

1 Double-digest RAD-sequencing: do wet and dry protocol

2 parameters impact biological results?

3

4 Tristan CUMER^{1*}, Charles POUCHON^{1*}, Frédéric BOYER¹, Glenn YANNIC¹,
5 Delphine RIOUX¹, Aurélie BONIN¹ and Thibaut CAPBLANCQ^{1#}

6 * co-first authors

7 ¹ *Univ. Grenoble Alpes, Univ. Savoie Mont-Blanc, CNRS, LECA, Grenoble, France*

8

9

10 # Corresponding author: Thibaut Capblancq ; Laboratoire d'Écologie Alpine, 2233 Rue
11 de la Piscine, 38041 Grenoble Cedex, France. Tel: +33 (0)4 76 51 36 71; e-mail:
12 thibaut.capblancq@univ-grenoble-alpes.fr

13

14 Running headline:

15 Impact of wet and dry protocol of ddRAD-seq

16

17 **ABSTRACT**

18

19 1. Next-generation sequencing technologies have opened a new era of research in genomics.
20 Among these, restriction enzyme-based techniques such as restriction-site associated DNA
21 sequencing (RADseq) or double-digest RAD-sequencing (ddRADseq) are now widely used in
22 many population genomics fields. From DNA sampling to SNP calling, both wet and dry
23 protocols have been discussed in the literature to identify key parameters for an optimal loci
24 reconstruction.

25 2. The impact of these parameters on downstream analyses and biological results drawn from
26 RADseq or ddRADseq data has however not been fully explored yet. In this study, we tackled
27 this issue by investigating the effects of ddRADseq laboratory (*i.e.* wet protocol) and
28 bioinformatics (*i.e.* dry protocol) settings on loci reconstruction and inferred biological signal
29 at two evolutionary scale using two systems: a complex of butterfly species (*Coenonympha*
30 *sp.*) and populations of Common beech (*Fagus sylvatica*).

31 3. Results suggest an impact of wet protocol parameters (DNA quantity, number of PCR
32 cycles during library preparation) on the number of recovered reads and SNPs, the number of
33 unique alleles and individual heterozygosity. We also found that bioinformatic settings (*i.e.*
34 clustering and minimum coverage thresholds) impact loci reconstruction (*e.g.* number of loci,
35 mean coverage) and SNP calling (*e.g.* number of SNPs, heterozygosity). We however do not
36 detect an impact of parameter settings on three types of analysis performed with ddRADseq
37 data: measure of genetic differentiation, estimation of individual admixture, and demographic
38 inferences. In addition, our work demonstrates the high reproducibility and low rate of
39 genotyping inconsistencies of the ddRADseq protocol.

40 4. Thus, our study highlights the impact of wet parameters on ddRADseq protocol with strong
41 consequences on experimental success and biological conclusions. Dry parameters affects loci
42 reconstruction and descriptive statistics but not biological conclusion for the two studied
43 systems. Overall, this study illustrates, with others, the relevance of ddRADseq for population
44 and evolutionary genomics at the inter- or intraspecific scales.

45

46 **Keywords**

47 ddRADseq, laboratory protocol, bioinformatics treatments

48

49 **INTRODUCTION**

50 For a decade, next-generation sequencing (NGS) technologies have opened a new era in the
51 large field of molecular ecology. In particular, the advances in sequencing capabilities have
52 deeply changed the field of population genetics, by providing tremendous amount of sequence
53 data/information (10 to 100 thousand markers) at a relatively low cost. (Andrews et al., 2014;
54 da Fonseca et al., 2016). Whole genome re-sequencing (WGR) methods, providing the
55 highest marker density among the current genomic methods, notably appear very useful to
56 investigate many questions in evolutionary biology and ecology (Fuentes-Pardo & Ruzzante,
57 2017). WGR has however a limited relevance for non-model species, because a reference
58 genome is not always available and because it requires considerable sequencing and
59 computing efforts (Fuentes-Pardo & Ruzzante, 2017). Stemming from these limitations,
60 reduced-representation sequencing methods have been developed. These approaches include
61 restriction-site associated DNA sequencing (RAD-sequencing), sequencing of transcribed
62 DNA from mRNA (RNA-sequencing), and whole-exome sequencing (WES). Overall,

63 reduced-representation sequencing methods allow accessing numerous homologous loci with
64 a great taxa coverage at a relatively low cost (Fuentes-Pardo & Ruzzante, 2017). Among these
65 methods, RAD-sequencing, or RADseq (Miller, Dunham, Amores, Cresko, & Johnson, 2007;
66 Baird et al., 2008), is certainly the most popular method to obtain thousands of single
67 nucleotide polymorphisms (SNPs) for non-model species (K. R. Andrews, Good, Miller,
68 Luikart, & Hohenlohe, 2016). The principle of RADseq is to use restriction enzymes to
69 subsample the genome of multiple individuals at homologous genomic locations (Miller et al.,
70 2007; Baird et al., 2008). The resulting DNA fragments are then sequenced and compared
71 among individuals to detect SNPs. Since its origin, this technique has been transformed into a
72 variety of related approaches (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012; S. Wang,
73 Meyer, Mckay, & Matz, 2012; Toonen et al., 2013; Campbell, Brunet, Dupuis, & Sperling,
74 2018). Among these, double-digest RADseq, or ddRADseq (Peterson et al., 2012), is highly
75 customizable as regards the final number of loci, depending on the choice of enzymes and
76 range of fragment size selected. The ddRADseq approach has been applied with success to
77 many purposes including population genetic studies (Kjeldsen et al., 2016; Black, Seears,
78 Hollenbeck, & Samollow, 2017; Sherpa, Rioux, Goindin, et al., 2018), phylogenetic
79 reconstructions (DaCosta & Sorenson, 2016; Vargas, Ortiz, & Simpson, 2017; Boubli et al.,
80 2018; Lee et al., 2018; Sherpa, Rioux, Pougnat-Lagarde, & Després, 2018), demographic
81 inferences (Capblancq, Després, Rioux, & Mavárez, 2015; Nunziata, Lance, Scott, Lemmon,
82 & Weisrock, 2017; Settepani et al., 2017; Elleouet & Aitken, 2018) and landscape genetic
83 analyses (Saenz-Agudelo et al., 2015; Johnson, Gaddis, Cairns, Konganti, & Krutovsky,
84 2017). Despite the recognized advantages of the ddRADseq technique, several limitations and
85 weaknesses arose in the literature (Davey et al., 2013; K. R. Andrews et al., 2016; Lowry et
86 al., 2017). The main concerns are related to both the wet laboratory and bioinformatic

87 procedures associated with the method (Puritz et al., 2014; Mastretta-Yanes et al., 2015;
88 Shafer et al., 2017).

89 The frequency and distribution of restriction sites in the genome vary considerably
90 depending on the species and the pair of enzymes considered (Herrera, Reyes-Herrera, &
91 Shank, 2015). Among wet lab specific aspects, the choice of enzymes is therefore critical for
92 appropriate genome subsampling through a ddRADseq procedure. For example, this choice
93 will influence the number of digested fragments, their location in the genome, and their size
94 distribution (Burns et al., 2017; Y. Wang et al., 2017). DNA quality also influences SNPs
95 recovery, because degraded (*i.e.* fragmented) DNA can greatly lower the efficiency of
96 restriction enzyme-based techniques, by inducing a loss of recovered fragments (Graham et
97 al., 2015). Amplification of ddRADseq or RADseq fragments during library preparation has
98 also been pointed out as a potential critical step (Davey et al., 2013; Mastretta-Yanes et al.,
99 2015). Indeed, non-homogeneous amplification of RAD fragments can lead to a substantial
100 loss of alleles due to unbalanced RAD fragments coverage (Andrews & Luikart, 2014;
101 Andrews et al., 2014; Puritz et al., 2014) . Furthermore, the number of PCR cycles during
102 library preparation is generally low to minimize PCR artifacts (such as PCR errors) in the
103 RAD tags sequences (Hohenlohe, Catchen, & Cresko, 2012; Peterson et al., 2012).

104 Another important concern about RAD-based methods is the bioinformatic treatment
105 of sequences and the reconstruction of RADseq loci (*e.g.* Shafer et al., 2017). The principle of
106 the RADseq technique relies on the identification of homologous loci among individuals. This
107 task is implemented by clustering single-copy loci according to a similarity threshold, which
108 is determined using either distance-based (*e.g.* STACKS; Catchen, Hohenlohe, Bassham,
109 Amores, & Cresko, 2013), or global alignment (*e.g.* pyRAD; Eaton, 2014) methods. In both
110 cases, stringent parameter settings will avoid the clustering of paralogs but can also split

111 highly divergent single-copy loci in different clusters (Catchen et al., 2013; Eaton, 2014).
112 Coverage is another important parameter for loci reconstruction. A minimum number of reads
113 is generally set in order to take into account an allele or not (Catchen et al., 2013). Defining a
114 high threshold value can induce a loss of alleles via insufficient coverage, while a too low
115 value will not discard rare sequences originating from PCR or sequencing errors (Paris,
116 Stevens, & Catchen, 2017). The influence of parameter settings on quality and quantity of
117 recovered fragments and SNPs has therefore been widely tested (Eaton, 2014; Mastretta-
118 Yanes et al., 2015; Paris et al., 2017; Rochette & Catchen, 2017; Shafer et al., 2017) and it
119 sometimes impacts downstream population genomics analyses (Ilut, Nydam, & Hare, 2014;
120 Harvey et al., 2015; Willis, Hollenbeck, Puritz, Gold, & Portnoy, 2017).

121 Finally, the impact of missing data, which are inherent to any genotyping technique,
122 has also been evaluated over the years (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013;
123 Gautier et al., 2013; Malinsky, Trucchi, Lawson, & Falush, 2018). Missing data can be due to
124 some extent to an experimental lack of reproducibility, but more frequently to polymorphism
125 in restriction sites. This polymorphism leads to allele drop-out (ADO) for the individuals
126 lacking the restriction site in one or two of the homologous chromosomes. ADO directly
127 influences the estimation of genetic variation and diversity (Davey et al., 2013; Gautier et al.,
128 2013; Cariou, Duret, & Charlat, 2016). It has been particularly investigated in phylogenetic
129 studies (Cariou, Duret, & Charlat, 2013; Eaton, 2014; DaCosta & Sorenson, 2016) because of
130 the direct correlation between ADO and the divergence time among lineages and species
131 (Cariou et al., 2013).

132 The scientific community has accumulated expertise about RAD-based methods over
133 the last decade to make better use of such techniques while some critical issues deserve
134 further investigations. Indeed, if the proximal consequences have been investigated (*e.g.*

135 number of SNPs or deficit in heterozygosity), the distal consequences (effect on genetic
136 structure estimation, demographic inferences, etc.) remain rarely explored in most cases. The
137 impact of the minimum coverage or the similarity threshold for sequences clustering has for
138 example already been largely discussed in the literature. However, if the community accepts
139 that these parameters directly influence the number and coverage of loci or the percentage of
140 missing data in the final genetic matrix (Catchen et al., 2013), their potential effects on the
141 biological signal unraveled by downstream analyses are not systematically tested and this
142 issue lacks empirical investigation (but see Mastretta-Yanes et al., 2015; Rodríguez-Ezpeleta
143 et al., 2016; Shafer et al., 2017). In addition, some of the wet laboratory procedures are
144 thought to be critical for the success of the experiment (Peterson et al., 2012; Mastretta-Yanes
145 et al., 2015), like the initial DNA quantity and number of PCR cycles, but they have never
146 been experimentally evaluated.

147 This study aims to examine these steps of the ddRADseq procedure, providing a novel
148 contribution to the literature, in an animal system at an interspecific level (in the butterfly
149 species complex of *Coenonympha*) and in plants at an intraspecific scale (in tree populations
150 of European/common beech, *Fagus sylvatica*). We first evaluate the impact of both initial
151 DNA quantity and number of PCR cycles on the experiment results, and the reproducibility of
152 our wet protocol by evaluating the percentage of genotyping inconsistencies and missing
153 fragments between replicates. Concerning the bioinformatic treatment, we investigate the
154 influence of the minimum coverage and the similarity threshold during the loci reconstruction
155 on three types of analyses based on ddRADseq data: genetic differentiation (evaluated using
156 F_{ST} estimation and Principal Component Analysis), genetic structure (genetic clustering) and
157 demographic inferences (Approximate Bayesian Computation method).

