RAD sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference

Alicia Mastretta-Yanes (AMY)[1], Nils Arrigo (N.Arrigo)[2], Nadir Alvarez (N.Alvarez)[2], Tove H. Jorgensen (THJ)[3], Daniel Piñero (DP)[4] and Brent C. Emerson (BCE)[1, 5]

Keywords: RADseq, *Stacks*, de novo assembly, optimization, error rate, replicates

Running title: Genotyping error and assembly of RAD data

[1] Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, 14 Norwich NR4 7TJ, UK

[2] Department of Ecology and Evolution, Biophore Building, University of Lausanne, 1015 Lausanne, Switzerland

[3] Department of Bioscience, Aarhus University, Aarhus, Denmark

[4] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado postal 70-275, Mexico, DF, 04510, Mexico

[5] Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), C/Astrofísico Francisco Sánchez 3, La Laguna, Tenerife, Canary Islands, 38206, Spain

Corresponding author: Alicia Mastretta-Yanes

e-mail: A.Yanes@uea.ac.uk

Abstract

Restriction site-associated DNA sequencing (RADseq) provides researchers with the ability to record genomic polymorphism across thousands of loci for non-model organisms, with the potential to revolutionise the field of molecular ecology. However, as with other genotyping methods, RADseq is prone to different sources of error that may have consequential effects for population genetic inferences, and that have, as yet, received limited attention in terms of the estimation and reporting of genotyping error rates. Here, we rely on sample replicates, under the expectation of genotype identity, to quantify genotyping error in the absence of a reference genome. We then use sample replicates to (1) optimize *de novo* assembly parameters of the program *Stacks,* by minimizing error and maximizing the retrieval of informative loci, and; (2) quantify error rates for loci, alleles and SNPs. As an empirical example we use a double digest RAD dataset of a non-model plant species, *Berberis alpina*, sampled from high altitude mountains of Mexico.

Introduction

Restriction site-associated DNA sequencing (RADseq) is a genotyping method created by Baird *et al.* (2008), it provides a reduced representation of a genome from which it is possible to analyse variation in DNA sequences. It has subsequently developed into a family of related approaches (reviewed by Davey *et al.* 2011) that allow subsampling of a genome at putatively homologous locations across many individuals to identify and type single nucleotide

51 polymorphisms (SNPs) in short DNA sequences. These approaches apply to non-
52 model organisms, regardless of genome size and previous genomic knowledge,
53 and thousands of loci for hundreds of individuals can potentially be sequenced
54 rapidly and at low cost. As a result, RADseq is increasingly being used across the
55 spectrum of evolutionary analysis, ranging from phylogenetic relationships
56 within a genus (e.g. Jones *et al.* 2013), to genome wide association studies to
57 identify of regions under selection (e.g. Hohenlohe *et al.* 2010; Parchman *et al.*
58 2012; Richards *et al.* 2013), through to ecological and conservation studies
59 (Narum *et al.* 2013).

60 Although the validity of RADseq data has been demonstrated, the method
61 suffers from genotyping errors, as all molecular markers do to some degree.
62 RADseq is prone to both technical and human sources of error (Table 1), similar
63 to some types of error that have been identified for other molecular markers
64 (e.g. Bonin *et al.* 2004) and inferences from next generation sequencing (Pool *et*
65 *al.* 2010; Gompert & Buerkle 2011). Wet lab procedures, parallel sequencing and
66 the properties of the species genome also contribute to error in several ways
67 (Table 1), leading to variance in: (a) the total reads per individual in a pool; (b)
68 the number of loci represented in each individual; (c) the read count of a locus;
69 and (d) the read counts of alternative alleles at polymorphic loci (Hohenlohe *et*
70 *al.* 2012). For example, differences in amplification success during the PCR step
71 may lead to variation of depth of coverage among loci and individuals,
72 potentially causing locus or allele dropouts (Supporting Information 1).

73 The consequences and statistical ways to account for error have been
74 widely discussed for several other molecular makers, from AFLPs and
75 microsatellites (Bonin *et al.* 2004; Pompanon *et al.* 2005; Price & Casler 2012) to

76 whole-genome sequence inferences (Pool *et al.* 2010; Gompert & Buerkle 2011).
77 Error can lead to incorrect biological conclusions, such as an artificial excess of
78 homozygotes (Taberlet *et al.* 1996), false departure from Hardy–Weinberg
79 equilibrium (Xu *et al.* 2002), overestimation of inbreeding (Gomes *et al.* 1999),
80 unreliable inferences about population structure (Miller *et al.* 2002), wrong
81 estimations of nucleotide diversity and incorrect inferences from the allele
82 frequency spectrum, such as demographic expansion from the confounding
83 influence of low frequency error-derived SNPs (Pool *et al.* 2010). These
84 potentially inaccurate inferences can be mitigated and accounted for if error
85 rates are reported (Bonin *et al.* 2004; Pompanon *et al.* 2005; Pool *et al.* 2010;
86 Davey *et al.* 2011; Price & Casler 2012) or incorporated into data analysis
87 (Gompert & Buerkle 2011; Gautier *et al.* 2013a). This is important for the
88 robustness of any study and meta-analysis. However, the quantification and
89 reporting of such errors has been largely overlooked by most recent RAD
90 studies.

91 In addition to errors introduced during wet lab and sequencing
92 procedures, errors arise during the bioinformatic processing of RADseq data
93 (Table 1). For instance, when RAD sequences are assembled into loci and alleles,
94 often using distance-based criteria, genotyping results will vary according to the
95 algorithm used (Davey *et al.* 2013) (note that we refer to a *locus* as a short DNA
96 sequence produced by clustering together unique RAD *alleles*; in turn, alleles
97 differ from each other by small number of *SNPs*). Several assembly and
98 genotyping tools have been recently released to handle RADseq data, such as
99 *RaPiD* (Willing *et al.* 2011), *RADtools* (Baxter *et al.* 2011), graph-based distance
100 clustering approaches (Peterson *et al.* 2012), *Stacks* (Catchen *et al.* 2011, 2013),

101 *Rainbow* (Chong *et al.* 2012) and *pyRAD* (Eaton 2013). It is to be expected that
102 different parameters and settings within a given tool will result in different
103 levels of assembly-related error. For instance, *Stacks* relies on a set of core
104 parameters (summarized in Table 2) to create sets of short-read sequences that
105 match (i.e. stacks) and then to curate and assemble them into genotyped loci per
106 individual. Catchen *et al.* (2013) have explored how variation in: 1) the minimum
107 number of raw reads required to form a stack (-*m*), 2) the mismatches allowed
108 between stacks (-*M*), 3) the maximum stacks per locus parameter (-
109 *max_locus_stacks*) and 4) modulating the assumed rate of sequencing errors
110 (using a bounded SNP calling model) affect the recovery of RAD loci within the
111 program. To do so, they ran the *de novo* pipeline using different parameter
112 values and compared the results to the expected loci based on a reference
113 genome. They concluded that the optimal values for these parameters will
114 depend upon the polymorphism of the genome being analysed, the amount of
115 sequencing error and the depth of sequencing performed. The authors
116 recommend to test a range of parameter values in order to optimize the analysis
117 of each RADseq dataset. However, their strategy to asses if true or erroneous loci
118 were assembled involved a reference genome, so therefore alternative criteria
119 are need for taxa where such reference is not available.

