Performance and precision of double digestion RAD (ddRAD) genotyping in large multiplexed datasets of marine fish species

Maroso, F, Hillen, JEJ, Pardo, BG, Gkagkavouzis, K, Coscia, I, Hermida, M, Franch, R, Hellemans, B, Van Houdt, J, Simionati, B, Taggart, JB, Nielsen, EE, Maes, G, Ciavaglia, SA, Webster, LMI, Volckaert, FAM, Martinez, P, Bargelloni, L and Ogden, R

http://dx.doi.org/10.1016/j.margen.2018.02.002

Performance and precision of double digestion RAD (ddRAD) genotyping in large multiplexed datasets of marine fish species

Maroso, F.[a], Hillen, J.E.J.[b], Pardo, B.G.[c], Gkagkavouzis, K.[d], Coscia, I.[b,e], Hermida, M.[c], Franch, R.[a], Hellemans, B.[b], Van Houdt, J.[f], Simionati, B.[g], Taggart, J.B.[h], Nielsen, E.E [i], Maes, G.[b, f, l], Ciavaglia, S.A.[m], Webster, L.M.I.[m], Volckaert, F.A.M.[b], Martinez, P[c], Bargelloni, L. [a], AquaTrace Consortium, Ogden, R.[n]

[a] Department of Compared Biomedicine and Food Science, University of Padova, 35020 Legnaro, ITALY

[b] Laboratory of Biodiversity and Evolutionary Genomics, University of Leuven, Ch. de Bériotstraat 32 box 2439, B-3000 Leuven, Belgium

[c] Departmento de Zoología, Genética y Antropología Física, Universidade de Santiago de Compostela, 27002, Lugo, Spain

[d] Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

[e] Current address: School of Environmental and Life Science, Rm 332, Peel building, University of Salford, Salford, M5 4WT, UK

[f] Department of Human Genetics, University of Leuven, O&N I Herestraat 49 - box 602, B-3000 Leuven, Belgium

[g] BMR Genomics, Via Redipuglia 21a, Padova, Italy

[h] Division of Environmental and Evolutionary Biology, School of Biology and Biochemistry, The Queen's University of Belfast, Belfast BT7 INN, Northern Ireland, U.K.

[i] National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, 8600 Silkeborg, Denmark

[l] Centre for Sustainable Tropical Fisheries and Aquaculture, Comparative Genomics Centre, College of Marine and Environmental Sciences, Faculty of Science and Engineering, James Cook University, Townsville, 4811 QLD, Australia

[m] Science and Advice for Scottish Agriculture, Roddinglaw Road, Edinburgh EH12 9FJ

[n] Royal (Dick) School of Veterinary Studies and the Roslin Institute, University of Edinburgh, Edinburgh EH25 9RG, UK

Key words: *ddRAD, European sea bass, GBS, gilthead sea bream, sequencing precision, turbot*

Corresponding author: francesco.maroso@studenti.unipd.it

Abstract

The development of Genotyping-By-Sequencing (GBS) technologies enables cost-effective analysis of large numbers of Single Nucleotide Polymorphisms (SNPs), especially in 'non-model' species. Nevertheless, as such technologies enter a mature phase, biases and errors inherent to GBS are becoming evident. Here, we evaluated the performance of double digest Restriction enzyme Associated DNA (ddRAD) sequencing in SNP genotyping studies including high number of samples. Datasets of sequence data were generated from three marine teleost species (>5,500 samples, >2.5x10^{12} bases in total), using a standardized protocol. A common bioinformatics pipeline based on STACKS was established, with and without the use of a reference genome. We performed analyses throughout the production and analysis of ddRAD data in order to explore (i) the loss of information due to heterogeneous raw read number across samples; (ii) the discrepancy between expected and observed tag length and

coverage; (iii) the performances of reference based vs. *de novo* approaches; (iv) the sources of potential genotyping errors of the library preparation/bioinformatics protocol, by comparing technical replicates. Our results showed use of a reference genome and *a posteriori* genotype correction improved genotyping precision. Individual read coverage was a key variable for reproducibility; variance in sequencing depth between loci in the same individual was also identified as an important factor and found to correlate to tag length. A comparison of downstream analysis carried out with ddRAD *vs* single SNP allele specific assay genotypes provided information about the levels of genotyping imprecision that can have a significant impact on allele frequency estimations and population assignment. The results and insights presented here will help to select and improve approaches to the analysis of large datasets based on RAD-like methodologies.

Introduction

The options for studying the genomic constitution of individuals and populations are increasing rapidly thanks to the development of powerful and accurate sequencing technologies that provide higher throughput at decreasing costs (Liu et al. 2012). Meanwhile, efficient reduced representation methods have been proposed to provide high sequence coverage for selected genomic regions, collectively named as Genotyping-By-Sequencing (GBS) technologies (Narum et al. 2013). One of these GBS methods, Restriction-site Associated DNA sequencing (RAD-seq) (Baird et al. 2008) has become particularly popular as it allows the cost-effective analysis of thousands of markers for tens/hundreds of individuals in a single sequencing lane. The original RAD protocol has also been modified to optimize throughput and ease of use, generating several alternative RAD-like methods (*e.g.* Peterson et al. 2012; Wang et al. 2012; and the review by Andrews et al. 2016).

As GBS technologies enter a more mature phase, biases and errors inherent to such methods are becoming apparent (Arnold et al. 2013) and comparative analysis of the most popular RAD-like protocols have addressed some of these subjects (Puritz et al. 2014). Two recent studies (DaCosta e Sorenson 2014; Mastretta-Yanes et al. 2015) focused specifically on genotyping issues relating to double digest Restriction enzyme Associated DNA (ddRAD) (Peterson et al. 2012). ddRAD is one of the most recently developed RAD variants, known for its relative flexibility and ease of use. In addition to the sources of error that also affect other methodologies, the authors recorded ddRAD-specific issues such as the recovery of restriction fragments shorter than expected, amplification bias toward GC-rich fragments, non-specific cutting by restriction enzymes, newly formed restriction enzyme sites and drop of fragment number due to loss of restriction sites.

Beyond laboratory-based assessments of variation in ddRAD performance, there is a need to better understand the risk of errors associated with the production and use of ddRAD data, which is becoming increasingly relied upon for population genetic inference. Unawareness of the presence of biased markers can indeed lead to artificial excess of homozygotes (Taberlet et al. 1996), false departure from Hardy–Weinberg equilibrium (Xu et al. 2002), overestimation of inbreeding (Gomes et al. 1999) and unreliable inferences about population structure that have the potential to distort research conclusions. As a consequence, natural resource management and policy can be seriously affected. In this study, we seek to expand the experimental evaluation of ddRAD by focusing on the performance of common bioinformatics approaches as applied to multiple, comparable, large ddRAD datasets of marine fish species. A technical evaluation focused on marine fish data is interesting due to some biological characteristics of this taxon, such as relatively high SNP frequency, that can further affect genotyping accuracy. The species analyzed in this study are the European sea bass (*Dicentrarchus labrax*), the gilthead sea bream (*Sparus aurata*) and the turbot (*Scophtalmus maximus*).

Available genomic resources are increasing for three species studied. Sea bass (Tine et al. 2014) and turbot (Figueras et al. 2016) genomes have already been published and a draft sea bream genome will soon be published (L. Bargelloni, personal communication) and was made available for this work. The three differ in the quality of their assembly, as indicated by the contig length (i.e. their respective N50 values, which is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N). However, they share similar genome size and can thus provide comparable results (Table 1). The use of species with different levels of genome sequence development permits assessing effects of the reference genome quality on approaches that use genomes to improve the performance of clustering

104 methods for RAD data (e.g. reference based analysis in STACKS).

105 In this study, we set out to examine how variation in ddRAD sequence datasets and the
106 application and quality of available reference genome sequences affect the consistency and
107 accuracy of resulting data, generated through commonly used analytical approaches. The
108 laboratory and bioinformatic pipeline used to generate the ddRAD datasets followed standard
109 published methods (see below) and has been summarized in a flowchart (Figure 1). The
110 performance of the ddRAD pipeline was evaluated at different stages in order to investigate the
111 causes and effects of variation in individual sample coverage, RAD-tag sequence length and
112 application and quality of reference genomes on the eventual accuracy and error rates of
113 individual genotyping. We specifically addressed the following questions:

114 (i) *Evaluation of sample representation within multiplexed libraries*. What is the
115 typical variation in terms of number of raw reads per sample when multiple
116 individuals (144 in our case) are multiplexed in a single sequencing lane?