158

159 **MATERIALS AND METHODS**

160 **Sampling**

161 This study is based on a total of 108 samples, including 58 individuals of *Fagus sylvatica*
162 (deciduous tree; genome size around 600 Mbp) and 50 individuals of *Coenonympha* sp. (three
163 butterfly species: *C. arcana*, *C. gardetta* and *C. macromma*; genome size around 300 Mbp).
164 Depending on the conditions and settings tested (Fig. 1), different numbers of individuals and
165 populations were used (see Table S1).

166

167 **Standard ddRADseq protocol**

168 A double-digested RAD experiment was conducted on individuals using a common protocol
169 for both the wet and dry parts of the procedure. The protocol was the same for all samples,
170 except for some parameter settings as described in the following sections: “*Setting tests for*
171 *the wet laboratory protocol*” and “*Tests on bioinformatics parameters*”.

172

173 **Library preparation** - DNA was extracted from one leaf for *Fagus sylvatica* samples and
174 from the complete thorax of individual *Coenonympha* butterflies using a DNeasy Blood &
175 Tissue Kit (QIAgen) following manufacturer’s instructions. For each individual, 200 ng of
176 genomic DNA were double-digested with 10 units of each *PstI* and *MspI* (New England
177 Biolabs Inc.) at 37°C for two hours in a final volume of 34µL, using the CutSmart buffer
178 provided with the enzymes. Digestion was further continued with ligation of the P1 (with
179 individual tags) and P2 adapters (see Peterson et al., 2012) by adding to each sample 10 units
180 of T4 DNA ligase (New England Biolabs Inc.), adapters P1 and P2 in 10-fold excess
181 (compared to the estimated number of restriction fragments), 1µL of 10mM ribo-ATP (New
182 England Biolabs Inc.) and once again two units of *PstI* and *MspI* enzymes. This simultaneous

183 digestion-ligation reaction was performed on a thermocycler using 60 cycles of a succession
184 of 2 min at 37°C for digestion and 4 min at 16°C for ligation. An equal volume of all the
185 digested-ligated fragment mixtures were pooled and purified using magnetic beads
186 (Agencourt AMPure XP of Beckman Coulter, or NucleoMag of Macherey Nagel) with a
187 DNA/beads ratio equal to 1/1.5. Fragments were size-selected in a range between 250 and 500
188 bp on agarose gel (1.6%) and excised bands purified with the QIAquick Gel Extraction Kit
189 (Qiagen). The ddRADseq library obtained was amplified independently eight times by PCR,
190 and the obtained PCR products were then pooled, in order to minimize the impact of potential
191 PCR errors. We used the following PCR mix: a final volume of 20 μ L containing 1 μ L of
192 DNA template, 10 mM of dNTPs, 10 μ M of each PCR primer (Peterson et al., 2012) and
193 2U/ μ L of *Taq* Phusion-HF (New England Biolabs Inc.); and the following PCR program: an
194 initial denaturation at 98°C for 30 seconds; 15 cycles of 98°C for 10 sec, 66°C for 30 sec and
195 72°C for 1 min; followed by a final extension at 72°C for 10 min. The amplified ddRADseq
196 library was purified with magnetic beads and sequenced on half a lane of an Illumina Hi-Seq
197 2500 sequencer (paired-end 2 x 125 bp, Fasteris SA).

198

199 **Bioinformatic treatment** - DNA sequences of *Fagus sylvatica* and *Coenonympha* sp. libraries
200 were used to call SNP genotypes (total: ~100 million reads). We developed a homemade
201 pipeline called ProcessMyRAD (fully available at <https://github.com/cumtr/pmr>) to
202 automatically perform the different steps leading from the raw reads to genotype data. To call
203 the genotypes, ProcessMyRAD relies on the STACKS pipeline (Catchen et al., 2013). To
204 reconstruct loci, the STACKS procedure needs to set three thresholds: the minimum number of
205 reads to consider an allele (**m**), the maximum number of mismatches allowed between two

206 alleles to reconstruct a locus (**M**), and the maximum number of mismatches allowed between
207 two individual loci to consider them as homologous (**n**).

208

209 **Setting tests for the wet laboratory protocol**

210 *Impact of initial DNA amount and number of PCR cycles* - We evaluated the impact of two
211 parameters on the experiment results: 1) the DNA quantity used for the initial
212 digestion/ligation step; and 2) the number of PCR cycles used to produce the final library.

213 For 10 samples of *Coenonympha* and 10 samples of *Fagus sylvatica* (Table S1), we
214 repeated the ddRADseq lab experiment three times for each sample with the standard protocol
215 described above but with different quantities of DNA during the first step: 50, 150 or 250 ng
216 of initial genomic DNA. Similarly, we used 10 samples of *Coenonympha* sp. and 10 samples
217 of *Fagus sylvatica* to repeat three times the ddRADseq lab experiment, with different
218 numbers of PCR cycles in the final step of the protocol: 10, 15 or 25 cycles. We then
219 sequenced the resulting libraries all together.

220 The sequences resulting from these tests were treated with the *STACKS* program and the
221 following clustering parameters: **m**=4, **M**=6, **n**=8 (based on the results of the section
222 “Bioinformatics tests” for **m** and **M**, and with **n**=**M**+2 to increase the number of inter-
223 individual matches), keeping only one SNP by ddRADseq fragment. To estimate the impact
224 of the wet lab treatment on alleles frequencies, we did not filter out alleles based on their
225 frequency. On the resulting genetic datasets, we determined the number of polymorphic
226 fragments, the mean fragment coverage, the number of SNPs in the fragments, the individual
227 heterozygosity, and the proportion of private alleles in individuals.

228

229 **Experimental reproducibility** - We assessed the reproducibility of our laboratory protocol by
230 repeating the experiment, since the digestion/ligation step, three times for 11 *Fagus sylvatica*
231 individuals. Each replicate of these triplets was processed with the same protocol and was
232 sequenced within the same Illumina sequencing run. All sequences obtained were treated
233 together with the ProcessMyRAD pipeline with **m**=4, **M**=6, **n**=8 and a minor allele frequency
234 of 0.1 (corresponding to at least three individuals to keep the allele). A locus was kept only if
235 sequenced in at least 50% of the 33 replicates.

236 For each replicate, the number of ddRADseq fragments, the mean fragment coverage, the
237 proportion of polymorphic fragments and the individual heterozygosity were estimated. These
238 different parameters were then compared among replicates to assess the intra and inter-
239 replicate variability. We also evaluated reproducibility by performing a Principal Component
240 Analysis on genetic data with the R-package *adegenet* (Jombart, 2008), and looking at the
241 distances between replicates in the PCA projection space. In addition, the replicates were used
242 to estimate the proportion of inconsistencies in our final genetic dataset. These inconsistencies
243 can take two different forms: errors of genotypes (due to PCR errors or ADO) or fragment
244 absence (due to a lack of reproducibility of the experiment). We measured the genotyping
245 inconsistency rate by identifying the proportion of loci with inconsistencies among the three
246 replicates. Then, by looking at the fragments in the three replicates, we could estimate the
247 proportion of “true” fragment absences, when the fragment was missing in all three replicates,
248 and the proportion of “false” fragment absences, when the fragment was missing in just one
249 or two of the replicates.

250
251 **Test on bioinformatic parameters**

252 **Impact of bioinformatic treatment** - We estimated the influence of the **m** (*ustacks*) and **M**
253 (*cstacks*) values on ddRADseq fragment reconstruction and downstream analyses.

254 For this purpose, we used the standard ddRADseq wet protocol described above on 30
255 individuals of *Fagus sylvatica* coming from three different populations (Sainte Beaume,
256 Digne, and Bauges, France; see Table S1) and 30 individuals of *Coenonympha* butterflies
257 from three different species (*C. arcania*, *C. gardetta* and *C. macromma*; see Table S1). We
258 then repeated the bioinformatic pipeline with different **m** values ranging from 1 to 15 and
259 different values of **M** ranging from 1 to 25. When the **m** value varied, **M** and **n** were fixed to
260 6. When the **M** value varied, **m** was fixed to 4 and **n** was equal to **M**. For all these tests, the
261 remaining steps of the procedure were exactly the same and the last step of genetic dataset
262 export was performed by keeping only one SNP by RAD tag fragment and without any
263 filtering on allelic frequency.

264 To estimate the influence of the **m** and **M** values on RAD tag fragment recovery, we
265 determined the number of reconstructed fragments, their mean coverage and the proportion of
266 polymorphic fragments for each value of **M** and **m** tested. We also evaluated the impact of
267 these parameters on population genetics results by performing, for all **m** and **M** values, some
268 of the most commonly used analyses using ddRADseq data (Capblancq et al., 2015; Kjeldsen
269 et al., 2016; Black et al., 2017; Nunziata et al., 2017; Settepani et al., 2017; Elleouet &
270 Aitken, 2018; Sherpa, Rioux, Pouget-Lagarde, et al., 2018), i.e. mean individual
271 heterozygosity, F_{ST} among populations (estimated with the *adegenet* R package (Jombart,
272 2008)), Principal Component Analysis (PCA, using the *adegenet* R package (Jombart, 2008)),
273 genetic structure with sNMF (using the *LEA* R package (Frichot & François, 2015)) and
274 evolutionary history reconstruction using Approximate Bayesian Computation (performed
275 with the *diyABC* program (Cornuet et al., 2014)). The results of these analyses were then
276 compared across the **m** and **M** ranges and with results from other population genetic studies

277 on the same species *Coenonympha* sp. In Capblancq et al., 2015, and *Fagus sylvatica* in
278 Capblancq et al. (in review).

279

280 **RESULTS**

281 **Influence of DNA quantity and number of PCR cycles**

282 **DNA quantity** - For all initial DNA quantity conditions, the library produced a mean of 3,172
283 fragments for *Coenonympha*, with a mean coverage of 24.6 reads per fragment, and 450
284 fragments for *Fagus sylvatica*, with a mean coverage of 19.5 reads per fragment. With 50 ng
285 or 150 ng of genomic DNA as template, similar numbers of fragments and SNPs were
286 recovered (around 3,500 fragments for *Coenonympha* and around 500 for *Fagus sylvatica*). If
287 fragment coverage varies (from 15 to 30 for *Coenonympha* and from 15 to 23 for *Fagus*
288 *sylvatica*), this does not have much impact on individual heterozygosity ($He \sim 0.125$ for
289 *Coenonympha* and $He \sim 0.25$ for *Fagus sylvatica*). Conversely, we noticed that using 250 ng of
290 DNA during the initial step of digestion/ligation could dramatically decrease fragment
291 recovery (divided by 1.5) and SNPs identification for some individuals (Fig. 2). Using 250 ng
292 of DNA also induced a greater variability among tested individuals (Fig. 2).

293

294 **Number of PCR cycles** - For all PCR conditions, the library produced a mean of 2,480
295 fragments for *Coenonympha* with a mean coverage of 23.9 reads per fragment and 520
296 fragments for *Fagus sylvatica* with a mean coverage of 23.3 reads per fragment. Again, the
297 results showed a great variability depending on the PCR settings. Increasing the number of
298 PCR cycles in the final library preparation had a positive effect on the number of fragments
299 and SNPs recovery. The mean number of fragments for *Fagus sylvatica* ranged from 41 for
300 10 cycles to 1,028 for 25 cycles. Similar results were obtained for *Coenonympha* sp. for

301 which the number of fragments varied from 350 to 4,500. The number of PCR cycles was also
302 directly correlated with individual heterozygosity and with the number of private alleles in the
303 individuals. For example, the individual heterozygosity increased from 0.09 to 0.27 for *Fagus*
304 *sylvatica* individuals when the number of cycles increased from 10 to 25. In the same way, the
305 number of private alleles doubled when the number of PCR cycles increased from 10 to 25 for
306 both *Coenonympha* sp. and *Fagus sylvatica* samples. Finally, 10 cycles of PCR lowered
307 substantially the number of fragments and SNPs as well as the number of private alleles in
308 the final genetic dataset.