120 Here, we show that sample replicates can be used not only to estimate
121 error rates, but also to optimize the de novo assembly of RADseq data. This is
122 particularly important for low coverage datasets, facilitating the recovery of
123 more loci than could otherwise be reliably achieved. The central premise is that
124 DNA replicates derived from the same sample should have the same genotype.
125 Thus, after running any *de novo* assembly pipeline with different combinations of

126 parameters, one can evaluate which settings optimise for a high number of loci
127 and low differences between replicate pairs (Supporting Information 1). To
128 demonstrate this, we use double-digest RADseq (Parchman *et al.* 2012; Peterson
129 *et al.* 2012) data generated from populations of *Berberis alpina,* a non-model
130 plant species growing in high altitude mountains of Mexico. We focus on the
131 program *Stacks,* an efficient and well documented software that is increasingly
132 being used by molecular ecologists, but the same principle could be applied to
133 other assembly and genotyping tools for RADseq data.
134
135 **Methods**
136
137 *Study system and sampling*
138 The example species of our analyses is *Berberis alpina* (Zamudio 2009a; b), a
139 diploid plant with a probable genome size of between 0.5 to 1.83 Gbp based on
140 values of related species (Rounsaville and Ranney, 2010). *Berberis alpina*
141 inhabits the Transmexican Volcanic Belt (TMVB), a biodiversity hotspot for
142 temperate forest plant species (Rzedowski 1978; Myers *et al.* 2000; Calderón de
143 Rzedowski & Rzedowski 2005) where many plants are rare and restricted to one
144 or a few mountain tops (Koleff *et al.* 2008).
145 *Berberis alpina* was sampled in the TMVB and nearby areas of the Sierra
146 Madre Oriental (SMOr) during September-October 2010 and April-May 2011
147 (Sampling localities: doi:XXXXXXX/dryad.XXXXXX). Populations were found in
148 only seven locations that represent the known distribution of the species within
149 the TMVB (Fig. 1). In the mountain Cerro San Andrés, collected samples were
150 later identified as the closely related species *B. moranensis* based upon

151 morphological differences (Zamudio *pers. com.*), but were retained for analysis.
152 Fresh young leaves of 6-25 specimens per mountain (depending upon
153 population sizes) were collected and kept on ice while transported to the
154 molecular ecology laboratory within the Instituto de Ecología, Universidad
155 Nacional Autónoma de México (UNAM). Herbarium specimens were prepared
156 and deposited within the Herbario Nacional in Mexico City. *Berberis pallida* and
157 *B. trifolia* collected in the TMVB in October 2012 were used as outgroups. Half
158 the tissue of each sample was stored at -80ºC at UNAM, with the remainder dried
159 in silica gel for transport to the University of East Anglia (UEA), England where
160 samples were maintained at -20ºC until extraction. Samples were collected with
161 SEMARNAT permit No. SGPA/DGGFS/712/2896/10.
162
163 *Molecular methods*
164 DNA extractions of *Berberis alpina* and *B. moranensis* were performed at UEA
165 using the Qiagen DNeasy Plant Mini Kit (69106). DNA extractions of outgroup
166 samples were performed at UNAM using a CTAB method (Vázquez-Lobo, 1996)
167 with fresh tissue. Seventy-five specimens of *B. alpina* and *B. moranensis* (6-10 per
168 population) plus three samples of *B. trifolia* and three of *B. pallida* (outgroup
169 species) were used to prepare double digest RAD libraries (Parchman *et al.*
170 2012; Peterson *et al.* 2012) using the enzymes EcoRI-HF and MseI, T4 DNA
171 Ligase and Phusion Taq from New England Biolabs. Supporting Information 2
172 contains the complete lab protocol, including reaction mixes and quality details
173 of the sequencing. Individual DNA extracts were randomly divided into three
174 groups (BERL1, BERL2, BERL3), each corresponding to a libraries sequenced in
175 an independent lane. Each group was comprised of 27 *Berberis* sp. samples and 5

176 replicates for a total of 32 barcoded (sequence tagged) individuals. For each of
177 the groups, the 5 replicates consisted in 4 intra-library (group) replicates and 1
178 inter-library replicate. Replicates had the same DNA source but were treated and
179 barcoded independently. Replicates were chosen randomly but included at least
180 one replicate per outgroup and population. Within each group of 32 barcoded
181 samples, positions on PCR plates were randomly selected. The digestion, ligation
182 and PCR steps were performed in the same plate for the three groups. Samples of
183 the same group were then pooled together and the size selection for all three
184 groups was performed in the same gel. Each of the three groups were sequenced
185 with single-end reads (100bp long), in a separate lane of an Illumina HiSeq2000
186 with the service provided by the Lausanne Genomic Technologies Facility,
187 Switzerland.
188
189 *Basic quality filtering and general bioinformatics pipeline*
190 All raw reads were trimmed to 84bp because a considerable quality drop was
191 found in the subsequent positions of BERL3. Quality filtering and demultiplexing
192 were performed with a custom PerL script equivalent to the *Stacks* program
193 *process_rad_tags* (this custom script was developed prior to the release of the
194 update of *process_rad_tags* that allows for single-end double digested data).
195 Demultiplexed data was then *de novo* assembled and genotyped using *Stacks* v.
196 1.01 (Catchen *et al.* 2013), first with the default settings and all samples as a
197 exploratory run, and then with the settings and subset of samples described
198 below for the experiments: (i) *exploratory analysis of Stacks assembly key*
199 *parameters and SNP calling model using replicates,* and; (ii) *effect of using*
200 *different parameters on the output information content and on the detection of*

