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

Anna Barbanti[1*], Hector Torrado[1,2*], Enrique Macpherson[2], Luca Bargelloni[3], Rafaella Franch[3], Carlos Carreras[1§], Marta Pascual[1§].

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

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

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

* These authors contributed equally to the manuscript

§ These author are considered senior authors

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

Abstract

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

Introduction

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

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

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

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

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

Methods

Samples

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

DNA extraction and library construction

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

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

Genotyping

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

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

Resampling analysis

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

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

Selective-base ligation simulation

We assessed the potential impact of reducing the number of loci by selective-base-ligation in population genomic analyses. We bioinformatically selected trimmed reads of the corresponding combination of nucleotides to simulate the use of customised adaptors for selective-base ligation on each combination of species and enzyme (Annex II). This type of ligation is usually performed in the laboratory by designing adaptors that will attach only to reads having the target base at both sticky ends (Figure 2). The simulation of a selective-base-ligation aims to test whether the processing of a proportionally lower number of loci per individual results in the same genetic differentiation as for the whole sample set. We removed from this analysis all samples that had a final mean depth per locus < 10 to eliminate errors given by low depth of coverage. For *D. puntazzo* no samples were removed, while for *C. caretta* 5 samples were removed from the AlfI sample set and 7 form CspCI sample set. We used a customized script simulating the effects of building 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 simulations aimed to select trimmed sequences by their first and last base and allocate them in separate folders. These selected sequences were then analysed with Stacks and loci were filtered with the same process as explained above for the whole dataset. We calculated the genetic differentiation between pairs of individuals using Prevosti distance with the R function prevosti.dist from the package 'poppr' 2.8.0 (Kamvar, Tabima, & Grünwald, 2014; Kamvar, Brooks, & Grünwald, 2015) for the dataset containing all combinations (NN) and for the two simulated selective-base-ligation datasets. The pairwise genetic distance matrixes among individuals for each selective-base-ligation subset were compared to the original NN matrix 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

Results

Library construction and loci identification in C. caretta

In *C. caretta* extracted DNA ranged from 17.3 to 133.5 ng/µl, and showed high level of degradation in 38% of the samples probably due to the bad condition of the tissue used (Table S1). After adaptor ligation and amplification by PCR we observed generally good results with AlfI but much lower amplification success with CspCI with 46% of faint bands, as assessed with gel electrophoresis (Tables S1). After purification, library DNA concentration was similar for the two enzymes ranging between 6.7 and 52.3 ng/µl. The mean number of reads per sample was higher for AlfI digested samples, $7.6x10^6$ reads per sample (max $10.1x10^6$, min $4.0x10^6$), than for CspCI digested samples, $6.6x10^6$ reads per sample (max $10.7x10^6$, min $2.6x10^6$) mostly because some samples had low number of reads (Table S1). The trimming process discarded all the sequences that were shorter than 34bp for AlfI and 32bp for CspCI or missed the chosen restriction site, with an average (±SE) lower loss per sample in AlfI (19.2±2.1%) than in CspCI (41.9±4.7%) (Table 1). After the loci calling, *C.caretta* showed higher total number of loci with AlfI (66907 loci) than CspCI (25416 loci). The mean number of loci retained after all filtering steps were slightly higher for AlfI (72.9±0.4%) than for CspCl (69.4±0.9%), although their final mean depth was smaller (Table 1).

Library construction and loci identification in D. puntazzo

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

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 AlfI in *C. caretta* and the number of loci with CspCl 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, AlfI 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 AlfI 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*

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

Protocol optimization

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

Discussion

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

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

The library quality after adaptor ligation and amplification was a good predictor of sequencing success. The electrophoresis gel after the library amplification of the loggerhead turtle clearly showed that AlfI resulted in a better amplification than CspCI, which failed to yield a clear band in 46% of individuals. Moreover, the sequencing success was poor for samples with faint amplification bands, which resulted in lower number of reads per individual and thus lower number of loci. We thus suggest discarding samples with poor library DNA quality to help optimising sequencing costs. In the case of the sharpsnout seabream, both enzymes showed good results after the amplification, although a few individuals yielded poorer amplification that 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, AlfI provided higher number of loci than
313 CspCI in both species as expected, since AlfI recognition sequence has six fixed nucleotides, while CspCI has
314 seven fixed nucleotides. Therefore, AlfI 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 then
320 when using AlfI in both species, despite poorer amplification success for CspCl 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 sharpsnout 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 AlfI 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.