309

310 **Reproducibility of the experiment and estimation of inconsistencies**

311 The library of the 11 *Fagus sylvatica* triplicates produced a mean of 7,547 fragments with a
312 mean coverage of 20.3 reads per fragments. The PCA performed on the complete genetic
313 dataset showed very consistent results across the 11 tested individuals (Fig. 3). Inter-
314 individual genetic variability was higher than inter-replicate genetic variability. All triplicates
315 clustered in the PCA plot and the different individuals could easily be differentiated.
316 Moreover, considering the eigenvalues, the 10 first axes retained most of the genetic variance
317 within the three replicates * 11 sample tests (92%). For example, PC1 strongly discriminates
318 individual VTX_H_83 from the rest of the sampling and PC2 differentiates individuals
319 SB_H_42 and BG_1_1. This suggests that each PC captured parts of inter-individual genetic
320 variability differentiating a particular individual from the remaining samples. Replicates did
321 not seem to add any substantial genetic variability that could have been caught by the PCA.

322 Across all replicates and individuals, the number of recovered fragments varied from
323 around 4,000 to 16,000; the mean coverage from 8 to 40; the proportion of polymorphic loci
324 from 0.15 to 0.38, and the individual heterozygosity from 0.27 to 0.37 (Fig. 3). For seven

325 individuals, the replicates returned almost exactly the same number of fragments, the same
326 ratio of polymorphic loci and the same individual heterozygosity. Four individuals showed
327 more contrasted results but with similar patterns across the different parameters we measured.
328 No association between the initial DNA concentration after extraction and the consistency of
329 ddRADseq results was observed (data not shown).

330 Regarding the estimation of genotype inconsistencies, the results were congruent
331 across the 11 tested individuals (Fig. 4). The maximum inconsistency rate was just above 4%
332 and the minimum is around 1.6%. Similarly, we obtained a good proportion of fragments
333 recovery among replicates. Between 66% and 90% of the ddRADseq fragments were found in
334 all three replicates (Fig. 4). The individuals with low fragment recovery rate were the exact
335 same ones that showed a great variability in the reproducibility experiment (see above). Some
336 fragments were missing for all three replicates, and the proportions of missing fragments were
337 pretty homogeneous across individuals, varying from 5% to 13%. Finally, for all samples, a
338 fair proportion of fragments (2 to 24%) was found in only one or two replicates, giving an
339 estimation of fragment loss not due to restriction site polymorphism across individuals but to
340 incomplete digestion, ligation, amplification or sequencing of these fragments.

341

342 **Influence of bioinformatic thresholds on the biological results**

343 The ddRADseq libraries used for bioinformatic tests produced very variable numbers of
344 fragments and coverage depending on the thresholds used along the analysis pipeline. With
345 the parameter values $\mathbf{m} = 4$ and $\mathbf{M} = 6$, we obtained a mean of 3,246 fragments for
346 *Coenonympha* individuals with a mean coverage of 40.16 reads per fragment and 11,018
347 fragments for *Fagus sylvatica* samples with a mean coverage of 25.18 reads per fragment.

348 The minimum coverage required to create a RAD tag locus during the first step of
349 *STACKS* procedure (**m**) had a direct influence in the number of fragments, the mean coverage
350 of the fragments and the number of SNPs identified in these fragments (Fig. 5). Furthermore,
351 the pattern was very similar for the *Coenonympha* sp. and *Fagus sylvatica* models. An
352 increase in **m** value was associated with a decrease in fragment recovery, the variation being
353 particularly important between **m** = 1 and **m** = 2. Similarly, an increase of **m** value was
354 associated with an increase in mean fragment coverage, ranging from 10x for a **m** value of 1
355 to 58x for a **m** value of 15 in *Coenonympha* sp. and from 5x for a **m** value of 1 to 30x for a **m**
356 value of 15 in *Fagus sylvatica*. The number of SNPs identified in the fragments was also
357 strongly affected by variation of the **m** threshold: the highest number of SNPs was reached for
358 **m** = 2 and the lowest number for **m** = 15. Regarding individual heterozygosity, there was a
359 slight increase when **m** increased. Heterozygosity ranged from 0.29 for **m** = 1 to 0.36 for **m** =
360 15 for *Fagus sylvatica* individuals, and from 0.13 to 0.15 for *Coenonympha* sp. samples.

361 Nevertheless, the **m** parameter did not seem to influence any of the downstream
362 population genetic analyses. No major difference among the **m** parameter settings was
363 observed for F_{ST} estimation among populations (Fig. 6), PCA and genetic clustering results
364 (Fig. 6, Fig S1 and Fig. S2) or demographic inferences (Fig. 7). Here again, the results were
365 similar for both the animal and plant models. While slight changes in F_{ST} values or PCA
366 scores were noticed when **m** varied, the populations remained differentiated in the same way
367 and strength (Fig. 6). For example, the F_{ST} values ranged from 0.28 to 0.33 between
368 *Coenonympha arcania* and *C. gardetta* but the ranking of F_{ST} values among the three pairs of
369 species did not change depending on **m** (Fig. 6). An increasing **m** seemed to slightly influence
370 the percentage of inertia retained by the first two PCs for both *Fagus sylvatica* and
371 *Coenonympha* sp. (from 18 to 20% of the genetic variance, see Fig. 6) but population

372 differentiation on the PCA was not affected. The Procrustes superimposition performed on the
373 first two axes of the PCAs returns correlation coefficients superior to 0.96 between each pair
374 of **m** values for *Coenonympha* and superior to 0.85 for *F. sylvatica* (Fig. S7). Regarding the
375 genetic structure (sNMF analysis), the number of K selected with the cross-entropy criterion
376 did not vary for *Coenonympha* samples and varied between 2, 3 and 4 for *Fagus sylvatica*
377 individuals (Fig. 6). This variation was due to very close values of cross-entropy for K = 2, 3
378 and 4 (Fig. S2). Similarly, the differentiation of species groups in the sNMF analysis
379 remained exactly the same across the range of **m** values (Fig. 6). The Procrustes
380 superimposition performed on the percentage of assignation to the three main clusters
381 obtained with sNMF returns correlation coefficients superior to 0.975 between each pair of **m**
382 values for *Coenonympha* and superior to 0.99 for *F. sylvatica* (Fig. S8). Finally, we did not
383 detect any influence of the **m** parameter on the estimations of population size, divergence time
384 or hybridization rate through ABC procedure (Fig. 7). All model parameters showed
385 approximately the same posterior distribution whatever the **m** value, with only a small
386 variation between the maximum and the minimum of the estimates across the **m** range (Table
387 S2).

388 The maximum number of mismatches accepted between two stacks of sequences to
389 merge two alleles in one locus (**M**) greatly influenced the number of recovered fragments, the
390 number of identified SNPs and individual heterozygosity (Fig. 5). When **M** varied from 1 to
391 25 the number of recovered ddRADseq fragments decreased from 12,000 to 10,000 for *Fagus*
392 *sylvatica* and from 3,500 to 3,000 for *Coenonympha* sp. On the opposite, the number of SNPs
393 identified increased rapidly for the first **M** values of the range (1-6) until a plateau was
394 reached around 3,800 fragments for *Fagus sylvatica* individuals. The influence of **M** on
395 individual heterozygosity was clear for **M** values between 1 and 6, for which heterozygosity

396 increased with **M**. For higher values of **M**, the relationship was less obvious and the variation
397 of individual heterozygosity did not seem to follow the variation of **M**.

398 In agreement with the results obtained for the **m** parameter, population genetic
399 analyses showed very consistent results across the range of tested **M** values. Again, no
400 substantial effect was observed for F_{ST} values among populations, PCA results, genetic
401 clustering results or demographic inferences when **M** value varied (Fig. 6, Fig. 7, Fig. S4 and
402 Fig. S5). **M** variation did not impact PCA, neither in terms of population differentiation, nor
403 in terms of percentage of inertia of the two first axes. The Procrustes superimposition
404 performed on the first two axes of the PCAs returns correlation coefficients superior to 0.95
405 between each pair of **m** values for *Coenonympha* and superior to 0.85 for *F. sylvatica* (Fig.
406 S7). For genetic structure however, we observed a slight change in the number of K selected
407 by the cross-entropy criterion. Here again, it is rather due to close cross-entropy values at K =
408 3 and K = 4 than to a real variation across the **M** range (Fig. S5). Even though, neither the
409 genetic grouping of individuals nor the percentage of assignation varied across the range of **M**
410 value (Fig. 6 and Fig. S8). Finally, all parameters inferred during the ABC analysis showed
411 very consistent distributions depending on the **M** value used for the sequence clustering (Fig.
412 7), with only a small variation between the maximum and minimum of the estimates across
413 the **M** range (Table S2).

414

415 DISCUSSION

416 Double-digest RAD-sequencing is a widely used technique to investigate population genetics
417 for a wide range of non-model organisms (Peterson et al., 2012; K. R. Andrews et al., 2016).
418 Multiple studies have speculated and tested the impact of different parameter settings on pre-
419 and post-sequencing procedures of RAD-based protocols, especially in loci reconstruction

420 summary statistics, i.e number of loci or SNPs, or heterozygosity (Davey et al., 2013; Gautier
421 et al., 2013; K. Andrews & Luikart, 2014; K. R. Andrews et al., 2014; Puritz et al., 2014;
422 Mastretta-Yanes et al., 2015; Burns et al., 2017; Rochette & Catchen, 2017; Y. Wang et al.,
423 2017; Willis et al., 2017). Nevertheless, only a handful of them linked those parameters to the
424 downstream biological interpretation (Mastretta-Yanes et al., 2015; Rodríguez-Ezpeleta et al.,
425 2016; Shafer et al., 2017; Malinsky et al., 2018). In addition, some pre-sequencing factors,
426 regularly pointed out as sensitive parts of the RADseq protocol (Hohenlohe et al., 2012;
427 Peterson et al., 2012), lack experimental testing on various models that would allow building
428 a solid knowledge of their influence in ddRAD data production. Our objectives here were to
429 test the impact of some wet laboratory and bioinformatic treatment settings on loci recovery
430 as well as on population genetics and demographic inferences.

431

432 **Pre-sequencing treatment** - Regarding the wet protocol, we focused on two factors for
433 which we did not find any proper evaluation in the literature: the initial DNA quantity and the
434 number of PCR cycles in the last step of library amplification.

435 The initial amount of DNA required in ddRADseq library preparation is a constraint
436 for small individuals or museum samples (Blair, Campbell, & Yoder, 2015; Shortt et al.,
437 2017). Here, we found that only a small amount of DNA template (*i.e.* ~50 ng; Fig. 2) was
438 required in our library preparation for the two tested systems, which could open the
439 possibility of using ddRADseq technique with low quantity DNA template from non-invasive
440 sampling in a conservation genomics context. If the initial protocol from Peterson et al.
441 (2012) already suggested to use less than 100 ng of DNA per individual, ddRADseq users
442 commonly process more than 200 ng and even up to 1 µg (Capblancq et al., 2015; Yang et al.,
443 2016; Burns et al., 2017; Sherpa, Rioux, Goindin, et al., 2018). Here we showed that enzyme

444 saturation can already happen at the digestion step with 250 ng. Obviously, the exact amount
445 of DNA that can be digested is directly dependent on the number of enzyme units used and
446 the number of restriction sites. We showed that even a 5-time variation (*i.e.* 250 ng instead of
447 50 or 150 ng in our study) increased the probability of experiment failure by reducing the loci
448 and the SNP recovery and by inflating the variability between samples (Fig. 2). This points
449 out the need to calibrate finely the amount of DNA and number of enzyme units to avoid a
450 dramatic loss of fragments.

451 The number of PCR cycles is another part of the experiment that has to be carefully
452 considered (Fig. 2). In their study, Davey et al., 2013 highlighted that PCR cycles introduced
453 GC biases in sequenced RAD libraries. For example, RAD loci with high GC content were
454 sequenced more often compared to RAD loci with low GC content for high numbers of PCR
455 cycles and the opposite was true for low numbers of PCR cycles (Davey et al., 2013). In
456 this study, our results showed that the number of SNPs and individual heterozygosity is
457 reduced with a low number of PCR cycles, while a high number of PCR cycles increases the
458 number of private alleles. This highlights the trade-off existing between a satisfactory
459 coverage, directly related to the number of PCR cycles, and the limitation of errors occurring
460 during PCR, because these can lead to very weak fragment coverage impeding loci
461 reconstruction (Hohenlohe et al., 2012). Usually the number of PCR cycles is set between 12
462 and 16 (Peterson et al., 2012; Capblancq et al., 2015; Yang et al., 2016; Burns et al., 2017).
463 Considering the important increase in individual heterozygosity and number of private alleles
464 observed with 25 PCR cycles, our results greatly support this practice. It would be interesting
465 to investigate further if the increase in private alleles and individual heterozygosity could be
466 simply overcome by stringent filtering on allele frequencies aiming at discarding low
467 frequency alleles in the population. Unfortunately, our sampling size did not allow such tests.