201 *genetic structuring*. Trimming, demultiplexing and *Stacks de novo* assembly were
202 performed using a computer cluster.
203
204 *Experiment 1. Exploratory analysis of Stacks assembly key parameters and SNP*
205 *calling model using replicates*
206 We explored the effect on error rates and amount of loci recovered of using
207 different *de novo* assembly conditions and SNP calling model settings within
208 *Stacks.* To do so we used eleven replicate pairs (those of the 15 replicated pairs
209 that yielded enough reads to have >50% of the mean number of loci in a first
210 exploratory analyses of the full dataset) to run *Stacks* multiple times with a range
211 of parameter values. For the assembly, the following key parameters were tested
212 with the values specified in parenthesis: the minimum number of raw reads
213 required to form a stack (–*m* 2 to 15), the number of mismatches between stacks
214 when processing an individual (-*M* 2 to 10), mismatches between loci when
215 building the catalog (-*n* 0 to 5) and maximum stacks per locus (–*max_locus_stacks*
216 2 to 6), varying only one parameter at a time, and fixing the rest to *m*=3, *M*=2,
217 *n*=0 and *max_locus_stacks*=3. The value of –*N* was always defined as *M*+2.
218 Regarding the SNP calling model, we compared the default (where error rate
219 varies freely) and the a bounded model, testing also several values (0.5, 0.25,
220 0.15, 0.1, 0.05 and 0.0056) for the upper bound (*sequencing error upper bound,* a
221 parameter used by the bounded model, Catchen *et al.* 2013). Note that values
222 >0.15 represent high and unrealistic levels of sequencing error. The minimum
223 was set to 0.0056 because this was the PhiX estimate of sequencing error for
224 BERL3 (which had the largest sequencing error of all lanes) at cycle 100 (instead
225 of 75 to compensate for a slight quality drop at 80-84 bp). As for the rest of the

226 settings, three different minimum coverage values (*m*=3, 4 and 10) were
227 explored and the remaining of the parameters were set to the values considered
228 to perform better in the assembly exploratory analyses (*M*=2, *N*=4, *n*=3,
229 *max_locus_stacks*=3, see results).

230 Outputs were then processed as detailed in *General processing of* Stacks
231 *outputs* (see below) and the results were analysed in R v. 2.15.1 (R. Core Team
232 2012) to estimate: (1) the number of output loci and SNPs; (2) locus, allele and
233 SNP error rates (as defined in *Error rates,* see below), and; (3) Euclidean distance
234 matrices among individuals to build neighbour joining (NJ) dendrograms (as a
235 visual way to examine if replicate pairs cluster together, as would be expected).

236

237 *Experiment 2. The effect of parameter values on output information content and*
238 *the detection of genetic structuring*

239 To examine the effect of using different *Stacks* settings on the full dataset (78
240 specimens) we ran *Stacks* with four *de novo* parameter profiles, namely: default,
241 optimal, near optimal and high coverage. The default values were *m*=3, *M*=2,
242 *N*=4, *n*=0, *max_locus_stacks*=3 and the default SNP calling model. The other
243 parameter profiles were given values that provided the highest number of loci
244 and SNPs at the lowest error rates in the exploratory analysis using the replicate
245 pairs (*M*=2, *N*=4, *n*=3, *max_locus_stacks*=3 and a SNP calling model with an upper
246 bound of 0.05, see results) but increasing the minimal coverage: *m*=3 (optimal),
247 *m*=4 (near optimal) and *m*=10 (high coverage). Note that we define optimal as
248 the profile that performed better in the assembly exploratory analyses for our
249 data, and thus optimal parameter values will vary for other RADseq data. Each
250 parameter profile was used to run *Stacks* with all samples of *B. alpina, B.*

251 *moranensis* populations (75), the three samples of the closest outgroup (*B.*
252 *trifolia*) and the replicates (14).

253 Outputs were then processed as detailed in *General processing of* Stacks
254 *outputs,* and locus, allele and SNP error rates (as defined in *Error rates*) were
255 estimated for each profile. Afterwards, only one sample of each replicate pair
256 was retained with the non-replicated samples to: (1) estimate an Euclidean
257 distance matrix based on the SNPs; (2) perform a principal coordinates analysis
258 (PCoA); (3) normalize the distance matrix and extract the distances between
259 individuals of the same population; and (4) run the population program of *Stacks*
260 to estimate F_{ST} between population pairs using only samples from *B. alpina* and
261 *B. moranensis* and the loci present in 80% of the samples.

262

263 *General processing of* Stacks *outputs*

264 *Stacks* outputs of experiments 1 and 2 were imported to a desktop computer,
265 where data was visualized and exported as allele and coverage matrices. These
266 matrices were then analysed with custom R scripts to: (1) estimate the number
267 of reads and coverage per locus, per individual and per lane; (2) filter data to
268 keep only those samples having more than 50% of the mean number of loci per
269 sample, and only those loci present in at least 80% of the barcoded samples; and
270 (3) output as plink format the loci and samples that passed the previous filter.
271 Further analyses were performed as described before for each experiment.

272

273

274 *Error rates*

275 Replicate pairs were used to estimate three error rates using custom R scripts
276 (github.com/AliciaMstt/RAD-error-rates): (1) locus error rates, corresponding
277 to missing data at the locus level and measured as the number of loci present in
278 only one of the samples of a replicate pair, divided by the total number of loci
279 found; (2) allele error rate, calculated as the number of allele mismatches
280 between replicate pairs, divided by the number of loci being compared; and (3)
281 SNP error rate, measured as the proportion of SNP mismatches between
282 replicate pairs.

283 Note that we refer to a locus as a short DNA sequence produced by
284 clustering together unique RAD alleles; in turn, alleles differ from each other by a
285 small number of SNPs. We define a missing locus as that absent in at least one
286 sample of a replicate pair, but present in any other sample of the dataset. In
287 addition to the locus error rate, we further examined the distribution of missing
288 data within replicate pairs by estimating: (1) the number of missing loci per
289 replicate pair; (2) the proportion of missing loci (number of missing loci per
290 replicate pair over the total); and (3) the percentage of missing loci of a given
291 sample replicate that were not the same missing loci in the other replicate
292 (proportion of missing loci different within a replicate pair). Supporting
293 Information 1 provides a diagram detailing the differences between replicates
294 estimated here.

295 The R scripts utilized here used the packages: adegenet_1.3-7 (Jombart
296 2008), ape_3.0-8 (Paradis *et al.* 2004), gtools_2.7.1 (Warnes *et al*. 2013),
297 multicore_0.1-7 (Urbanek 2011) and stringr_0.6.2 (Wickham 2012). All custom R
298 and Perl scripts and *Stacks* jobs used in this study are available in Dryad
299 repository. doi:XXXXXXX/dryad.XXXXXX

300

301 **Results and discussion**

302

303 *RAD sequencing output and coverage*

304 An average of 1,632,914 reads per sample were obtained after demultiplexing,
305 with no major differences between lanes or populations. Full details of
306 sequencing output are provided in Supporting Information 2. In a first
307 exploratory analysis (using *Stacks* default settings and post-filtering the data
308 with the >50% and 80% criteria described in the basic quality exploration
309 section), fifteen out of the 96 samples had too few reads and therefore did not
310 pass the filter of sharing >50% of the mean number of loci with the rest of the
311 samples. Among these were the interlibrary replicate sequenced in lane BERL1
312 (PeB01_ir1) and one sample of a replicate pair (MaB21). Also, a strong lane effect
313 associated with lane BERL3 was found. Samples sequenced from this lane were
314 found to cluster together within a NJ dendogram, while the samples from BERL1
315 and BERL2 were intermixed, clustering typically by geography. The source of the
316 lane effect was determined to be a single SNP found in position 70 of many reads,
317 which was then identified as an artefact by the sequencing facility. Deleting
318 position 70 in all the demultiplexed reads removed the lane effect.

