

2018

These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists

Shannon J. O'Leary

Jonathan B. Puritz
University of Rhode Island, jpuritz@uri.edu

Stuar C. Willis

Christopher M. Hollenbeck

David S. Portnoy

Follow this and additional works at: https://digitalcommons.uri.edu/bio_facpubs

Citation/Publisher Attribution

O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Mol Ecol*. 2018;27:1–14. <https://doi.org/10.1111/mec.14792> Available at: <https://doi.org/10.1111/mec.14792>

This Article is brought to you by the University of Rhode Island. It has been accepted for inclusion in Biological Sciences Faculty Publications by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu. For permission to reuse copyrighted content, contact the author directly.

These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists

The University of Rhode Island Faculty have made this article openly available.
Please let us know how Open Access to this research benefits you.

This is a pre-publication author manuscript of the final, published article.

Terms of Use

This article is made available under the terms and conditions applicable towards Open Access Policy Articles, as set forth in our [Terms of Use](#).

1 **Title:**

2 These aren't the loci you're looking for: Principles of effective SNP filtering for molecular
3 ecologists

4 **Authors:**

5 Shannon J. O'Leary*, Jonathan B. Puritz, Stuart C. Willis, Christopher M. Hollenbeck, David S.
6 Portnoy

7 *contact corresponding author: shannon.j.oleary@gmail.com

8 **Abstract:**

9 Sequencing reduced-representation libraries of restriction-site associated DNA (RADseq)
10 to identify single nucleotide polymorphisms (SNPs) is quickly becoming a standard
11 methodology for molecular ecologists. Because of the scale of RADseq data sets, putative loci
12 cannot be assessed individually, making the process of filtering noise and correctly identifying
13 biologically meaningful signal more difficult. Artifacts introduced during library preparation
14 and/ bioinformatic processing of SNP data can create patterns that are incorrectly interpreted as
15 indicative of population structure or natural selection. Therefore, it is crucial to carefully
16 consider types of errors that may be introduced during laboratory work and data processing, and
17 how to minimize, detect, and remove these errors. Here, we discuss issues inherent to RADseq
18 methodologies that can result in artifacts during library preparation and locus reconstruction,
19 resulting in erroneous SNP calls and ultimately, genotyping error. Further, we describe steps that
20 can be implemented to create a rigorously filtered data set consisting of markers accurately
21 representing independent loci and compare the effect of different combinations of filters on four
22 RAD data sets. Finally, we stress the importance of publishing raw sequence data along with
23 final filtered data sets in addition to detailed documentation of filtering steps and quality control
24 measures.

25 **1 The Rise of RAD**

26 Advances in sequencing technology coupled with increases in computational power have
27 resulted in a shift towards genome-scale data analysis, for which data sets typically consist of
28 thousands to tens-of-thousands of loci. At the same time, bioinformatic pipelines have become
29 more user-friendly and accessible to scientists without extensive backgrounds in bioinformatics
30 or programming. As a result, new analytical methods are rapidly being developed for studies
31 assessing levels of population structure and genomic diversity, identifying and mapping
32 quantitative trait loci (QTL), and screening for F_{ST} outliers putatively indicative of selection,
33 Increasingly, restriction site-associated DNA sequencing (RADseq)-derived single nucleotide
34 polymorphisms (SNPs) are becoming the molecular marker of choice. RADseq methods are
35 time- and cost-efficient techniques that utilize restriction enzymes to generate DNA fragments
36 from which thousands of SNPs can be identified using next-generation sequencing. This set of
37 methods does not require a fully sequenced reference genome as loci can be reconstructed *de*
38 *novo* from sequencing reads, greatly widening the types of organisms that can be studied beyond
39 traditional model species (Miller et al. 2007; Baird et al. 2008; Davey & Blaxter 2010). In
40 addition to the original RADseq protocol (Miller et al. 2007), ddRAD (Peterson et al. 2012),
41 ezRAD (Toonen et al. 2013) and 2b-RAD (Wang et al. 2012) are commonly applied techniques.
42 Despite differences between RADseq techniques and more traditional approaches, typically
43 limited to data sets consisting of mitochondrial and/or nuclear loci (e.g. 10 – 100 microsatellite
44 loci) all are unified by the assumption that the final data set consists of markers that each
45 represent a single locus and that these loci are unlinked (freely-recombining), a condition that
46 must be met when allele and genotype frequencies are being used to infer biological processes.

47 Recent reviews have summarized differences between individual RADseq techniques,
48 compared their respective advantages and disadvantages, and pointed out some potential sources
49 of genotyping error that can lead to biased datasets (Andrews et al. 2014; Puritz, et al. 2014).
50 More effort, however, is required to establish widely-accepted protocols to detect and remove
51 putative markers that in reality do not represent single loci, identify and correct erroneous SNP
52 calls, and assess genotyping error (but see Ilut et al. 2014; Li & Wren 2014; Mastretta-Yanes et
53 al. 2015). For other commonly used molecular markers such as AFLPs and microsatellites,
54 sources of genotyping error (e.g. allelic dropout, null alleles, stuttering) and best-practice
55 methods to efficiently detect and correct for them are well established (Bonin et al. 2004), and

56 standards of reporting regarding data quality control have been formalized. Currently, published
57 RADseq studies report (and practice) a wide array of data filtering and error detection procedures
58 after variant calling, but many publications underreport quality control methods, making it
59 difficult for the reader to assess data quality.

60 Generating SNP data sets using RADseq approaches involves three general steps: library
61 preparation, bioinformatic processing, and filtering for data quality. It is important to realize that
62 error potentially resulting in artifacts downstream can be introduced at any of these steps. The
63 introduction of some error during technical stages is unavoidable; therefore, it is important to
64 employ quality control steps that allow for the identification and reduction of error before the
65 dataset is analyzed. Here, we briefly review and make recommendations on how to limit and
66 detect common sources of technical artifacts during library preparation and bioinformatic
67 processing and suggest a set of filtering strategies that can be employed to create a robust data
68 set consisting of markers representing physically unlinked, correctly reconstructed loci (Table 1).
69 Further, we apply different combinations of suggested filters to several RAD data sets and
70 discuss the effectiveness of different filtering strategies.

71 **2 Minimizing artifacts associated with library preparation**

72 The goal of library preparation for a typical RADseq experiment is to consistently sample
73 the same set of fragments with sufficient coverage to correctly identify all alleles present at each
74 locus across all individuals within and across sequencing runs. In this context, ‘library’ refers to
75 a set of RADseq fragments isolated from a given number of individuals that are barcoded and
76 sequenced together on a single lane. Common technical artifacts introduced during library
77 preparation include (1) coverage effects, (2) locus drop-in/drop-out, (3) PCR artifacts, and (4)
78 library effects. Another common artifact, allele dropout, causes alleles to systematically remain
79 unsampled due to physical properties of the genome, i.e. cut-site or length polymorphisms.
80 Because allele dropout has a biological origin, it should be considered a biological artifact that
81 cannot be technically mitigated but rather can only be managed during bioinformatic processing
82 (discussed in detail in section 4.3). In contrast, technical artifacts are associated with technical
83 choices made by researchers and thus can be limited by careful planning during library
84 preparation, as discussed below.

