

1 **Helping decision making for a reliable and cost-effective 2b-RAD sequencing and**
2 **genotyping analyses in non-model species**

3

4

5 Anna Barbanti^{1*}, Hector Torrado^{1,2*}, Enrique Macpherson², Luca Bargelloni³, Raffaella Franch³, Carlos
6 Carreras^{1§}, Marta Pascual^{1§}.

7

8 1.Department of Genetics, Microbiology and Statistics and IRBio, University of Barcelona, Diagonal 643,
9 08028 Barcelona, Spain.

10 2.Center for Advanced Studies of Blanes (CEAB-CSIC). C/ d'accés a la Cala St. Francesc, 14, 17300 Blanes,
11 Girona, Spain.

12 3.Department of Comparative Biomedicine and Food Science. University of Padova, Viale dell'Universita' 16,
13 I-35020 Legnaro, Italy.

14

15 * These authors contributed equally to the manuscript

16 § These author are considered senior authors

17

18 Keywords: Conservation genomics, high-throughput sequencing, *Caretta caretta*, *Diplodus puntazzo*,
19 genotyping-by-sequencing, sequencing simulations

20 **Abstract**

21 High-throughput sequencing has revolutionized population and conservation genetics. RAD sequencing
22 methods, such as 2b-RAD, can be used on species lacking a reference genome. However, transferring
23 protocols across taxa can potentially lead to poor results. We tested two different IIB enzymes (AlfI and CspCI)
24 on two species with different genome sizes (the loggerhead turtle *Caretta caretta* and the sharpshout
25 seabream *Diplodus puntazzo*) to build a set of guidelines to improve 2b-RAD protocols on non-model
26 organisms while optimising costs. Good results were obtained even with degraded samples, showing the
27 value of 2b-RAD in studies with poor DNA quality. However, library quality was found to be a critical
28 parameter on the number of reads and loci obtained for genotyping. Resampling analyses with different
29 number of reads per individual showed a trade-off between number of loci and number of reads per sample.
30 The resulting accumulation curves can be used as a tool to calculate the number of sequences per individual
31 needed to reach a mean depth ≥ 20 reads to acquire good genotyping results. Finally, we demonstrated that
32 selective-base ligation does not affect genomic differentiation between individuals, indicating that this
33 technique can be used in species with large genome sizes to adjust the number of loci to the study scope, to
34 reduce sequencing costs and to maintain suitable sequencing depth for a reliable genotyping without
35 compromising the results. Finally, we provide a set of guidelines to improve 2b-RAD protocols on non-model
36 organisms with different genome sizes, helping decision-making for a reliable and cost-effective genotyping.

37

38 **Introduction**

39 High-throughput sequencing technologies have revolutionized the fields of population and conservation
40 genetics during the last ten years by providing easy access to genomic data from virtually any taxonomic
41 group (Andrews & Luikart, 2014; Bellin et al., 2009; Davey & Blaxter, 2011; Hudson, 2008). Many studies have
42 explored the potential of genomic analysis to address a variety of questions, such as population structuring
43 (Girault, Blouin, Vergnaud, & Derzelle, 2014), inbreeding depression (Hoffman et al., 2014), local adaptation
44 (Savolainen, Lascoux, & Merilä, 2013) or hybridization (Hohenlohe, Amish, Catchen, Allendorf, & Luikart,
45 2011). Restriction site associated techniques (RAD) are based on massive sequencing after enzymatically
46 reducing the fraction of the genome being analysed and can identify and score thousands of genetic markers,
47 randomly distributed across the genome in many individuals simultaneously (Davey & Blaxter, 2011;
48 Pecoraro et al., 2016). The advantage of these methodologies is that they can be carried out with no or
49 limited previous sequence knowledge, since RAD tags can be analysed using pipelines for *de novo* loci
50 identification if a reference genome is not available (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013;
51 Davey & Blaxter, 2011; Hapke & Thiele, 2016; Lu, Glaubitz, Harriman, Casstevens, & Elshire, 2012). These
52 methods allow parallel and multiplexed sample sequencing of tag libraries, with a rapid and very cost-
53 effective procedures resulting in high genome coverage (Baird et al., 2008; Pecoraro et al., 2016). The RAD
54 marker approach has the flexibility to assay different number of markers depending on the restriction
55 enzyme of choice (Baird et al., 2008).

56
57 Many studies focusing on population structure in non-model organisms have implemented different RAD
58 technologies, such as RADseq (e.g. Lim et al., 2017; Xu et al., 2014), ddRAD (e.g. Lavretsky, DaCosta, Sorenson,
59 McCracken & Peters, 2019; Portnoy et al., 2015), GBS (e.g. Carreras et al., 2017; Hess et al., 2015), and 2b-
60 RAD (e.g. Boscari et al., 2019; Galaska, Sands, Santos, Mahon, & Halanych, 2017). By shifting the realms of
61 genomics from laboratory-based studies of model species towards studies of natural populations of
62 ecologically well-characterized organisms, researchers can now start to address important ecological and
63 evolutionary questions on a scale and precision that, only a few years ago, was unrealistic (Ekblom & Galindo,

64 2011). As for all genotyping-by-sequencing methodologies, the mean number of reads per locus (mean depth
65 of coverage) is crucial to consider reliable the quality of markers and their genotypes (Sims, Sudbery, Ilott,
66 Heger, & Ponting, 2014). Some recent population studies prioritised the number of sequenced individuals
67 over depth of coverage or used improved bioinformatics pipelines to extract information from low coverage
68 data (Buerkle & Gompert, 2013; Maruki & Lynch, 2014). However, when depth is generally low, statistical
69 uncertainty of individual sequence data is high and calling of genotypes is difficult (Maruki & Lynch, 2017).
70 Although probabilistic genotyping methods are thought to overcome shortcomings of low-depth sequencing
71 data, they may behave unpredictably when compared to high-depth data (Hendricks et al. 2018). Thus, any
72 analysis involving individual genotypes is going to be limited by coverage (Chow, Anderson & Shedlock, 2019).
73 For this reason, RAD sequencing techniques and laboratory protocols should be adjusted to target enough
74 sequencing depth to obtain reliable genotypes while optimising sequencing costs.

75

76 2b-RAD is a RAD methodology that uses IIB restriction endonucleases, which cleave genomic DNA upstream
77 and downstream of the target sites producing 32-34 bp fragments (Wang, Meyer, McKay, & Matz, 2012). This
78 method is simple and provides a cost-effective alternative to existing reduced representation genotyping
79 methods, allowing its use in routine experimental laboratory (Baird et al., 2008; Luo et al., 2017; Wang et al.,
80 2012). One of the most interesting features of 2b-RAD is that the number of loci/marker density can be
81 adjusted by using selective adaptors (Wang et al., 2012) to reduce the number of expected markers and
82 increase the coverage per locus for a given sequencing effort. This RAD sequencing technique has been used
83 to identify candidate genes associated with specific traits (Luo et al., 2017), to construct ultra-high density
84 genetic maps (Fu, Liu, Yu, & Tong, 2016), to identify genomic regions under selection in population genetic
85 studies (Pecoraro et al., 2016), and to perform genomic prediction for relevant traits in agricultural species
86 (Palaikostas, Ferraresso, Franch, Houston & Bargelloni, 2016). It has also been extended to microbial
87 ecology (Pauletto et al., 2016).

88

89 In this paper, we provide a protocol for laboratory and bioinformatic analyses to optimise studies using 2b-
90 RAD sequencing on different non-model organisms. We focused our study on the sharpsnout seabream
91 *Diplodus puntazzo* Walbaum, 1792 and the loggerhead turtle *Caretta caretta* Linnaeus, 1758 characterized
92 by very different genome sizes. This study aims to unveil key elements to adapt library building of non-model
93 organisms to best profit from this genomic method. Specifically, we focused our analyses on five main
94 objectives. 1) Assess the effect of initial DNA quality and concentration on sequencing results. 2) Evaluate
95 the performance of different IIB enzymes (i.e. Alfl and CspCl) on genomic library construction in the two
96 species. 3) Calculate the optimum number of raw sequences needed per each combination of species and
97 enzyme in order to achieve the maximum number of loci with an optimum depth per locus for a correct
98 genotyping. 4) Assess if selective base ligation protocols have an impact on genetic differentiation among
99 individuals. 5) Set guidelines for new population genomic studies on non-model organisms. Our study
100 provides useful information for future studies on non-model species with different genome sizes, helping
101 decision-making to obtain a reliable and cost-effective genotyping.