117 (ii) *Tag length and coverage*. Is there any difference between the expected and observed
118 length of analyzed tags? Does any relationship exist between tag length and depth
119 of coverage?

120 (i) *De novo and reference-based genotyping using STACKS.* What is the effect of
121 different clustering approaches (e.g. *de novo* vs reference-based*, a posteriori*
122 genotyping correction) on the number of markers identified?

123 (ii) *Genotyping precision and error rates*. What are the effects of the variables described
124 above on the number of mismatches between technical replicates?

125 Based on these insights we suggest approaches which can help to mitigate the identified risks
126 of error in ddRAD analysis. Finally, the potential effect of genotyping imprecision on

downstream analysis was evaluated using comparative data between ddRAD and single SNP allele specific genotyping, focusing on how genotyping errors could affect allele frequency and population assignment.

Material and Methods

Samples and library preparation

Specimens of European sea bass, gilthead sea bream and turbot were collected in the context of the European Union's FP7 funded project 'AQUATRACE' (KBBE 311920). The entire sample set included more than 5,581 specimens (2,128 European sea bass, 2,156 gilthead sea bream and 1,297 turbot) from the species' distribution range, some of which were collected specifically for the project (years 2013-2014, from now on referred to as "fresh" samples), while others had been collected earlier ("archived" samples) (Supplementary Material Table). For fresh samples, fin clips were preserved separately in 95% ethanol at 4°C until genomic DNA (gDNA) extraction. Samples were extracted either with Invisorb® DNA tissue HTS 96 kit (Stratec biomedical) or with a standard NaCl isopropanol precipitation protocol (Cruz et al. 2016). Extracted DNA samples were then classified as "high", "mid" or "low" quality according to the level of degradation assessed with agarose gel electrophoresis (see Supplementary material).

The same ddRAD protocol, with minor modifications, was used for the three species. The library preparation followed the original guidelines of Peterson et al. (2012), with some modifications that facilitate the screening of large number of individuals (see Supplementary Material for details), and was carried out in three different laboratories within the AquaTrace consortium, each focusing on a single species: the sea bass at the Laboratory of Biodiversity and Evolutionary Genomics, University of Leuven, sea bream at the Department of Compared

Biomedicine and Food Science, University of Padova and turbot at the Departmento de Zoología, Genética y Antropología Física, Universidade de Santiago de Compostela. To promote a common standardized approach, staff from the three laboratories completed a hands-on training course in library preparation at the Institute of Aquaculture, Stirling, where the modified ddRAD protocol originated. Multiple ddRAD libraries were prepared for each species (sea bream n=14; sea bass n=14; turbot n=9). Each library comprised 144 samples, and in all the libraries the same three or four control samples for each species were included, to enable cross-library comparisons and mismatch rates between replicates to be assessed. In particular, four sea bream specimens (SAC3, SAC4, SAC5 and SAC6 from Sardinia, Italy); three sea bass specimens (DLTY40, from the Central Mediterranean Sea; DLM44, from the Atlantic and DLFF1, from a European broodstock); and four turbot specimens (SMFF1, SMFF2 and SMFF3 from a Spanish broodstock; SMNS32 from North Sea's wild population) were used.

Sequence data analysis – standard pipeline

The following approach to sequence data analysis was used for all datasets as the basis for subsequent comparative analysis. Raw data were filtered to retain only high quality reads, using STACKS 1.28 (J. Catchen et al. 2013; J. M. Catchen et al. 2011) *process_radtags* program, which allows simultaneous quality filtering and sample demultiplexing. After barcode removal (5-7 bases), the sequences were 3' end-trimmed to a standard 90 nucleotides length. Each read was then analyzed to assess sequence quality. Briefly, a 3-base sliding window (STACKS' option –w) was used to parse each read and where the average phred score of three consecutive bases was lower than 20 (STACKS' option –s) the entire read was discarded.

STACKS was also used for clustering reads and for SNP discovery, following standard *de novo* and reference based pipelines, well described in the program website

173 (http://catchenlab.life.illinois.edu/stacks/). In our case parameter –m (minimum number of
174 reads to call a stacks) was set to four and –M (maximum number of mismatches between reads
175 to be considered as part of the same cluster) was set to five, according to the suggestion of
176 Mastretta-Yanes (2015), and considering the longer reads of our study (due to concatenation).
177 For the *de novo* approach, reads from primer P1 were concatenated with the reverse
178 complement sequence of reads from primer P2, obtaining 180 bp *pseudo-contigs*. This approach
179 was used to create longer sequence tags which reduces the risk of over-merging (i.e. clustering
180 together tags coming from different genomic regions) by keeping the information about relative
181 proximity of Read 1 and Read 2. As an added benefit, this approach allowed to be fully aware
182 of linkage issues. Since reference based approach require reads to be mapped against a reference,
183 we used the software package BOWTIE (Langmead et al. 2009), considering read pairing in the
184 alignment process. We kept only read pairs that matched a single genomic position.
185 When building the RAD-tag catalog a maximum number of five mismatches between tags was
186 set. For the reference-based approach, clustering was based on mapping position. Within a stack
187 containing variant reads (i.e. potential SNPs), the following call thresholds were used: rare
188 variant frequency (rvf) <0.01 = homozygote; 0.01<rvf<0.1 = 'genotype unknown'; rvf >0.1 =
189 heterozygote called. *rxstacks,* STACKS' component that corrects genotypes on the basis of
190 population information, was also implemented for comparison. Finally, we used the algorithm
191 implemented in STACKS' *populations* step to retain only individual loci represented with at
192 least 10 reads per individual sample and genotyped in at least 80% of the samples analyzed.
193 This is an important step when the genotypes of multiple individuals need to be compared, as
194 only shared loci provide useful information for genetic analysis.

195 *Analysis of the pipeline*

Here, we describe the methods used to assess the pipeline based on the four issues described in the Introduction (Figure 1).

(i) Evaluation of sample representation within multiplexed libraries

Considering the number of samples multiplexed and the average output of the sequencing platform/chemistry (180 M reads), approximately 1.3 M reads per sample are theoretically expected. However, even if initial DNA quantification is accurate and input DNA is equal among samples, subsequent library preparation steps may alter individual representation within the library resulting in variability in inter-sample sequencing effort. To investigate sample read homogeneity in libraries with up to 144 pooled individuals, we first established a threshold number of reads per sample against which to filter individual sample data. A threshold of 150 k reads was chosen as a minimum to accept an individual sample for downstream data processing, based on an expected number of 7,000 stacks per sample (estimated from *in-silico* analysis) and an average coverage of 20x. This threshold was used in the analysis of the sequencing output for all available ddRAD data including more than 5,000 samples.

To identify the factors correlated with fewer reads, we tested the correlation between number of reads (above or below the threshold) and variables such as "DNA quality", whether a sample was "fresh" or "archived", "individual sample collector" (i.e. the project partner that collected the sample), and "index barcode" (different length/sequence barcodes could perform differently in the amplification or sequencing by synthesis steps), testing the effect of each variable under a Generalized Linear Model (GLM), as implemented in R 3.2.3 library function Rcmdr (Team 2013; Fox 2005). Chi-squared tests were applied to check association between tested variables.

For the analysis described further on, only replicate samples with sufficient read numbers were used.

(ii) Tag length and coverage

To understand whether the length of the RAD-tags ('tag length') observed corresponded to the expected length (i.e. the 'insert length' from size selection minus adapter length) and to investigate association between tag length and coverage, we extracted fragment length and DNA sequences of ddRAD-tags from BOWTIE alignment results. Data on coverage depth was extracted for each single locus of each sample, separately. To allow comparison between samples with different average coverage, standardized coverage depth was obtained by dividing locus specific values by the average coverage across all loci for each sample. Similarly, when comparing the distribution of the number of tags with different lengths, 10 bp bins were used and the relative number of tags was calculated dividing the number of tags of a certain length bin by the average number of tags across all the bins. A Wilcoxon signed-rank test, as implemented in R 3.2.3 library Rcmdr (Team 2013; Fox 2005), was used to test for differences between distributions from the three datasets.