85 2.1 Coverage effects: DNA quality, quantity and restriction digestion

86 RADseq methods, with the possible exception of recently developed hybrid enrichment
87 methods (Schmid et al. 2017; Suchan et al. 2016), require high molecular weight DNA to ensure
88 consistent digestion using restriction enzymes. Compared to other molecular markers, RADseq
89 protocols also require greater amounts of DNA (up to 500ng), and while there is some flexibility
90 in how much DNA is used, lower starting amounts of DNA increase the risks of low quality data.
91 Inconsistent digestions can be due to partially degraded DNA, inhibitors present in the reaction
92 (usually left over from extraction), and star-activity of the enzymes (i.e. cleavage of
93 noncanonical recognition sequences). This is problematic because it inhibits consistent recovery
94 of all fragments and produces downstream variance in coverage and/or missing data among loci
95 within and between libraries (Graham et al. 2015). To help ensure consistent digestions,
96 researchers should use high fidelity versions of restriction enzymes and perform trial digestions
97 to determine adequate concentrations and sufficient digestion times. Quality control measures
98 such as running digested samples on a fragment analyzer or agarose gel can be implemented to
99 compare digestion results. Unit definitions for enzymes and standard protocols are generally
100 based on the digestion of purified *Lambda* phage DNA; therefore, it is often advisable to use
101 more enzyme than manufacturer guidelines suggest. In addition, purifying genomic DNA before
102 digestion can remove inhibitors (e.g. phenol or pigments) carried over from extraction.

103 When read depth per locus per individual (hereafter ‘coverage’) is insufficient, alleles
104 may not be detected. Coverage effects may occur when initial DNA quality differs among
105 individuals or standardization of the amount of DNA prior to pooling is inconsistent resulting in
106 an unequal distribution of sequenced reads among individuals and loci. The use of high
107 sensitivity quantification kits, and standardization of DNA quantity prior to enzyme digestion
108 and again prior to adapter ligation can help to mitigate this issue. Similarly, pooling too many
109 individuals on a sequencing lane can result in systematic low read depth across all samples and
110 loci. This can be avoided by reducing the number of individuals per sequencing lane or by
111 adjusting the size selection window and enzyme(s) used to decrease the number of targeted
112 fragments. For loci affected by coverage effects, false homozygote calls will result in biased
113 allele frequency estimates which may cause genomic diversity to be underestimated, F_{ST} and
114 effective population size to be incorrectly estimated, and an increase in false positives/negatives
115 in F_{ST} -outlier tests (Arnold et al. 2013; Gautier et al. 2012).

116 2.2 Locus drop-in/drop-out due to size selection

117 Size selection is a crucial step for ensuring the consistent sampling of the same set of
118 fragments across ddRAD libraries. The magnitude of the variance in the distribution of fragment
119 lengths between libraries is dependent on the method used for size selection (Puritz et al. 2015).
120 Two commonly employed methods are manual gel cutting and automated (e.g. Pippin Prep) size
121 selection. While the latter is expected to increase the accuracy and precision of size selection,
122 there can still be inconsistencies caused by factors including salt concentration of the loaded
123 samples, and variable ambient laboratory temperature that can result in changes in the size
124 distribution of eluted fragments. Size selection anomalies can therefore result in fragments
125 dropping-in or out of the targeted size window for individually prepared libraries. To ensure
126 consistent fragment recovery it is important to make sure that both means and variances of
127 fragment size distributions are similar across runs. Because small fragments may be amplified
128 preferentially, libraries with wider variances may have suboptimal coverage for larger fragments
129 as compared to libraries with less variance even if means are similar. Thus, it is important to
130 implement quality control steps to determine whether the selected fragments fall into the
131 expected distribution given the targeted size window. For example, a fragment analyzer or high-
132 resolution electrophoresis gel can be used to determine the actual length of the fragments
133 retained in each library prior to sequencing.

134 2.3 PCR Artifacts

135 With the exception of proposed PCR-free protocols (e.g. ezRAD; Toonen et al. 2013),
136 and protocols performing PCR before size selection (Elshire et al. 2011), the final step of library
137 preparation is PCR amplification, during which artifacts may also be introduced. These can be
138 classified as (1) PCR error, including PCR chimeras, heteroduplexes, and *Taq* polymerase error
139 that could be exponentially propagated during PCR cycling, and (2) PCR bias, i.e. the
140 preferential amplification of shorter fragments and those with higher GC content. PCR artifacts
141 can be minimized by using high fidelity polymerase and high annealing temperatures to limit
142 copy error, reducing the number of cycles to minimize PCR bias, and providing sufficient
143 extension time based on fragment size. Additionally, several authors have recommended the
144 incorporation of barcodes with degenerate bases to aid in detection and removal of PCR
145 duplicates (Tin et al. 2015; Schweyenet al. 2014), i.e. reads stemming from the same fragment
146 template, which artificially increase read depth and therefore increase confidence in a SNP call

147 despite not actually representing independent observations. Finally, multiple reactions can be
148 completed with fewer cycles and combined into a final product to further mitigate PCR error and
149 bias.

150 *2.4 Library effects*

151 One of the principal benefits of reduced representation sequencing techniques is the
152 reproducibility of the library preparation process. In theory, repeating the process with the same
153 restriction enzymes and size selection window should consistently yield the same set of
154 fragments. In practice however, subtle differences between experiments, frequently beyond the
155 control of the researcher, can result in a situation where different sets of fragments are sequenced
156 and/or coverage differs greatly among libraries ('library effects'). Library effects can be caused
157 by a number of factors including differences in reagents and protocols used, ambient laboratory
158 temperature, poor accuracy and/or precision of size selection, and differences in DNA pool
159 quality and/or concentration (Bonin et al. 2004). While not all library effects can be avoided,
160 measures can be implemented to reduce the impact of library effects and identify markers most
161 severely affected.