102

103 **Methods**

104 *Samples*

105 We analysed two species with approximately three-fold different genome sizes. We consider the sharpsnout
106 seabream (*Diplodus puntazzo*) genome size to be similar to that of *Diplodus anularis* (0.9Gb), its closest
107 relative's sequenced genome (www.genomesize.com). The loggerhead turtle (*Caretta caretta*) genome size
108 was considered to be similar to the genome of *Chelonia mydas* (Wang et al., 2013), which measures 2,24Gb.
109 Juveniles of *D. puntazzo* were collected in Blanes (N=12) and Xabia (N=12) (Spain) during recruitment using
110 hand nets (Figure 1). Samples of *C. caretta* were taken from bycaught juveniles at the foraging ground off
111 Valencia (Spain) (N=9) (Figure 1) and from dead hatchlings at the nesting ground west of Sirte (Lybia) (N=14)
112 (Clusa et al., 2018). We also added a sample collected from a live female turtle nesting in Pulpí (Spain) as
113 positive control (Carreras et al., 2018). Consequently, our study included 24 samples per species. All samples
114 were stored in 96% ethanol.

115

116 *DNA extraction and library construction*

117 Genomic DNA was extracted using Qiagen® Qiamp blood and tissue kit following the manufacturer's
118 protocol. DNA concentration was measured with Nanodrop® or Qubit®, and DNA degradation assessed in 1%
119 agarose gels. This information was recorded to be used in further statistical analysis. We coded the level of
120 degradation as 'high' if the DNA was mostly located at the bottom of the run in the agarose gel or the smear
121 intensity increased in direction top-to-bottom, and as 'low' if the DNA was mostly located at the top of the
122 gel or the smear intensity faded in direction top-to-bottom. When possible we included samples that
123 presented degraded DNA or low DNA concentration to test 2b-RAD efficiency for population genomics, as
124 DNA degradation is a common issue when sampling non-model organisms (*e.g.* marine turtle studies
125 targeting stranded individuals or dead embryos found after excavation of nests). A total of 24 individual
126 libraries were constructed with each enzyme per species. Individual libraries were prepared adjusting the
127 protocol from Wang et al. (2012). In brief, the construction of 2b-RAD libraries consisted of four main steps
128 (for detailed protocol, see Annex I). i) Genomic DNA was digested by a IIB restriction enzyme providing short

129 (32-34 bp) sequences. Each individual sample was digested with *AlfI* and *CspcI* enzymes separately. ii) During
130 ligation, adaptors were attached to the sticky ends of the digested sequences. This step is crucial in the library
131 preparation process because at this point, adaptors can be customised to attach to any sticky end or to attach
132 only to sticky ends with specific sequences, based on the last two bases of the adaptor. For this study we
133 used degenerated bases (5'-NN-3') for our adaptors (Figure 2). iii) In the amplification step, barcodes and
134 Illumina primers were attached to the adaptors and sequences were amplified by PCR. At the end of this step
135 the resulting fragment is expected to measure ~165 bp. Library products were run through a 1.8% agarose
136 gel to check amplification success. The library DNA quality of each sample was coded as 'good' when the
137 band of the agarose gel was bright or 'bad' when it was faint (Figure S1). iv) Purification was performed using
138 magnetic beads to remove primers and sequences longer and shorter than 165 bp. At the end of this step,
139 2b-RAD libraries were ready to be sequenced. The DNA concentration of purified libraries was quantified
140 using a Real Time PCR. The 48 libraries of each species were pooled for SR50 single read sequencing (one
141 species per lane) with a HiSeq 4000 Illumina at the DNA Technologies and Expression Analysis Cores at the
142 UC Davis Genome Center.

143

144 *Genotyping*

145 Sequences were processed using customized scripts (Annex II). First, raw sequences were trimmed to
146 eliminate ligation adaptors and then cut down to the same length (*i.e.* 32bp for *CspCI* and 34bp for *AlfI*).
147 Processed sequences were used for genotyping using the STACKS vs 1.47 pipeline (Catchen, Amores,
148 Hohenlohe, Cresko, & Postlethwait, 2011; Catchen et al., 2013). To construct a loci catalogue we used Stacks
149 function *denovo_map.pl* setting the following parameters: a minimum depth of three reads to consider a
150 stack within an individual ($m = 3$), up to three mismatches allowed between stacks (putative alleles) to merge
151 them into a putative locus within an individual ($M = 3$), and two mismatches allowed between stacks (putative
152 loci) during construction of the catalogue ($n = 2$). Individual genotypes were outputted as haplotype loci VCF
153 files. We used 5 main filters to process loci found in our samples. We removed individual genotypes based
154 on less than 5 reads, loci present in less than 70% of individuals, loci with outlier values of mean depth across

155 all individuals (those above the upper whisker of the R 'boxplot' , corresponding to 1.5 times the interquartile
156 range from the data), loci with a major allele frequency higher than 99% and loci out of Hardy-Weinberg
157 equilibrium (HWE) in at least one of the populations. In the case of *C. caretta* HWE was considered only for
158 Libya, since Valencia is a feeding aggregation of individuals from different populations, and thus deviations
159 from HWE are expected (Clusa et al, 2014). Filtering was performed with VCFtools vs 1.12 (Danecek et al.,
160 2011), with the exception of loci with a major allele frequency higher than 99%, which were identified by the
161 function isPoly from the package 'adegenet' (Jombart, 2008) and the assessment of HWE, computed with the
162 function hw.test from the package 'pegas' (Paradis, 2010) in R (R Core Team, 2018). We performed linear
163 regression and Wilcoxon-Mann-Whitney test in R to assess whether initial and library DNA concentrations,
164 initial DNA degradation and library quality influenced the number of total sequences and the final number of
165 loci of each sample.

166

167 *Resampling analysis*

168 We used bioinformatic simulations for each species and enzyme to obtain several sample sets, each one
169 presenting a different number of reads per individual. We used a customised script to create new sample
170 sets with different number of reads per sample by performing a random selection with replacement of the
171 real data up to different target numbers of raw reads per sample (Annex II). We performed 10 iterations for
172 each target number. Target numbers varied for each species to accommodate the data points to the expected
173 accumulation curve results for the different genome sizes. For *D. puntazzo* we simulated 0.5, 1, 2, 4, 8 and
174 10 million raw reads per sample for CspCI and Alfl enzymes. For *C. caretta* we simulated 4, 8, 12, 16 and 20
175 million raw reads per sample for each enzyme. Each resampled set underwent the same process of loci
176 identification and filtering as explained above with the exception of the filter removing loci out of Hardy-
177 Weinberg equilibrium. This filter was not applied because loci genotyping could be biased in the low depth
178 datasets, artificially creating loci out of H-W equilibrium, since resampling was done with replacement. For
179 this reason, this technique should not be used to artificially increase locus depth for a proper genotyping, as
180 these genotyping errors are going to persist in the extended datasets. We calculated the formula that best

181 fitted the accumulation curve for each species and enzyme and plotted the curve with R package 'ggplot2'
182 (Wickham, 2016). We calculated the number of reads per individual needed to obtain a mean depth of
183 coverage of 20x, since this threshold of quality is used in 2b.RAD studies (Resh, Galaska & Mahon, 2018;
184 Whelan et al., 2019). We also estimated values for a coverage of 25x (Warmuth & Ellegren, 2019) to evaluate
185 if with higher coverage we can detect an improvement in the number of total loci.