319 In general, mean coverage per locus was low (increasing the min.
320 coverage *-m* from 3 to 10 produced a substantially lower number of loci, Fig. 2
321 and Table 3). Coverage is the main filter to distinguish sequencing error from
322 real variation. However, if coverage is generally low, a high filter threshold for
323 coverage can lead to allele dropout, which in turn becomes genotyping error.
324 Assembling and genotyping a low coverage RAD-seq dataset like that of *Berberis*

325 is thus challenging, and may lead researchers to keep only a small fraction of the
326 loci and alleles that have high coverage for all samples which, as shown below,
327 may not be the most reliable data. Many RADseq datasets may have low
328 coverage, particularly for species for which exact genome size is unknown, or if a
329 study design aims for more individuals or loci to increase the accuracy of
330 population genetic parameters (Buerkle & Gompert 2013).
331
332 *Exploratory analysis of Stacks assembly parameters and SNP calling model using*
333 *replicates*
334 We ran *Stacks* with 11 replicate pairs (22 samples). After filtering the output so
335 that all samples shared >50% of the mean number of loci per sample, most
336 assembly parameter profiles recovered 19-20 samples and only runs with $n \geq 3$
337 recovered all 22. The samples that were not recovered for some of the parameter
338 profiles explored for *Stacks* either had a small number of reads relative to other
339 samples, or belonged to the more distant outgroup (*B. pallida, OutBs*). These
340 samples did not share >50% of the mean number of loci with the remainder of
341 the dataset and thus were excluded by the filtering step. When both samples of a
342 replicate pair passed filtering, they clustered together in the NJ dendogram
343 (Supporting Information 3), with two exceptions: (1) the interlibrary replicate
344 (PeB01) pair clustered together in only 18 of 36 parameter profiles tested, and in
345 the remaining analyses it formed a paraphyletic group with other samples from
346 the same population, and (2) one replicate pair (AjB21) did not cluster together
347 in 9 occasions, clustering instead with other populations. Why only these two
348 samples were affected does not seem to be related to their number of reads, so
349 maybe it was due to contamination (AjB21) or a slight lane effect (PeB01).

350 Importantly, the parameter profiles at which incorrect clustering occurred were
351 high values for minimal coverage (-*m*) and the number of mismatches between
352 loci when processing an individual (-*M*). This suggests that setting –*m* too high
353 can lead to locus/allele dropout large enough to cause incorrect inferences of
354 individual differentiation. Why setting -*M* to high values causes differences
355 between replicates is less evident, but it is likely related to the formation of
356 nonsensical loci (*Catchen* et al. 2013). That not all replicates pairs clustered
357 together indicates that differentiation among individuals should be interpreted
358 with care, and that this occurred only with some parameter values highlights
359 that assembly settings can be tuned to minimize differences between replicates.

360 Across all the explored parameter profiles, the number of loci recovered
361 ranged from ~200 to >5,000 (Fig. 2a), the number of SNPs ranged from ~200 to
362 >8,000 (Fig. 2b), and the total number of missing loci ranged from 50 to >500
363 (Fig. 3a). In general the parameters that control the minimal coverage (-*m*) and
364 number of mismatches allowed between loci when building the catalog of loci (-
365 *n*) contributed most to the variance of the amount of data (Fig. 2a) and missing
366 loci produced (Fig. 3a and 3b).

367 A key source of variation between replicate pairs is that the identity of
368 most (>70%) of the missing loci in a given replicate is not the same in its
369 corresponding pair (Fig. 3c), which leads to a locus error rate typically >10%
370 (Fig. 3d) regardless of the parameter values used. As these differences are
371 between samples from the same DNA source that were processed together, it
372 seems that stochastic PCR and sequencing sampling events are the main source
373 of heterogeneous coverage among loci.

374 Allele error rates ranged from ~5% to >15%, depending on the

375 parameter profile used to execute *Stacks* (Fig 4a). Allele mismatches between
376 replicates can be caused by allelic dropout, or by the acceptance of error-based
377 variation during assembly. Similarly, the SNP error rate ranged from ~2% to
378 12% (Fig. 4b). Again, the most important differences were related to changes in
379 –*m* and –*n*. Increased values of –*m* decreased the allele error rate, but not to a
380 level below 10%, and at a cost of yielding fewer loci. Similarly, the SNP error rate
381 was reduced from ~7% at *n*=0 to ~2.5% at *n*=3.

382 The parameter –*m* controls the total number of raw reads per sample to
383 create a stack, so the higher it is set, the lower is the probability that there will be
384 enough reads per locus to assemble an allele. Setting -*m* to a higher value could
385 also result in genuine alleles being considered as secondary reads (reads that are
386 not used to assemble reference alleles and that are set aside), and as a
387 consequence treated as sequencing errors (see *Stacks* documentation for further
388 details). For the *Berberis* dataset, the danger of labelling stacks with concurrent
389 sequencing errors is reduced by the fact that the data was run in three different
390 lanes with a randomized sample design.

391 The parameter –*n* modulates the maximum number of mismatches
392 allowed between loci when building the catalog (list of all loci and alleles in the
393 population created by *Stacks*). So, if *n*=0, there would be loci represented
394 independently across individuals that are in reality homologous alleles of the
395 same locus. When *n*>0, *Stacks* uses the consensus sequence from each locus to
396 attempt to merge loci (*Stacks* documentation). Increasing -*n* may have resulted
397 in significant error reduction for the *Berberis* data set because replicates
398 involved samples from geographically isolated populations and outgroups,
399 conditions that would be expected to result in loci that exhibit fixed differences

400 among populations. By merging fixed alleles into a single locus the allele error
401 rate decreased, probably because the chances of assembling the same true alleles
402 in both replicates increased. A potential negative consequence of a high value of
403 –*n* is the creation of erroneous loci, which can be assembled for reasons such as
404 the acceptance of short sequencing/PCR error-based stacks, and the clustering of
405 repetitive sequence regions, (Catchen et al. 2013). However, the locus error rate
406 did not vary significantly when –*n* varied from 0 to 5 (Fig. 3d), so it seems that
407 the erroneous loci that were potentially created have less weight than the error
408 reduction benefits gained from increasing the value of –*n*.