162 The most effective ways to decouple the putative biological signal from patterns
163 introduced by library effects are by (1) randomly allocating individuals from different treatments
164 or geographic localities across libraries and (2) including technical replicates (repeated samples)
165 across libraries (Meirmans 2015). Randomizing samples across libraries broadly diminishes the
166 chances that artifactual signal will be confused as a biologically meaningful pattern, while also
167 allowing for downstream identification and removal of library effects. By performing a PCA, or
168 similar analysis, with data grouped by library and identifying and examining those markers most
169 associated with axes discriminating libraries, library effects can be mediated by removing biased
170 loci (Figure 1). When studies incorporate multiple libraries prepared at different times, under
171 different conditions and sequenced on multiple lanes, including a subset of individuals across
172 libraries ('technical replicates') should be standard practice. Incorporating these technical
173 replicates enables a direct comparison of genotypes across libraries, allowing for the
174 identification of loci that are consistently sampled with sufficient coverage to identify both
175 alleles, as well as loci exhibiting systematic genotyping errors. Implementing randomization of
176 individuals and including technical replicates during the library preparation stage is crucial for
177 identifying library effects during bioinformatic processing and data filtering.

178 **3 Minimizing artifacts associated with bioinformatics**

179 During bioinformatic processing of RADseq data in the absence of a fully sequenced and
180 assembled genome, reads are first clustered into contigs (contiguous sequence alignments) with
181 the goal that each contig should represent a single locus. Second, reads are clustered or aligned at
182 each reconstructed locus to identify and call SNPs for each individual. Artifacts most commonly
183 introduced at this stage are (1) clustering errors, i.e. the chosen values for the parameters of the
184 clustering algorithm result in under-splitting or over-splitting of putative loci and (2) artifactual
185 SNPs resulting from mapping errors or failure to identify PCR or sequencing error.

186 *3.1 Clustering error*

187 One of the main advantages of RADseq methods is the fact that SNPs can be identified
188 *de novo*, i.e. without a draft genome. The critical step in generating markers that accurately
189 represent these loci is the clustering of sequences into contigs that each represent a single locus
190 (Ilut et al. 2014). Several pipelines for marker reconstruction exist, including *Stacks* (Catchen et
191 al. 2013), *PyRAD* (Eaton 2014), *dDocent* (Puritz et al. 2014), and *AfrRAD* (Sovic et al. each of
192 which differs slightly in the strategies and methods employed. While the algorithmic details of
193 each pipeline are different, they all make the assignment of putative homology (orthology) of
194 fragments based on the number of mismatches or percent similarity. Efficacy of this technique
195 requires that the maximum divergence among alleles at a given locus is smaller than the
196 minimum divergence among loci (Ilut et al. 2014). Under-splitting occurs when sequence
197 similarity thresholds are too low such that multiple loci are combined into a single cluster
198 forming multi-locus contigs. The formation of multi-locus contigs will occur more frequently
199 with paralogs, repetitive elements and otherwise superficially similar sequences in the genome.
200 These multi-locus contigs can inflate the mean estimated heterozygosity. Conversely, over-
201 splitting occurs when sequence similarity thresholds are too high, causing alleles of the same
202 locus to be split into two or more contigs. Over-splitting results in deflation of mean estimated
203 heterozygosity. Picking similarity thresholds that result in no over- or under-splitting is not
204 possible because every genome contains elements that will suffer over- or under-splitting at
205 every threshold selected (Ilut et al. 2014). However, it is generally better to err on the side of
206 under-splitting, because methods to identify and remove multi-locus contigs are more effective
207 than those for identifying over-split loci (Ilut et al. 2014; Mastretta-Yanes et al. 2015; Willis et
208 al. 2017). In addition, understanding differences between bioinformatic pipelines is critical to

209 properly clustering the data. For example, Puritz et al. (*in prep*) found that rates of over-splitting
210 vary between *dDocent*, *PyRAD*, *Stacks*, and *AfrRAD* across various combinations of parameters.
211 Because effective thresholds for clustering will depend on the bioinformatic pipeline and vary by
212 organism, enzyme(s), and dataset, researchers should test parameters to identify values where
213 over-splitting is minimized.

214 3.2 Artifactual SNPs

215 Artifactual SNPs, those that do not exist in the actual genome but are called from the
216 mapped reads, may be the result of erroneous read clustering/mapping, PCR error, and/or
217 sequencing error. Because the rate of sequencing error varies by platform employed, chemistry
218 and read length, the typical user cannot control all error introduced at this stage, therefore, it is
219 important to account for sequencing error during bioinformatic analysis. FASTQ-format
220 sequence reads include PHRED-scale quality scores indicating the probability of a base call
221 being correct. The quality score, Q , equals $-10 \log_{10} P$, with P being the probability of a base-
222 calling error; for example, $Q = 30$ corresponds to the expectation that 1 in 1000 base-calls will be
223 incorrect, i.e. the probability of a correct base call is 99.9%. Quality scores can be used during
224 bioinformatic processing to trim low-quality sections from the beginnings and/or ends of reads or
225 to eliminate reads entirely, failure to do so can affect mapping quality downstream and/or
226 introduce artifactual SNPs. Similarly, library effects may be introduced at this stage if sequence
227 data is not carefully assessed for quality (especially at the 3' and 5' ends) and properly trimmed.
228 A PHRED-like quality score is also used by several variant callers, including *freebayes* and
229 GATK (Depristo et al. 2011; Garrison & Marth 2012), to determine the probability of a SNP call
230 being real or artifactual.

231 4 Filtering SNP data

232 Despite attempts to limit the introduction of technical artifacts during library preparation
233 and bioinformatic processing, SNP data sets require rigorous filtering because the inclusion of
234 only a few incorrectly genotyped loci in a data set can create a significant, misleading signal
235 (Davey et al. 2013; Li & Wren 2014; Meirmans 2015; Puritz et al. 2014). This is especially
236 important for F_{st} -outlier detection to determine loci potentially under selection because signal
237 caused by genotyping error is likely to stand out in pattern and magnitude from the signal
238 produced by the background SNP data (Hendricks et al. 2018; Xue et al. 2009). Full post-
239 processing exploration of each dataset should include an evaluation of the quality of each locus

240 and individual, the confidence in both SNP calls and genotypes, and whether specific loci are
241 likely to be multi-locus contigs. This should involve generating frequency distributions of
242 parameters including missing data per locus and individuals, read depth, and heterozygosity to
243 determine appropriate threshold values for these parameters. In addition, the comparison of
244 multiple filtered data sets generated using different parameter values provides guidance for
245 which combinations of thresholds retain the most loci while minimizing artifacts.

246 Beyond identifying parameters and threshold values that best identify and remove
247 specific types of artifacts, other important considerations include the order in which filters are
248 applied, whether individual genotypes should be selectively coded as missing (e.g. due to
249 insufficient coverage) or entire loci removed, whether to remove specific SNPs or entire SNP-
250 containing contigs, and whether threshold values should be applied across the entire data set or
251 separately across biologically meaningful groups, e.g. geographic sampling locations or, to
252 mitigate library effects, separately across individuals grouped e.g. by library/sequencing lane.
253 Additionally, every data set will be unique in terms of the number and quality of
254 samples/sequencing runs, and differences in the protocols employed (e.g. enzyme combinations,
255 targeted coverage, etc.); this means that individual data sets will differ in terms of missing data,
256 coverage, etc. Therefore, while certain parameters should always be considered during filtering,
257 the exact steps employed, and the applied thresholds will be specific to each data set.