186

187 *Selective-base ligation simulation*

188 We assessed the potential impact of reducing the number of loci by selective-base-ligation in population
189 genomic analyses. We bioinformatically selected trimmed reads of the corresponding combination of
190 nucleotides to simulate the use of customised adaptors for selective-base ligation on each combination of
191 species and enzyme (Annex II). This type of ligation is usually performed in the laboratory by designing
192 adaptors that will attach only to reads having the target base at both sticky ends (Figure 2). The simulation
193 of a selective-base-ligation aims to test whether the processing of a proportionally lower number of loci per
194 individual results in the same genetic differentiation as for the whole sample set. We removed from this
195 analysis all samples that had a final mean depth per locus < 10 to eliminate errors given by low depth of
196 coverage. For *D. puntazzo* no samples were removed, while for *C. caretta* 5 samples were removed from the
197 Alfl sample set and 7 from CspCl sample set. We used a customized script simulating the effects of building
198 libraries with adaptors ending in 5'-WN-3' (W = A and T) or 5'-SN-3' (S = G and C) instead of 5'-NN-3'. These
199 simulations aimed to select trimmed sequences by their first and last base and allocate them in separate
200 folders. These selected sequences were then analysed with Stacks and loci were filtered with the same
201 process as explained above for the whole dataset. We calculated the genetic differentiation between pairs
202 of individuals using Prevosti distance with the R function prevosti.dist from the package 'poppr' 2.8.0
203 (Kamvar, Tabima, & Grünwald, 2014; Kamvar, Brooks, & Grünwald, 2015) for the dataset containing all
204 combinations (NN) and for the two simulated selective-base-ligation datasets. The pairwise genetic distance
205 matrixes among individuals for each selective-base-ligation subset were compared to the original NN matrix
206 with a Mantel test using Genalex v6.503 (Peakall & Smouse, 2012), then for each matrix we ran a Principal

207 Coordinate Analysis (PCoA) to evaluate whether individuals maintained the same clustering pattern among
208 subsets, using the same program. To detect the eventual decrease of heterozygosity in the subsets compared
209 to their original set of loci we calculated individual observed heterozygosities for the three datasets with
210 VCFtools and used R to perform a Kruskal-Wallis test for each species and enzyme.

211

212 **Results**

213 *Library construction and loci identification in C. caretta*

214 In *C. caretta* extracted DNA ranged from 17.3 to 133.5 ng/ μ l, and showed high level of degradation in 38% of
215 the samples probably due to the bad condition of the tissue used (Table S1). After adaptor ligation and
216 amplification by PCR we observed generally good results with Alfl but much lower amplification success with
217 CspCl with 46% of faint bands, as assessed with gel electrophoresis (Tables S1). After purification, library DNA
218 concentration was similar for the two enzymes ranging between 6.7 and 52.3 ng/ μ l. The mean number of
219 reads per sample was higher for Alfl digested samples, 7.6×10^6 reads per sample (max 10.1×10^6 , min 4.0×10^6),
220 than for CspCl digested samples, 6.6×10^6 reads per sample (max 10.7×10^6 , min 2.6×10^6) mostly because some
221 samples had low number of reads (Table S1). The trimming process discarded all the sequences that were
222 shorter than 34bp for Alfl and 32bp for CspCl or missed the chosen restriction site, with an average (\pm SE)
223 lower loss per sample in Alfl ($19.2 \pm 2.1\%$) than in CspCl ($41.9 \pm 4.7\%$) (Table 1). After the loci calling, *C. caretta*
224 showed higher total number of loci with Alfl (66907 loci) than CspCl (25416 loci). The mean number of loci
225 retained after all filtering steps were slightly higher for Alfl ($72.9 \pm 0.4\%$) than for CspCl ($69.4 \pm 0.9\%$), although
226 their final mean depth was smaller (Table 1).

227

228 *Library construction and loci identification in D. puntazzo*

229 In *D. puntazzo* starting concentrations ranged from 22.3 to 43.1 ng/ μ l and none of the samples was degraded.
230 Adaptor ligation and amplification yielded successful amplifications with both enzymes although 17% of the
231 samples digested with CspCl had faint bands (Table S2). After purification, library DNA concentration was
232 slightly higher for Alfl ranging between 13.6 and 109.63 ng/ μ l. As for *C. caretta* the sequencing of Alfl in *D.*
233 *puntazzo* resulted in slightly higher mean number of reads per sample than for CspCl (Table 1). After the loci
234 calling and filtering higher number of loci were also found for *D. puntazzo* for Alfl (84382 loci) than for CspCl
235 (31111 loci). The mean number of loci retained after all filtering steps was similar for Alfl ($90.6 \pm 0.1\%$) than
236 for CspCl ($90.8 \pm 0.1\%$), although their final mean depth was almost double in the latter (Table 1).

237

238 *Quality predictors of sequencing success*

239 In the two species analysed and for both restriction enzymes the number of raw reads was significantly
240 correlated to the final number of loci (Table 2). For *D. puntazzo*, initial DNA concentration, DNA degradation
241 and library DNA quality had no significant effect in the number of raw reads or number of loci. However, for
242 CspCI in *C. caretta*, the initial DNA concentration showed a significant impact on number of reads and loci,
243 and on library concentration (Table 2). The library DNA concentration explained sequencing success in both
244 species since the regression between library DNA concentration and the number of reads and loci was
245 significant in most cases, with the exception of Alfl in *C. caretta* and the number of loci with CspCI in *D.*
246 *puntazzo* (Table 2). The impact of DNA degradation on sequencing success was only assessed in *C. caretta*
247 since in *D. puntazzo* DNA had initial good quality (Tables S1 and S2). Interestingly, initial DNA degradation
248 was not a good predictor of neither the number of reads nor loci (Table 2). However, library DNA quality and
249 thus amplification success assessed in an agarose gel significantly increased the number of raw reads and
250 final number of loci (Table 2).

251

252 *Resampling analysis*

253 We simulated the sequencing of different target number of reads per sample set and we obtained the total
254 number of loci and mean depth for each simulation (Figure 3, Table S3). In all simulations, the mean depth
255 of coverage was highly correlated to the number of reads per individual with an $R^2 > 0.99$. Based on the
256 accumulation curve (Figure 3) we estimated the mean number of reads per individual and the corresponding
257 number of loci for two mean depth of coverage, 20x and 25x (Table 3). For both species, Alfl needed a much
258 higher number of reads per individual than CspCI to reach the desired coverage of 20x, due to the higher
259 number of loci obtained with this enzyme. We found that, using a coverage of 25x, the total number of final
260 loci improved in Alfl by 4% and by 7% for *D. puntazzo* and *C. caretta* respectively, and by 9% in CspCI for both
261 species.

262

263 *Selective-base ligation simulation*

264 The selective-base ligation subsets obtained from *C. caretta* retained between 22.2% and 31.5% of the total
265 loci from their original sample sets (Table S4). In *D.puntazzo* the amount of loci retained was more variable
266 between the two tested subsets (Table S4), ranging from 19.8% to 43.4%. In this species we also found that
267 for CspCI enzyme the subsets presented lower coverage than the original set, which could be a consequence
268 of the base composition of the regions where this enzyme is cutting and related with the characteristics of
269 the genomes that make the results species specific (Seetharam & Stuart, 2013). Mantel tests in both species
270 showed high correlation between the pairwise genetic distances among individuals assessed with all loci and
271 assessed with a selective base ligation, for both CspCI and Alfl enzymes (Figure 4). This was also reflected in
272 the PCoA, as *C. caretta* samples do not have the exact same pattern among subsets whereas *D. puntazzo*
273 patterns match perfectly despite the lower number of loci retained in the different datasets (Figure S2). The
274 Kruskal-Wallis test showed no significant differences in observed heterozygosity among any of the subsets
275 and the original set of loci for both species and enzymes (Table S5).

276

277 *Protocol optimization*

278 We used the results obtained from these simulations to refine the laboratory protocol for 2b-RAD libraries
279 preparation and sequencing. In fact, given the mean value of depth of coverage, the optimum number of loci
280 and the size of the studied species genome, we can calculate the number of samples to be sequenced in one
281 lane to optimize costs without compromising the results. To facilitate the decision-making process, based on
282 our results, we constructed a flowchart (Figure 5) and a set of guidelines (Box 1) to help future studies design
283 the most efficient and cost effective protocol to reach their goals.

284

285

286 **Discussion**

287 In this study, we have shown that 2b-RAD protocol provides efficient results even with degraded samples
288 and we demonstrated how this protocol can be optimised for population genomics of non-model species
289 with different genome sizes. To prove this point, we analysed the sharpsnout seabream *D. puntazzo* and the
290 loggerhead turtle *C. caretta* with two different enzymes, Alfl and CspCl, and performed bioinformatic
291 simulations. Our simulations allow estimating the mean number of reads needed per individual to obtain a
292 reliable genotyping and the corresponding expected number of loci. Moreover, our results indicate that
293 selective-base ligation can be used without compromising pairwise genetic distances among individuals.

294

295 In the case of the loggerhead turtle, where several samples had highly degraded DNA, we found that the
296 quality of the initial DNA did not affect the number of raw reads nor the final number of loci, for both
297 enzymes. In fact, the DNA short length for proper IIB enzyme functioning (i.e. 32-34bp digested fragment)
298 reduces the probability of missing loci even in highly degraded samples. This is a highly valuable characteristic
299 of 2b-RAD methodology, since not all studies can easily access high quality samples. For instance, marine
300 turtle genetic studies usually rely on sampling of stranded individuals (Clusa et al., 2016) or dead embryos
301 found at nests after excavation (Clusa et al., 2018), due to the complexity of their behaviours and the paucity
302 of individuals. In such cases, a genomic protocol capable of providing optimal results with degraded samples
303 is invaluable.