409 Regarding the SNP calling model, reducing the upper bound increases the
410 chance of calling true heterozygous loci instead of wrongly labelling them as
411 homozygous loci with sequencing error (Catchen et al. 2013). For the *Berberis*
412 data, differences in genotyping errors were found only after decreasing the
413 upper bound down to 0.0056 in the runs of *m*=3 and *m*=4 (Supporting
414 Information 4), such that the allele error rate decreased from >5% down to
415 approximately 2.5%. However, this increased the SNP error rate from ~2.5 % to
416 7%. Thus, for the *Berberis* dataset, it seems better to leave the upper bound of
417 the SNP calling model to a relatively high value. Finally, there were no
418 differences in loci error rate between the SNP calling models (Supporting
419 Information 4b).

420 In summary, for the *Berberis* dataset the parameter values that seemed to
421 both increase the number of loci and reduce the SNP and allele error rates were
422 *m*=3, *M*=2, *N*=4, *n*=3, *max_locus_stacks*=3 and a SNP calling model with an upper
423 bound of 0.05.

424

425 *Effect of using different parameters on the output information content and on*
426 *detection of genetic structuring*
427 The four combinations of *Stacks* settings (optimal, near optimal, default and high
428 coverage) used to process the full dataset differed in the number of recovered
429 loci, number of SNPs and error rates (Table 3). Among the four combinations, the
430 optimal profile generated the highest number of RAD-loci (6,292) and SNPs
431 (11,057) and had the lowest allele (5.9%) and SNP (2.4%) error rates, although
432 the locus error rate (17%) was high (Table 3). The smallest locus error rate was
433 found with the high coverage setting (8.8%, Table 3), but this parameter profile
434 produced the highest allele and SNP error rates (8.7% and 5.7% respectively)
435 and the smallest number of loci and SNPs (292 and 502, Table 3).
436 The SNP error rate is important for population genetic and
437 phylogeographic analyses. If SNP error is high within a given dataset, less of the
438 observed genetic variation would be explained by the geographic origin of the
439 samples, and noise will contribute more to the genetic distance between
440 individuals. From a drift-mutation-migration equilibrium perspective,
441 individuals sampled from the same geographic region should be expected to be
442 genetically more similar in datasets with smaller SNP error rates. As a simple
443 way to test this, the genetic distances between individuals of the same sampling
444 locality were compared among the four combinations of *Stacks* settings explored
445 here. As expected, the data with the smallest SNP error rate (optimal profile)
446 systematically produced shorter genetic distances between individuals of the
447 same sampling locality, if compared to the other three parameter profiles (Fig.
448 5).
449 To be of relevance for population genetics and phylogeographic analyses,

450 molecular markers must not only have minimal noise, but also provide
451 meaningful variation (Zhang & Hare 2012; Price & Casler 2012). The *Berberis*
452 data produced by the optimal parameter profile resulted in substantial genetic
453 variation, 80% of which was explained by the first two axes of the PCoA, which
454 clustered samples by population origin (Supporting Information 5). The same
455 axes of the PCoAs produced with the data from the high coverage and default
456 parameter profiles explained only 47% and 57%, respectively (Table 3,
457 Supporting Information 5). Also, the mean value of the pairwise F_{ST} matrix was
458 higher for the data produced by the optimal parameter profile (0.19) compared
459 to the default (0.07) or any of the other *Stacks* settings examined (Table 3). This
460 is congruent with simulations that show that low coverage datasets with a larger
461 sample of sites in the genome would yield more accurate and precise population
462 genetics parameter estimates (Buerkle & Gompert 2013).

463 Assembling *Berberis* data *de novo*, with the optimal parameter profile,
464 maximised the number of informative SNPs and minimized the error that
465 increases intra-population variation (Fig. 5), adding further support for optimal
466 assembly parameter values (test a range and choose those that both increase the
467 number of output loci and reduce the SNP and allele error rates) yielding reliable
468 data. We advice researchers to include replicates and follow the same principles
469 of analysis presented here, regardless of the *de novo* assembly tool. In the case of
470 RADseq datasets already produced without DNA replicates, we recommend the
471 exploration of a range of parameter values to maximize the amount of SNPs
472 recovered and minimize the genetic dissimilarity between individuals from the
473 same sampling locality. This recommendation should be used as a starting point
474 and with care, as locality may be the wrong metric to use when minimizing

475 genetic dissimilarity in some cases (e.g. hybrid zones, breeding areas).

476

477 *Error rate implications and recommendations for RADseq analyses*

478 Next-generation sequencing methods applied to population genetic inference
479 need to account not only for sequencing error, but also for assembly error and
480 missing data (Pool *et al.* 2010; Davey *et al.* 2011). Including DNA replicates in the
481 preparation of RADseq libraries (see below for some recommendations)
482 improves the characterisation of error derived from different sources (Table 1)
483 and provides the ability to partition error into locus, allele and SNP rates. High
484 locus error rates, such as the >10% error for all combinations of parameters
485 evaluated for *B. alpina* (Fig. 3d, Table 3) can be accommodated as missing data
486 and mitigated by appropriate statistical corrections (Pool et al. 2010; Davey *et al.*
487 2011), as is possible to do with principle components analysis, principal
488 coordinates analysis and STRUCTURE (Pritchard et al. 2000). However, incorrect
489 SNP calling and allelic dropout are more problematic if data analyses are to be
490 performed under the assumption that genotypes are known with complete
491 certainty. Allele error can affect both allele frequency estimates and the accurate
492 discrimination of different genotypes (Bonin et al. 2004), and SNP error will
493 inflate nucleotide diversity and skew the SNP Frequency Spectrum toward rare
494 SNPs (Johnson & Slatkin 2008; Pool et al. 2010), affecting the meaningful
495 biological interpretation of the data. Excitingly, as population genomics and next-
496 generation sequencing technology and analytical tools further develop, genotype
497 uncertainty could be incorporated into data analysis itself (Buerkle & Gompert
498 2013), using Bayesian hierarchical models and genotype probabilities rather
499 than genotypes *per se* (Gompert & Buerkle 2011; Buerkle & Gompert 2013;

500 Gautier et al. 2013a). But for this to account not only for sequencing error, but
501 the full range of sources that may affect RADseq (Table 1), DNA replicates should
502 be included to allow for error rate estimation.

503 The estimation of genotyping error is affected by sample size, as
504 exemplified by the variance of error rate estimation across replicates for the
505 *Berberis* data (Fig. 3, Fig. 4 and Table 3). Including multiple replicates is thus
506 useful, but there is no minimum number for RADseq studies. For *B. alpina,* we
507 aimed to replicate ~15% of the samples, but as some samples failed we achieved
508 11%. The number of replicates for a given study will be a function of the final use
509 of the data, the targeted coverage depth, and the precision in error rate
510 estimation needed. Replicates should be randomly chosen while also broadly
511 representing important data features such as geography and taxonomy. In the
512 case of geographic sampling, we would recommend the inclusion of at least one
513 sample replicate per sampling location.