258 To illustrate the effects of various filtering strategies and parameter thresholds, we
259 employed six different filtering schemes (FS) across four different data sets (Hollenbeck et al.
260 2018; O’Leary et al. 2018; Portnoy et al. 2015; Puritz et al. 2016). All data sets were created
261 using the *dDocent* pipeline and differ in terms of the focal organism, type of reference used to
262 map reads, the type of reads and the number of libraries sequenced (Table 1). The red snapper
263 data (Puritz et al. 2016) set consists of previously published data that has been recalled against a
264 fully sequenced draft genome consisting of large contigs (154,064 contigs; N50 = 233,156 bp;
265 total length 1.23 Gb) while the other three were assembled *de novo* as previously published. For
266 all FS, we first filtered genotypes, loci and individuals. Because most researchers analyze
267 datasets of biallelic SNPs, as a final step we decomposed multi-nucleotide variants and retained
268 only SNPs. Details of full FS are available in Table 2 and fully annotated scripts for filtering are
269 available at <https://github.com/sjoleary/SNPFILT>. The results of these FS are discussed in the
270 following sections to illustrate suggested filters.

271 *4.1 Low quality loci versus low quality individuals*

272 Filtering parameters used to identify loci and individuals that did not sequence well
273 include genotype call rate per locus (i.e. proportion of individuals a locus is called in) and
274 missing data per individual, as well as genotype depth and the mean depth per locus, i.e. mean
275 number of reads at a given locus across individuals. For data sets characterized by high levels of
276 missing data (e.g. red snapper, Figure 2), applying hard thresholds can result in retaining little to
277 no loci in the filtered data set. For example, for the red snapper data set, setting hard cut-offs
278 retaining only loci with genotype call rates >95% and individuals with <25% missing data, leads
279 to a final data set of only 10 SNPs on 3 contigs in 262 individuals (raw data set contains
280 1,106,387 SNPs on 25,168 contigs for 282 individuals, Table 3).

281 As an alternative strategy, starting with low cutoff values for missing data (per locus and
282 individual) and iteratively and alternately increasing them may result in more high-quality loci
283 and individuals being retained. For example, in the red snapper data set, first removing low
284 confidence genotypes by filtering for minimum genotype read depth >5, SNP quality score >20,
285 minor allele count >3, minimum mean read depth per locus >15 changes the distribution of
286 missing data per locus and individual and decreases the mean missing data from approximately
287 75% to 35% (Compare Figure 2A, B with C, D). Then iteratively increasing the stringency of
288 allowed missing data (final threshold values of a 95% genotype call rate and 25% allowed
289 missing data per individual) results in 9,478 – 12,056 SNPs on 1,626 – 1,680 contigs and 187 –
290 189 individuals being retained (Table 3), depending on the FS outlined in Table 2. This occurs
291 because poor quality individuals tend to deflate genotype call rates in otherwise acceptable loci,
292 and poor-quality loci increase missing data in otherwise acceptable individuals. Applying an
293 iterative filtering strategy consistently results in more loci and individuals being retained overall,
294 even in data sets consisting of individuals sequenced on a single sequencing lane for which the
295 initial distributions of missing data per locus and individuals are more favorable (Figure 3). For
296 example, after removing low confidence loci from the flounder data set as described above and
297 then setting a hard cutoff for a genotype call rate of >95% and allowed missing data per
298 individual of <25% results in a data set consisting of 15,682 SNPs on 3,802 contigs over 170
299 individuals, while iterative filtering results in data sets consisting of 18,663 – 24,103 SNPs on
300 4,789 – 5,341 contigs over 164 – 167 individuals (Table 3).

301 *4.2 Confidence in SNP identification*

302 The ability to filter loci depends on the pipeline used to reconstruct and genotype loci and
303 the set of parameters reported. As previously mentioned, variant callers such as report PHRED-
304 like quality scores for variants (SNPs) indicating the confidence in the SNP call being correct.
305 Similarly, users can set a minimum genotype depth below which genotypes are coded as missing
306 to determine the minimum number of reads that need to be present at each locus to be confident
307 that false homozygotes are excluded from the data (for further discussion see section 4.3).

308 Further, users often choose to set a minor allele count to remove potentially artifactual
309 SNP calls. For example, a minor allele count of three requires an allele to be observed in at least
310 two individuals (homozygote and heterozygote). It is common practice to assume that loci with a
311 minor allele frequency $< 5\%$ are not informative at a population level and to remove them from
312 data sets. Unfortunately, this strategy will remove true rare alleles from the data set that could be
313 informative in understanding patterns of connectivity and local adaptation. Because minor and
314 private alleles can be vital to accurately drawing inferences about past demographic events (e.g.
315 genetic bottlenecks), elucidating fine-scale population structure, understanding patterns of local
316 adaptation, and analyzing shifts in frequency spectra (Cubry et al. 2017; O'Connor et al. 2015;
317 Slatkin 1985), being able to distinguish between true minor alleles and genotyping error would
318 allow for better analysis of data sets. Carefully applying the filters as discussed in this section
319 can allow users to make this distinction, as illustrated by comparing the difference between data
320 sets created using specific filters before and after applying a minor allele count threshold.

321 *4.3 Confidence in genotypes: allele dropout/coverage effects*

322 While artifactual SNPs as described above will result in genotyping error (individuals
323 called heterozygous for alleles that do not exist), genotyping error at real SNPs may also occur.
324 Allele dropout and coverage effects can lead to unsampled alleles and individuals incorrectly
325 genotyped as homozygotes. Whereas coverage effects can be technically mitigated by setting a
326 target number of read per-individual, per-locus based on the total number of reads expected on
327 each sequencing lane and the number of fragments expected, allele dropout is an unavoidable
328 artifact of using restriction enzymes and size selection during library preparation. For targeted
329 fragments to be amplified and sequenced, adapters must be correctly ligated to the “sticky” ends
330 left by the enzymes, but polymorphisms may occur in the enzyme recognition site (cut-site
331 polymorphisms) resulting in alleles that are not cut by the restriction enzymes. Similarly, length

332 polymorphisms (insertion-deletions, or “indels”) may result in allele dropout when alleles fall
333 outside of the selected size window. In either case, the result is allele-specific sequencing failure.