(iii) De novo and reference-based genotyping using STACKS

In order to understand how the alignment to a reference genome influences SNP genotyping, we obtained individual genotypes using both *de novo* and reference-based analysis in STACKS. Since we expected *de novo* approach to detect also tags that are not contained in the reference genome, we wanted to evaluate the amount of *de novo* tags that could be found in the genome. In order to do this, RAD-tags resulting from *de novo* analysis (180 bp long) were subsequently split in two (in order to reconstitute the original 90 bp tags) and mapped against the reference genome using BOWTIE, with the same parameters used while aligning reads for reference based analysis. Under both *de novo* and reference-based analysis, results were compared with and without the final step in *rxstacks*. Statistical differences between approaches were tested with

242 chi-squared tests.

243 *(iv) Genotyping precision and error rates*

244 To investigate the level of reproducibility across different bioinformatic approaches we
245 examined the level of consistency among scored SNP genotypes within the sets of nine to 15
246 replicated samples for each species. The most frequent genotypes were considered as the
247 "correct" ones, and mismatches were counted for each locus in each sample to estimate
248 genotyping error.

249 When comparing results from different approaches, statistical significance was tested using
250 either on-line applications (e.g. Kruskall-Wallis: http://vassarstats.net) or the Rcmdr library for
251 R 3.2.3 (Team 2013; Fox 2005). A first global analysis was carried out to assess the effect of
252 several parameters ("coverage", "genome reference" mapping, "*rxstacks* correction",
253 percentage of high-quality reads) on mismatch rate across the entire dataset. Individual
254 mismatch rates were classified either as a binary outcome (0 for values lower than the overall
255 median mismatch rate, 1 for those equal or greater), or grouped into quartiles for a finer
256 evaluation of the effects of different explanatory factors. In both cases, either a Generalized
257 Linear Model (used with binary outcome) or Ordinal Linear Regression (used with samples
258 grouped into quartiles) were used to detect the most influential variables. The same statistical
259 approach was then implemented, within each dataset, across single specimens, to look more
260 into detail at individual-specific features that could affect genotyping quality and to avoid
261 dataset-specific biases and errors. This additional analysis was possible thanks to the large
262 number of replicates available for each species and the standardization of library preparation
263 technique and bioinformatics protocols. Lastly, mismatch rates were analyzed across loci, to
264 check the expectation that, within each "species+strategy" dataset (e.g. sea bream+*de novo* or

turbot+reference based) loci with lower average coverage also showed higher mismatch rates.

Assessing the impact of genotyping accuracy

The impact of genotyping imprecisions on downstream applications varies depending on the type of analysis carried out. Since the principal applications of the project data were the analysis of genetic structure, based on allele frequency, and the development of traceability tools, based on population assignment, we evaluated the impact of variation in genotype scoring on these applications. To do this, we used a reduced set of highly informative markers (14 for sea bass, 15 for sea bream and 18 for turbot) genotyped with both ddRAD and with a single SNP allele-specific assay (KASP). Using these approaches we genotyped 22, 25 and 22 samples of sea bass, sea bream and turbot, respectively. Comparison of the two genotype datasets was conducted at the following three levels: a) *Genotype data:* a simple analysis of genotype mismatch between data from the two approaches analysed in the same individual fish, with missing data differentiated from observed differences in genotype; b) *Allele frequency data: th*e impact of individual genotyping mismatches on allele frequencies was assessed by testing for statistical significance (Student's T-tests) between allele frequencies across all loci, with genotype differences not differentiated from missing data in their effect on allele frequencies; c) *Individual assignment data: a*ssignment was conducted using Geneclass2 software (Piry et al. 2000). Individual assignment scores (%) output from GeneClass2 were used to assess differences in assignment of individual genotypes produced using the two methods and significance was assessed using Student's T-tests. Reference data for population assignment consisted of a larger set of more than 900 wild and farmed samples for the three species genotyped with ddRAD.

Results

The first part of the study addressed the loss of analytical power in terms of number of samples filtered due to unequal representation of individuals within libraries; it was based on a data set of more than 5,581 samples, in which the replicate individuals were included.

(i) Evaluation of sample representation within multiplexed libraries

As indicated by high values of standard deviation (in particular for turbot), variation in the number of raw reads among individuals within species was very high. In fact, 129 samples (71 sea bass, 16 sea bream and 42 turbot) were represented by less than 1,000 reads and three samples (all in turbot dataset) had more than 5,000,000 reads. Using the threshold of 150,000 raw reads, 6.8% of sea bass samples, 8.1% of sea bream samples and 16.0% of turbot samples were discarded. After quality filter was applied, an average of 74.5%±10.8% reads remained available for further analysis. After filtering, the average number of high quality reads was similar across species, 687,426±447,701 in European sea bass, 614,099±406,018 in gilthead sea bream and 610,703±707,152 in turbot. Regression analysis indicated that better quality DNA resulted in higher number of high quality reads ($t= -11.4$ $p<0.001$); similarly, "fresh" samples had a higher amount of high quality reads than "archived samples" ($t= -3.1$ $p<0.005$). "DNA quality" of individual samples was neither significantly associated with species ($X^2=4.6$ $p>0.25$), nor with fresh/archived condition ($X^2=3.1$ $p>0.25$). The DNA of 129 samples showing less than 1,000 reads were all of good quality, which means that inaccurate quantification or pipetting errors are probably what caused this strong under-representation.

After filtering and quality checking, the final number of replicated samples available for downstream analysis was 111: 43 sea bream samples (11 replicates for SAC3, 11 for SAC4, 10 for SAC5 and 11 for SAC6) genotyped across 11 independent libraries, 34 sea bass samples (5 replicates for DLCTY_40, 14 for DLT_1 and 15 for DLM_44) genotyped across 15 libraries

311 and 34 turbot samples (9 replicates for SMFF1, 8 for SMFF2, 9 for SMFF3 and 8 for SMNS32)
312 genotyped across 9 libraries.

313 *(ii) Tag length and coverage*

314 On average across species, 78.4% of the reads were successfully mapped on the reference
315 genomes and mapping rates ranged from 71.3% uniquely mapped reads in sea bream to 85.4%
316 in sea bass.

317 Average tag length across datasets was 288.9±110.5 bp. Most of the tags (79.5%) were 100-
318 380 bp. In addition, substantial fractions (21.1% sea bream, 24.5% sea bass, 15.9% turbot) of
319 analyzed RAD-tags were shorter than 190 bp (the minimum size expected according to the
320 library construction protocol) (Figure 2). Paired-tests between datasets suggested that size
321 distribution was not significantly different across species (Wilcoxon signed-rank test, bream-
322 bass p=0.803, bream-turbot p=0.865, bass-turbot p=0.984).

323 Although average coverage depth per locus differed among datasets for the three species (157
324 ± 94 for sea bass, 248 ± 126 for sea bream, 700 ± 544 for turbot), relative coverage was evenly
325 distributed (Wilcoxon signed-rank test, bream-bass p=0.697, bream-turbot p=0.865, bass-turbot
326 p=0.689) with respect to RAD-tag length (Figure 3). Significant ($p<0.01$) positive linear
327 correlations between length and coverage were also found for fragments in the range from 100
328 to 250 bp (Spearman rho=0.903 in sea bream, 0.957 in sea bass and 0.918 in turbot). Fragments
329 longer than 250 bp showed significant ($p<0.01$) negative linear correlation between length and
330 coverage (Spearman rho=-0.969 in sea bream; -0.968 in sea bass, -0.952 in turbot). No
331 significant correlation between GC content of fragments and coverage depth was observed.

332 *(iii) De novo and reference-based genotyping using STACKS*

333 The number of independent RAD-tags identified varied depending on the approach. In all cases

the number of tags found by the reference genome-based approach was much lower than that found with the *de novo* approach (up to 5.5 times, in turbot dataset) (Table 2). However, when a filter was applied to retain only tags shared by at least 80% of samples analyzed, higher proportion was retained for reference-based analysis (on average 44.9%±19.7%) than *de novo* analysis (on average 9.1%±6.0%). This made that in most cases the final number of retained tags was higher using the reference-based approach. Similarly, a higher number of SNPs was observed in the reference-based approach after filtering. The application of the genotype correction implemented in *rxstacks* reduced the number of tags by different extents: a minimum of 63% of total tags were retained in the turbot reference-based analysis and a maximum of 99.6% in the sea bass *de novo* analysis. The proportion of SNPs retained was comparable, ranging from 56.9% to 99.8% in turbot (reference-based) and sea bass (*de novo*), respectively. Mapping tags from *de novo* analysis against the reference genomes produced 11,121 matches for sea bass (28.3% of *de novo* RAD tags); 11,650 for sea bream (23.0% of *de novo* RAD tags) and 7,889 for turbot (6.8% of *de novo* RAD tags). These figures are in agreement with the relative length of the genomes utilized (Table 1), while the lower than expected difference between sea bass and sea bream results can be explained by the lower quality of the bream assembly, as indicated by the N50 value.

