MOLECULAR ECOLOGY

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

Journal:	*Molecular Ecology*
Manuscript ID	Draft
Manuscript Type:	Opinion
Date Submitted by the Author:	n/a
Complete List of Authors:	O'Leary, Shannon; Texas A&M Corpus Christi, Life Science Department Puritz, Jonathan; University of Rhode Island, Biological Sciences Willis, Stuart; Texas A&M University Corpus Christi, Life Sciences Hollenbeck, Chris ; Scottish Oceans Institute Portnoy, David; Texas A&M University Corpus Christi, Life Sciences
Keywords:	Conservation Genetics, Ecological Genetics, Landscape Genetics, Molecular Evolution, Population Genetics - Empirical, Population Ecology

Title:

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

Authors:

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

Abstract:

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

1 The Rise of RAD

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

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

RADseq studies report (and practice) a wide array of data filtering and error detection procedures after variant calling, and many publications underreport quality control methods, making it difficult for the reader to assess data quality.

Generating SNP data sets using RADseq approaches involves three general steps: library preparation, bioinformatic processing, and filtering for data quality; it is important to realize that error can be introduced at any of these steps. The introduction of some error during technical stages is unavoidable; therefore, it is important to employ quality control steps that allow for the identification and reduction of error before the dataset is analyzed. Here, we briefly review common sources of technical artifacts during library preparation and bioinformatic processing, make recommendations on how to limit and detect them, and suggest a set of filtering strategies that can employed to create a robust data set consisting of markers representing physically unlinked, correctly reconstructed loci.

2 Minimizing artifacts associated with library preparation

The goal of library preparation for a typical RADseq experiment is to consistently sample the same set of fragments with sufficient coverage to correctly identify all alleles present at each locus across all individuals within and across sequencing runs. In this context, 'library' refers to a set of RADseq fragments isolated from a given number of individuals that are barcoded and sequenced together. Common technical artifacts introduced during library preparation include coverage effects, locus drop-in/drop-out, PCR artifacts and library effects. Another common artifact, allele dropout, causes alleles to systematically remain unsampled due to physical properties of the genome, i.e. cut-site or length polymorphisms. Because allele dropout has a biological origin, it should be considered a biological artifact and can only be managed during bioinformatic processing of the data set. By contrast, technical artifacts are associated with technical choices made by researchers and thus can be limited by careful planning during library preparation.

2.1 Coverage effects: DNA quality, quantity and restriction digestion

RADseq methods, with the possible exception of recently developed hybrid enrichment methods (Suchan *et al.* 2016; Schmid *et al.* 2017), require high molecular weight DNA to ensure consistent, complete digestion using restriction enzymes. Compared to other molecular markers, RADseq protocols also require greater amounts of DNA (up to 500ng), and while there is some flexibility in how much DNA is used, lower starting amounts of DNA increase the risks of low

quality data. Incomplete digestions can be due to partially degraded DNA, inhibitors present in the reaction (usually left over from DNA extraction), and star-activity of the enzymes (i.e. cleavage of noncanonical recognition sequences). Incomplete digestion is problematic because it inhibits consistent recovery of all fragments and produces downstream variance in coverage and/or missing data among loci within and between libraries (Graham *et al.* 2015). To help ensure complete digestions, researchers should use high fidelity restriction enzymes and perform trial digestions to determine adequate concentrations and sufficient digestion times. Unit definitions for enzymes and standard protocols are generally based on the digestion of purified *Lambda* phage DNA; therefore, it is often advisable to use more enzyme than manufacturer guidelines suggest. In addition, purifying genomic DNA before digestion can remove inhibitors (e.g. phenol or pigments) carried over from extraction.

When coverage is insufficient, alleles may not be detected. Coverage effects may occur when initial DNA quality differs among individuals or standardization of the amount of DNA prior to pooling is inconsistent. The use of high sensitivity quantification kits, and standardization of DNA quantity prior to enzyme digestion and again prior to adapter ligation can help to mitigate this issue. In addition, pooling too many individuals in a library can result in systematic low coverage across all samples and loci. This can be avoided by reducing the number of individuals per library or by adjusting the size selection window and enzyme(s) used to decrease the number of targeted fragments. For loci affected by coverage effects, false homozygote calls will result in biased allele frequency estimates which may cause genomic diversity to be underestimated, F_{ST} and effective population size to be incorrectly estimated, and an increase in false positives/negatives in F_{ST}-outlier tests (Gautier *et al.* 2012; Arnold *et al.* 2013).

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

Size selection is a crucial step for ensuring the consistent sampling of the same set of fragments across ddRAD libraries. The magnitude of the variance in the distribution of fragment lengths between libraries is dependent on the method used for size selection (Puritz *et al.* 2015). Two commonly employed methods are manual gel cutting and automated (e.g. Pippin Prep) size selection. While the latter is expected to increase the accuracy and precision of size selection, there can still be inconsistencies caused by factors including salt concentration of the loaded samples, and variable ambient laboratory temperature that can result in changes in the size

distribution of eluted fragments. Size selection anomalies can therefore result in fragments dropping-in or out of the targeted size window for individually prepared libraries. Thus, it is important to implement quality control steps to determine whether the selected fragments fall into the expected distribution given the targeted size window. For example, a fragment analyzer or high-resolution electrophoresis gel can be used to determine the actual length of the fragments retained in each library prior to sequencing.