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

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

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

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 ≥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.

Conclusions

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

Acknowledgements

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

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

References

Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D., & Dudaniec, R. Y. (2018). The search for loci under selection: trends, biases and progress. *Molecular Ecology*, 27(6), 1342-1356.

Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis. *Molecular ecology*, 23(7), 1661-1667.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, *3*(10), e3376.

Bellin, D., Ferrarini, A., Chimento, A., Kaiser, O., Levenkova, N., Bouffard, P., & Delledonne, M. (2009). Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC genomics*, *10*(1), 555.

Boscari, E., Abbiati, M., Badalamenti, F., Bavestrello, G., Benedetti-Cecchi, L., Cannas, R., ... & Fraschetti, S. (2019). A population genomics insight by 2b-RAD reveals populations' uniqueness along the Italian coastline in Leptopsammia pruvoti (Scleractinia, Dendrophylliidae). *Diversity and Distributions*, 25, 1101-1117.

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology*, 22(11), 3028-3035.

Carreras, C., Ordóñez, V., Zane, L., Kruschel, C., Nasto, I., Macpherson, E., & Pascual, M. (2017). Population genomics of an endemic Mediterranean fish: differentiation by fine scale dispersal and adaptation. *Scientific reports*, 7, 43417.

Carreras, C., Pascual, M., Tomás, J., Marco, A., Hochscheid, S., Castillo, J. J., ... & Cardona, L. (2018). Sporadic nesting reveals long distance colonisation in the philopatric loggerhead sea turtle (*Caretta caretta*). *Scientific reports*, 8(1), 1435.

Casso M., Turon X., Pascual M. 2019. Single zooids, multiple loci: independent colonisations revealed by population genomics of a global invader. *Biological Invasions*. https://doi.org/10.1007/s10530-019-02069-8

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, genomes, genetics*, 1(3), 171-182.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140.

Chow, J. C., Anderson, P. E., & Shedlock, A. M. (2019). Sea turtle population genomic discovery: Global and locus-specific signatures of polymorphism, selection, and adaptive potential. *Genome biology and evolution*, 11(10), 2797-2806.

Clusa, M., Carreras, C., Pascual, M., Gaughran, S. J., Piovano, S., Giacoma, C., ... & Maffucci, F. (2014). Fine-scale distribution of juvenile Atlantic and Mediterranean loggerhead turtles (*Caretta caretta*) in the Mediterranean Sea. *Marine biology*, 161(3), 509-519.

Clusa, M., Carreras, C., Pascual, M., Gaughran, S. J., Piovano, S., Avolio, D., ... & Aguilar, A. (2016). Potential bycatch impact on distinct sea turtle populations is dependent on fishing ground rather than gear type in the Mediterranean Sea. *Marine biology*, 163(5), 122.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in genetics*, 6, 208.

Lavretsky, P., DaCosta, J. M., Sorenson, M. D., McCracken, K. G., & Peters, J. L. (2019). ddRAD-seq data reveal significant genome-wide population structure and divergent genomic regions that distinguish the mallard and close relatives in North America. *Molecular ecology*, 28, 2594-2609.

Lim, H. C., Gawin, D. F., Shakya, S. B., Harvey, M. G., Rahman, M. A., & Sheldon, F. H. (2017). Sundaland's east–west rain forest population structure: variable manifestations in four polytypic bird species examined using RAD-Seq and plumage analyses. *Journal of biogeography*, 44(10), 2259-2271.

Lu, F., Glaubitz, J., Harriman, J., Casstevens, T., & Elshire, R. (2012). TASSEL 3.0 Universal Network Enabled Analysis Kit (UNEAK) pipeline documentation. *White Paper*, 2012, 1-12.