(iv) Genotyping precision and error rates

Our analysis suggested that "*rxstacks* correction" and "coverage" significantly affected the level of accuracy in the comparison of different approaches, regardless the species. In particular, lower mismatch rate were recorded when *rxstacks* was implemented and when coverage depth per sample was higher. However, variation in mismatch rates were found between different species datasets (Table 3); they were apparently linked with differences in species-specific

coverage, which varied significantly both for *de novo* RAD-tags (Kruskall-Wallis test, H=15.27 $p<0.001$) and reference-based ones (Kruskall-Wallis test, H=30.74 $p<0.0001$). To overcome biases linked to species-specific differences, more specific tests were carried out within single datasets. In fact, additional factors were found to be significantly affecting mismatch rate. In addition to "*rxstacks* correction", also "library", "reference-mapping" and "sample" (only in the turbot database) showed significant correlations. At species level, "Coverage" showed a significant correlation in two out of three datasets (sea bream ($p<0.05$) and turbot ($p<0.001$)). Nevertheless, across loci (i.e. within "single species+strategy" dataset) no significant correlation between mismatch and coverage was found.

Impact of genotyping accuracy on population assignment

The percentage of samples with at least one different genotype observed in sea bass, sea bream and turbot was 31.5% (seven samples), 52% (13 samples) and 36% (eight samples), respectively, reflecting a total level of genotyping variation of 3.7% in sea bass (nine allelic differences), 4.8% in sea bream (eighteen allelic differences) and 3.0% in turbot (12 allelic differences). Resulting allele frequencies differed significantly at one locus in sea bream, that displayed the largest single allele frequency difference between the two genotype datasets (11.5%). A similar single locus deviation was observed in sea bass, despite the overall difference being non-significant. The effect of genotyping mismatch on individual assignment was lower; only for turbot, was one individual assigned to different populations of origin with the two genotype datasets. On closer inspection, the two genotypes for this sample differed at three alleles observed at two loci. The two genotypes at first locus were alternate homozygotes, whereas the other discrepancy was between heterozygous and homozygous genotypes at another locus. Further analysis of the sample based on population exclusion testing revealed that neither population of origin (wild or farmed) was

excluded using the two genotype datasets meaning that misassignment would be possible in this scenario. When this individual was excluded from the sample dataset, there was no significant difference in quantitative assignment scores between genotypes from RAD and Kasp approaches. Neither sea bream nor sea bass exhibited discrepancies in population assignment, nor in assignment scores, obtained with different sets of genotype data.

Discussion

The aim of the present work was to quantify the level of genetic information that can be obtained with ddRAD approach, net of information loss during bioinformatic processing; and to evaluate the performance of different bioinformatics approaches on the number of markers detected and the precision of the genotype calling. The use of large datasets of marine fish species and the application of the same approaches as those used in real case studies make our results informative on the practical application of this technique.

(i) Evaluation of sample representation within multiplexed libraries

The first step in which genotyping information is lost is quality filtering, required to obtain reliable results with NGS analysis (Minoche, Dohm, and Himmelbauer 2011; Bokulich et al. 2013). The filtering used in this work was stricter than the default of STACKS filtering, and probably stricter than many of the filtering approaches used in population genetics studies. However in this study our filtering process did not result in the loss of many samples; where samples were removed this was mostly due to low initial read depth resulting from unequal sequencing effort.

Relaxing the STACKs filtering parameters would be one method of retaining more reads, however this would risk increasing genotyping error. An alternative approach, given that low

403 sequence quality is typically concentrated at the read ends, would be to employ further trimming
404 of all reads prior to commencing STACKS analysis. This should have the effect of retaining
405 more reads during subsequent STACKS filtering based on phred-scores. Nevertheless, this
406 procedure still causes loss of potentially high quality genetic information and a more efficient
407 approach would be to trim only those reads affected by low quality instead of trimming every
408 read to the same extent. This would only be possible if the downstream SNP caller program
409 allowed for different length reads (unlike STACKS).
410 One of the main advantages of RAD techniques is the possibility of multiplexing many
411 individuals in the same sequencing run thanks to individual sample barcoding. However, as the
412 number of multiplexed individual samples increases, the chance to have poorly represented
413 samples increases as well (Baird et al. 2008; Peterson et al. 2012), causing lower coverage and
414 in the worst case, too few reliably genotyped or false homozygote excess for a number of
415 individuals. In particular, the combination of samples at different quality/concentration, rather
416 than the quality itself of single samples, is the influencing variable (I.e. even using the exact
417 same starting DNA, result might vary in relation with the other samples genotyped in the same
418 library). The threshold at 150,000 raw reads used here is much lower than the expected average
419 number of reads per individual (1.3 millions) and may not be appropriate for other species. In
420 fact, it should be set taking into consideration the number of expected tag and the desired
421 average coverage depth. However, "losing" a certain amount of samples (up to 16% in our case)
422 needs to be considered when planning a ddRAD sequencing project, even when significant
423 effort was given to equalize DNA input under library preparation.
424 Not surprisingly, DNA quality was a good predictor of poorly performing samples (Graham et
425 al. 2015). Gel-based quality analysis essentially reflects the level of DNA degradation, that can

426 be caused by many factors that act before or after extraction. In our specific case, pre-extraction
427 factors are probably the most relevant, as extraction and post-extraction protocols were the same
428 for all the samples. Ethanol has been recognized as a good media for long term tissue storage
429 (Gillespie et al. 2002; Dawson, Raskoff, and Jacobs 1998), and it is easily available and not
430 hazardous. Nevertheless, Seutin, White, and Boag (1991) reported that ethanol conservation
431 can decrease DNA yield and cause significant degradation to the extracted DNA, that can be
432 reduced by keeping samples refrigerated as soon as possible after sampling. DNA from long-
433 term stored specimens might have some additional features reducing the efficiency in library
434 preparation. Therefore, when selecting the DNA samples to be pooled as part of the same library,
435 it is advisable to avoid mixing samples of heterogeneous DNA quality as well as mixing "fresh"
436 with "archived" specimens. When this is not possible (e.g. for those projects that use only one
437 or few sequencing pools), an upward correction for the starting amount of DNA of poor quality
438 samples and DNA from "archived" samples might be considered. However, further analysis is
439 necessary to better understand how this procedure should be applied.

440 *(ii) Tag length and coverage*

441 Accuracy and consistency in size selection is not easily achievable, but tag length distribution
442 was not significantly different across species in our study. From this point of view, the period
443 of training of the personnel proved to be effective in order to have consistent results.
444 Nevertheless, tags shorter than 190 bp were retained in our analysis, which was unexpected
445 considering that size selection step was implemented. Indeed, low accuracy has been
446 documented in particular for manual *vs* automated gel band extraction (Puritz 2015). A similar
447 result was found by DaCosta e Sorenson (2014), who recovered tags down to a length of 10 bp.
448 In our case, the purification steps performed at the very end of the library preparation protocol,

should eliminate most inserts shorter than 200 bp, that translates into RAD tags of 75 bp, after removing adapters. It is important to notice that, considering the 100 bp paired-end sequencing protocol used, all the analyzed fragments shorter than 190 bp are affected by Read1-Read2 overlapping of the final parts of the sequences, potentially causing SNP duplication, redundant data and a waste of sequencing effort that further lower the actual power of ddRAD technique. Improvement in size selection step is fundamental to optimize the performance of the ddRAD technique.

Davey et al. (2013), using data from a *Caenorhabditis elegans* RAD library, found a strong positive correlation between fragment length and coverage depth. In other published ddRAD studies, such as DaCosta e Sorenson (2014), the relationship between coverage and length was similar to our work. Tags with different lengths show variable coverage within individual samples. This means that additional care should be taken when multiplex size is calculated, in order to achieve a desired minimum depth of coverage across loci. According to our results, loci in the shortest and longest length range will be underrepresented if coverage was calculated just by dividing the number of individual reads by the number of expected loci. Upward correction in the number of reads per individual should be applied to obtain minimum coverage also for loci in short and long fragments.