334 Allele dropout cannot be avoided by optimizing standard laboratory procedures, but can
335 be accounted for during filtering by removing genotypes below a certain threshold of minimum
336 reads, and by identifying loci with high variance in read depth among individuals (Cooke et al.
337 2016; Davey et al. 2013). Low coverage can result in false homozygotes because the number of
338 reads may not be high enough to successfully call both alleles. Loci can be filtered based on a
339 threshold of minimum mean depth per locus and users can code individuals’ genotypes at
340 specific loci as missing if they fall below a minimum depth threshold that reflects the number of
341 reads required to confidently call homozygotes. This increases the confidence in individual
342 genotypes, and results in the removal of loci that consistently have genotypes not called with
343 high confidence across individuals. Unfortunately, during filtering it is difficult to distinguish
344 between allele dropout and coverage effects because they create similar patterns of missing data,
345 variance in depth and excess homozygosity. In both cases, failure to remove potentially affected
346 loci causes the introduction of false homozygotes and may result in biased estimates of
347 population genetic parameters based on allele frequencies and heterozygosity (DaCosta &
348 Sorenson 2014; Gautier et al. 2012), though the magnitude of this bias will vary depending on
349 the magnitude of the true biological signal in the data.

350 Hence, it is important to consider the statistical model being used for variant calling, and
351 how the model relates to read depth. For example, *freebayes* and GATK (Depristo et al. 2011;
352 Garrison & Marth 2012) are Bayesian callers that integrate data across all samples when
353 determining genotypes, meaning lower read-depth genotypes can be called with greater
354 accuracy. This is in contrast to genotyping models implemented in STACKS or PyRAD
355 (Catchen et al. 2011; Eaton 2014) which genotype individuals one at a time without the ability to
356 integrate data across samples until genotyping is completed. Finally, when deviations from
357 Hardy-Weinberg proportions are not expected, χ^2 tests of Hardy-Weinberg expectations for
358 individual loci within demes can also indicate heterozygote deficits that may indicate allele
359 dropout.

360 *4.4 Identification of multi-locus contigs*

361 Multi-locus contigs can be identified by assessing distributions of read depth, excess
362 heterozygosity, and the number of haplotypes observed per each individual at each marker (Ilut

363 et al. 2014; Li & Wren 2014; Willis et al. 2017). In general, total or mean read depth per locus
364 should be approximately normally distributed. Loci with coverage falling well above this
365 distribution may be reads clustered or mapped from multiple loci. Loci with excess coverage are
366 best identified by generating a frequency distribution of coverage and choosing thresholds, for
367 example, two times the mode (Willis et al. 2017) or the 90th quantile
368 (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters; Figure 4). Appropriate
369 thresholds will vary between data sets and species. Because fixed or near-fixed differences may
370 exist between non-orthologous loci, multi-locus contigs often have an excess number of
371 heterozygotes (Hohenlohe et al. 2011; Willis et al. 2017). *VCFtools* (Danecek et al. 2011)
372 provides a statistical framework for assessing heterozygote-excess via a χ^2 test of Hardy-
373 Weinberg expectations for VCF files. Finally, reads in multi-locus contigs often exhibit more
374 than two haplotypes per individual, and therefore loci can be removed based on a threshold for
375 the number of individuals with excess haplotypes (Ilut et al. 2014, Willis et al. 2017). While each
376 of these filters applied alone may catch many or even the majority of multi-locus contigs, the
377 most effective strategy to remove multi-locus contigs appears to be applying each filter in
378 parallel and removing markers flagged by any of the three filters (Willis et al. 2017).

379 4.5 INFO-flag filtering of vcf files

380 *Freebayes* and other multi-sample variant callers create annotated output files (VCF-
381 files) containing additional data pertaining to individual SNPs, coded as “INFO”-flags. Using
382 utilities such as *VCFtools* (Danecek et al. 2011), the suite of tools from *vcflib*
383 (<https://github.com/vcflib/vcflib>), and simple PERL and BASH scripting, it is possible to create
384 custom filters based on these flags. Li (2014) investigated false heterozygote calls on a SNP data
385 set generated from a haploid genome and estimated that the raw data set contained one erroneous
386 call in 10 – 15 kb. After implementing a set of filters based on the INFO-flags, the genotyping
387 error rate was reduced to one in 100 – 200 kb. The INFO-flag filters include allele balance,
388 mapping quality ratio, reads mapped as proper pairs, strand bias, and the relationship of read
389 depth to quality score.

390 Allele balance (AB) compares the number of reads for the reference allele to the number
391 of reads for the alternate allele across heterozygotes. The expected allele balance is 0.5; large
392 deviations may indicate false heterozygotes due to coverage effects, multi-locus contigs, or other
393 artifacts. Figure 5 shows AB for a raw data set, and for data sets that have been filtered for low

394 quality genotypes, loci and individuals. In both unfiltered and filtered data sets, loci with
395 high/low AB are present, indicating that problematic loci will remain unless AB is explicitly
396 filtered for.

397 Reads supporting either allele in a heterozygote should have similar mapping quality
398 values, the ratio of mapping quality between alleles, therefore, should be approximately one. The
399 mapping quality of a read is the probability of a given read mapping similarly well to another
400 location in the reference; reads stemming from paralogous or multi-copy loci should therefore
401 have reduced mapping quality, as they will map similarly well to multiple locations in the
402 reference. Hence, systematically large discrepancies between the mapping quality for reads
403 supporting the reference and alternate alleles at a SNP may be indicative of read-mapping errors,
404 due to repetitive elements, paralogs, or multi-locus contigs. Users should remove loci where
405 reads supporting the alternative allele have a substantially lower mapping quality compared to
406 reads supporting the reference allele. For example, *dDocent_filters*
407 (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters), a companion script to
408 the dDocent pipeline, suggests a lower threshold of 0.25 (Figure 6). Similarly, reads supporting
409 the reference allele are expected to have high mapping quality scores thus limiting how much
410 higher the mapping quality of reads supporting the alternative allele can become. Therefore, high
411 ratios only occur when mapping quality of reads supporting the reference allele are low, resulting
412 in a need for an upper threshold value (default 1.75 for *dDocent_filters* Figure 6). Users are
413 encouraged to assess their data sets to identify appropriate cut-offs. Standard filtering steps do
414 not remove all loci with biased mapping quality ratios (Figure 6). As mentioned in section 4.2,
415 assessing mapping quality ratios has the added benefit that it can help to identify minor alleles
416 that are not true alleles (Figure 6B), allowing researchers to retain true minor alleles that may
417 contain an important biological signal.

418 For paired-end libraries, artifacts can also be identified by examining the properly paired
419 status of reads and potential strand bias. The forward and reverse reads of a known pair should
420 always map to the same contig; improper read pairing, in which forward and reverse reads of a
421 known pair map to different contigs, indicates mapping anomalies such as multi-copy or
422 improperly assembled loci. Strand bias describes the relationship between forward and reverse
423 reads and SNP-calls at a given locus. For most paired-end RADseq libraries, the forward and
424 reverse reads do not overlap because the actual RAD fragments will be too long. For example, a

425 350 bp RAD fragment characterized with 125 bp pair-end reads will have 100 bp of
426 uncharacterized, intervening sequence. Therefore, a given SNP should only be apparent on either
427 the forward or reverse read. Calls of the same SNP in in both forward and reverse reads often
428 indicate mapping anomalies. However, the implications of this criterion depend on read length
429 and fragment length, and therefore the expected overlap of paired reads in a given data set.