468 Furthermore, we showed that, when properly calibrated, the protocol is greatly
469 reproducible. The triplicates used for 11 individuals of *Fagus sylvatica* showed very close
470 proximities in genetic PCA and most of them showed similar numbers of fragments and
471 coverage (Fig. 3). Such results were consistent with those of Mastretta-Yanes et al. (2015), on
472 which most replicate pairs were clustered together in neighbour-joining dendograms. Even if
473 there are several steps during ddRADseq laboratory experiment that could lack
474 reproducibility to some extent (*e.g.* digestion/ligation, range size selection, amplification by
475 PCR), our results were robust across replicates. Combined with a low rate of genotyping
476 inconsistency and missing fragments (Fig. 4), our results illustrate that ddRADseq is an
477 accessible method with some key parameters that have to be finely tuned to gain in robustness
478 and reproducibility. Our testing procedure does not claim to cover all parameters that could
479 influence the ddRADseq method but points at key information about lab protocols and gives
480 clues to optimize the technique.

481

482 **Post-sequencing treatment** - Concerning the dry protocol, an important part of the RAD-
483 based sequencing literature pertains to the bioinformatic treatment of sequences and to loci
484 reconstruction (Mastretta-Yanes et al., 2015; Paris et al., 2017; Rochette & Catchen, 2017; Y.
485 Wang et al., 2017). These studies highlight the impact of clustering thresholds (*e.g.* **M** and **m**
486 parameters of the *STACKS* software procedure) on bioinformatic results and summary
487 statistics. Indeed, these thresholds have been shown to influence the number of recovered loci,
488 coverage, number of identified SNPs and error rates (Mastretta-Yanes et al., 2015; Paris et al.,
489 2017; Rochette & Catchen, 2017; Y. Wang et al., 2017). In agreement with previous works,
490 we found a substantial impact of the minimum coverage (**m**) and clustering (**M**) thresholds on
491 ddRADseq loci recovery and reconstruction during the bioinformatic process. The minimum

492 coverage imposes a minimum number of reads to consider an allele. Many alleles are
493 expected to be lost with a high **m**, while a low **m** can give too much importance to very rare
494 sequences, and thus potentially to sequencing or PCR errors (Catchen et al., 2013). The
495 similarity threshold (**M**) determines the minimum sequence homology to consider that two
496 sequences are variants of the same locus. Choosing a too high **M** can wrongly impede the
497 clustering of different alleles of the same locus, while a too low value could lead to the
498 merging of paralog regions of the genome (Catchen et al., 2013).

499 We could have expected that such variation in loci reconstruction and SNPs
500 identification would influence, to some extent, the population genetic analyses performed
501 with most ddRADseq datasets. However, our results suggest that the bioinformatic treatment
502 has only a marginal influence on population genetic results. Indeed, no change in genetic
503 differentiation, clustering or demographic inferences were detected neither at the inter-species
504 level for the animal model nor at the intraspecific level for the plant model (Fig 6, 7, S7 and
505 S8).

506 Moreover, despite a reduced number of individuals, the results of this study are
507 congruent with previous results obtained with a larger sampling for both *Coenonympha* and
508 *Fagus sylvatica* (Capblancq et al., 2015; Capblancq, unpublished data). Such patterns may be
509 explained by the large amount of information generated by the ddRADseq method (10 or 100
510 of thousands of SNPs). A potential “false” signal due to genotyping inconsistencies at some
511 loci seems negligible compared to the abundance of “true” signal provided by most of the
512 RAD loci. Similar observations have also been made for another species model, *i.e.*..
513 Galapagos sea lion (*Zalophus wollebaeki*) in Shafer et al. (2017), and similarly, several
514 studies have demonstrated that clustering parameters have no impact on phylogenetic
515 reconstruction (Herrera et al., 2015; Hou et al., 2015; Lee et al., 2018).

516 Finally, combined with the results from Malinsky et al. (2018) showing that non-
517 random data missingness due to a batch (i.e library) effect had no impact on downstream
518 analyses of the genetic structure. These findings moderate the message from the literature,
519 which commonly presents the bioinformatics treatment as a key parameter of ddRADseq loci
520 reconstruction. If this step indeed has an influence on loci recovery (*i.e* summary statistics), it
521 only has a weak impact on the biological signal resulting from population genetic analyses, at
522 least for the two models tested in this study.

523

524 **ACKNOWLEDGEMENTS**

525 The authors belong to the DivAdapt and MALBIO research teams at the Laboratoire
526 d'Écologie Alpine (LECA). We would like to thank Nadir Alvarez and Marianne Gagnon for
527 their useful comments on the manuscript. TCa acknowledges support from the ANR project
528 APPATS (ANR-15-CE02-0004) and all authors acknowledge the LECA for funding support.
529 The LECA is part of the Labex OSUG@2020 (ANR10 LABX56). Most of the computations
530 presented in this paper were performed using the CIMENT infrastructure, which is supported
531 by the Auvergne-Rhône-Alpes region (GRANT CPER0713 CIRA)

532

533

534

535 **REFERENCES**

536 Andrews, K., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis.
537 *Mol. Ecol.*, 23(7), 1661–1667.

538 Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the
539 power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17(2),
540 81–92. doi:10.1038/nrg.2015.28

541 Andrews, K. R., Hohenlohe, P. A., Miller, M. R., Hand, B. K., Seeb, J. E., & Luikart, G. (2014).
542 Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. *Molecular
543 Ecology*, 23(24), 5943–5946. doi:10.1111/mec.12964

544 Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity
545 and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*,
546 22(11), 3179–3190. doi:10.1111/mec.12276

547 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A.
548 (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, 3(10), 1–7. doi:10.1371/journal.pone.0003376

549 Black, A. N., Sears, H. A., Hollenbeck, C. M., & Samollow, P. B. (2017). Rapid genetic and
550 morphologic divergence between captive and wild populations of the endangered Leon
551 Springs pupfish, *Cyprinodon bovinus*. *Molecular Ecology*, 26(8), 2237–2256.
552 doi:10.1111/mec.14028

553 Blair, C., Campbell, C. R., & Yoder, A. D. (2015). Assessing the utility of whole genome amplified
554 DNA for next-generation molecular ecology. *Molecular Ecology Resources*, 15(5), 1079–
555 1090. doi:10.1111/1755-0998.12376

556 Boubli, J. P., da Silva, M. N. F., Rylands, A. B., Nash, S. D., Bertuol, F., Nunes, M., ... Hrbek, T.
557 (2018). How many pygmy marmoset (*Cebuella* Gray, 1870) species are there? A taxonomic
558 re-appraisal based on new molecular evidence. *Molecular Phylogenetics and Evolution*,
559 120(October 2017), 170–182. doi:10.1016/j.ympev.2017.11.010

560 Burns, M., Starrett, J., Derkarabetian, S., Richart, C. H., Cabrero, A., & Hedin, M. (2017).
561 Comparative performance of double-digest RAD sequencing across divergent arachnid
562 lineages. *Molecular Ecology Resources*, 17(3), 418–430. doi:10.1111/1755-0998.12575

563 Campbell, E. O., Brunet, B. M. T., Dupuis, J. R., & Sperling, F. A. H. (2018). Would an RRS by any
564 other name sound as RAD? *Methods in Ecology and Evolution*, 9(9), 1920–1927.
565 doi:10.1111/2041-210X.13038

566 Capblancq, T., Després, L., Rioux, D., & Mavárez, J. (2015). Hybridization promotes speciation in
567 *Coenonympha* butterflies. *Molecular Ecology*, 24(24). doi:10.1111/mec.13479

568 Cariou, M., Duret, L., & Charlat, S. (2013). Is RAD-seq suitable for phylogenetic inference? An in
569 silico assessment and optimization. *Ecology and Evolution*, 3(4), 846–852.
570 doi:10.1002/ece3.512

571 Cariou, M., Duret, L., & Charlat, S. (2016). How and how much does RAD-seq bias genetic diversity
572 estimates? *BMC Evolutionary Biology*, 16(1), 1–8. doi:10.1186/s12862-016-0791-0

573 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis
574 tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140.
575 doi:10.1111/mec.12354

576 Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., ... Estoup, A.
577 (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences
578 about population history using single nucleotide polymorphism, DNA sequence and
579 microsatellite data. *Bioinformatics (Oxford, England)*, 30(8), 1187–1189.
580 doi:10.1093/bioinformatics/btt763

581 da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Marety, L.,
582 ... Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches
583 for non-model organisms. *Marine Genomics*, 30, 3–13. doi:10.1016/j.margen.2016.04.012

584 DaCosta, J. M., & Sorenson, M. D. (2016). DdRAD-seq phylogenetics based on nucleotide, indel, and
585 presence-absence polymorphisms: Analyses of two avian genera with contrasting histories.
586 *Molecular Phylogenetics and Evolution*, 94(August), 122–135.
587 doi:10.1016/j.ympev.2015.07.026

588 Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special
589 features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, 22(11),
590 3151–3164. doi:10.1111/mec.12084

591 Eaton, D. A. R. (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses.
592 *Bioinformatics*, 30(13), 1844–1849. doi:10.1093/bioinformatics/btu121

593 Elleouet, J. S., & Aitken, S. N. (2018). Exploring Approximate Bayesian Computation for inferring
594 recent demographic history with genomic markers in nonmodel species. *Molecular Ecology*
595 *Resources*, 18(3), 525–540. doi:10.1111/1755-0998.12758

596

597 Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association
598 studies. *Methods in Ecology and Evolution*, 6(8), 925–929. doi:10.1111/2041-210X.12382

599 Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for
600 conservation biology: Advantages, limitations and practical recommendations. *Molecular
601 Ecology*, 26(20), 5369–5406. doi:10.1111/mec.14264

602 Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., ... Estoup, A. (2013). The
603 effect of RAD allele dropout on the estimation of genetic variation within and between
604 populations. *Molecular Ecology*, 22(11), 3165–3178. doi:10.1111/mec.12089

605 Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., ... Somers, C.
606 M. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing
607 (RADSeq). *Molecular Ecology Resources*, 15(6), 1304–1315. doi:10.1111/1755-0998.12404

608 Harvey, M. G., Judy, C. D., Seeholzer, G. F., Maley, J. M., Graves, G. R., & Brumfield, R. T. (2015).
609 Similarity thresholds used in DNA sequence assembly from short reads can reduce the
610 comparability of population histories across species. *PeerJ*, 3, e895. doi:10.7717/peerj.895

611 Herrera, S., Reyes-Herrera, P. H., & Shank, T. M. (2015). Predicting RAD-seq Marker Numbers
612 across the Eukaryotic Tree of Life. *Genome Biology and Evolution*, 7(12), 3207–3225.
613 doi:10.1093/gbe/evv210

614 Hohenlohe, P. A., Catchen, J., & Cresko, W. A. (2012). Population Genomic Analysis of Model and
615 Nonmodel Organisms Using Sequenced RAD Tags. In *Data Production and Analysis in
616 Population Genomics* (pp. 235–260). Humana Press, Totowa, NJ. doi:10.1007/978-1-61779-
617 870-2_14

618 Hou, Y., Nowak, M. D., Mirré, V., Bjoråa, C. S., Brochmann, C., & Popp, M. (2015). Thousands of
619 RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus Diapensia
620 (Diapensiaceae). *PLoS ONE*, 10(10), 1–15. doi:10.1371/journal.pone.0140175

621 Ilut, D. C., Nydam, M. L., & Hare, M. P. (2014). Defining loci in restriction-based reduced
622 representation genomic data from nonmodel species: Sources of bias and diagnostics for
623 optimal clustering. *BioMed Research International*, 2014. doi:10.1155/2014/675158

624 Johnson, J. S., Gaddis, K. D., Cairns, D. M., Konganti, K., & Krutovsky, K. V. (2017). Landscape
625 genomic insights into the historic migration of mountain hemlock in response to holocene
626 climate change. *American Journal of Botany*, 104(3), 439–450. doi:10.3732/ajb.1600262

627 Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers.
628 *Bioinformatics*, 24(11), 1403–1405. doi:10.1093/bioinformatics/btn129

629 Kjeldsen, S. R., Zenger, K. R., Leigh, K., Ellis, W., Tobey, J., Phalen, D., ... Raadsma, H. W. (2016).
630 Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population
631 variability across its range. *Conservation Genetics*, 17(2), 337–353. doi:10.1007/s10592-015-
632 0784-3

633 Lee, K. M., Kivelä, S. M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., & Mutanen, M. (2018).
634 Information Dropout Patterns in Restriction Site Associated DNA Phylogenomics and a
635 Comparison with Multilocus Sanger Data in a Species-Rich Moth Genus. *Systematic Biology*,
636 syy029.