(iii) De novo and reference-based genotyping using STACKS

The possibility to use RAD techniques in species without genomic resources (i.e. *de novo* approach) has been highlighted as one of the method's biggest advantages (Willing et al. 2011; Pegadaraju et al. 2013). However, we showed that using a reference genome improves RAD genotyping performance, i.e. better precision and higher number of shared markers. With reference based approach, only reads correctly mapped against the genome are used. Hence, the

quality of reference-based analysis is also dependent on the quality of the assembly used. In particular, N50 seemed to better predict mapping percentage compared to average contig length. Turbot shows the longest average contig length, but ranked second in terms of positive mapping matches, in agreement with N50 ranking (Table 2). J. Catchen et al. (2013) showed that in threespined stickleback *de novo* approach yielded a higher number of tags (42,300) than the reference based one (37,600), mostly due to loss of loci that could not be mapped against the reference genome (>4,700). Likewise, in our analysis, using the genome as a reference returned a lower number of tags compared to the *de novo* approach (Table 3). In any case, the number of *de novo*-based tags that mapped correctly to the reference genome was in good agreement with the number of tags identified by the reference based analysis. The larger number of *de novo* ddRAD tags might then be explained in part by the incomplete mapping of reads against the reference genome as in the case of threespined stickleback. This seems reasonable considering that the reference sequences used represent only a portion of the entire genomes of the species. Indeed, compared to the genome lengths estimated from the c-values, from 70% (turbot) to 85%(sea bream) of the entire genomes is represented in the references. So, at least part of the *de novo* loci found could be real fragments of the genomes, coming from regions that have been more difficult to sequence and assemble so far. A second possibility is that a fraction of tags, which STACKS identified as separate "loci" in the *de novo* analysis, is likely represented by divergent alleles of the same locus. However, STACKS controls for such phenomenon through the –M parameter and, in the present study, a less conservative value (–M=5) than the default one (–M=2) was set for all species. More likely, *de novo* approach might include some "spurious" loci at individual level. In support of this hypothesis, a filter that exclude loci shared by less than 80% of individuals, filtered out most of *de novo* loci. The origin

of these tags is difficult to find but some sources can be the presence of exogenous DNA, e.g. from viral/ bacterial contaminants or from other species. In fact, blasting the tags that did not present matches with sea bream genome showed that around 20% of these tags could come from virus, bacteria, human or from other species analyzed in the laboratory (data not shown). Using a filter that prunes poorly shared loci would exclude these RAD. In any case, when using a *de novo* approach, a further filter based on alignment of tags with potential contaminant species should be implemented as it require little bioinformatic effort and reduces a potential source of background noise in the results. In addition, we cannot exclude the presence of sequencing errors introduced with amplification in library preparation and sequencing steps. While we cannot exclude that these sequences can provide useful information or could be used as dominant markers (Fu et al, 2013), we recognize that they need to be studied more in detail to understand their origin and whether they can have bad effects on certain downstream applications (i.e. those requiring the use of markers shared by a percentage of individuals). Without deeper knowledge of the origin of these sequences, it is therefore advisable to use the above mentioned filters to reduce source of bias in the final filtered datasets. In general, even if in the form of a draft, a reference genome should allow more efficient SNP detection.

(iv) Genotyping precision and error rates

Genotyping reproducibility across technical replicates is one of the most important test to evaluate genotyping methods. A first analysis on over 100 replicates over the three species datasets, showed that "coverage" represented a significant explanatory variable for differences in mismatch rates. In fact, sea bass' technical replicates, which were characterized by a significantly lower coverage, also showed lower precision than the two other species. The effect of reduced coverage also appears to be affecting samples characterized by a high DNA quality.

518 Davey et al. (2011) suggested at least 30x average coverage depth for reference genome-based
519 analysis and at least 60x coverage depth for *de novo* analysis in order to obtain a complete
520 coverage of all restriction sites in a genome. In addition, Fountain (2016), based on Mendelian
521 inheritance incompatibilities, showed that genotyping errors decreased with increasing
522 coverage from 5x to 30x in both reference based and *de novo* datasets. In the present study, the
523 average coverage for all the three species was higher than that suggested in the studies
524 mentioned, but also the variability across loci was high (36x-386x in sea bass, 31x-2840x in
525 sea bream and 69x-2731x in turbot), which might influence the outcome in term of mismatch
526 rates. However, we couldn't find any significant correlation between mismatch rate and average
527 locus coverage when analyzing results within single datasets (i.e. "species+strategy").
528 The same analysis showed that the SNPs in the reference-based tags are more consistently
529 genotyped than *de novo* ones in both turbot and sea bream. The positive effect of using a
530 reference genome on genotyping reproducibility is an additional one to the advantage of
531 avoiding inflation of tag number described above. More reproducible genotypes are also
532 obtained when *a posteriori* genotype correction was implemented. Both approaches (reference-
533 based analysis and *a posteriori* correction) come at a price as the total number of tags/SNPs
534 analyzed gets reduced, so its use should be considered to obtain more reliable data according to
535 the aims of a particular project. Other issues deserve care when using genetic markers. On of
536 the most important is allelic dropout, that can lead to errors in population statistics, due to biased
537 heterozygosity estimates (Gautier et al., 2013). Some approaches have been proposed to detect
538 loci affected by allelic dropout, such as analysis of coverage, based on the expectation that
539 mutations within the restriction site would result in fewer reads being generated for one allele,
540 thus creating a bi-modal read depth distribution across loci (Cooke et al., 2016). Nevertheless,

541 a preliminary analysis of coverage distribution of our data didn't show the bi-modal distribution
542 expected with the high coverage obtained. This may be due to the fact that coverage also varies
543 with tag length, swamping any signal from allelic drop-out.
544 *Impact of genotyping accuracy on population assignment*
545 The average level of discrepancy between the test datasets used for the comparison was slightly
546 higher than the genotyping error recorded with the different ddRAD bioinformatic approaches,
547 but allowed us to detect the putative threshold at which genotyping imprecision starts affecting
548 downstream analysis, since we recorded both samples with and without assignment or allele
549 frequency deviations. In any case, levels of mismatch between replicates higher than those
550 found in this study (and approaching the threshold identified as causing deviation in
551 downstream analysis) are commonly found when using RAD genotyping (e.g. Forsström et al.
552 2017; Pecoraro et al. 2016). According to our results, the effect of sub-4 % genotyping
553 differences on allele frequency is not significant, while if variation increased (e.g. 4.8 % in sea
554 bream), the resulting allele frequencies were significantly different. Such findings are clearly
555 dependent on factors such as sample size and distribution of variation over loci, but do provide
556 an indication of the point at which genotype variation impacts allele frequency. Population
557 assignment was only affected when levels of genotyping variation were higher; the only
558 discrepancy being recorded for turbot where, despite displaying the lowest overall genotype
559 variation of the three species, one sample with three allele differences (8.3% of 36 alleles), was
560 assigned to different source populations. Again, such results will depend on which loci
561 displayed genotype variation and cannot be used in isolation to define threshold errors.
562 However, the finding does indicate the potential for realistic levels of genotype error to result
563 in significant changes to diagnostic results if not accounted for when evaluating the accuracy

of downstream applications.

Conclusions

Application of new genotyping techniques is rapidly increasing as they potentially allow more accurate, easier and less expensive population genetic analysis of any species. However, several issues might affect the quality of the results. In the present study, it was demonstrated that some factors, i.e. DNA fragmentation and archived-fresh samples, affect the throughput in terms of percentage and absolute number of high quality sequence reads in ddRAD datasets. Similarly, actual fragment length and coverage can differ from expectations, leading to redundant loci and loci with too low coverage. Although RAD has been proven to be applicable on non-model species, the use of a preliminary draft genome sequence increase genotyping performance enabling to obtain higher numbers of loci shared between multiplexed individuals. We highlight the critical importance of introducing replicate individuals among samples to assess the performance of the approach used and we demonstrate how variation in genotype datasets can potentially impact the results of downstream population genetic applications. Our results are useful for setting up genotyping project and for considering the features that can affect genotyping throughput and precision.