304 The library quality after adaptor ligation and amplification was a good predictor of sequencing success. The
305 electrophoresis gel after the library amplification of the loggerhead turtle clearly showed that Alfl resulted
306 in a better amplification than CspCl, which failed to yield a clear band in 46% of individuals. Moreover, the
307 sequencing success was poor for samples with faint amplification bands, which resulted in lower number of
308 reads per individual and thus lower number of loci. We thus suggest discarding samples with poor library
309 DNA quality to help optimising sequencing costs. In the case of the sharpsnout seabream, both enzymes
310 showed good results after the amplification, although a few individuals yielded poorer amplification that
311 resulted in significantly lower number of loci, as observed also in the loggerhead turtle were the difference

312 in library quality with the two enzymes was even greater. Moreover, Alfl provided higher number of loci than
313 CspCI in both species as expected, since Alfl recognition sequence has six fixed nucleotides, while CspCI has
314 seven fixed nucleotides. Therefore, Alfl is expected to have a greater density of restriction sites across any
315 genome than CspCI, and potentially yield more loci as observed in the kissing bug *Rhodnius ecuadoriensis*
316 (Hernandez-Castro et al. 2017).

317

318 Obtaining more loci, though, reduces depth of coverage per locus for the same mean number of reads per
319 individual. As expected, when using CspCI enzyme our sample sets showed higher values of mean depth than
320 when using Alfl in both species, despite poorer amplification success for CspCI in the loggerhead turtle. A low
321 mean depth per locus leads to less accurate genotype calling and thus higher percentage of missing data
322 across loci (Casso, Turon & Pascual, 2019; Maruki & Lynch, 2017; Chow et al., 2019), and for this reason a
323 good depth coverage is important to consider data reliable. Since library construction and sequencing
324 produces a variable number of reads per locus, a mean depth of 20x would guarantee that the minimum of
325 five reads per genotype is consistently achieved across most loci for each sample. This would result in fewer
326 genotypes lost and thus more loci retained over all samples. Our simulations on resampling analyses, allowed
327 the construction of the accumulation curve relating the number of reads per sample and the resulting
328 number of loci as well as the linear correlation between the mean depth per locus and the number of reads
329 per individual. Based on the combination of these two functions the number of individuals to be sequenced
330 in one lane can be calculated easily, simplifying decision-making and analysis design for optimizing population
331 genomic studies at the lowest cost. The amount of reads per individual required by the sharpshout seabream
332 would allow including a fair number of individuals per lane for each enzymes, since both yielded good library
333 DNA quality across samples. However, in the case of the loggerhead turtle, only Alfl enzyme should be used
334 according to library DNA quality. In this case, the amount of reads needed to achieve an adequate coverage
335 would be very large and the number of loci obtained very high, due to the size of the genome. Under these
336 circumstances, the number of individuals of loggerhead turtle to be included in one sequencing lane would
337 be too small and not affordable by most research groups.

338

339 The difference between the two species is mostly related to the crucial role played by the genome size.
340 Species with large genomes will likely produce more loci (due to a greater number of regions yielding the
341 enzyme recognition site) and would need a greater sequencing effort to reach the suitable number of reads
342 per sample for an adequate genotyping. Using a selective-base ligation the number of individuals can be
343 adjusted to the needs of the study considering the number of loci projected by the accumulation curve. Our
344 simulations of customized adaptors with selective base ligation, which extremities would end in –WN or –
345 SN, proved that this type of reduction in the number of loci does not affect genetic differentiation between
346 pairs of individuals. Therefore, the use of a selection of sequences for each sample instead of the whole set,
347 would allow reducing costs by fitting more samples in one lane without compromising overall genetic
348 differentiation. In both species we found that the subsets from the simulated selective-base ligation had a
349 proportionally similar lower number of raw sequences and final loci than the original sets (~25%). However,
350 some differences were observed according to the base and enzyme used in each species suggesting that the
351 species' genome base composition may affect the outcome. Nonetheless, the high levels of correlation that
352 we found between the subsets and the original sets, regardless of the number of loci retained, indicate that
353 they are reliable sources of information. In fact, the slightly lower correlation in genetic distances of *C. caretta*
354 and its differences in PCoAs patterns among subsets were probably a consequence of the bigger genome size
355 of the species, resulting in a lower coverage. This type of selective ligation would be particularly interesting
356 in the case of species with large genomes such as *C. caretta*. Considering the size of this species genome
357 (2.24Gb) and referring to our resampling simulation, we would need 13.5-17.4 million reads per sample to
358 achieve 20x-25x of coverage, therefore only 20-25 samples could be sequenced in the same lane of a platform
359 providing 340 million reads per run as in the present study. A selective-base ligation would allow reducing
360 the costs of sequencing while ensuring good loci coverage, without influencing the outcome. In fact, since
361 the selective-base ligated set would need only ~25% of the original set, between 3.4 and 4.4 million reads
362 per sample are expected to reach the adequate coverage (Warmuth & Ellegren, 2019). Therefore, as much
363 as 78-100 samples could fit in the same Illumina lane, greatly reducing costs without compromising genetic

364 differentiation between individuals. Nevertheless, the number of loci required for a study depends on the
365 scope, the type of analysis performed, and the target species. For instance, selective-base ligation would be
366 less powerful for studies aiming to identify adaptation, since the probability of finding candidate genes can
367 decrease when analysing only a small fraction of the genome (Ahrens et al., 2018).

368

369 Finally, we show that 2b-RAD methodologies can be reliable even for degraded DNA samples. Following our
370 set of guidelines, researchers can optimize effort, time, and sequencing cost of 2b-RAD library building for
371 non-model species while maintaining good sequencing depth for a proper genotyping (Box 1, Figure 5).

372

BOX 1

Guidelines for the optimisation of a 2b-RAD protocol with non-model species.

- Use 2b-RAD instead of other RAD sequencing techniques if you have degraded samples.
- If the target species has a big genome size, consider performing a selective-base ligation to retain 20-40% of total loci.
- If the species genome is small, proceed without selective base-ligation.
- Test different IIB enzymes with the target species.
- Use library quality and concentration as predictors of sequencing success.
- Sequence the test samples with conservative conditions to obtain good coverage.
- Calculate an accumulation curve in a preliminary analysis with the test samples to identify the number of reads needed per individual and the total number of loci corresponding to a coverage $\geq 20x$.
- If the total number of loci is adequate for the selected type of study, proceed to sequence the rest of your samples to obtain the mean number of reads needed according to the curve.
- If the total number of loci is too high for the selected study, use a selective base ligation for library building to reduce the amount of loci.
- The number of samples to be sequenced in the same lane is a trade-off between the number of reads per individual, the number of reads provided per lane and available budget.
- If the total number of loci is adequate but the cost of sequencing is over budget, use a selective base ligation for further 2b-RAD library building to reduce the amount of reads needed per sample and therefore fit more samples in one lane.

373

374

375 *Conclusions*

376 Genomic population studies are increasing in species without reference genomes that rely on restriction-site
377 associated DNA sequencing techniques, although some protocols require good quality DNA. Moreover,
378 transferring protocols across taxa can potentially lead to poor results, such as low number of recovered
379 markers or inadequate genotyping due to differential genomic features. Researchers working with species
380 with large genome sizes or needing lower number of markers can adjust the number of loci by performing
381 selective-base ligation, allowing the sequencing of a larger number of samples, without altering genomic
382 differentiation between individuals as observed by our simulations. The optimal number of samples per lane
383 can, therefore, be adjusted as a trade off with the desired target number of loci and the species genome size
384 for an adequate mean depth of coverage for a correct genotyping. Our results and guidelines aim to improve
385 2b-RAD protocols on non-model organisms with different genome sizes, helping initial decision-making for a
386 reliable, faster and cost-effective genotyping for population genomic studies.

387

388

389

390 **Acknowledgements**

391 M. Babucci (University of Padova) for helping with the initial trimming and preliminary analysis of sequenced
392 data and J. Abril (University of Barcelona) for helping with script building for the simulation analysis with
393 replacement. L. Cardona (University of Barcelona), J.Tomás (University of Valencia) and A.A. Hamza (Alfateh
394 University and University Malaysia Terengganu) for providing tissue samples of *Caretta caretta* from the
395 Valencian feeding ground and the Libyan rookery.