637 Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A.
638 (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA
639 sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17(2), 142–152.
640 doi:10.1111/1755-0998.12635

641 Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure:
642 Population Inference from RADseq Data. *Molecular Biology and Evolution*, 35(5), 1284–
643 1290. doi:10.1093/molbev/msy023

644 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015).
645 Restriction site-associated DNA sequencing, genotyping error estimation and de novo
646 assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15(1),
647 28–41. doi:10.1111/1755-0998.12291

648 Miller, M., Dunham, J., Amores, a, Cresko, W., & Johnson, E. (2007). Genotyping using restriction
649 site associated DNA (RAD) markers. *Genome Research*, 17, 240–248.
650 doi:10.1101/gr.5681207

651 Nunziata, S. O., Lance, S. L., Scott, D. E., Lemmon, E. M., & Weisrock, D. W. (2017). Genomic data
652 detect corresponding signatures of population size change on an ecological time scale in two
653 salamander species. *Molecular Ecology*, 26(4), 1060–1074. doi:10.1111/mec.13988

654 Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks.
655 *Methods in Ecology and Evolution*, 8(10), 1360–1373. doi:10.1111/2041-210X.12775

656 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest
657 RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and
658 non-model species. *PLoS ONE*, 7(5). doi:10.1371/journal.pone.0037135

659 Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014).
660 Demystifying the RAD fad. *Molecular Ecology*, 23(24), 5937–5942. doi:10.1111/mec.12965

661 Rochette, N. C., & Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using
662 Stacks. *Nature Protocols*, 12(12), 2640–2659. doi:10.1038/nprot.2017.123

663 Rodríguez-Ezpeleta, N., Bradbury, I. R., Mendibil, I., Álvarez, P., Cotano, U., & Irigoien, X. (2016).
664 Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers:
665 effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology
666 Resources*, 16(4), 991–1001. doi:10.1111/1755-0998.12518

667 Saenz-Agudelo, P., Dibattista, J. D., Piatek, M. J., Gaither, M. R., Harrison, H. B., Nanninga, G. B., &
668 Berumen, M. L. (2015). Seascape genetics along environmental gradients in the Arabian
669 Peninsula: insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, 24(24),
670 6241–6255. doi:10.1111/mec.13471

671 Settepani, V., Schou, M. F., Greve, M., Grinsted, L., Bechsgaard, J., & Bilde, T. (2017). Evolution of
672 sociality in spiders leads to depleted genomic diversity at both population and species levels.
673 *Molecular Ecology*, 26(16), 4197–4210. doi:10.1111/mec.14196

674 Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W.
675 (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream
676 population genetic inference. *Methods in Ecology and Evolution*, 8(8), 907–917.
677 doi:10.1111/2041-210X.12700

678 Sherpa, S., Rioux, D., Goindin, D., Fouque, F., François, O., & Després, L. (2018). At the Origin of a
679 Worldwide Invasion: Unraveling the Genetic Makeup of the Caribbean Bridgehead
680 Populations of the Dengue Vector *Aedes aegypti*. *Genome Biology and Evolution*, 10(1), 56–
681 71. doi:10.1093/gbe/evx267

682 Sherpa, S., Rioux, D., Pougnet-Lagarde, C., & Després, L. (2018). Genetic diversity and distribution
683 differ between long-established and recently introduced populations in the invasive mosquito
684 *Aedes albopictus*. *Infection, Genetics and Evolution*, 58(August), 145–156.
685 doi:10.1016/j.meegid.2017.12.018

686 Shortt, J. A., Card, D. C., Schield, D. R., Liu, Y., Zhong, B., Castoe, T. A., ... Pollock, D. D. (2017).
687 Whole Genome Amplification and Reduced-Representation Genome Sequencing of
688 *Schistosoma japonicum* Miracidia. *PLOS Neglected Tropical Diseases*, 11(1), e0005292.
689 doi:10.1371/journal.pntd.0005292

690 Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., &
691 Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model
692 organisms. *PeerJ*, 1, e203. doi:10.7717/peerj.203

693 Vargas, O. M., Ortiz, E. M., & Simpson, B. B. (2017). Conflicting phylogenomic signals reveal a
694 pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae:
695 *Diplostephium*). *New Phytologist*, 214(4), 1736–1750. doi:10.1111/nph.14530

696 Wang, S., Meyer, E., Mckay, J. K., & Matz, M. V. (2012). 2b-RAD: A simple and flexible method for
697 genome-wide genotyping. *Nature Methods*, 9(8), 808–810. doi:10.1038/nmeth.2023

698 Wang, Y., Cao, X., Zhao, Y., Fei, J., Hu, X., & Li, N. (2017). Optimized double-digest genotyping by
699 sequencing (ddGBS) method with highdensity SNP markers and high genotyping accuracy for
700 chickens. *PLoS ONE*, 12(6), 1–19. doi:10.1371/journal.pone.0179073

701 Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD
702 loci: an efficient method to filter paralogs and account for physical linkage. *Molecular*
703 *Ecology Resources*, 17(5), 955–965. doi:10.1111/1755-0998.12647

704 Yang, G. Q., Chen, Y. M., Wang, J. P., Guo, C., Zhao, L., Wang, X. Y., ... Guo, Z. H. (2016).
705 Development of a universal and simplified ddRAD library preparation approach for SNP
706 discovery and genotyping in angiosperm plants. *Plant Methods*, 12(1), 1–17.
707 doi:10.1186/s13007-016-0139-1

708

709

710 DATA ACCESSIBILITY

711 All data and scripts will be uploaded on Dryad and publicly accessible after acceptance.

712

713 AUTHOR CONTRIBUTIONS

714 Thibaut Capblancq, Frédéric Boyer, Tristan Cumer and Charles Pouchon conceived and
715 planned the study. TCa, CP, and Délphine Rioux carried out the lab experiment. TCa and TCu
716 run the majority of the analysis with the help of CP and FB. TCu, CP, FB, Glenn Yannic,
717 Aurélie Bonnin and TCa contributed to the interpretation of the results. TCa and TCu took the
718 lead in writing the manuscript and all authors provided critical feedback and helped shape the
719 analyses and manuscript.

720

721 FIGURES AND TABLES

722

723 **Figure 1:** Overview of the experimental design used in this study.

724

725 **Figure 2:** Impacts of initial DNA quantity (bottom) and number of PCR cycles (top) on
726 ddRADseq results. For each condition and each biological model (*Coenonympha* sp. or *Fagus*
727 *sylvatica*), boxplots present the number of recovered fragments, the mean coverage of these
728 fragments, the number of identified SNPs, the individual heterozygosity and the number of
729 private alleles in individuals. A log-transformation was performed on the results in order to
730 simplify the comparison of the two models.

731

732 **Figure 3:** Reproducibility of the experimental wet protocol. The PCA on the left shows the
733 inter-replicate genetic variability in comparison with inter-individual variability for *Fagus*
734 *sylvatica* individuals. Each three-replicate group is circled by an ellipse. Boxplots on the right
735 show the variation of the number of fragments, the mean coverage of the fragments, the
736 proportion of polymorphic fragments and the individual heterozygosity within the three
737 replicates of each individual.

738

739 **Figure 4:** Left: boxplot of the genotyping inconsistency rate within *Fagus sylvatica*
740 individuals. Right: proportion of loci found either in all the replicates (white), in only one or
741 two replicates (grey) or in none of the three replicates (black).

742

743 **Figure 5:** Impact of the *ustacks* thresholds **m** and **M** on the number of fragments, the mean
744 coverage of the fragments, the number of SNPs, and the mean individual heterozygosity.

745

746 **Figure 6:** Impact of *ustacks* thresholds **m** and **M** in F_{ST} between pairs of populations, genetic
747 differentiation and structure (PCA and sNMF results) for *Fagus sylvatica* and *Coenonympha*
748 sp. individuals. Only the results for **m** = 1, 4 and 15 and **M** = 1, 6 and 25 are shown (see Fig.
749 S1 to S6 for the complete results). Ellipses in the PCA distinguish the different populations or

750 species, sNMF results are shown for $K = 3$ which is the best number of clusters in almost all
751 cases according to the cross-entropy criterion.

752
753 **Figure 7:** Impact of *ustacks* thresholds \mathbf{m} and \mathbf{M} in demographic inferences obtained the
754 ABC procedure. The boxplot summarizes the values of the selected parameters in the 1000
755 simulations closest to the observed dataset. The parameters include divergence times (t1 and
756 t2) and effective population size (N1, N2, N3) for *Fagus sylvatica* and *Coenonympha*
757 populations, and hybridization contribution (ra) for *Coenonympha*.

758

759

760 SUPPLEMENTARY MATERIAL

761

762 **Figure S1:** Impact of the bioinformatic threshold \mathbf{m} (ranging from 1 to 15) on a genetic PCA
763 of *Fagus sylvatica* and *Coenonympha* sp. samples.

764

765 **Figure S2:** Impact of the bioinformatic threshold \mathbf{m} (ranging from 1 to 15) on genetic
766 clustering (sNMF method) of *Fagus sylvatica* and *Coenonympha* sp. samples.

767

768 **Figure S3:** Impact of the bioinformatic threshold \mathbf{m} (ranging from 1 to 15) on Euclidean
769 distances between PCA origin and loci scores in PC1 vs PC2 space, for *Fagus sylvatica* and
770 *Coenonympha* sp. samples.

771

772 **Figure S4:** Impact of the bioinformatic threshold \mathbf{M} (ranging from 1 to 25) on a genetic PCA
773 of *Fagus sylvatica* and *Coenonympha* sp. samples.

774

775 **Figure S5:** Impact of the bioinformatic threshold \mathbf{M} (ranging from 1 to 25) on genetic
776 clustering (sNMF method) of *Fagus sylvatica* and *Coenonympha* sp. samples.

777

778 **Figure S6:** Impact of the bioinformatic threshold \mathbf{M} (ranging from 1 to 25) on Euclidean
779 distances between PCA origin and loci scores in PC1 vs PC2 space, for *Fagus sylvatica* and
780 *Coenonympha* sp. samples.

781

782 **Figure S7:** Procrustes superimposition of PCA results for a range of \mathbf{M} and \mathbf{m} values and for
783 both the *Coenonympha* and *Fagus sylvatica* models. The two first axes of the PCA were kept
784 to do the Procrustes superimposition among the different \mathbf{M} and \mathbf{m} values. The distribution of
785 pairwise correlation coefficients between sets of coordinates resulting from the procrustes
786 superimposition are shown for each case.

787

788 **Figure S8:** Procrustes superimposition of sNMF results for a range of \mathbf{M} and \mathbf{m} values and
789 for both the *Coenonympha* and *Fagus sylvatica* models. The individual percentages of
790 assignation to the three clusters obtained with sNMF analyses at $K = 3$ were kept to do the
791 Procrustes superimposition among the different \mathbf{M} and \mathbf{m} values. The distribution of pairwise
792 correlation coefficients between sets of assignation scores resulting from the Procrustes
793 superimposition are shown for each case.