Data deposition

Raw sequencing data are available at NCBI, with accession numbers SAMN7145243-7145512.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme under agreement no. KBBE-311920 (AQUATRACE). Samples used in this study were collected by the AQUATRACE partners: Laboratory of Biodiversity and Evolutionary Genomics (LBEG), University of Leuven (European sea bass); Departmento de

587 Genética, Universidad de Santiago de Compostela (turbot); Sabina De Innocentiis, ISPRA
588 (gilthead sea bream). JH benefited from a PhD scholarship from Flanders Innovation &
589 Entrepreneurship (IWT). GK was funded by a PhD scholarship of Onassis Foundation.

590 **References**

591 Andrews, Kimberly R., Jeffrey M. Good, Michael R. Miller, Gordon Luikart, e Paul A.
592 Hohenlohe. 2016. «Harnessing the power of RADseq for ecological and evolutionary
593 genomics». *Nature Reviews Genetics* 17 (2): 81–92.
594 Arnold, B., R. B. Corbett-Detig, D. Hartl, e K. Bomblies. 2013. «RADseq underestimates
595 diversity and introduces genealogical biases due to nonrandom haplotype sampling».
596 *Molecular Ecology* 22 (11): 3179–90.
597 Baird, Nathan A., Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver,
598 Zachary A. Lewis, Eric U. Selker, William A. Cresko, e Eric A. Johnson. 2008.
599 «Rapid SNP discovery and genetic mapping using sequenced RAD markers». *PloS*
600 *one* 3 (10).
601 Bokulich, Nicholas A., Sathish Subramanian, Jeremiah J. Faith, Dirk Gevers, Jeffrey I.
602 Gordon, Rob Knight, David A. Mills, e J. Gregory Caporaso. 2013. «Quality-filtering
603 vastly improves diversity estimates from Illumina amplicon sequencing». *Nature*
604 *methods* 10 (1): 57–59.
605 Catchen, Julian, Paul A. Hohenlohe, Susan Bassham, Angel Amores, e William A. Cresko.
606 2013. «Stacks: an analysis tool set for population genomics». *Molecular ecology* 22
607 (11): 3124–40.
608 Catchen, Julian M., Angel Amores, Paul Hohenlohe, William Cresko, e John H. Postlethwait.
609 2011. «Stacks: building and genotyping loci de novo from short-read sequences». *G3:*
610 *Genes, Genomes, Genetics* 1 (3): 171–82.
611 Cooke, T. F., Yee, M. C., Muzzio, M., Sockell, A., Bell, R., Cornejo, O. E., ... & Kenny, E. E.
612 (2016). «GBStools: a statistical method for estimating allelic dropout in reduced

representation sequencing data». *PLoS genetics*, *12*(2), e1005631.

Cruz, Vanessa P., Manuel Vera, Belén G. Pardo, John Taggart, Paulino Martinez, Claudio Oliveira, e Fausto Foresti. 2016. «Identification and validation of single nucleotide polymorphisms as tools to detect hybridization and population structure in freshwater stingrays». *Molecular Ecology Resources*.

DaCosta, Jeffrey M., e Michael D. Sorenson. 2014. «Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol».

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). «Genome-wide genetic marker discovery and genotyping using next-generation sequencing». *Nature reviews. Genetics*, *12* (7): 499

Davey, John W., Timothée Cezard, Pablo Fuentes-Utrilla, Cathlene Eland, Karim Gharbi, e Mark L. Blaxter. 2013. «Special features of RAD Sequencing data: implications for genotyping». *Molecular Ecology* 22 (11): 3151–64.

Dawson, Mike N., Kevin A. Raskoff, e David K. Jacobs. 1998. «Field preservation of marine invertebrate tissue for DNA analyses». *Molecular marine biology and biotechnology* 7 (2): 145–52.

Figueras, Antonio, Diego Robledo, André Corvelo, Miguel Hermida, Patricia Pereiro, Juan A. Rubiolo, Jèssica Gómez-Garrido, Laia Carreté, Xabier Bello, e Marta Gut. 2016. «Whole genome sequencing of turbot (Scophthalmus maximus; Pleuronectiformes): a fish adapted to demersal life». *DNA Research*, dsw007.

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). «Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates». *Molecular ecology resources*, *16* (4): 966-978

636 Forsström, T., Ahmad, F., & Vasemägi, A. (2017). «Invasion genomics: genotyping-by-
637 sequencing approach reveals regional genetic structure and signatures of temporal
638 selection in an introduced mud crab». *Marine Biology*, *164*(9), 186
639 Fox, John. 2005. «Getting started with the R commander: a basic-statistics graphical user
640 interface to R». *Journal of statistical software* 14 (9): 1–42.
641 Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., ... & Estoup, A.
642 (2013). «The effect of RAD allele dropout on the estimation of genetic variation
643 within and between populations». *Molecular Ecology*, *22* (11), 3165-3178.
644 Gillespie, John W., Carolyn JM Best, Verena E. Bichsel, Kristina A. Cole, Susan F. Greenhut,
645 Stephen M. Hewitt, Mamoun Ahram, Yvonne B. Gathright, Maria J. Merino, e Robert
646 L. Strausberg. 2002. «Evaluation of non-formalin tissue fixation for molecular
647 profiling studies». *The American journal of pathology* 160 (2): 449–57.
648 Gomes, I., A. Collins, C. Lonjou, N. S. Thomas, J. Wilkinson, M. Watson, e N. Morton. 1999.
649 «Hardy–Weinberg quality control». *Annals of human genetics* 63 (6): 535–38.
650 Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S.,
651 Manzon, R. G., Martino, J. A., Pierson, T., Rogers, S. M., Wilson, J. Y. and Somers,
652 C. M. 2015. Impacts of degraded DNA on restriction enzyme associated DNA
653 sequencing (RADSeq). *Molecular Ecology Resources* 15: 1304–1315.
654 Langmead, Ben, Cole Trapnell, Mihai Pop, e Steven L. Salzberg. 2009. «Ultrafast and
655 memory-efficient alignment of short DNA sequences to the human genome». *Genome*
656 *biol* 10 (3): R25.
657 Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, e Maggie
658 Law. 2012. «Comparison of next-generation sequencing systems». *BioMed Research*

659 *International* 2012.
660 Mastretta-Yanes, A., Nils Arrigo, Nadir Alvarez, Tove H. Jorgensen, D. Piñero, e B. C.
661 Emerson. 2015. «Restriction site-associated DNA sequencing, genotyping error
662 estimation and de novo assembly optimization for population genetic inference».
663 *Molecular ecology resources* 15 (1): 28–41.
664 Minoche, André E., Juliane C. Dohm, e Heinz Himmelbauer. 2011. «Evaluation of genomic
665 high-throughput sequencing data generated on Illumina HiSeq and genome analyzer
666 systems». *Genome Biol* 12 (11): R112.
667 Narum, Shawn R., C. Alex Buerkle, John W. Davey, Michael R. Miller, e Paul A. Hohenlohe.
668 2013. «Genotyping-by-sequencing in ecological and conservation genomics».
669 *Molecular Ecology* 22 (11): 2841–47.
670 Pecoraro, C., Babbucci, M., Villamor, A., Franch, R., Papetti, C., Leroy, B., & Murua, H.
671 (2016). «Methodological assessment of 2b-RAD genotyping technique for population
672 structure inferences in yellowfin tuna (Thunnus albacares)». *Marine genomics*, *25*, 43-
673 48
674 Pegadaraju, Venkatramana, Rick Nipper, Brent Hulke, Lili Qi, e Quentin Schultz. 2013. «De
675 novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site
676 Associated DNA) approach». *BMC genomics* 14 (1): 556.
677 Peterson, Brant K., Jesse N. Weber, Emily H. Kay, Heidi S. Fisher, e Hopi E. Hoekstra. 2012.
678 «Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and
679 Genotyping in Model and Non-Model Species». *PLoS ONE* 7 (5): e37135.
680 doi:10.1371/journal.pone.0037135.
681 Piry, Sylvain, Alapetite, A., Cornuet, J. M., Paetkau, D., Baudouin, L., & Estoup, A. (2004).

«GENECLASS2: a software for genetic assignment and first-generation migrant detection». *Journal of heredity*, *95*(6), 536-539

Puritz, Jonathan B. 2015. «Fishing for Selection, but Only Catching Bias: Examining Library Effects in Double-Digest RAD Data in a Non-Model Marine Species». In *Plant and Animal Genome XXIII Conference*. Plant and Animal Genome.