430 Finally, the relationship between SNP quality score and read depth should be assessed;
431 these measures should be positively correlated, because, theoretically, increasing read depth
432 should decrease the likelihood of false homozygous calls (Li & Wren 2014). Users may choose
433 to apply a general threshold value for the ratio of locus quality to read depth and/or apply a
434 separate SNP quality score threshold value for loci with high read depth. For example,
435 *dDocent_filters* (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters), a
436 companion script to the dDocent pipeline, implements this by considering SNPs with a depth $>$
437 mean + 1 standard deviation as high coverage and then removing high coverage SNPs for which
438 the quality score is less than two times the read depth (Figure 7, Li & Wren 2014).

439 **5. Physical linkage**

440 After filtering, most RADseq data sets will generally contain sets of SNPs located on the
441 same contig. SNPs located within a few hundred base pairs of each other are generally physically
442 linked (Hohenlohe et al. 2012; Miyashita & Langley 1988), whereas most commonly used
443 analyses assume that all genetic markers are independent, of course, due to the fact that RAD
444 methods randomly sample the genome it is possible that selected fragments are linked as well
445 and users should, where appropriate, test for linkage disequilibrium between loci to avoid biasing
446 results. Treating physically linked SNPs as independent markers provides biased results,
447 including false signals of population structure. A common method to remove this bias is to retain
448 only one SNP from each contig (“thinning”). This is an appropriate strategy but one that reduces
449 the information content of a given marker if multiple SNPs are contained on a single contig.
450 Another way to deal with physical linkage is to infer haplotypes for each contig based on the
451 combination of filtered SNPs within paired reads (Willis et al. 2017). This strategy will produce
452 the same number of markers as thinning, but many markers will be multi-allelic, therefore,
453 haplotyping manages physical linkage while preserving the total information content of the data
454 set.

455 **6. Conclusions & outlook (on the importance of reproducible research)**

456 With the shift from data sets consisting of markers for tens to hundreds of microsatellite
457 loci to several thousand SNP-containing loci, bioinformatic processing has become the only
458 viable means of ensuring data quality. If careful quality control is implemented, RAD methods
459 are a powerful instrument in the molecular ecologist's tool box to assess levels population
460 structure and connectivity and local adaptation in non-model species for which genomic
461 resources might not (yet) be available. Many studies currently report very few details pertaining
462 to quality control methods applied to the output from SNP calling pipelines beyond very basic
463 filtering, frequently limited to the removal of markers and/or individuals with low coverage or
464 high levels of missing data. Enabling this under-reporting is a lack of clear quality control
465 standards. Nevertheless, it is incumbent upon the authors to document data preparation and
466 quality control steps and make these available to the scientific community along with raw data
467 sets to ensure that data analyses are transparent and fully reproducible (Leek & Peng 2015; Peng
468 2014).

469 Here, we have provided a discussion of several of the places that errors and artifacts may
470 be introduced into RADseq datasets and provided recommendations for how to minimize, detect,
471 and account for these artifacts from laboratory through bioinformatic and filtering stages. We
472 hope that these recommendations facilitate discussion about standardization of quality control in
473 RAD-based population genomics data sets. While a detailed description of each filtering step
474 would exhaust available space for the methods section of a manuscript, researchers should
475 include detailed procedures in the supplementary material and deposit custom script(s) in public
476 data or code repositories (e.g. Portnoy et al. 2015; Puritz et al. 2016; O'Leary et al. 2018).
477 Further, platforms such as GitHub (<http://github.com>) allow for convenient archiving as well as
478 assigning DOIs (digital object identifiers) to make code citable. A description of processing
479 should accompany data sets archived in readily interpretable formats, along with the associated
480 meta-data, and consist of the tools (name and version) and exact parameters used for processing.
481 In addition to making data analysis fully transparent and reproducible, this will allow developed
482 approaches to be applied to other data sets and facilitate the development of new and better
483 approaches in the application of genomics to molecular ecology.

484 **Acknowledgements**

485 We would like to thank John R. Gold for his role supporting this work and members of
486 the marine genomics working group at Texas A&M Corpus Christi for many helpful suggestions.
487 JP would like to thank the participants/organizers of the Bioinformatics for Adaptation Genomics
488 class from 2014-2017 for their help with testing various aspects of SNP filtering. The example
489 data sets used for this study were generated using funding from the National Marine Fisheries
490 Service (National Oceanographic and Atmospheric Administration) Marfin Award #
491 NA12NMF4330093 and Sea Grant Award # NA10OAR4170099; Texas Parks and Wildlife and
492 the U.S. Fish and Wildlife Service through a Wildlife & Sport Fish Restoration State Wildlife
493 Grant (Subcontract 5624, CFDA# 15.634) and by the College of Science and Engineering at
494 Texas A&M University-Corpus Christi.

495 **References**

- 496 Andrews, K. R., Hohenlohe, P. A., Miller, M. R., Hand, B. K., Seeb, J. E., & Luikart, G. (2014).
497 Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. *Molecular*
498 *Ecology*. doi:10.1111/mec.12964
- 499 Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates
500 diversity and introduces genealogical biases due to nonrandom haplotype sampling.
501 *Molecular Ecology* 22(11), 3179–3190. doi:10.1111/mec.12276
- 502 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson,
503 E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers.
504 *PLoS ONE*, 3(10), 1–7. doi:10.1371/journal.pone.0003376
- 505 Bonin, A., Bellemain, E., Eidesen, P. B., Pompanon, F., Brochmann, C., & Taberlet, P. (2004).
506 How to track and assess genotyping errors in population genetics studies. *Molecular*
507 *Ecology*, 13(11), 3261–3273. doi:10.1111/j.1365-294X.2004.02346.x
- 508 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An
509 analysis tool set for population genomics. *Molecular Ecology* 22(11), 3124–3140.
510 doi:10.1111/mec.12354
- 511 Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks :
512 Building and Genotyping Loci De Novo From Short-Read Sequences. *G3:*
513 *Genes/Genomes/Genetics*, 1(3), 171–182. doi:10.1534/g3.111.000240
- 514 Cooke, T. F., Yee, M. C., Muzzio, M., Sockell, A., Bell, R., Cornejo, O. E., ... Kenny, E. E.
515 (2016). GBStools: A statistical method for estimating allelic dropout in reduced
516 representation sequencing data. *PLoS Genetics*, 12(2), e1005631.
517 doi:10.1371/journal.pgen.1005631
- 518 Cubry, P., Vigouroux, Y., & François, O. (2017). The Empirical Distribution of Singletons for
519 Geographic Samples of DNA Sequences. *Frontiers in Genetics*, 8, 139.
520 doi:10.3389/fgene.2017.00139