396 This work was supported by the project CTM2017-88080 from the Ministerio de Ciencia, Innovación y
397 Universidades, Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER). CC
398 and MP are part of the research group SGR2017-1120, and EM of SGR2017-378 (Catalan Government). HT
399 was supported by a PhD scholarship funded by the Spanish Ministry of Science, Innovation and Universities
400 (FPU15/02390). AB was supported by grant 2017 FI_B 00997 of the Catalan Government-AGAUR.

401

402

403 **References**

- 404 Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D., & Dudaniec, R. Y. (2018). The search
405 for loci under selection: trends, biases and progress. *Molecular Ecology*, 27(6), 1342-1356.
406
- 407 Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis.
408 *Molecular ecology*, 23(7), 1661-1667.
409
- 410 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008).
411 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS one*, 3(10), e3376.
412
- 413 Bellin, D., Ferrarini, A., Chimento, A., Kaiser, O., Levenkova, N., Bouffard, P., & Delledonne, M. (2009).
414 Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model
415 species. *BMC genomics*, 10(1), 555.
416
- 417 Boscari, E., Abbiati, M., Badalamenti, F., Bavestrello, G., Benedetti-Cecchi, L., Cannas, R., ... & Frascchetti, S.
418 (2019). A population genomics insight by 2b-RAD reveals populations' uniqueness along the Italian coastline
419 in *Leptopsammia pruvoti* (Scleractinia, Dendrophylliidae). *Diversity and Distributions*, 25, 1101-1117.
420
- 421 Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low
422 should we go? *Molecular ecology*, 22(11), 3028-3035.
423
- 424 Carreras, C., Ordóñez, V., Zane, L., Kruschel, C., Nasto, I., Macpherson, E., & Pascual, M. (2017). Population
425 genomics of an endemic Mediterranean fish: differentiation by fine scale dispersal and adaptation. *Scientific*
426 *reports*, 7, 43417.
427
- 428 Carreras, C., Pascual, M., Tomás, J., Marco, A., Hochscheid, S., Castillo, J. J., ... & Cardona, L. (2018). Sporadic
429 nesting reveals long distance colonisation in the philopatric loggerhead sea turtle (*Caretta caretta*). *Scientific*
430 *reports*, 8(1), 1435.
431
- 432 Casso M., Turon X., Pascual M. 2019. Single zooids, multiple loci: independent colonisations revealed by
433 population genomics of a global invader. *Biological Invasions*. <https://doi.org/10.1007/s10530-019-02069-8>
434
- 435 Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and
436 genotyping loci de novo from short-read sequences. *G3: Genes, genomes, genetics*, 1(3), 171-182.
437
- 438 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for
439 population genomics. *Molecular ecology*, 22(11), 3124-3140.
440
- 441 Chow, J. C., Anderson, P. E., & Shedlock, A. M. (2019). Sea turtle population genomic discovery: Global and
442 locus-specific signatures of polymorphism, selection, and adaptive potential. *Genome biology and evolution*,
443 11(10), 2797-2806.
444
- 445 Clusa, M., Carreras, C., Pascual, M., Gaughran, S. J., Piovano, S., Giacoma, C., ... & Maffucci, F. (2014). Fine-
446 scale distribution of juvenile Atlantic and Mediterranean loggerhead turtles (*Caretta caretta*) in the
447 Mediterranean Sea. *Marine biology*, 161(3), 509-519.
448
- 449 Clusa, M., Carreras, C., Pascual, M., Gaughran, S. J., Piovano, S., Avolio, D., ... & Aguilar, A. (2016). Potential
450 bycatch impact on distinct sea turtle populations is dependent on fishing ground rather than gear type in the
451 Mediterranean Sea. *Marine biology*, 163(5), 122.
452

453 Clusa, M., Carreras, C., Cardona, L., Demetropoulos, A., Margaritoulis, D., Rees, A. F., ... & Aguilar, A. (2018).
454 Philopatry in loggerhead turtles *Caretta caretta*: beyond the gender paradigm. *Marine Ecology Progress*
455 *Series*, 588, 201-213.

456

457 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. (2011). The
458 variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158.

459

460 Davey, J., & Blaxter, M. L. (2011). RADSeq: next-generation population genetics. *Briefings in Functional*
461 *Genomics*, 9, 108.

462

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

465

466 Fu, B., Liu, H., Yu, X., & Tong, J. (2016). A high-density genetic map and growth related QTL mapping in bighead
467 carp (*Hypophthalmichthys nobilis*). *Scientific reports*, 6, 28679.

468

469 Galaska, M. P., Sands, C. J., Santos, S. R., Mahon, A. R., & Halanych, K. M. (2017). Geographic structure in the
470 Southern Ocean circumpolar brittle star *Ophionotus victoriae* (Ophiuridae) revealed from mt DNA and single-
471 nucleotide polymorphism data. *Ecology and evolution*, 7(2), 475-485.

472

473 Girault, G., Blouin, Y., Vergnaud, G., & Derzelle, S. (2014). High-throughput sequencing of *Bacillus anthracis*
474 in France: investigating genome diversity and population structure using whole-genome SNP discovery. *BMC*
475 *genomics*, 15(1), 288.

476

477 Hapke, A., & Thiele, D. (2016). GIBPS: a toolkit for fast and accurate analyses of genotyping-by-sequencing
478 data without a reference genome. *Molecular ecology resources*, 16(4), 979-990.

479

480 Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., ... & Sethuraman, A. (2018).
481 Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, 11(8),
482 1197-1211.

483

484 Hernandez-Castro, L. E., Paterno, M., Villacís, A. G., Andersson, B., Costales, J. A., De Noia, M., ... & Llewellyn,
485 M. S. (2017). 2b-RAD genotyping for population genomic studies of Chagas disease vectors: *Rhodnius*
486 *ecuadoriensis* in Ecuador. *PLoS neglected tropical diseases*, 11(7), e0005710.

487

488 Hess, J. E., Campbell, N. R., Docker, M. F., Baker, C., Jackson, A., Lampman, R., ... & Wildbill, A. J. (2015). Use
489 of genotyping by sequencing data to develop a high-throughput and multifunctional SNP panel for
490 conservation applications in Pacific lamprey. *Molecular Ecology Resources*, 15(1), 187-202.

491

492 Hoffman, J. I., Simpson, F., David, P., Rijks, J. M., Kuiken, T., Thorne, M. A. ... & Dasmahapatra, K. K. (2014).
493 High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the*
494 *National Academy of Sciences*, 201318945.

495

496 Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD
497 sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope
498 cutthroat trout. *Molecular ecology resources*, 11, 117-122.

499

500 Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*,
501 24(11), 1403-1405.

502

503 Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations
504 with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2, e281.