Luo, X., Shi, X., Yuan, C., Ai, M., Ge, C., Hu, M., ... & Yang, X. (2017). Genome-wide SNP analysis using 2b-RAD sequencing identifies the candidate genes putatively associated with resistance to ivermectin in *Haemonchus contortus*. *Parasites & vectors*, 10(1), 31.

Maruki, T., & Lynch, M. (2014). Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics*, 197(4), 1303-1313.

Maruki, T., & Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, 7(5), 1393-1404.

Palaiokostas, C., Ferraresso, S., Franch, R., Houston, R.D., & Bargelloni, L. (2016). Genomic prediction of resistance to pasteurellosis in gilthead sea bream (*Sparus aurata*) using 2b-RAD sequencing. *G3 (Bethesda)*, 8;6(11), 3693-3700. doi:10.1534/g3.116.035220

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3), 419-420.

Pauletto, M., Carraro, L., Babbucci, M., Lucchini, R., Bargelloni, L., & Cardazzo B. (2016). Extending RAD tag analysis to microbial ecology: a comparison between MultiLocus Sequence Typing and 2b-RAD to investigate Listeria monocytogenes genetic structure. *Molecular Ecology Resources,* 16(3):823-35. doi: 10.1111/1755-0998.12495.

Peakall, R. & Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, 28, 2537-2539.

Pecoraro, C., Babbucci, M., Villamor, A., Franch, R., Papetti, C., Leroy, B., ... & Murua, H. (2016). Methodological assessment of 2b-RAD genotyping technique for population structure inferences in yellowfin tuna (*Thunnus albacares*). *Marine genomics*, *25*, 43-48.

Portnoy, D. S., Puritz, J. B., Hollenbeck, C. M., Gelsleichter, J., Chapman, D., & Gold, J. R. (2015). Selection and sex-biased dispersal in a coastal shark: the influence of philopatry on adaptive variation. *Molecular ecology*, *24*(23), 5877-5885.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

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

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

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

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

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

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

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

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

Data accessibility

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

Author contributions

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

Tables

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

Species	*C. caretta*		*D. puntazzo*	
Enzyme	AlfI	CspcI	AlfI	CspcI
TR	7.6±0.3x10^6	6.6±0.4x10^6	7.1±0.3x10^6	6.5±0.3x10^6
TMR	6.2±0.4x10^6	4.2±0.5x10^6	5.3±0.2x10^6	4.3±0.2x10^6
IL	48740±1489	17811±1010	75971±130	27989±40
FL	35576±1124	12455±732	68978±115	25421±27
RL	72.9±0.4%	69.4±0.9%	90.6±0.1%	90.8±0.1%
FMD	11.5±0.7	19.3±2.4	29.2±1.4	52.2±2.3

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 samples with bad initial DNA quality or low library DNA quality.

