Junior O, Kirst M, Marco de Lima M, Faria DA, Pappas Jr GJ (2011) SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. *BMC Plant Biology*, **11**, 65.

Gupta PK, Rustgi S, Mir GG (2008) Array-based high-throughput DNA markers for crop improvement. *Heredity*, **101**, 5-18.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41,** 95-98.

Haffray P, Vandeputte M, Petit V, Pincent C, Chatain B, Chapuis H, Meriaux JC, Coudurier B, Quillet E, Dupont-Nivet M (2012). Minimizing maternal effect in salmonid families mixed since eyed stages and a posteriori DNA-pedigreed. *Livestock Science*, **150**, 170-178.

Haffray P, Bugeon J, Rivard Q, Quittet B, Puyo S, Allamelou J.M., Vandeputte M., Dupont-Nivet M (2013). Genetic parameters of in-vivo prediction of carcass, head and fillet yields by internal ultrasound and 2D external imagery in large rainbow trout (*Oncorhynchus mykiss*). In press in *Aquaculture*.

Haffray P, Genestout L, Barbotte L, Lapègue L (2013) GigADN : Tentative de mise au point du contrôle de filiation avec des marqueurs ADN chez l'huître creuse. Rapport final FranceAgrmer, 27pp.

Harrang E, Lapègue S, Morga B, Bierne N (2013) A high load of non-neutral amino-acid polymorphisms explains high protein diversity despite moderate effective population

size in a marine bivalve with sweepstakes reproduction. *Genes Genomes Genetics*, **3**, 333-341.

Hedgecock D, Gaffney PM, Goulletquer P, Guo X, Reece K, Warr G (2005) A case for sequencing the Pacific oyster genome. *Journal of Shellfish Research*, **24**, 429–441.

Hubert S, Hedgecock D (2004) Linkage maps of DNA microsatellite markers for the Pacific oyster *Crassostrea gigas*. *Genetics*, **168**, 351-362.

Huvet A, Fabioux C, McCombie H, Lapègue S, Boudry P (2004) Natural hybridization between genetically differenciated populations of *Crassostrea gigas* and *C. angulata* highlighted by sequence variation in flanking regions of a microsatellite locus. *Marine Ecology Progress Series*, **272**, 141-152.

Hyten DL, Song Q, Choi I-Y et al. (2008) High-Throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoritical and Applied genetics*, **116**, 945-952.

Jermstad KD, Eckert AJ, Wegrzyn JL *et al.* (2011) Comparative mapping in pinus: sugar pine (*Pinus lambertiana* Dougl.) and loblolly pine (*Pinus taeda* L.). *Tree Genetics and Genomes*, **7**, 457-468.

Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*, **28**, 1676-1681.

Lallias D, Boudry P, Lapègue S, King J, Beaumont AR (2010) Strategies for the retention of high genetic variability in European flat oyster (*Ostrea edulis*) restoration programmes. *Conservation Genetics*, **11**, 1899-1910.

Lepoittevin C, Frigerio J-M, Garnier-Géré P *et al.* (2010) *In vitro* versus *in silico* detected SNPs for the development of a genotyping array: what can we learn from a non-model species ? *Plos ONE*, **5**, e11034.

Lin CH, Yeakley JM, McDaniel TK, Shen R (2009) Medium- to high-throughput SNP genotyping using VeraCode Microbeads. In: Bugert P, ed. DNA and RNA Profiling in Human Blood. New York: Humana Press. Pp 129-142.

Loridon K, Burgarella C, Chantret N et al. (2013) Single-nucleotide polymorphism discovery and diversity in the model legume *Medicago truncatula*. *Molecular Ecology Resources*, **13**, 84-95.

Morga B, Arzul I, Faury N, Segarra A, Chollet B, Renault T (2011) Molecular responses of *Ostrea edulis* haemocytes to an in vitro infection with Bonamia ostreae. *Developmental & Comparative Immunology*, **35**, 323-333.

Morga B, Renault T, Arzul I (2012) New insights in flat oyster *Ostrea edulis* resistance against the parasite Bonamia ostreae. *Fish & Shellfish Immunology*, **32**, 958-968.

Ohashi J, Tokunaga K (2003) Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *Journal of Human Genetics*, **48**, 487-491.

Pavy N, Pelgas B, Beauseigle S et al. (2008) Enhancing genetic mapping of complex genomes through the design of higly-multiplexed SNP arrays : application to the large and unsequenced genomes of white spruce and black spruce. *BMC Genomics*, **9**, 17.

Primmer CR, Borge T, Lindell J, Saetre GP (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, **11**, 603-612.

Pritchard VL, Abadia-Cardoso A, Garza JC (2012) Discovery and characterization of a large number of diagnostic markers to discriminate *Oncorhynchus mykiss* and *O. clarkii*. *Molecular Ecology Resources*, **12**, 918-931.

Renault T, Faury N, Barbosa-Solomieu V, Moreau K (2011) Suppression substractive hybridisation (SSH) and real time PCR reveal differential gene expression in the

Pacific cupped oyster, *Crassostrea gigas*, challenged with Ostreid herpesvirus 1. *Developmental and comparative immunology*, **35**, 725-35.