- 521 DaCosta, J. M., & Sorenson, M. D. (2014). Amplification Biases and Consistent Recovery of
522 Loci in a Double-Digest RAD-seq Protocol. *PLoS ONE*, 9(9), e106713.
523 doi:10.1371/journal.pone.0106713
- 524 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R.
525 (2011). The variant call format and VCFtools. *Bioinformatics* 27(15) 2156–2158.
526 doi:10.1093/bioinformatics/btr330
- 527 Davey, J. L., & Blaxter, M. W. (2010). RADseq: Next-generation population genetics. *Briefings*
528 *in Functional Genomics*, 9(5–6), 416–423. doi:10.1093/bfpg/elq031
- 529 Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013).
530 Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*
531 22(11), 3151–3164. doi:10.1111/mec.12084
- 532 DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., ... Daly, M. J.
533 (2011). A framework for variation discovery and genotyping using next-generation DNA
534 sequencing data. *Nature Genetics*, 43(5), 491–501. doi:10.1038/ng.806
- 535 Eaton, D. A. R. (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses.
536 *Bioinformatics*, 30(13), 1844–1849. doi:10.1093/bioinformatics/btu121
- 537 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell,
538 S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity
539 species. *PLoS ONE*, 6(5), e19379. doi:10.1371/journal.pone.0019379
- 540 Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.
541 *Plos One*, 11(3), e0151651. doi:arXiv:1207.3907 [q-bio.GN]
- 542 Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., ... Estoup, A. (2012).
543 The effect of RAD allele dropout on the estimation of genetic variation within and between
544 populations. *Molecular Ecology* 22(11), 3165–3178. doi:10.1111/mec.12089
- 545 Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., ... Somers,
546 C. M. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing
547 (RADSeq). *Molecular Ecology Resources*, 15(6), 1304–1315. doi:10.1111/1755-
548 0998.12404
- 549 Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., ... Luikart,
550 G. (2018). Title: Recent advances in conservation and population genomics data analysis.
551 *Evolutionary Applications*. doi:10.1111/eva.12659
- 552 Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-
553 generation RAD sequencing identifies thousands of SNPs for assessing hybridization
554 between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, 11(SUPPL.
555 1), 117–122. doi:10.1111/j.1755-0998.2010.02967.x
- 556 Hohenlohe, P. A., Bassham, S., Currey, M., & Cresko, W. A. (2012). Extensive linkage
557 disequilibrium and parallel adaptive divergence across threespine stickleback genomes.
558 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587), 395–
559 408. doi:10.1098/rstb.2011.0245
- 560 Hollenbeck, C. M., Portnoy, D. S., Puritz, J. B., Samollow, P., & Gold, J. R. (2018). Fine-scale
561 population structure and genomic islands of divergence in a coastal marine fish, red drum
562 (*Sciaenops ocellatus*). *Molecular Ecology*.

- 563 Ilut, D. C., Nydam, M. L., & Hare, M. P. (2014). Defining loci in restriction-based reduced
564 representation genomic data from nonmodel species: Sources of bias and diagnostics for
565 optimal clustering. *BioMed Research International 2014*. doi:10.1155/2014/675158
- 566 Leek, J. T., & Peng, R. D. (2015). Opinion: Reproducible research can still be wrong: Adopting a
567 prevention approach. *Proceedings of the National Academy of Sciences*, 112(6), 1645–1646.
568 doi:10.1073/pnas.1421412111
- 569 Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-
570 coverage samples. *Bioinformatics*. doi:10.1093/bioinformatics/btu356
- 571 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C.
572 (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de
573 novo assembly optimization for population genetic inference. *Molecular Ecology*
574 *Resources*, 15(1) 28–41. doi:10.1111/1755-0998.12291
- 575 Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them.
576 *Molecular Ecology* 24(July), 3223–3231.
- 577 Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and
578 cost-effective polymorphism identification and genotyping using restriction site associated
579 DNA (RAD) markers. *Genome Research*, 17(2) 240–248. doi:10.1101/gr.5681207
- 580 Miyashita, N., & Langley, C. H. (1988). Molecular and phenotypic variation of the white locus
581 region in *Drosophila melanogaster*. *Genetics*, 120(1), 199–212.
- 582 O'Connor, T. D., Fu, W., Mychaleckyj, J. C., Logsdon, B., Auer, P., Carlson, C. S., ... Akey, J.
583 M. (2015). Rare Variation Facilitates Inferences of Fine-Scale Population Structure in
584 Humans. *Molecular Biology and Evolution*, 32(3), 653–660. doi:10.1093/molbev/msu326
- 585 O'Leary, S. J., Hollenbeck Christopher M. Vega, R. R., Gold, J. R., & Portnoy, D. S. (2018).
586 Comparative genomics as a tool for restoration enhancement and culture of southern
587 flounder, *Paralichthys lethostigma*. *BMC Genomics*.
- 588 Peng, R. D. (2014). Reproducible Research in Computational Science. *Science*, 1226(2011).
589 doi:10.1126/science.1213847
- 590 Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest
591 RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and
592 non-model species. *PLoS ONE*, 7(5). doi:10.1371/journal.pone.0037135
- 593 Portnoy, D. S., Puritz, J. B., Hollenbeck, C. M., Gelsleichter, J., Chapman, D., & Gold, J. R.
594 (2015). Selection and sex-biased dispersal: the influence of philopatry on adaptive variation.
595 *PeerJ*, 1–20. doi:10.7287/peerj.preprints.1300v1
- 596 Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2016). Fine-scale partitioning of genomic variation
597 among recruits in an exploited fishery: Causes and consequences. *Scientific Reports*, 6(1),
598 36095. doi:10.1038/srep36095
- 599 Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent : a RADseq, variant-calling
600 pipeline designed for population genomics of non-model organisms. *PeerJ* 2, e431.
601 doi:10.7717/peerj.431
- 602 Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2015). Fishing for Selection, but Only Catching
603 Bias: Examining Library Effects in Double-Digest RAD Data in a Non-Model Marine
604 Species. In *Plant and Animal Genome XXIII Conference*.