514 Regarding recommendations to reduce error rate, as has been suggested
515 for other marker systems (e.g. Bonin *et al.* 2004, Pompanon *et al.* 2005), good lab
516 practice and experimental design will help to minimize error rate. In the case of
517 RADseq data, locus and allele recovery depend on the level of coverage of reads
518 for each allele, locus and individual, but as shown here large numbers of markers
519 can be recovered reliably from low coverage datasets (down to ~7x, as the mean
520 for BERL1 here). Thus, given budget limitations, coverage depth may be
521 sacrificed for increased sampling for both the number of individuals, and sites in
522 the genome, both of which can provide better estimates of population genetic
523 parameters (Buerkle & Gompert 2013). However, studies that require very low
524 error rates, or that are rightly recovering paralogs, should consider increasing

525 the coverage up to 60x (Davey *et al.* 2011). We have shown that error rates can
526 also be reduced at the bioinformatics stage by assembling data and calling SNPs
527 with an optimal combination of parameters. However, in a similar way to how
528 the number of variable loci needed for a RADseq study depends upon the
529 biological question (Catchen et al. 2011; Hohenlohe et al. 2012; Peterson et al.
530 2012), the acceptable error rate will also be study specific. In the case of *Berberis*
531 *alpina*, the quantification of RAD allele and SNP error rates found for the
532 optimized *Stacks* settings (5.9% and 2.4%, respectively, Table 3) provides
533 reassurance that the geographic structuring of genetic variation is biologically
534 meaningful, but would warn against more fine-scale analyses of individual
535 relatedness if differences between individuals fell within the error rate
536 threshold.

537 We have demonstrated that the use of sample replicates is an important
538 procedure for both the estimation of error rates and for the optimization of *de*
539 *novo* assemblage and genotyping parameters. We focused on *Stacks,* but the
540 same principle can be applied to the other assembly tools for RADseq data. Thus,
541 we suggest that the use of replicates for RADseq studies should be encouraged,
542 and we consider it pertinent to extend Crawford *et al.*'s (2012) call for more
543 transparent reporting of genotyping error to RADseq data.

544

545 **Acknowledgements**

546 We thank Subject Editor Alex Buerkle, Brant Faircloth and one anonymous
547 referee for their constructive comments on an earlier version of the manuscript;
548 Zamudio for help identifying *Berberis* specimens and taxonomical discussion; L.
549 Figueroa, C. Berney, T. Wyss and A. Brelsford for labwork assistance; O. Trejo for

assistance with sampling permits and and SMG, JRPPK, JJRL, AOM, ROF, SSF, RAF, TSA, JAA, FDRG, FQB y MJLF for help during fieldwork. Part of the analyses were carried out on the High Performance Computing (HPC) Cluster supported by the Research and Specialist Computing Support service at UEA. This work has been generously supported by CONACYT doctorate scholarship to AMY (213538), by CONACYT grant to DP (178245) and by a SSE Rosemary Grant Student Research Award to AMY.

References

Baird NA, Etter PD, Atwood TS *et al.* (2008a) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE*, **6**, e19315.

Bonin A, Bellemain E, Bronken Eidesen P *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Calderón de Rzedowski G, Rzedowski J (2005) *Flora Fanerogámica del Valle de México*. Instituto de Ecología A.C. y Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Pátzcuaro, Michoacán, México.

573 Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks:
574 Building and Genotyping Loci De Novo From Short-Read Sequences. *G3:*
575 *Genes, Genomes, Genetics*, **1**, 171–182.
576 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an
577 analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–
578 3140.
579 Chong Z, Ruan J, Wu C-I (2012) Rainbow: an integrated tool for efficient
580 clustering and assembling RAD-seq reads. *Bioinformatics*, **28**, 2732–2737.
581 Crawford LA, Koscinski D, Keyghobadi N (2012) A call for more transparent
582 reporting of error rates: the quality of AFLP data in ecological and
583 evolutionary research. *Molecular Ecology*, **21**, 5911-5917.
584 Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD
585 Sequencing data: implications for genotyping. *Molecular Ecology*, **22**,
586 3151–3164.
587 Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker
588 discovery and genotyping using next-generation sequencing. *Nature*
589 *Reviews Genetics*, 12, 499–510.
590 Eaton DAR (2013) PyRAD: assembly of de novo RADseq loci for phylogenetic
591 analyses. *bioRxiv*. doi: 10.1101/001081
592 Gautier M, Foucaud J, Gharbi K *et al.* (2013a) Estimation of population allele
593 frequencies from next-generation sequencing data: pool-versus
594 individual-based genotyping. *Molecular Ecology*, **22**, 3766–3779.
595 Gautier M, Gharbi K, Cezard T *et al.* (2013b) The effect of RAD allele dropout on
596 the estimation of genetic variation within and between populations.
597 *Molecular Ecology*, **22**, 3165–3178.

598 Gomes I, Collins A, Lonjou C *et al.* (1999) Hardy-Weinberg quality control. *Annals*
599 *of human genetics*, **63**, 535–538.
600 Gompert Z, Buerkle CA (2011) A Hierarchical Bayesian Model for Next-
601 Generation Population Genomics. *Genetics*, **187**, 903–917.
602 Grundemann D, Schomig E (1996) Protection of DNA during preparative agarose
603 gel electrophoresis against damage induced by ultraviolet light.
604 *BioTechniques*, 21, 898–903.
605 Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel
606 Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS*
607 *Genet*, **6**, e1000862.
608 Hohenlohe PA, Catchen J, Cresko WA (2012) Population Genomic Analysis of
609 Model and Nonmodel Organisms Using Sequenced RAD Tags. *Methods in*
610 *molecular biology (Clifton, N.J.)*, **888**, 235–260.
611 Johnson PLF, Slatkin M (2008) Accounting for Bias from Sequencing Error in
612 Population Genetic Estimates. *Molecular Biology and Evolution*, **25**, 199–
613 206.
614 Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic
615 markers. *Bioinformatics*, **24**, 1403–1405.
616 Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history
617 of Xiphophorus fish and their sexually selected sword: a genome-wide
618 approach using restriction site-associated DNA sequencing. *Molecular*
619 *Ecology*, **22**, 2986–3001.
620 Koleff P, Soberón J, Arita HT *et al.* (2008) Patrones de diversidad espacial en
621 grupos selectos de especies. In: *Capital natural de México*, pp. 323–364.
622 CONABIO, México.