505
506 Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population
507 genetic data with emphasis on clonality. *Frontiers in genetics*, 6, 208.
508
509 Lavretsky, P., DaCosta, J. M., Sorenson, M. D., McCracken, K. G., & Peters, J. L. (2019). ddRAD-seq data reveal
510 significant genome-wide population structure and divergent genomic regions that distinguish the mallard
511 and close relatives in North America. *Molecular ecology*, 28, 2594-2609.
512
513 Lim, H. C., Gawin, D. F., Shakya, S. B., Harvey, M. G., Rahman, M. A., & Sheldon, F. H. (2017). Sundaland's
514 east–west rain forest population structure: variable manifestations in four polytypic bird species examined
515 using RAD-Seq and plumage analyses. *Journal of biogeography*, 44(10), 2259-2271.
516
517 Lu, F., Glaubitz, J., Harriman, J., Casstevens, T., & Elshire, R. (2012). TASSEL 3.0 Universal Network Enabled
518 Analysis Kit (UNEAK) pipeline documentation. *White Paper*, 2012, 1-12.
519
520 Luo, X., Shi, X., Yuan, C., Ai, M., Ge, C., Hu, M., ... & Yang, X. (2017). Genome-wide SNP analysis using 2b-RAD
521 sequencing identifies the candidate genes putatively associated with resistance to ivermectin in *Haemonchus*
522 *contortus*. *Parasites & vectors*, 10(1), 31.
523
524 Maruki, T., & Lynch, M. (2014). Genome-wide estimation of linkage disequilibrium from population-level
525 high-throughput sequencing data. *Genetics*, 197(4), 1303-1313.
526
527 Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes,*
528 *Genomes, Genetics*, 7(5), 1393-1404.
529
530 Palaiokostas, C., Ferrareso, S., Franch, R., Houston, R.D., & Bargelloni, L. (2016). Genomic prediction of
531 resistance to pasteurellosis in gilthead sea bream (*Sparus aurata*) using 2b-RAD sequencing. *G3 (Bethesda)*,
532 8;6(11), 3693-3700. doi:10.1534/g3.116.035220
533
534 Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach.
535 *Bioinformatics*, 26(3), 419-420.
536
537 Pauletto, M., Carraro, L., Babbucci, M., Lucchini, R., Bargelloni, L., & Cardazzo B. (2016). Extending RAD tag
538 analysis to microbial ecology: a comparison between MultiLocus Sequence Typing and 2b-RAD to investigate
539 *Listeria monocytogenes* genetic structure. *Molecular Ecology Resources*, 16(3):823-35. doi: 10.1111/1755-
540 0998.12495.
541
542 Peakall, R. & Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for
543 teaching and research—an update. *Bioinformatics*, 28, 2537-2539.
544
545 Pecoraro, C., Babbucci, M., Villamor, A., Franch, R., Papetti, C., Leroy, B., ... & Murua, H. (2016).
546 Methodological assessment of 2b-RAD genotyping technique for population structure inferences in yellowfin
547 tuna (*Thunnus albacares*). *Marine genomics*, 25, 43-48.
548
549 Portnoy, D. S., Puritz, J. B., Hollenbeck, C. M., Gelsleichter, J., Chapman, D., & Gold, J. R. (2015). Selection and
550 sex-biased dispersal in a coastal shark: the influence of philopatry on adaptive variation. *Molecular*
551 *ecology*, 24(23), 5877-5885.
552
553 R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical
554 Computing, Vienna, Austria. URL <https://www.R-project.org/>.
555

556 Resh, C. A., Galaska, M. P., & Mahon, A. R. (2018). Genomic analyses of Northern snakehead (*Channa argus*)
557 populations in North America. *PeerJ*, 6, e4581.
558

559 Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. *Nature Reviews*
560 *Genetics*, 14(11), 807.
561

562 Seetharam, A.S. and Stuart, G.W. (2013). Whole genome phylogeny for 21 *Drosophila* species using predicted
563 2b-RAD fragments. *PeerJ* 1:e226 <https://doi.org/10.7717/peerj.226>
564

565 Sims, D., Sudbery, I., Illott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key
566 considerations in genomic analyses. *Nature Reviews Genetics*, 15(2), 121.
567

568 Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-
569 wide genotyping. *Nature methods*, 9(8), 808.
570

571 Warmuth, V. M., & Ellegren, H. (2019). Genotype-free estimation of allele frequencies reduces bias and
572 improves demographic inference from RADSeq data. *Molecular Ecology Resources*, 19, 586–596.
573

574 Whelan, N. V., Galaska, M. P., Siple, B. N., Weber, J. M., Johnson, P. D., Halanych, K. M., & Helms, B. S. (2019).
575 Riverscape genetic variation, migration patterns, and morphological variation of the threatened Round
576 Rocksnail, *Leptoxis ampla*. *Molecular Ecology*, 28(7), 1593-1610.
577

578 Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
579

580 Xu, P., Xu, S., Wu, X., Tao, Y., Wang, B., Wang, S., ... & Li, G. (2014). Population genomic analyses from low-
581 coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. *The Plant Journal*, 77(3), 430-
582 442.
583

584 **Data accessibility**

585 Raw reads from all individuals, including information of location of all samples, will be stored in a SRA
586 Bioproject upon acceptance. All customised scripts (.sh files) can be found in the Supplementary File
587 customised_scripts.zip.

588 **Author contributions**

589 *AB, HT, EM, CC and MP conceived and designed the study. AB and HT did the laboratory analysis with inputs*
590 *from LB and RF. AB and HT conducted the data analysis. AB and HT wrote the manuscript with input from all*
591 *authors.*

592 **Tables**

593 **Table 1. Summary of sequencing outcome.** Mean (\pm SE) values per individual are given for TR: total number
 594 of reads, TMR: number of trimmed reads, IL: initial number of loci, FL: final number of loci after filtering, RL:
 595 percentage of loci retained after filtering, and FMD: final mean depth of coverage per locus.

Species Enzyme	<i>C. caretta</i>		<i>D. puntazzo</i>	
	Alfi	Cspcl	Alfi	Cspcl
TR	7.6 \pm 0.3 \times 10 ⁶	6.6 \pm 0.4 \times 10 ⁶	7.1 \pm 0.3 \times 10 ⁶	6.5 \pm 0.3 \times 10 ⁶
TMR	6.2 \pm 0.4 \times 10 ⁶	4.2 \pm 0.5 \times 10 ⁶	5.3 \pm 0.2 \times 10 ⁶	4.3 \pm 0.2 \times 10 ⁶
IL	48740 \pm 1489	17811 \pm 1010	75971 \pm 130	27989 \pm 40
FL	35576 \pm 1124	12455 \pm 732	68978 \pm 115	25421 \pm 27
RL	72.9 \pm 0.4%	69.4 \pm 0.9%	90.6 \pm 0.1%	90.8 \pm 0.1%
FMD	11.5 \pm 0.7	19.3 \pm 2.4	29.2 \pm 1.4	52.2 \pm 2.3

596

597

598 **Table 2. Statistical analyses of potential quality predictors.** In bold are shown significant p-values after FDR correction. na: tests not available due to insufficient
 599 samples with bad initial DNA quality or low library DNA quality.

600

Explanatory variable	Response Variable	Test	<i>Caretta caretta</i>				<i>Diplodus puntazzo</i>			
			<i>CspCl</i>		<i>Alfi</i>		<i>CspCl</i>		<i>Alfi</i>	
			F or W	p value	F or W	p value	F or W	p value	F or W	p value
Raw reads	Final loci	Linear Regression	17.7	0.000	30.4	0.000	4.7	0.041	34.4	0.000
Initial DNA concentration	Raw reads	Linear Regression	15.4	0.001	2.7	0.115	0.5	0.469	0.4	0.522
	Final loci	Linear Regression	5.2	0.032	2.1	0.159	0.0	0.959	1.5	0.236
	Library DNA concentration	Linear Regression	15.8	0.001	0.0	0.986	3.2	0.087	0.3	0.611
Initial DNA degradation	Raw reads	Wilcoxon-Mann-Whitney	60.0	0.682	61.0	0.726	na	na	na	na
	Final loci	Wilcoxon-Mann-Whitney	44.0	0.170	45.0	0.194	na	na	na	na
	Library DNA concentration	Wilcoxon-Mann-Whitney	34.0	0.048	44.0	0.174	na	na	na	na
Library DNA concentration	Raw reads	Linear Regression	14.2	0.001	2.0	0.174	22.6	0.000	6.3	0.020
	Final loci	Linear Regression	20.3	0.000	3.2	0.086	0.4	0.559	6.2	0.021
Library DNA quality	Raw reads	Wilcoxon-Mann-Whitney	19.0	0.002	13.0	0.037	na	na	25.0	0.261
	Final loci	Wilcoxon-Mann-Whitney	12.5	0.001	6.0	0.005	na	na	9.0	0.018