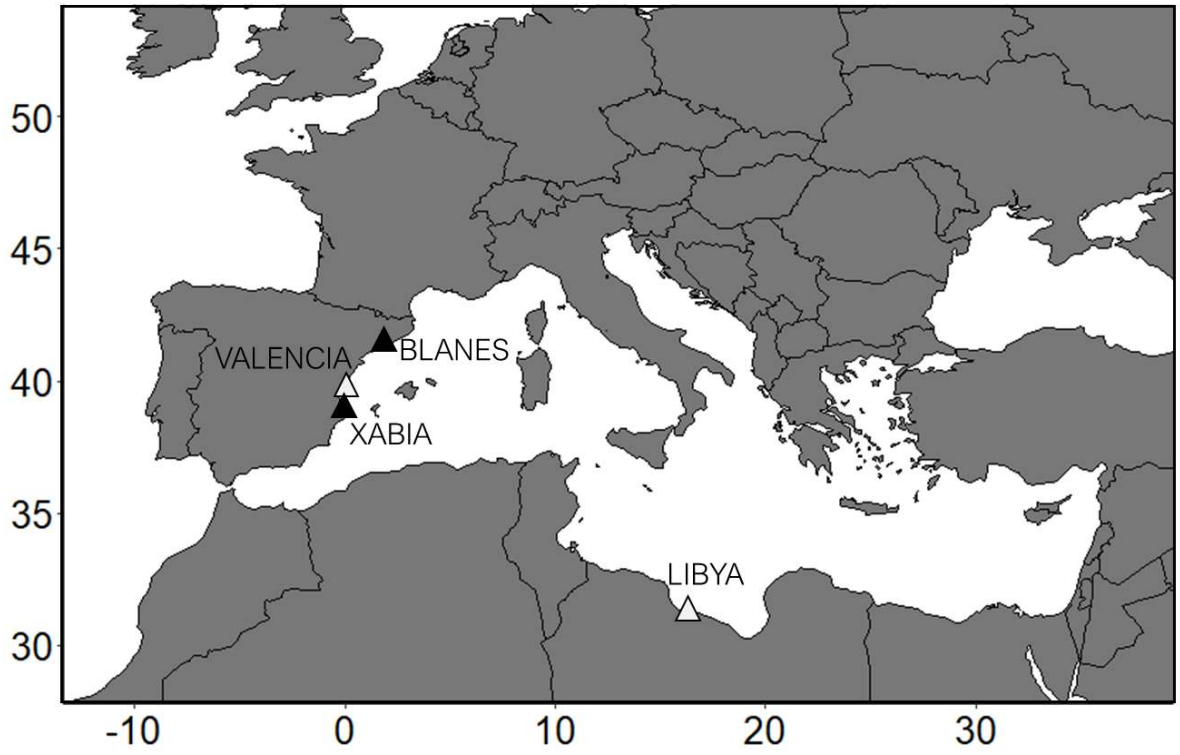
			Caretta caretta				***Diplodus puntazzo***			
			CspCI		*AlfI*		*CspCI*		*AlfI*	
Explanatory variable	**Response Variable**	**Test**	**F or W**	**p value**	**F or W**	**p value**	**F or W**	**p value**	**F or W**	**p value**
Raw reads	Final loci	Linear Regression	17.7	**0.000**	30.4	**0.000**	4.7	**0.041**	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**

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

		Caretta caretta		*Diplodus puntazzo*	
		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