- 605 doi:10.6084/m9.figshare.1287474.v3
- 606 Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014).
607 Demystifying the RAD fad. *Molecular Ecology*. doi:10.1111/mec.12965
- 608 Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., & Alvarez, N. (2017).
609 HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract
610 genomic information from ancient DNA. *Methods in Ecology and Evolution*.
611 doi:10.1111/2041-210X.12785
- 612 Schweyen, H., Rozenberg, A., & Leese, F. (2014). Detection and removal of PCR duplicates in
613 population genomic ddRAD studies by addition of a degenerate base region (DBR) in
614 sequencing adapters. *The Biological Bulletin* 227(2), 146–60. doi:10.1086/BBLv227n2p146
- 615 Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution*, 39(1), 53–65.
616 doi:10.2307/2408516
- 617 Sovic, M. G., Fries, A. C., & Gibbs, H. L. (2015). AfrRAD: a pipeline for accurate and efficient
618 de novo assembly of RADseq data. *Molecular Ecology Resources*, 15(5), 1163–1171.
619 doi:10.1111/1755-0998.12378
- 620 Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., ... Alvarez,
621 N. (2016). Hybridization capture using RAD probes (hyRAD), a new tool for performing
622 genomic analyses on collection specimens. *PLoS ONE*, 11(3), e0151651.
623 doi:10.1371/journal.pone.0151651
- 624 Tin, M. M. Y., Rheindt, F. E., Cros, E., & Mikheyev, A. S. (2015). Degenerate adaptor
625 sequences for detecting PCR duplicates in reduced representation sequencing data improve
626 genotype calling accuracy. *Molecular Ecology Resources*, 15(2), 329–336.
627 doi:10.1111/1755-0998.12314
- 628 Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R.,
629 & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model
630 organisms. *PeerJ*, 1, e203. doi:10.7717/peerj.203
- 631 Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible
632 method for genome-wide genotyping. *Nature Methods*, 9(8), 808–810.
633 doi:10.1038/nmeth.2023
- 634 Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping
635 RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular
636 Ecology Resources*. doi:10.1111/1755-0998.12647
- 637 Xue, Y., Zhang, X., Huang, N., Daly, A., Gillson, C. J., Macarthur, D. G., ... Tyler-Smith, C.
638 (2009). Population differentiation as an indicator of recent positive selection in humans: an
639 empirical evaluation. *Genetics*, 183(3), 1065–77. doi:10.1534/genetics.109.107722

640 **Data Accessibility**

641 Annotated scripts for filtering are available at <https://github.com/sjoleary/SNPFILT> along with
642 information to obtain versions of published data sets used to illustrate filtering principle set forth
643 in this manuscript.

644 **Figures and Tables**

645 **Figure 1:** Library effects (adapted from Puritz *et al.* 2015). PCA of RAD data set combining four
646 libraries (yellow squares, red diamonds, blue triangles, green circles) before (A) and after (B)
647 correcting for library effects by removing affected markers.

648 **Figure 2:** Missing data per locus and individual (indv), respectively for unfiltered red snapper
649 data set (A, B) and after coding genotypes with <5 reads as missing and removing low quality
650 loci with SNP quality score <20 and minimum mean depth <15 reads (C, D). Red dashed line
651 indicates mean proportion of missing data.

652 **Figure 3:** Missing data per locus and individual, respectively for unfiltered southern flounder
653 data set (A, B) and after coding genotypes with <5 reads as missing and removing low quality
654 loci with SNP quality score <20 and minimum mean depth <15 reads (C, D). Red dashed line
655 indicates mean proportion of missing data.

656 **Figure 4:** Distribution of mean depth per locus across all loci for red snapper data set after
657 removing low confidence/quality loci (minimum genotype depth >3, SNP quality score >20,
658 minor allele count >3, mean minimum depth across all individuals >15), and iterative filtering of
659 missing data to final threshold of genotype call rate >95% and allowed missing data per
660 individual <25%. Blue dotted line indicates 95% percentile (123.5) and red dashed line 2x the
661 mode (156) as potential cut-offs to remove loci with excessively high depth indicative of multi-
662 locus contigs following Willis *et al.* (2017).

663 **Figure 5:** Allele balance in heterozygous genotypes (proportion of reads corresponding to the
664 reference allele) for (A) unfiltered red drum data set, (B) data set with genotype read depths <3
665 reads coded as missing and loci with SNP quality score <20, mean depth <15 reads and/or >30%
666 missing data removed, and (C) data set filtered as (B) and loci with a minor allele count <3
667 removed in addition. Except for minor sampling error, reference and alternate allele should be
668 supported by the same number of reads, i.e. allele balance should be 0.5 (red dashed line); values
669 away from this indicate potential anomalies. The blue dotted lines indicate default cut-off values
670 of 0.2 and 0.8 implemented in dDocent_filters

671 (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters).

672 **Figure 6:** Ratio of mean mapping quality scores for the reference and alternate allele for
673 southern flounder data set. (A) Genotypes with <5 reads have been coded as missing and loci
674 with SNP quality score <20, mean read depth <15 reads, >30% missing data and/or and minor

675 allele count of <3 removed; (B) same data set without applying minor allele count filter. Red
676 dashed line indicates loci with mapping quality ratio of 1, i.e. the further away the larger the
677 discrepancy between the mapping quality of the reference and alternate allele. Blue dashed lines
678 indicate cut-off values for ratio of mean mapping quality score of 0.25 and 1.75 (alternate to
679 reference allele) as implemented in dDocent_filters
680 (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) to remove loci with high
681 discrepancy of mapping quality for the alleles of a given locus (indicated in red below the dashed
682 line).

683 **Figure 7:** Comparison of SNP quality score and total depth per locus for the bonnethead shark
684 data set. Vertical blue dashed line identifies loci with high depth (mean + 1 standard deviation).
685 Loci with a quality score <2x the depth at that locus are below the diagonal blue dashed line
686 (indicated in red).

687 **Table 1:** Overview of described potential issues in raw RAD data sets, their causes, and
688 strategies for technical and bioinformatic mitigation

689 **Table 2:** Detailed description of six different filtering schemes applied to example data sets, the
690 order of the rows indicates the order in which filters we applied. Applied filters are designed to
691 remove loci with low confidence SNP calls (minimum genotype read depth (minDP), SNP
692 quality score (qual), mean read depth per locus across all individuals (meanDP), minor allele
693 count (mac), missing data (allowed missing data per individual (imiss), genotype call rate
694 (number of individuals that have been called for a given locus (geno)) and INFO-filters as
695 described in the manuscript.

696 **Table 3:** Comparison of the number of SNPs, contigs (cont) and individuals (indv) in the raw
697 data sets and number (proportion) retained in each data set for six different filtering schemes
698 (FS) as described in Table 2.

699 **Supplementary Information**

700 **Table S1:** Comparison of four published ddRAD data sets compiled using the dDocent pipeline.
701 (A) Comparison of sequencing type used to create reference and call genotypes, the number of
702 combined libraries, approximate genome size, and enzymes used to fragment DNA. All data sets
703 were run on the Illumina platform to obtain either paired end (PE) or single end (SE) reads.