601

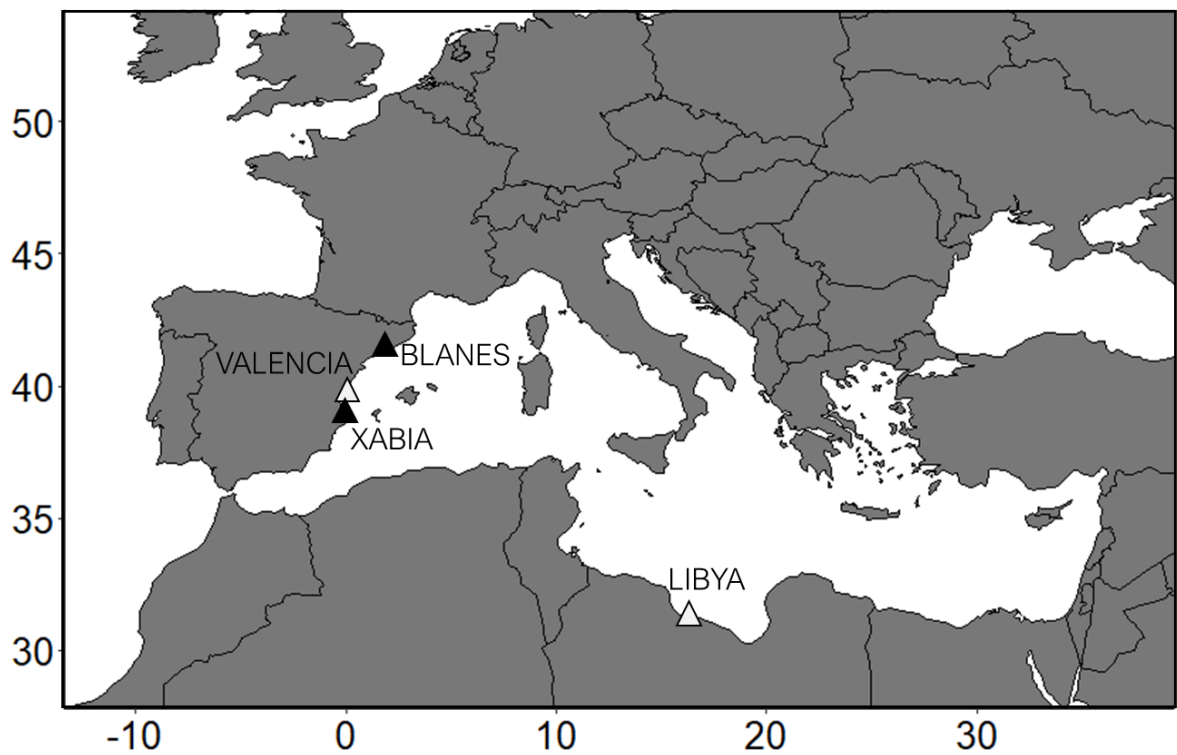
602

603 **Table 3. Estimated number of loci and reads needed to obtain different mean depth per locus as**
 604 **derived from the accumulation curve.** The table shows the number of reads per individual and the
 605 total number of loci per set corresponding to a mean depth of coverage of 20x and 25x for each species
 606 and enzyme.
 607

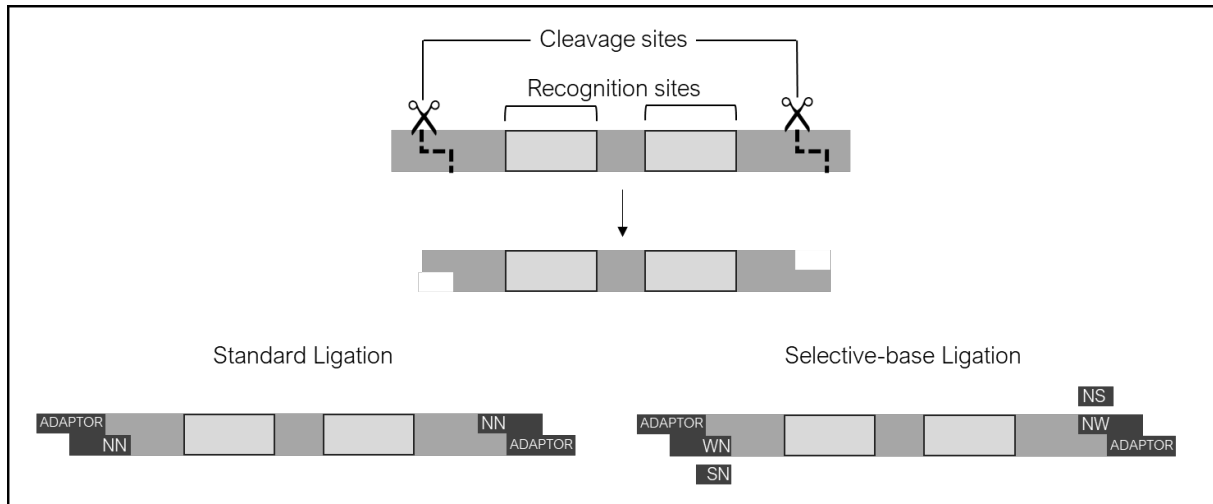
		<i>Caretta caretta</i>		<i>Diplodus puntazzo</i>	
		Alfi	CspCI	Alfi	CspCI
20x	Reads (10 ⁶)	13.5	6.1	3.5	1.7
	Loci	142910	49588	68079	22225
25x	Reads (10 ⁶)	17.4	7.9	4.6	2.2
	Loci	152998	53842	70571	24173

608

609 **Figures**
610



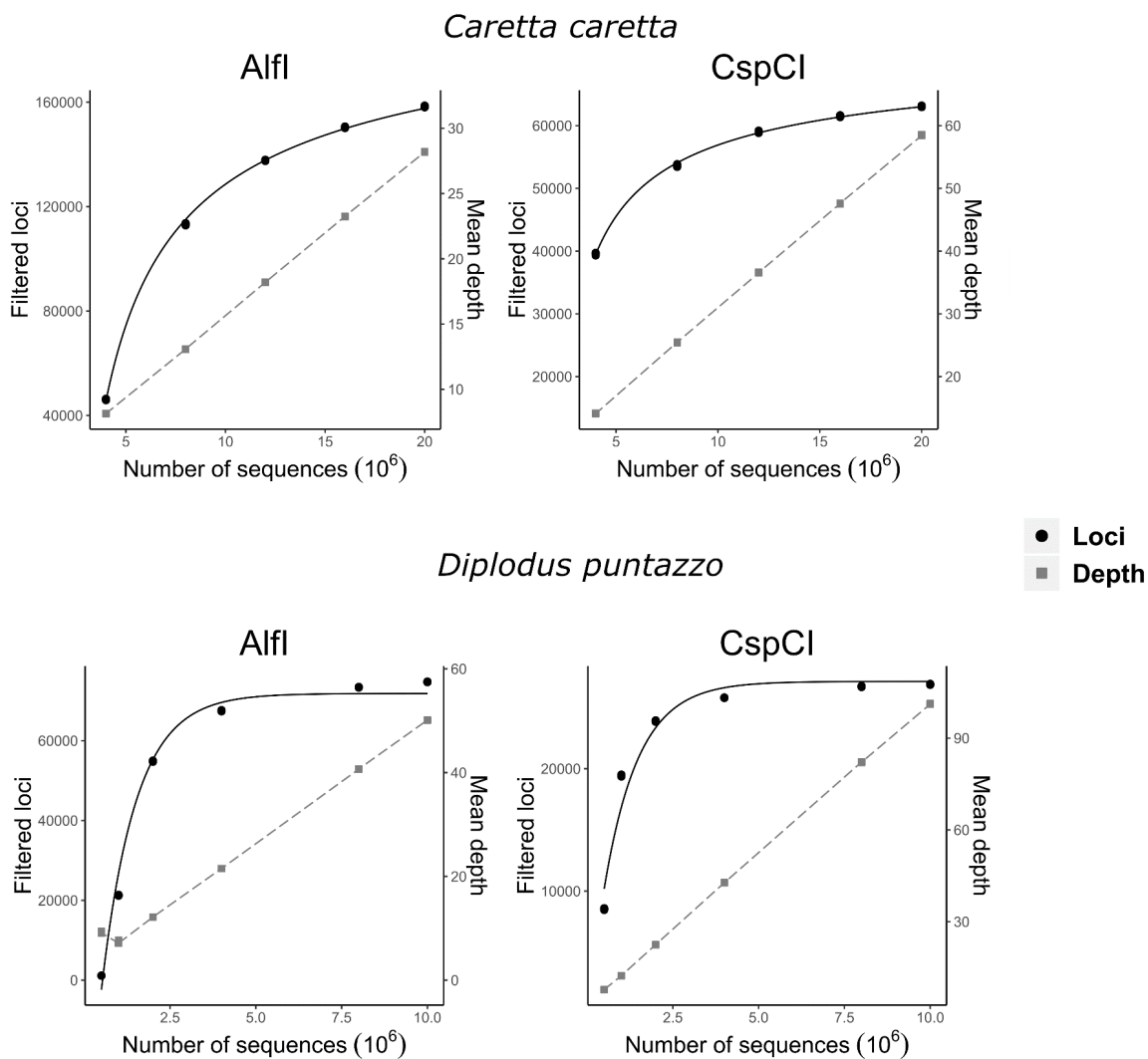
611 **Figure 1. Sampling sites.** White triangles show sampling sites for *C. caretta*, Libya is a nesting ground
612 while Valencia is a foraging ground. Black triangles show sampling sites for *D. puntazzo*.
613