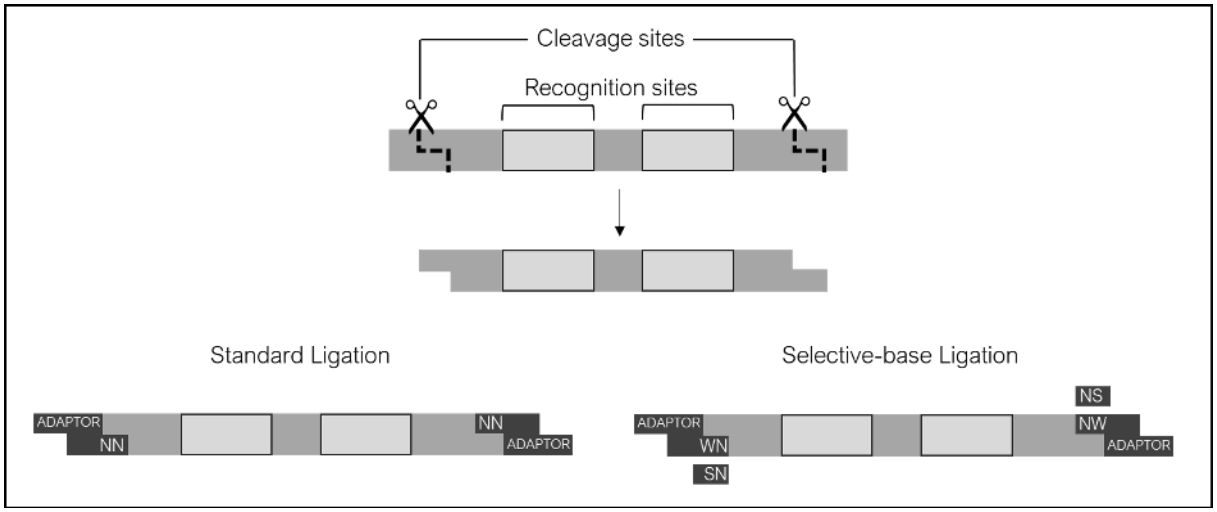
609 **Figures**

610



611

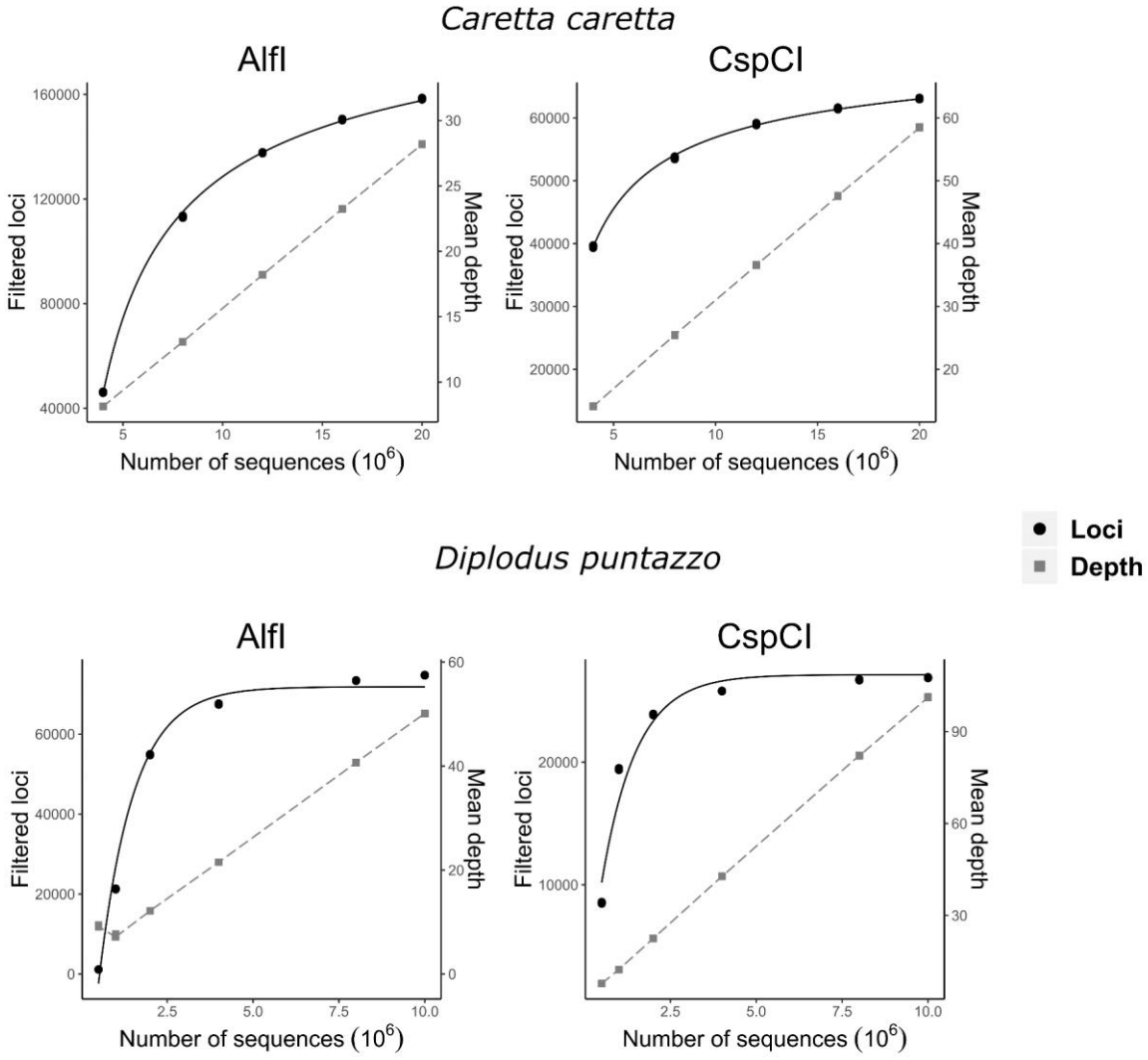
612 **Figure 1. Sampling sites.** White triangles show sampling sites for *C. caretta*, Libya is a nesting ground
613 while Valencia is a foraging ground. Black triangles show sampling sites for *D. puntazzo*.