623 Loman NJ, Misra RV, Dallman TJ et al. (2012) Performance comparison of
624 benchtop high-throughput sequencing platforms. *Nature Biotechnology*,
625 **30**, 434–439.
626 Meacham F, Boffelli D, Dhahbi J et al. (2011) Identification and correction of
627 systematic error in high-throughput sequence data. *BMC Bioinformatics*,
628 **12**, 451.
629 Miller CR, Joyce P, Waits LP (2002) Assessing Allelic Dropout and Genotype
630 Reliability Using Maximum Likelihood. *Genetics*, **160**, 357–366.
631 Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000)
632 Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.
633 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-
634 by-sequencing in ecological and conservation genomics. *Molecular*
635 *Ecology*, **22**, 2841–2847.
636 Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and
637 Evolution in R language. *Bioinformatics*, **20**, 289–290.
638 Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association
639 genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**,
640 2991–3005.
641 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest
642 RADseq: An Inexpensive Method for De Novo SNP Discovery and
643 Genotyping in Model and Non-Model Species. *PLoS ONE*, **7**, e37135.
644 Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes,
645 consequences and solutions. *Nature reviews. Genetics*, **6**, 847–859.
646 Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference
647 from genomic sequence variation. *Genome Research*, **20**, 291–300.

Price DL, Casler MD (2012) Simple regression models as a threshold for selecting AFLP loci with reduced error rates. *BMC Bioinformatics*, **13**, 268.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, **155**, 945–959.

R. Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards PM, Liu MM, Lowe N *et al.* (2013) RAD-Seq derived markers flank the shell colour and banding loci of the Cepaea nemoralis supergene. *Molecular Ecology*, **22**, 3077–3089.

Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic acids research*, 38, D234–236.

Rounsaville TJ, Ranney TG (2010) Ploidy Levels and Genome Sizes of Berberis L. and Mahonia Nutt. Species, Hybrids, and Cultivars. *HortScience*, **45**, 1029–1033.

Rzedowski J (1978) *Vegetación de México*. Limusa, México.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable Genotyping of Samples with Very Low DNA Quantities Using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Urbanek S (2011) *multicore: Parallel processing of R code on machines with multiple cores or CPUs*.

Vázquez-Lobo A (1996) Filogenia de hongos endófitos del género Pinus: Implementación de técnicas moleculares y resultados preliminares. Sc.

Bach. Dissertation. Facultad de Ciencias, Universidad Nacional Autónoma de México, México

Warnes GR, Bolker B, Lumley and T (2013) *gtools: Various R programming tools*.

Wickham H (2012) *stringr: Make it easier to work with strings*.

Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

Xu J, Turner A, Little J, Bleecker ER, Meyers DA (2002) Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? *Human genetics*, **111**, 573–574.

Zamudio S (2009a) Familia Berberidaceae. *Flora del Bajio y regiones adyacentes*, **Fasículo 163**, 1–40.

Zamudio S (2009b) Notas sobre el Género Berberis (Berberidaceae) en México. *Acta Botánica Mexicana*, **87**, 31–70.

Zhang H, Hare MP (2012) Identifying and reducing AFLP genotyping error: an example of tradeoffs when comparing population structure in broadcast spawning versus brooding oysters. *Heredity*, **108**, 616–625.

Data and code availability

Raw RADseq data Sequence Read Archive (SRA) accession SRP035472. Sampling information, custom scripts and jobs with settings used to run *Stacks*, output data to compare error rates and population differentiation: doi:XXXXX/dryad.XXXXXX. Main R functions updated and versioned: https://github.com/AliciaMstt/RAD-error-rates.

697

698 **Supporting information**

699 S1. Summary of ddRAD labwork reaction mixes used and the characteristics of

700 the resulting libraries

701 S2. Dendograms obtained in the replicates analyses of *Stacks* parameters

702 S3. Allele and SNP error rates for the SNP calling model analyses

703 S4. PCoA for the each of the four *Stacks* parameter profiles tested

704

705 **Author Contributions**

706 Designed research: AMY, BCE, DP, N. Alvarez and THJ. Provided samples: AMY

707 and DP. Performed labwork and data analyses: AMY. Contributed analytical

708 tools: AMY and N. Arrigo. Wrote the paper: AMY, N. Arrigo, BCE. All authors

709 contributed to results discussion and manuscript edition.

710

711

712

713

714

715 Figure 1. Surveyed mountains for *B. alpina* within the Sierra Madre Oriental (1-3)
716 and the Transmexican Volcanic Belt (4-17). Populations where *B. alpina* was
717 found are El Zamorano (Za), Nevado de Toluca (To), Ajusco (Aj), Tlaloc (Tl),
718 Iztaccihuatl (Iz), La Malinche (Ma) and Cofre de Perote (Pe). *B. moranensis* was
719 found on Cerro San Andrés (An). As outgroups *B. pallida* (black stars) and *B.*
720 *trifolia* (white star) where sampled.

721 Figure 2. Total number of a) RAD-loci and b) SNPs obtained using different values
722 on Stacks core parameters. In each run only one parameter varied and the rest
723 were set to *m*=3, *M*= 2, *n*=0 and *max_locus_stacks* (mx.lcs) = 3 and *N*= *M*+2.

724 Figure 3. Effect on missing data obtained using different values on *Stacks* core
725 parameters. In each run only one parameter varied (shown in x axis) and the rest

726 were set as explained in Fig. 2. a) total number of missing loci; b) proportion of
727 missing loci relative to the total, c) proportion of missing loci different within a
728 replicate pair and d) locus error rate.

729 Figure 4. Effect on a) the allele error rate and b) the SNP error rate of using
730 different values on Stacks core parameters. In each run only one parameter
731 varied (shown in x axis) and the rest were set as explained in Fig. 2.

732 Figure 5. Effect on the genetic distance between individuals of the same
733 population of analyzing the full dataset with different Stacks profiles: default and
734 the settings that were considered to perform better in the exploratory parameter
735 analyses but using a range of values for –m: m=10 (high coverage), m=4 (near
736 optimal) and m=3 (optimal).

737

738

739

740

741

742

743
744
745

746 **Table 1. Potential causes of genotyping error for RADseq data at different levels**
747

Source	Reason	Reference*
Common to any molecular makers		
Technical	Errors related inversely to the quality of reagents and equipment, and to the organization of the laboratory in different rooms to avoid contaminations.	A
Human mistake	Sample swaps, pipetting errors, confusion in the data entry or during concentration measurements	A
Wet lab (mostly RAD specific)		
Enzyme sensitivities to DNA quality and quantity	Digestion and PCR efficiency may be irregular among samples, which can underrepresent some restriction fragments	A
Pooling concentrations	Samples with higher concentration can be overrepresented in the sequencing output if they are not pooled in equimolar amounts	B, C
Shearing step (single digest)	Sonication shears DNA of different lengths with different efficiencies. This biases read depth at RAD loci depending on restriction fragment length	D
PCR errors	PCR error gets amplified and can appear in multiple reads resembling an alternative allele at a locus. PCR error may differ among samples depending on reaction conditions and experimental design	E
PCR bias	PCR amplification can occur more readily on one allele or barcode, biasing their representation. Differences in amplification success leads to variation of coverage among loci and individuals, potentially resulting in allelic dropout or non-representation of some loci	A, C, E
Size selection (double digest)	Different fragments may be selected if more than one excision is performed. Imprecise size selection can include fragments of lengths relatively distant from the size-selection	C