614

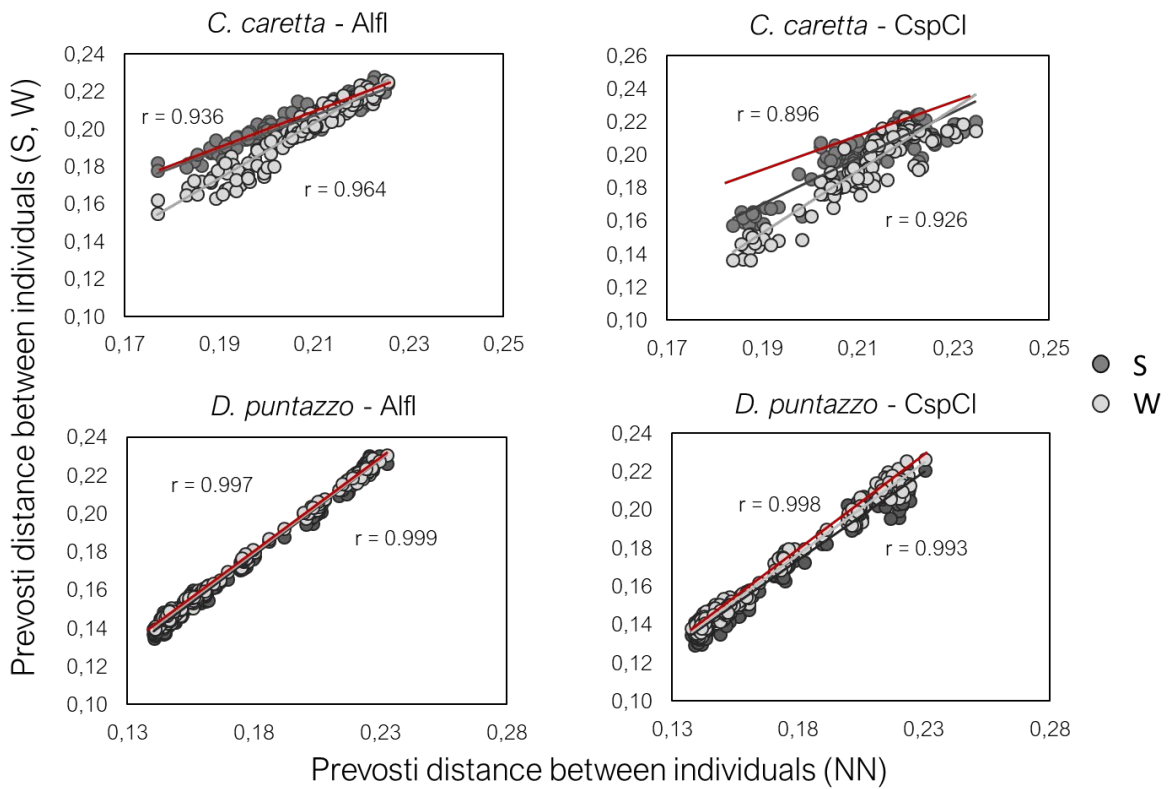
615 **Figure 2. Selective-base ligation.** In 2b-RAD protocol, after IIB enzyme digestion, specific fragments
 616 can be selected to reduce the density of markers to be amplified by designing customised adaptors
 617 with one fully degenerated base (N) and one partially degenerated base (S = G and C bases, W = A and
 618 T bases).

619



620

621 **Figure 3. Accumulation curves resulting from the resampling analysis.** The graphs show the number of final loci (circles) and the mean depth per locus (squares) obtained after filtering, for *C. caretta* (top)
622 and *D. punctazzo* (bottom).
623



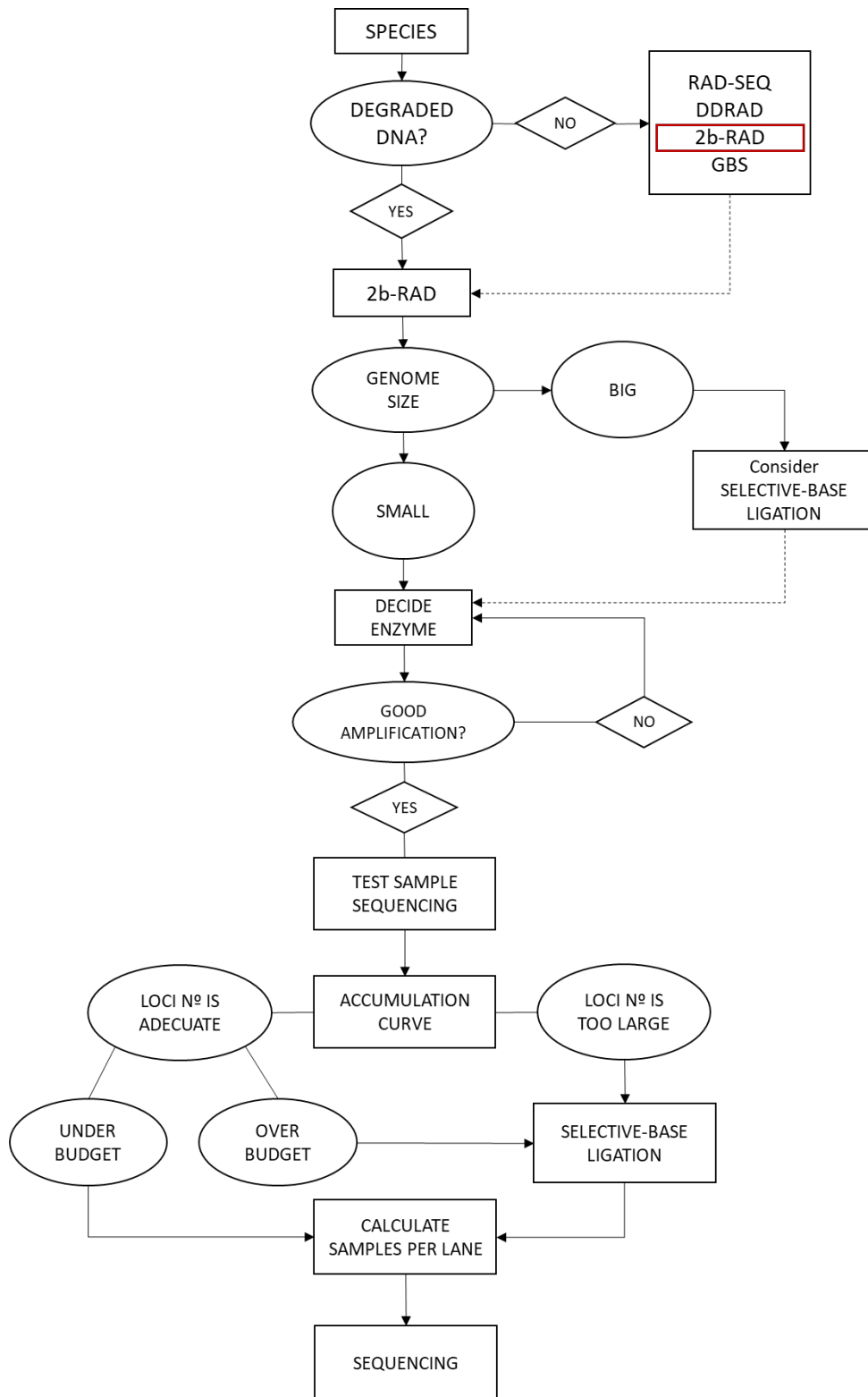
624

625

626 **Figure 4. Mantel test of genetic differentiation between selective-base subsets and original sets.** X-
 627 axes show Prevosti distance between pairs of individuals for each one of the four original sample sets
 628 (with fully degenerated bases –NN-). Y-axis show Prevosti distance between the same pairs of
 629 individuals for subsets obtained from bioinformatic simulations of selective base ligation (either –SN-
 630 or –WN-) for each species and enzyme. Dark grey shows genetic differentiation for S (G and C bases)
 631 subsets and light grey for W (A and T) subsets. Correlation coefficient (r) is given for each test above
 632 the lines for S and below for W. The red line represents the expected correlation function when no
 633 deviation in genetic distances is found in the selective-base subsets compared to NN.

634

635



636

637 **Figure 5. Flowchart for 2b-RAD laboratory protocol.** This flowchart is meant to aid decision making
 638 for 2b-RAD laboratory protocols when studying non-model species. Together with the guidelines listed
 639 above this chart aims to make 2b-RAD studies not only easier but also more cost-effective.