Rohfritsch A, Bierne N, Boudry P, Heurtebise S, Cornette F, Lapègue S (2013). Population genomics shed light on the demographic and adaptive histories of European invasion in the Pacific oyster, *Crassostrea gigas*. *Evolutionary Applications*, doi: 10.1111/eva.12086.

Rosenberg NA, Li LM, Ward L, et al. (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**, 1402-1422.

Rostoks N, Ramsay L, Mackenzie K *et al.* (2006) recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proceedings of the National Academy of Sciences*, **103**, 18656-18661.

Rozen, S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132,** 365-86.

Sanchez CC, Smith TPL, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 599.

Sauvage C, Bierne N, Lapègue S, Boudry P (2007) Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. *Gene*, **406**, 13–22.

Sauvage C (2008) Development of molecular markers associated with resistance to summer mortality in oyster *Crassostrea gigas* - QTL approach. PhD Thesis, University of La Rochelle, France, 272 pp.

Sauvage C, Boudry P, de Koning D-J, Haley CS, Heurtebise S, Lapègue S (2010) Quantitative Trait Loci for resistance to summer mortality and OsHV1 load in the Pacific oyster (*Crassostrea gigas*). *Animal Genetics*, **41**, 390-399.

Seddon JM, Parker HG, Ostrander EA, Ellengren H (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology*, **14**, 503-511.

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11 (Suppl.1)**, 1-8.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adapatation genomics : the next generation. *Trends in Ecology and Evolution*, **25**, 705-712.

Taris N, Boudry P, Bonhomme F, Camara MD, Lapègue S (2009) Mitochondrial and nuclear DNA analysis of genetic heterogeneity among recruitment cohorts of the European flat oyster *Ostrea edulis*. *Biological Bulletin*, **217**, 233-241.

Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, et al. (2008) Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomis*, **9**, 450.

Yan J, Yang X, Shah T *et al.* (2010) High-Throughput SNP genotyping with the GoldenGate assay in maize. *Molecular Breeding*, **25**, 441-451.

Zhang LS, Guo XM (2010). Development and validation of single nucleotide polymorphism markers in the eastern oyster *Crassostrea virginica* Gmelin by mining ESTs and resequencing. *Aquaculture*, **302**, 124-129.

Zhang G, Fang X, Guo X et al. (2012) The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, **490**, 49-54.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, and Cregan PB (2003) Single-nucleotide polymorphisms in soybean. *Genetics*, **163,**1123-1134.

Data Accessibility

Lists and characteristics of the 384 SNPs of the Pacific oyster array: Table S1

Lists and characteristics of the 384 SNPs of the European flat oyster: Table S2

Alignments of *in vitro* and *in silico* sequences used to select the SNPs to be included in the arrays: DRYAD entry doi:10.5061/dryad.jr233.

Author Contributions

SL conceived the protocols and wrote the first draft of the manuscript. EH, SH, EF, CD, PG, MB, CK participated in the choice of the *in silico* and *in vitro* SNPs. SL, SH, CD performed the SNP genotyping, and SL, LG, LB, RM, PH participated in testing the power of the panel assignment through simulations. All authors participated in the writing of the manuscript.

Figure Legends

Fig. 1 Examples of clustering results obtained for the *Crassostrea gigas* SNP array. Each dot represents a single oyster sample from a subset of 936 *C. gigas* samples and four SNPs. The y axis (Norm R) is the normalized sum of intensities of the two dyes involved, and the x axis (Norm Theta) represents the normalized *Theta* between 0 (homozygous for allele A) and 1 (homozygous for allele B). (a) Pattern considered as successful and polymorphic with the three clusters (AA in red, AB in purple, BB in blue). (b) Atypical pattern with three heterozygous clusters, which was nevertheless

considered successful because the two clusters corresponded to samples from three different populations and the different clusters were well separated. (c) Pattern considered as successful but monomorphic. (d) Pattern showing ambiguous clustering with a low intensity, considered as a failed SNP.

Fig. 2 Illumina Functionality scores of the SNPs from the *C. gigas* and *O. edulis* arrays according to their *in vitro* or *in silico* origin.

Fig. 3 Percentages of SNPs that were successfully genotyped in relation to their Illumina Functionality score for each array.

Fig. 4 Allele frequency spectrum for the 232 and 234 polymorphic SNPs genotyped in *C. gigas* and *O. edulis* populations, respectively.

Fig. 5 Distribution of minimum LOD score (minLOD) and difference between minimum LOD score and the LOD score for the next most likely sire-dam pair (Delta). Correct (green dots) and incorrect (red dots) assignments of male-female parent pairs are plotted for analyses with both parents present in the genotyped samples. Also plotted are the minLOD and Delta obtained when one parent was not present (dark blue dots), and when neither parent was present (light blue dots).

Supporting Information

Table S1 List and characteristics of the 384 SNPs of the Pacific oyster array.

Table S2 List and characteristics of the 384 SNPs of the European flat oyster array.

Fig. S1 Percentage of SNPs for which the genotyping failed, or that were considered as polymorphic or monomorphic according to their *in vitro* or *in silico* origin and for each array.

Fig. S2 Box plot representation of minor allele frequency (MAF) in a subset of four *C. gigas* populations (LAF = La Fumée, France; MAR = Marennes, France; ARE = Arcachon, France; DAN = Danemark) for a panel 203 SNPs that were polymorphic in this subset.