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
622 of final loci (circles) and the mean depth per locus (squares) obtained after filtering, for *C. caretta* (top)
623 and *D. puntazzo* (bottom).

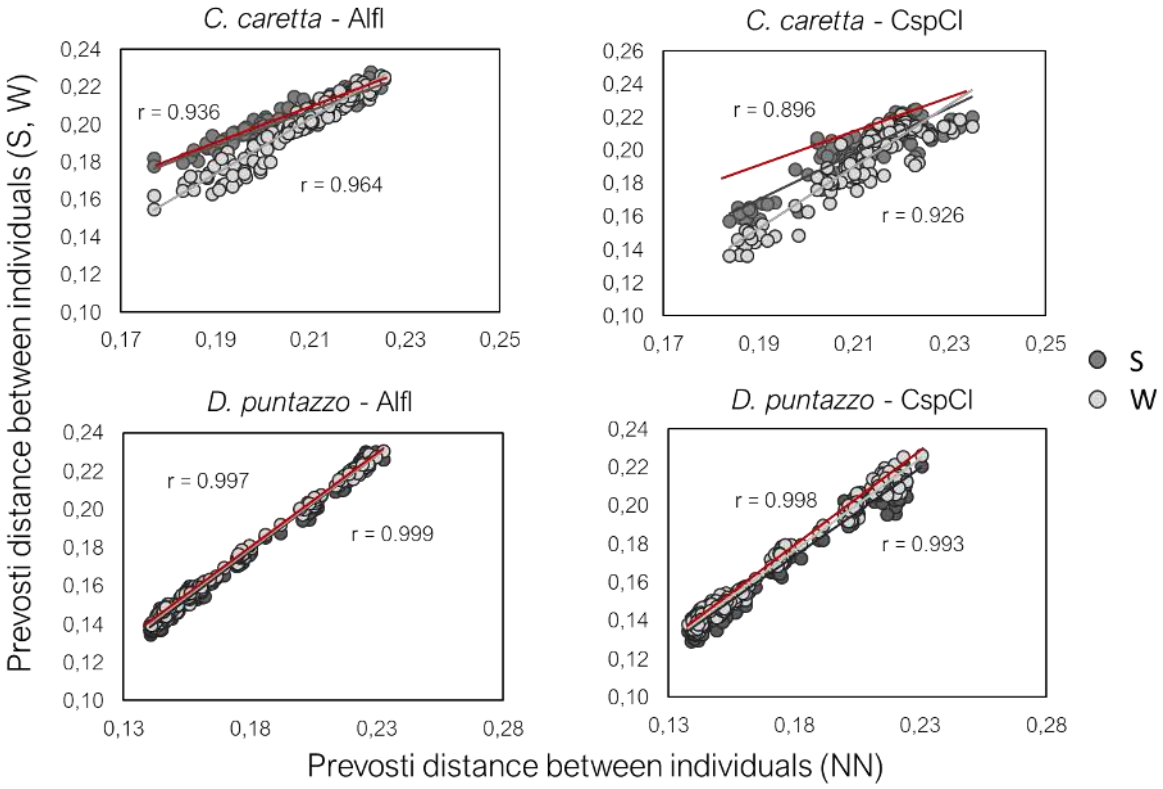


Figure 4. Mantel test of genetic differentiation between selective-base subsets and original sets. X-axes show Prevosti distance between pairs of individuals for each one of the four original sample sets (with fully degenerated bases –NN-). Y-axis show Prevosti distance between the same pairs of individuals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