Puritz, Jonathan B., Mikhail V. Matz, Robert J. Toonen, Jesse N. Weber, Daniel I. Bolnick, e Christopher E. Bird. 2014. «Demystifying the RAD fad». *Molecular ecology* 23 (24): 5937–42.

Seutin, Gilles, Bradley N. White, e Peter T. Boag. 1991. «Preservation of avian blood and tissue samples for DNA analyses». *Canadian Journal of Zoology* 69 (1): 82–90.

Taberlet, Pierre, Sally Griffin, Benoît Goossens, Sophie Questiau, Valérie Manceau, Nathalie Escaravage, Lisette P. Waits, e Jean Bouvet. 1996. «Reliable genotyping of samples with very low DNA quantities using PCR». *Nucleic acids research* 24 (16): 3189–94.

Team, R. Core. 2013. «R: A language and environment for statistical computing».

Tine, Mbaye, Heiner Kuhl, Pierre-Alexandre Gagnaire, Bruno Louro, Erick Desmarais, Rute ST Martins, Jochen Hecht, Florian Knaust, Khalid Belkhir, e Sven Klages. 2014. «European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation». *Nature communications* 5.

Wang, Shi, Eli Meyer, John K. McKay, e Mikhail V. Matz. 2012. «2b-RAD: A Simple and Flexible Method for Genome-Wide Genotyping». *Nature Methods* 9 (8): 808–10. doi:10.1038/nmeth.2023.

Willing, Eva-Maria, Margarete Hoffmann, Juliane D. Klein, Detlef Weigel, e Christine Dreyer. 2011. «Paired-end RAD-seq for de novo assembly and marker design without

705 available reference». *Bioinformatics* 27 (16): 2187–93.
706 Xu, Jianfeng, Aubrey Turner, Joy Little, Eugene R. Bleecker, e Deborah A. Meyers. 2002.
707 «Positive results in association studies are associated with departure from Hardy-
708 Weinberg equilibrium: hint for genotyping error?» *Human genetics* 111 (6): 573–74.
709

Figure 1 Flowchart of the analysis pipeline followed in this study, indicating the results evaluated in order to understand the performances of ddRAD sequencing technique

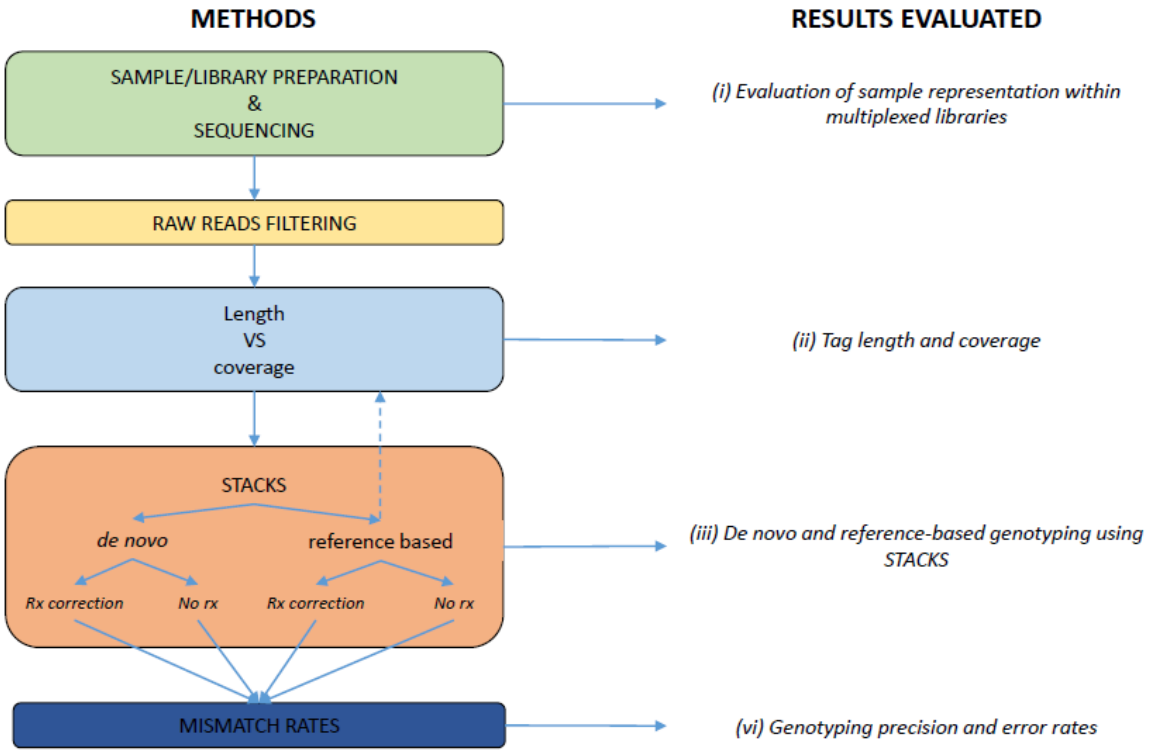


Figure 2: Graph of fragment length vs number of fragments in European sea bass (square), gilthead sea bream (diamond) and turbot (triangle). The graph is based on the reference-based analysis, as only for this it was possible to obtain information about fragments' length. Dash vertical line indicates the limit under which pair-end tags present overlapping between Read 1 and Read 2.

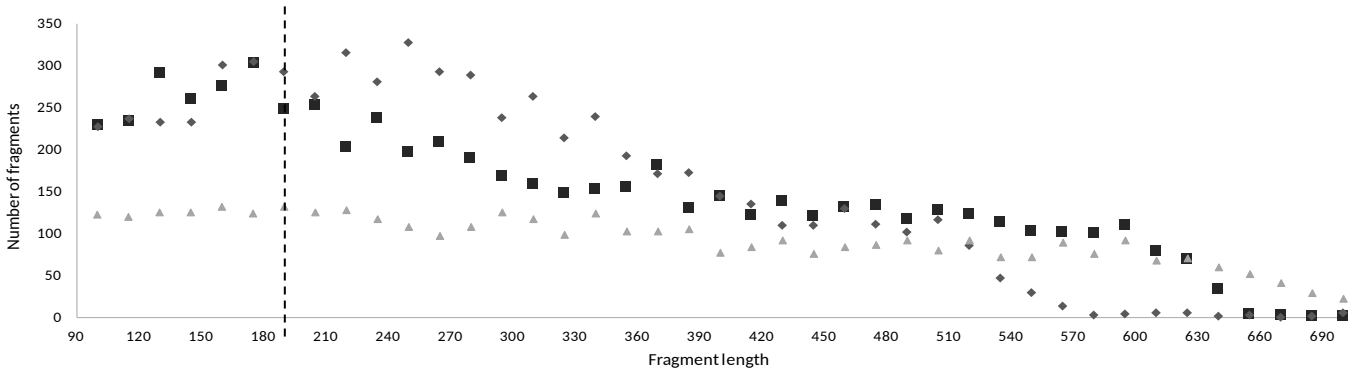


Figure 3: Graph of fragment length vs coverage depth in European sea bass (squares), gilthead sea bream (diamonds) and turbot (triangles). The graph is based on the reference-based analysis, as only here it was possible to obtain information about fragments' length. Coverage is expressed as relative to specific average coverage, in order to account for difference between species in average coverage depth. Trend lines were calculated as polynomial, third order for sea bass (solid line, R2=0.70), sea bream (dash, R2=0.93), turbot (point, R2=0.89)

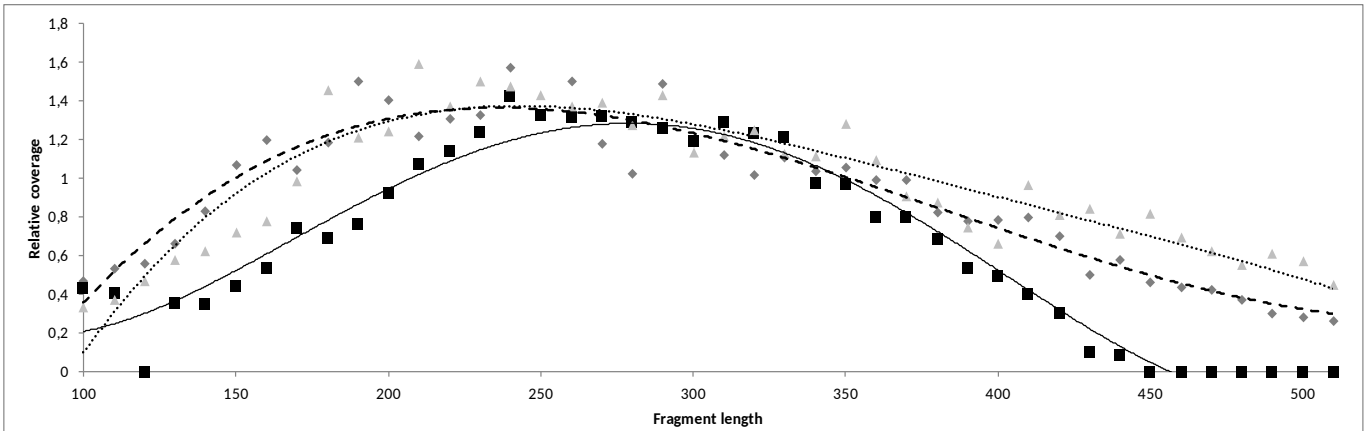


Table 1 Details of the genome resources used for European sea bass, gilthead sea bream and turbot.