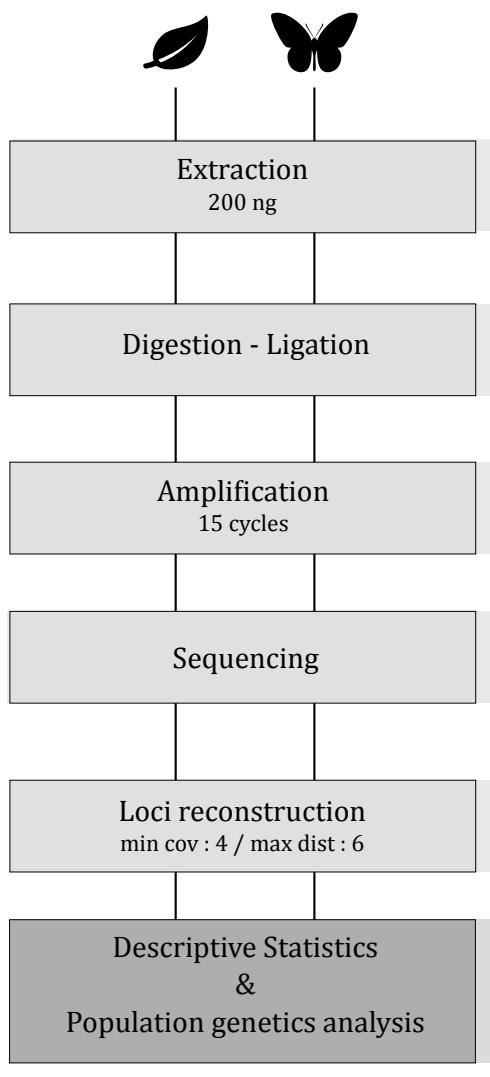
794

795 **Table S1:** Summary of the samples used in each part of this study.

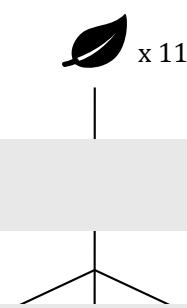
796

797 **Table S2:** Variation of parameters estimation during the ABC procedure for a range of **M** and
798 **m** values and for both the *Coenonympha* and *Fagus sylvatica* models. The minimum and
799 maximum estimation across all **M** or **m** values, and the percentage of variation between them
800 are given for each scenario parameter.

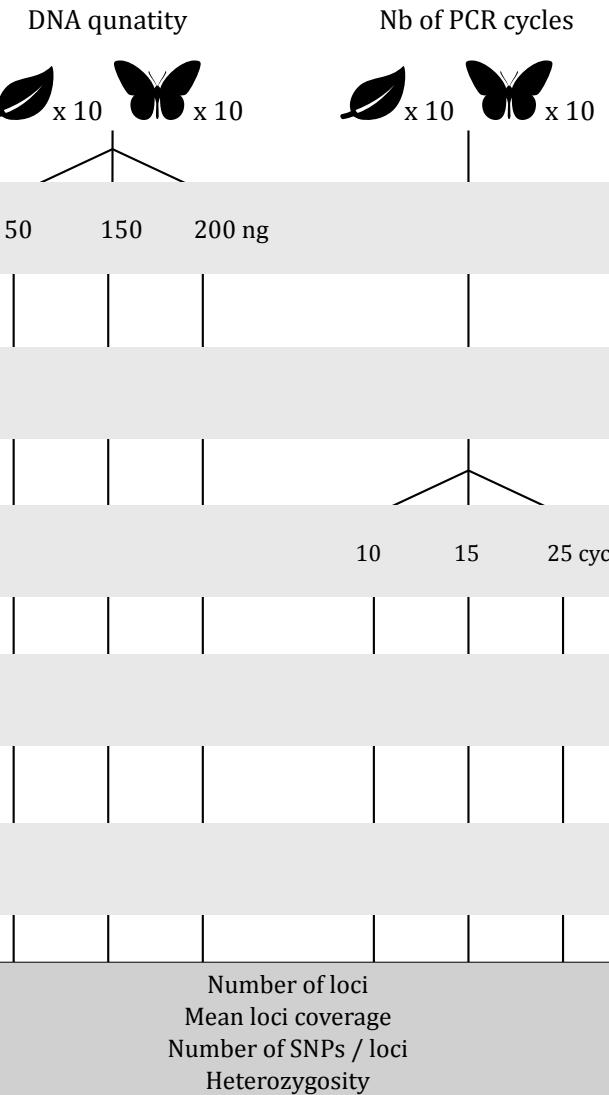
Standard ddRAD
Protocole



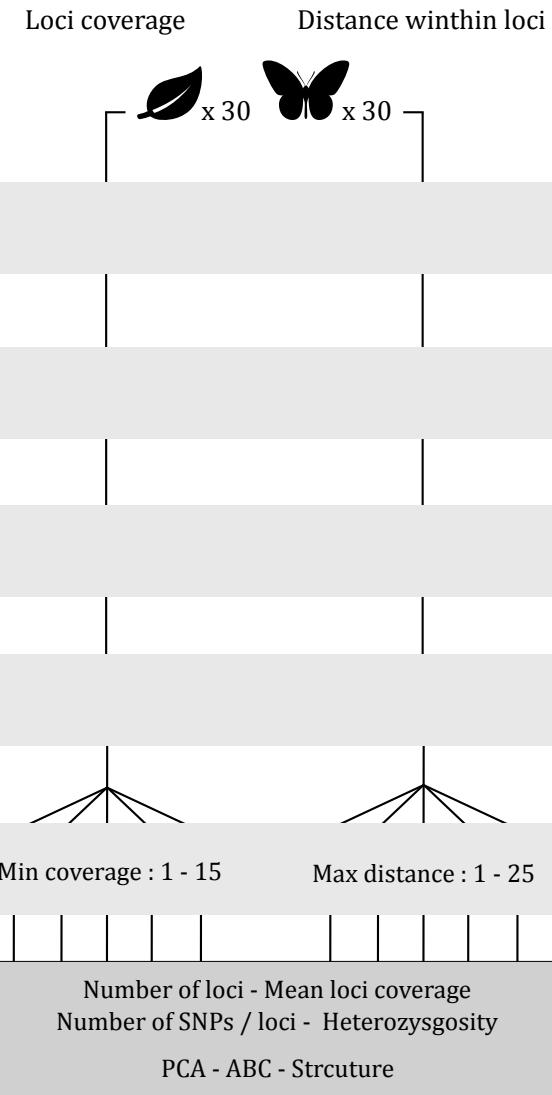
Replicability



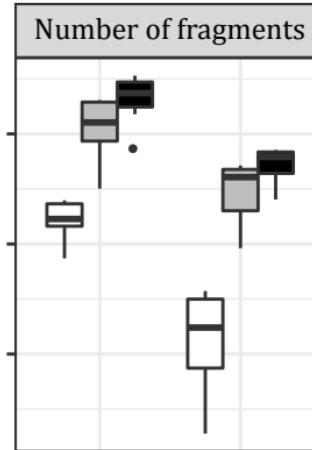
Wet protocol settings



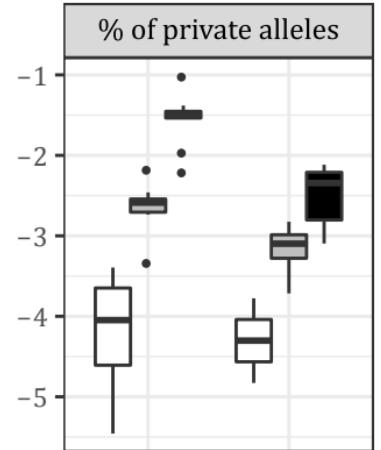
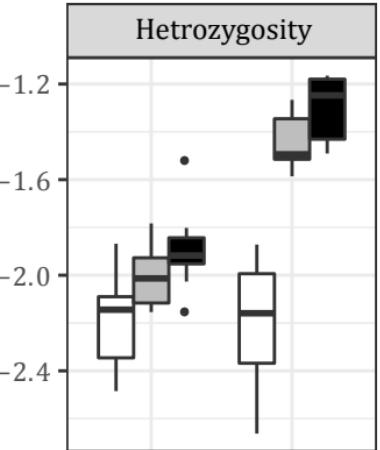
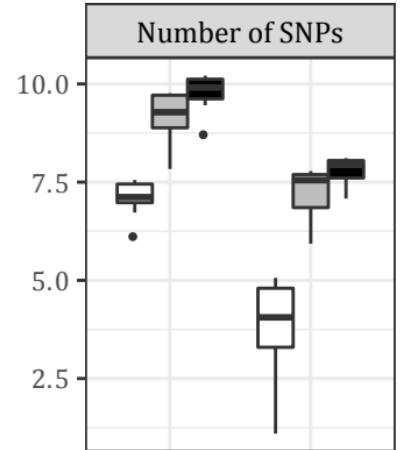
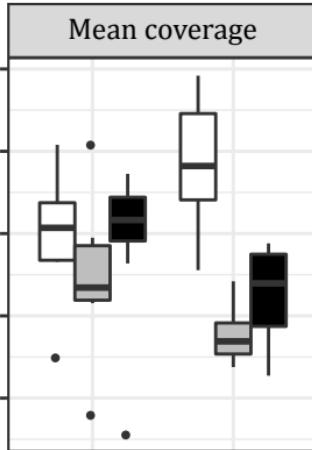
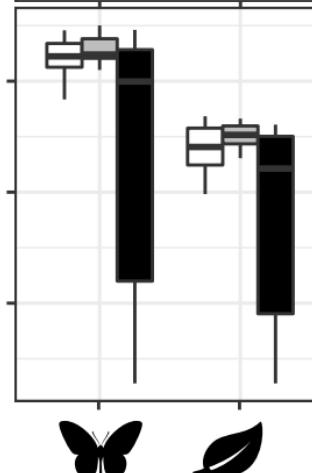
Dry protocol settings