	target mean	
Exposure to UV light	Can produce fragmentation (that could lead to locus/allele dropouts) and mutation of DNA strands (that introduces non-biological variation).	I
Next Generation Sequencing (NGS)		
Sequencing error	NGS introduces sequencing error (0.1–1.0% per nucleotide) into many of the reads. This can vary across samples, RAD sites and positions in the reads for each site.	E, K
Sequencing sampling	The sampling process of a heterogeneous library inherent in NGS introduces sampling variation in the number of reads observed across RAD sites as well as between alleles at a single site	E
Barcode error	PCR or sequencing errors at the barcode of a fragment can reduce the number of reads obtained for it	E
Genome intrinsic		
GC content	RAD loci with high GC content are sequenced at higher depths compared to RAD loci with low GC content, this increases with high number of PCR cycles	D
Variation in the restriction site	If there is variation in the restriction site, for a diploid organism one allele will be cut by the restriction enzyme and the other will not when truly it exists	D, G
DNA methylation	For some restriction enzymes digestion is impaired or blocked by methylated DNA. The same gene could or could not be methylated in different individuals or tissues.	J
Bioinformatic		
Handling variation in coverage	Coverage is the main filter to distinguish real variation from sequencing errors, repetitive regions and duplicates. But if there is coverage heterogeneity among samples and alleles, or if the general coverage is low, setting the filters with minimal coverage values too high can lead to allele drop. Setting it too low, however, can lead to incorrect SNP calls.	E, D, F
Mismatches parameters	The number of mismatches allowed to distinguish alleles and loci in an individual and in the population influences the number of real alleles and loci that are retrieved	F

Fragment length	Alleles will drop out as restriction fragment length decreases because RAD loci from short restriction fragments have low read depths. The efficacy of different bioinformatics tools at dealing with this varies.	D
Paralogs and repetitive regions handling	Paralogous regions with similar sequences can be erroneously assembled together	E
Presence of indels	Stacks and RADtools are unable to handle indels, therefore indel-containing loci are not clustered together, while they can be recovered by RaPiD and graph-based clustering approaches	C, D
Mapping using a reference genome	Mapping of alleles that are different from the reference genome is less probable than for a reference-matching allele, causing a bias in allele frequency toward the allele found in the reference sequence. It may additionally reduce the number of SNPs discovered and bias estimates of nucleotide diversity toward smaller values	H

* References: A) Bonin *et al.* 2004; B) Baird *et al.* 2008; C) Peterson *et al.* 2012, D) Davey *et al.* 2013; E) Hohenlohe *et al.* 2012; F) Catchen *et al.* 2013, G) Gautier *et al.* 2013b and H) Pool *et al.* 2010, I) Grundemann & Schomig 1996 , J) Roberts *et al*. 2010, K) Meacham *et al*. 2011 and Loman *et al*. 2012.

Table 2. Role of *Stacks* core parameters in loci assembly and potential sources of genotyping error

Parameter	How it affects assembly and genotyping error *
minimum number of identical, raw reads required to create a stack (–*m*) default 3	Reads with convergent sequencing errors are likely to be erroneously labelled as stacks if –*m* is too low. True alleles will not be recorded and will drop out if –*m* is too high. It can decrease the genotyping error by distinguishing real loci from PCR and sequencing error, but increase error by calling a heterozygous locus as homozygous when minimum coverage is set too high and one of the alleles is therefore dropped.
number of mismatches	If -*M* is too low some real loci will not be formed, and their alleles will be treated as different loci

allowed between loci when processing a single individual (-*M*) default 2	(undermerging), if it is too large repetitive sequences will chain together and form large nonsensical loci (overmerging).
number of mismatches allowed between loci when building the catalog (-*n*) default 0	When *n* = 0 there would be loci represented independently across individuals that truly are the same locus, and if *n* > 0 Stacks uses the consensus sequence from each locus to attempt to merge loci. This is important for population studies where monomorphic or fixed loci could exist in different individuals. Merging fixed alleles as a single locus can increase the chances of assembling real alleles, and therefore decrease the allele error rate. However, erroneous loci will be created if the –*n* value is too high.
maximum number of stacks at a single de novo locus (–*max_locus_stacks*) default 3	The expectation for nonrepetitive genomic regions is that a monomorphic locus will produce a single stack because the two sequences on the two homologous chromosomes are identical and thus indistinguishable. In contrast, a polymorphic locus will produce two stacks representing alternative alleles. Confounding cases that may arise from short, sequencing error-based stacks or from repetitive sequences, where hundreds of loci in the genome may collapse to a single putative locus. –max_locus_stacks allows to identify and blacklist confounding cases.
SNP calling model	In the default SNP calling model the error parameter is allowed to vary freely, whilst in a bounded-error model the boundary value is substituted if the maximum-likelihood value of ε exceeds a lower or upper bound. One consequence is that reducing the upper bound increases the chance of calling heterozygous a locus. This allows to balance the tolerance for false positive vs. false negative rates in calling genotypes, which in turn influences the genotyping error.

755 * Parameters explanation as in Catchen *et al.* 2013 and *Stacks* documentation, effect on genotyping error as discussed here.
756

757

758

759 **Table 3. Information content, error rates and efficacy to detect structuring of genetic variation for the full dataset processed**
760 **with different *Stacks* parameter settings.**
761

	optimal	*near optimal*	*high coverage*	*default*
Number of RAD-loci	6292	2449	292	4554
Total number of SNPs	11057	4353	502	7736
Mean read coverage per sample	10.32 (SD 4.16)	15.30 (SD 5.9)	58.92 (SD 21.9)	11.50 (SD 4.65)
Mean locus error rate	0.1738 (SD 0.103)	0.1657 (SD 0.100)	0.0882 (SD 0.088)	0.1590 (SD 0.094)
Mean allele error rate	0.0592 (SD 0.013)	0.0599 (SD 0.010)	0.0879 (SD 0.023)	0.0841 (SD 0.017)
Mean SNP error rate	0.0243 (SD 0.006)	0.0321 (SD 0.006)	0.0578 (SD 0.019)	0.0423 (SD 0.010)
Variation explained by first two axes of PCoA*	80(39)%	82(34)%	47(22)%	57(32)%
Mean of F_{ST} pairwise matrix*	0.19(0.07)	0.15(0.04)	0.03(0.01)	0.07(0.04)

762 * Results outside parenthesis were obtained using all the samples of the dataset, and the value inside parenthesis corresponds to the
763 results if excluding the samples from El Zamorano and the outgroup. El Zamorano (*B. alpina* population from SMOr) was excluded
764 because it explained as much variation as the *B. trifolia* outgroup (Supporting Information 5).