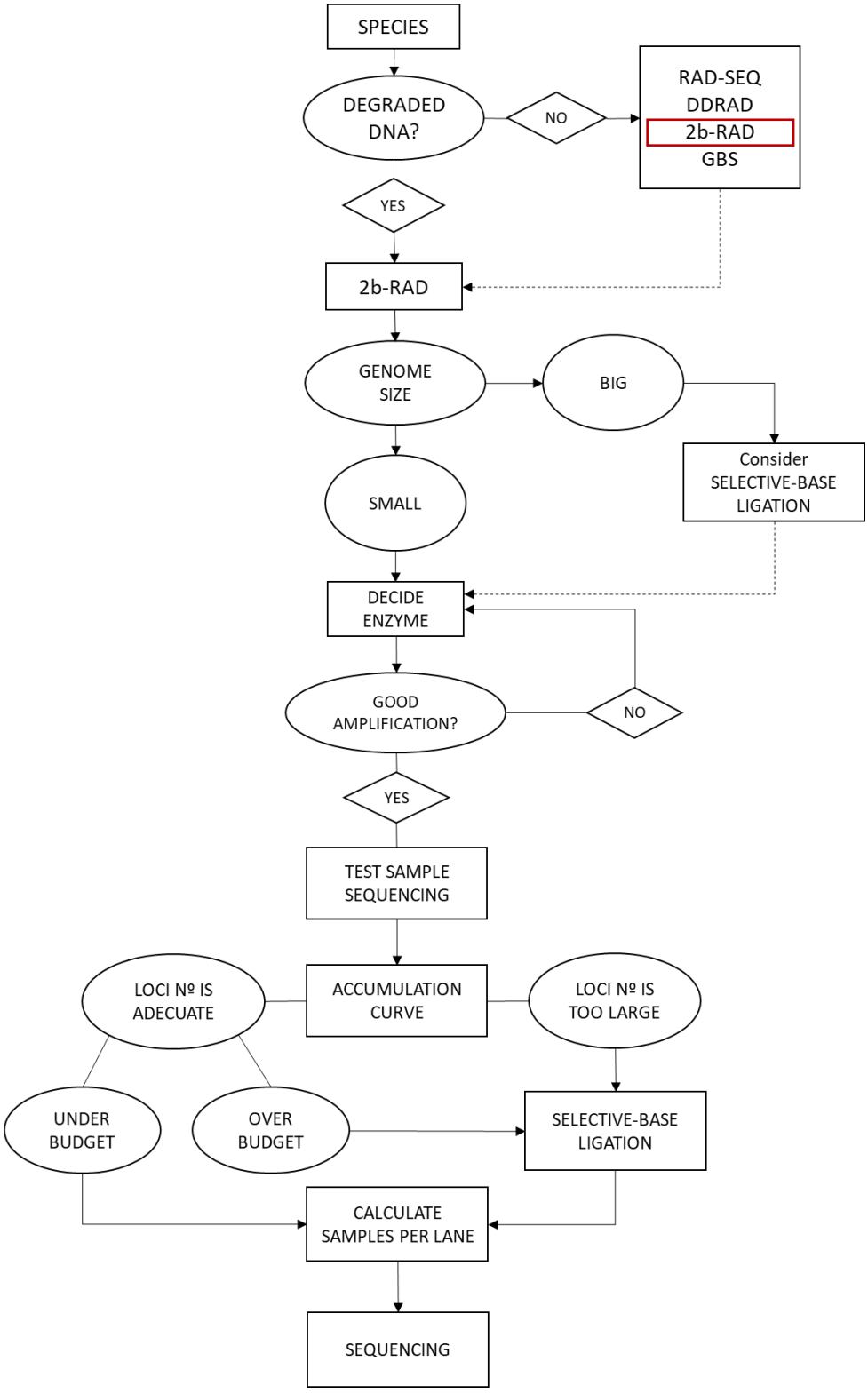


Figure 5. Flowchart for 2b-RAD laboratory protocol. This flowchart is meant to aid decision making for 2b-RAD laboratory protocols when studying non-model species. Together with the guidelines listed above this chart aims to make 2b-RAD studies not only easier but also more cost-effective.