Number of PCR cycles

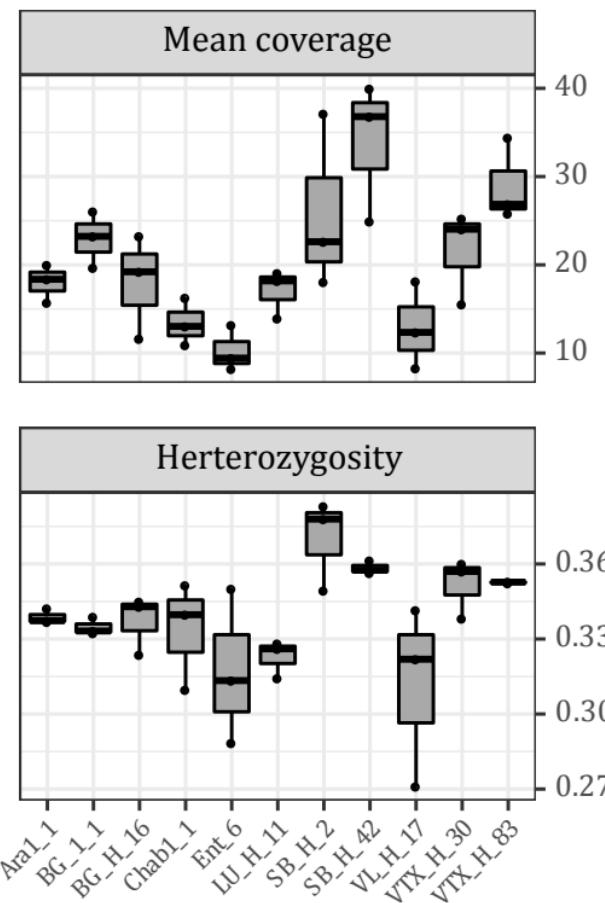
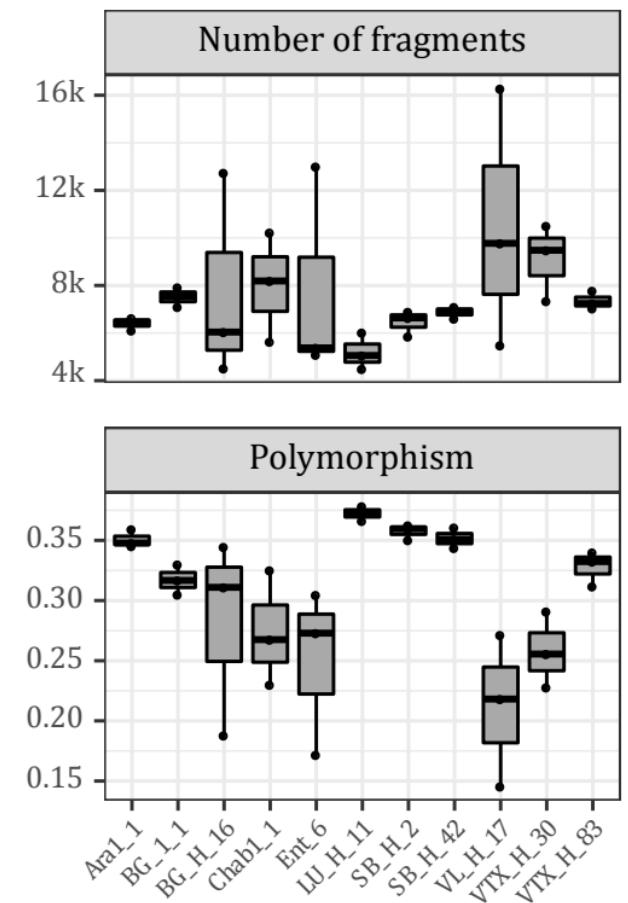
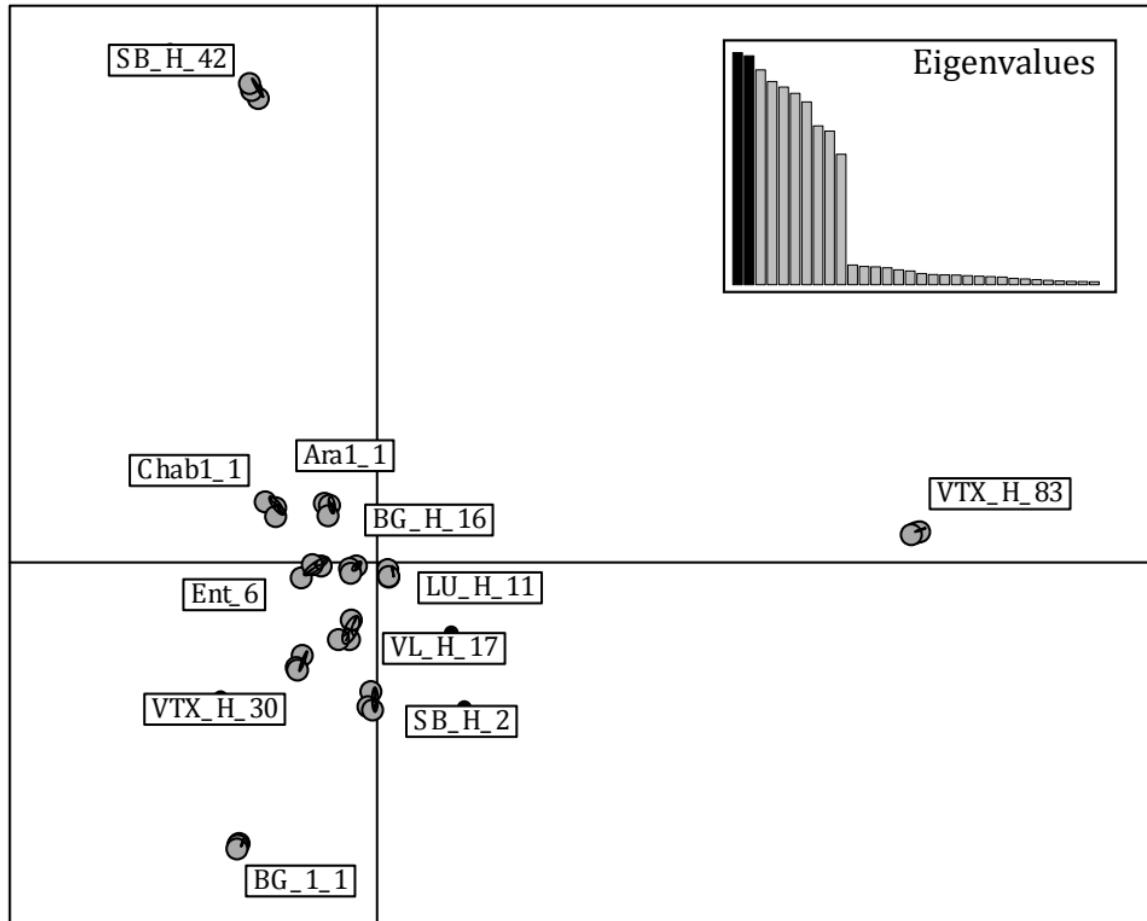


DNA quantity

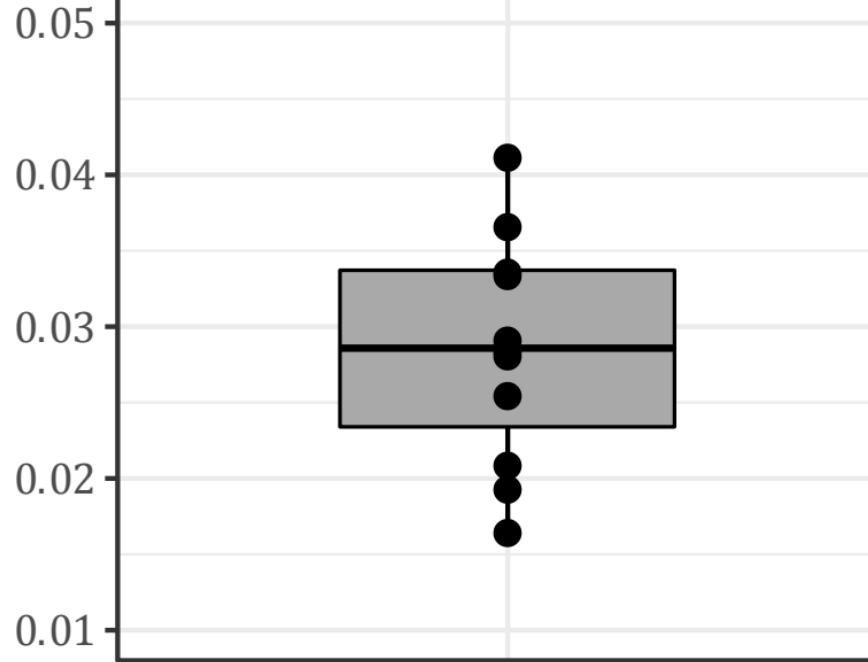


10
15
25

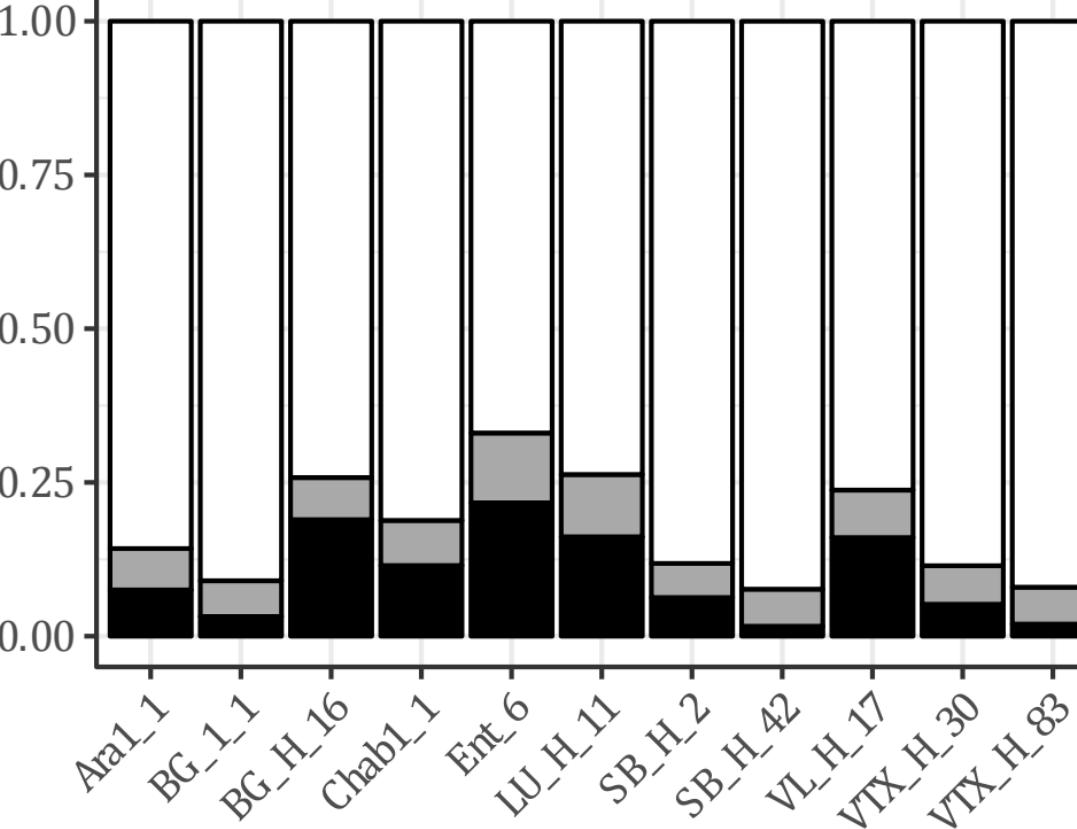
50 ng
150 ng
250 ng



Genotype mismatches



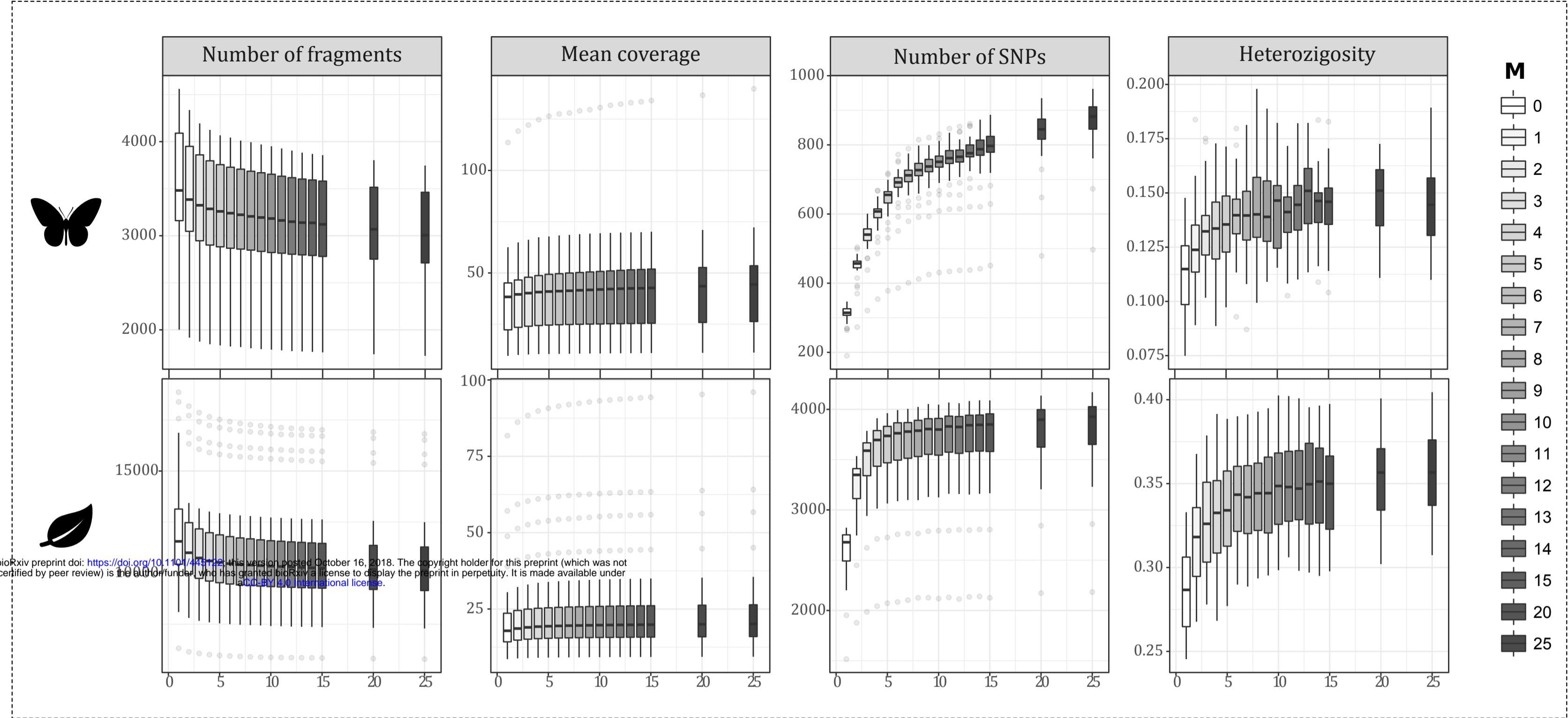
Missing data proportion among replicates