Species	Length (Mbp)	N° of contigs	Average contig length	N50 (kbp)	Reference
European sea bass	668.3	37,783	17,687	62	Tine et al., 2014
Gilthead sea bream	770.3	259,783	2,965	13.35	Bargelloni et al., unpublished
Turbot	544.2	16,463	33,058	31.2	Figueras et al. 2016

Table 2 Summary of the STACKS' analyses on European sea bass, gilthead sea bream and turbot using de novo and reference based approaches. Application of the correction sub-program rxstacks is indicated under column 'Correction'. SNP frequency is calculated as the number of base pairs analyzed (180 bp x number of tags for the de novo approach; 90 bp x number of tags for the reference based approach) and the SNPs detected. 'Tags 80%' indicates the number of tags after filtering for those shared by at least 80% of individuals analyzed.

Species	Type of analysis	Correction	Tags	SNPs	SNP freq	Tags 80%	Average coverage
European sea bass	*de novo*	No correction	19,672	16,342	216.7	3,246	111.0 ± 65.9
		rxstacks	19,595	15,612	225.9	1,347	101.51 ± 59.6
	reference based	No correction	13,458	3,013	402.0	4,913	156.8 ± 94.3
		rxstacks	13,379	3,007	400.4	1,764	153.9 ± 92.9
Gilthead sea bream	*de novo*	No correction	25,322	39,842	114.4	3,913	151.5 ± 72.0
		rxstacks	24,257	31,790	137.3	2,353	89.3 ± 48.3
	reference based	No correction	13,659	5,161	238.2	7,091	247.7 ± 126.4
		rxstacks	12,293	4,388	252.1	5,796	109.9 ± 52.6
Turbot	*de novo*	No correction	58,171	26,635	393.1	1,674	272.1 ± 226.8
		rxstacks	56,320	21,582	469.7	1,631	157.3 ± 150.2

reference based	No correction	8,887	2,530	316.1	4,175	700.9 ± 544.6
	rxstacks	5,595	1,440	346.7	4,106	255.4 ± 230.3

Table 3 Summary of mismatch analysis on European sea bass, gilthead sea bream and turbot using de novo and reference based approaches. Values are given as average or median percentage of genotypes that differ from the consensus (most frequently recorded) genotype over the total number of genotypes analyzed (number of individuals analyzed x number of SNPs). Application of correction subroutine rxstacks is indicated under column 'Correction'.

Species	Type of analysis	Correction	Average % of mismatches	Median % of mismatches
Sea bass	*de novo*	No correction	2.9	0.9
		rxstacks	2.9	0.9
	reference based	No correction	1.9	0.5
		rxstacks	1.7	0.4
Sea bream	*de novo*	No correction	0.7	0.3
		rxstacks	1.3	0.3
	reference based	No correction	0.2	0.2
		rxstacks	0.1	0.1
Turbot	*de novo*	No correction	0.5	0.2
		rxstacks	0.6	0.1
	reference based	No correction	0.4	0.2
		rxstacks	0.3	0.1

Performance and precision of double digestion RAD (ddRAD) genotyping in multiplexed datasets of marine fish species

Supplementary Material

Detailed library preparation protocol

Each group used biochemical consumables from the same manufacturers and were supplied with custom barcoded ddRAD adapters mixes, sourced from the same original stocks prepared at the Institute of Aquaculture, Stirling.

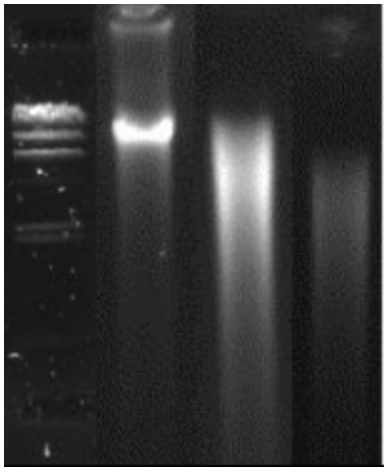
The original protocol of Peterson et al. (2012) involved processing each sample separately (i.e. restriction digestion, adapter ligation, fragment size selection, PCR amplification and purification, quantitation) prior to pooling into a single library for sequencing. A modified protocol (described in detail elsewhere; Palaikostas et al. 2014; Manousaki et al. 2016), which was more convenient for screening large numbers of individuals, was used for this project. The methodology allowed for pooling of samples after the adapter ligation step, which greatly reduced the number of manipulations required, ensured consistent size selection within libraries and reduced construction time to two to three working days. Library preparation began with basic qualitative and quantitative assessment of extracted DNA samples. DNA quality was evaluated by gel electrophoresis (0.8% agarose 0.5x TAE) and concentration was accurately measured by fluorimetry with each sample being finally diluted to 7 ng/μL in 5 mM Tris pH 8.5. For a library (144 samples), individual DNA samples (21 ng) were first simultaneously digested with *Sbf*I (recognition site CCTGCA'GG) and *Sph*I (recognition site GCATG'C) restriction enzymes, at 37° during 45 minutes. An adapter mix comprising individual-specific barcoded combinations of P1 (*Sbf*I-compatible) and P2 (*Sph*I-compatible) HPLC purified adapters (compatible with Illumina sequencing chemistry) were then added / ligated. Adapters were designed such that adapter– genomic DNA ligations did not reconstitute RE sites, residual RE activity limiting concatemerization of genomic fragments. Each adapter included an inline five- or seven-base barcode, allowing for post-sequencing identification of individuals (P1-P2 combinatorial barcoding). The ligation reactions were terminated by heat inactivation and all 144 samples combined in a single pool.

Following column purification of the pooled sample, DNA fragments in the range of 320 bp to 590 bp were size selected by agarose gel electrophoresis, followed by gel-based column purification. The eluted size-selected DNA template was then PCR amplified (14 cycles, 400 uL volume), column purified down to a 50 uL volume and then subjected to a further clean-up using an equal volume of AMPure magnetic beads (Perkin-Elmer, UK) (used in sea bream and turbot), to maximize removal of small fragments (less than ca. 200 bp). The final library was eluted in c.20 µL10 mM Tris pH 8.5.

Libraries were sequenced on Illumina HiSeq 2500 sequencers with pair-end (PE) 100 base option to allow sequencing of both barcodes at the Genomics Core of the University of Leuven, Belgium (sea bass and sea bream) and BMR S.r.l, Padova, Italy (turbot).

DNA quality from agarose gel electrophoresis

Figure Example of "high" (a), "mid" (b) and "low" (c) quality DNA taken from agarose gel of DNA samples used in the study. On the leftmost well run 1 kb ladder.



References

Manousaki, T., Tsakogiannis, A., Taggart, J. B., Palaiokostas, C., Tsaparis, D., Lagnel, J., & Tsigenopoulos, C. S. (2016). Exploring a Nonmodel Teleost Genome Through RAD Sequencing—Linkage Mapping in Common Pandora, Pagellus erythrinus and Comparative Genomic Analysis. *G3: Genes| Genomes| Genetics*, *6*(3), 509-519.

47 Palaiokostas, C., Bekaert, M., Khan, M. G., Taggart, J. B., Gharbi, K., McAndrew, B. J., & Penman, D. J. (2015).
48 A novel sex-determining QTL in Nile tilapia (Oreochromis niloticus). *BMC genomics*, *16*(1), 171.
49 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an
50 inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*,
51 *7*(5), e37135.