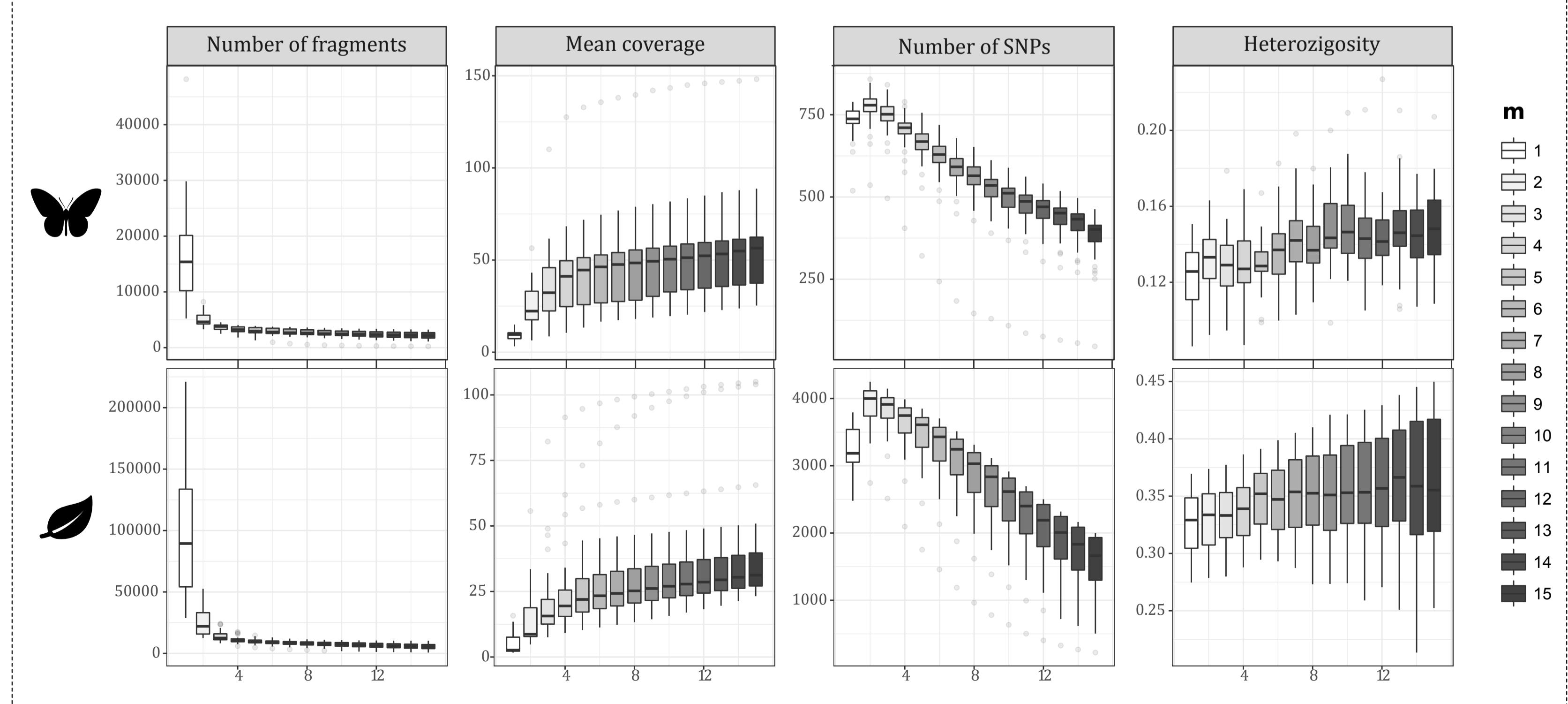
SNP present in:

- All (3) replicates
- 1 or 2 replicates
- 0 replicate

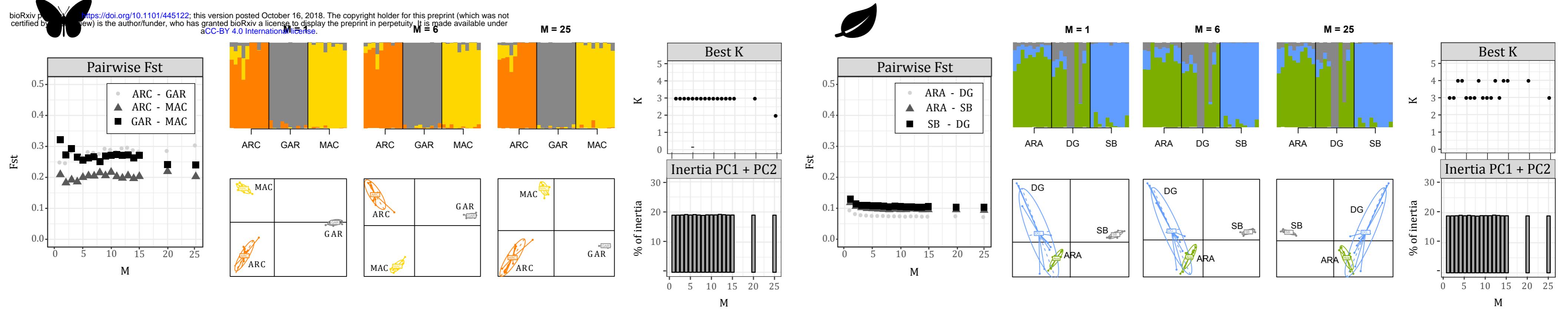
Similarity threshold (M)



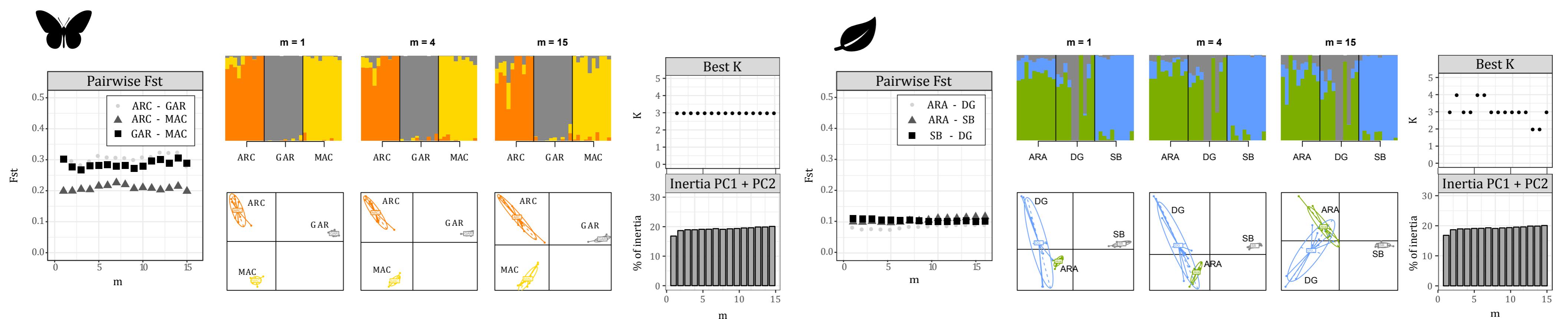
Minimum coverage threshold (m)



Similarity threshold (M)



Minimum coverage threshold (m)



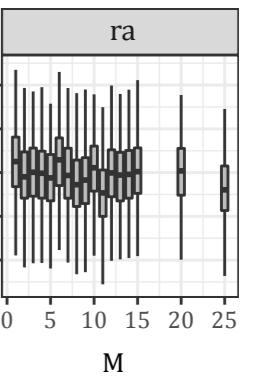
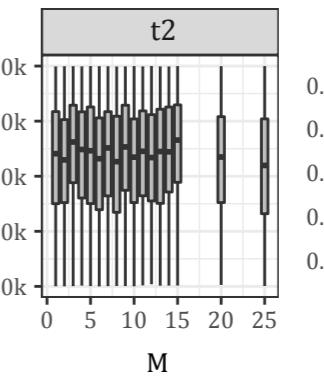
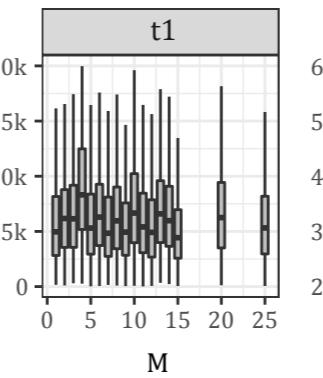
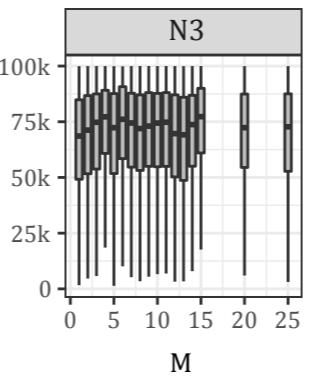
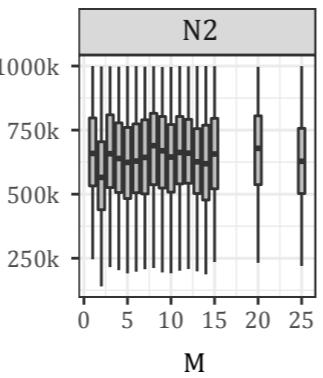
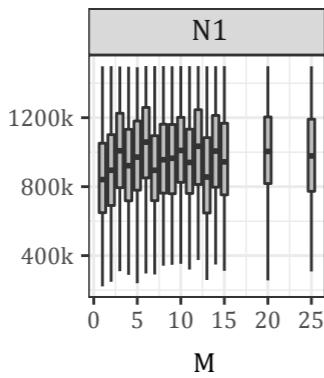
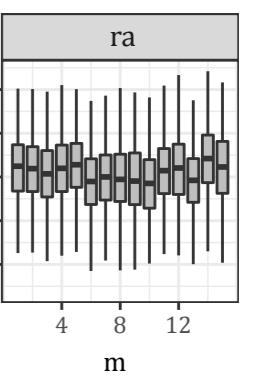
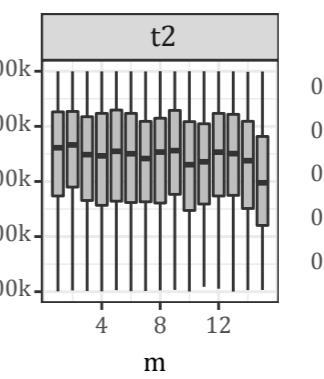
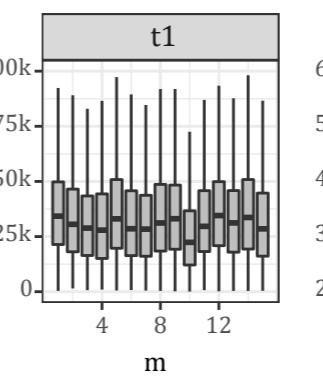
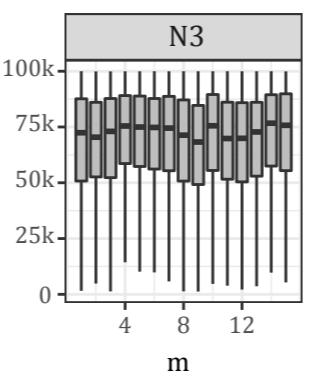
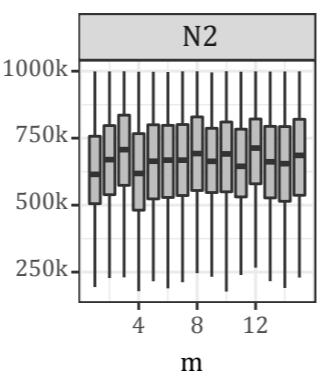
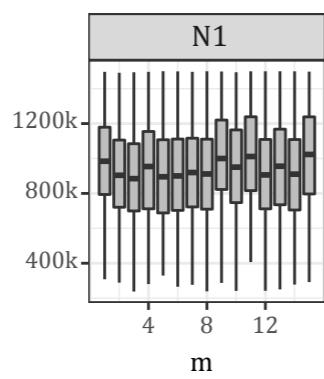


Time
+ t2
+ t1

N1

N3

N2



Time
+ t2
+ t1

N1

N3

N2