2.3 PCR Artifacts

With the exception of proposed PCR-free protocols (e.g. ezRAD; Toonen *et al.* 2013), the final step of library preparation is PCR amplification, during which artifacts may also be introduced. These can be classified as (1) PCR error, including PCR chimeras, heteroduplexes, and *Taq* polymerase error that could be exponentially propagated during PCR cycling, and (2) PCR bias, i.e. the preferential amplification of shorter fragments and those with higher GC content. PCR artifacts can be minimized by using high fidelity polymerase and high annealing temperatures to limit copy error, reducing the number of cycles to minimize PCR bias, and providing sufficient extension time based on fragment size. Additionally, multiple reactions can be completed with fewer cycles and combined into a final product to further mitigate PCR bias and artifacts.

2.4 Library effects

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

The most effective ways to decouple the putative biological signal from patterns introduced by library effects are by (1) randomly allocating individuals from different treatments

or geographic localities across libraries and (2) including technical replicates (repeated samples) across libraries (Meirmans 2015). Randomizing samples across libraries broadly diminishes the chances that artifactual signal will be confused as a biologically meaningful pattern, while also allowing for downstream identification and removal of library effects. By performing a PCA, or similar analysis, with data grouped by library and identifying and examining those markers most associated with axes discriminating libraries, library effects can be mediated by removing biased loci (**Figure 1**). Incorporating technical replicates enables a direct comparison of genotypes across libraries allowing for the identification of systematic genotyping error and artifactual markers. While these analytical steps are performed during data filtering (described below), they critically depend on planning during the library preparation stage.

3 Minimizing artifacts associated with bioinformatics

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

3.1 Clustering error

One of the main advantages of RADseq methods is the fact that SNPs can be identified *de novo*, i.e. without a draft genome. The critical step in generating markers that accurately represent these loci is the clustering of sequences into contigs that each represent a single locus (Ilut *et al.* 2014). Several pipelines for marker reconstruction exist, including *Stacks* (Catchen *et al.* 2013), *PyRAD* (Eaton 2014), *dDocent* (Puritz *et al.* 2014a), and *AftrRAD* (Sovic *et al.* 2015), each of which differs slightly in the strategies and methods employed. While the algorithmic details of each pipeline are different, they all make the assignment of putative homology (orthology) of fragments based on the number of mismatches or percent similarity. Efficacy of this technique requires that the maximum divergence among alleles at a given locus is smaller than the minimum divergence among loci (Ilut *et al.* 2014). Under-splitting occurs when sequence similarity thresholds are too low such that multiple loci are combined into a single cluster forming multi-locus contigs. The formation of multi-locus contigs will occur more

frequently with paralogs, repetitive elements and otherwise superficially similar sequences in the genome. These multi-locus contigs can inflate the mean estimated heterozygosity. Conversely, over-splitting occurs when sequence similarity thresholds are too high, causing alleles of the same locus to be split into two or more contigs. Over-splitting results in deflation of mean estimated heterozygosity. Picking similarity thresholds that result in no over- or under-splitting is not possible because every genome contains elements that will suffer over- or under-splitting at every threshold selected (Ilut *et al.* 2014). However, it is generally better to err on the side of under-splitting, because methods to identify and remove multi-locus contigs are more effective than those for identifying over-split loci (Ilut *et al.* 2014; Mastretta-Yanes *et al.* 2015; Willis *et al.* 2017). In addition, understanding differences between bioinformatic pipelines is critical to properly clustering the data. For example, Puritz et al. (*in prep*) found that rates of over-splitting vary between *dDocent, PyRAD, Stacks,* and *AftrRAD* across various combinations of parameters. Because effective thresholds for clustering will depend on the bioinformatic pipeline and vary by organism, enzyme(s), and dataset, researchers should test parameters to identify values where over-splitting is minimized.

3.2 Artifactual SNPs

Artifactual SNPs, those that do not exist in the actual genome but are called from the mapped reads, may be the result of erroneous read mapping/clustering, PCR error, and/or sequencing error. Because the rate of sequencing error varies by platform employed, chemistry and read length, the typical user cannot control all error introduced at this stage, therefore it is important to account for sequencing error during bioinformatic analysis. FASTQ-format sequence reads include PHRED-scale quality scores indicating the probability of a base call being correct. The quality score, Q, equals -10 $\log_{10} P$, with P being the probability of a base-calling error; for example, $Q = 30$ corresponds to the expectation that 1 in 1000 base-calls will be incorrect, i.e. the probability of a correct base call is 99.9%. Quality score can be used during bioinformatic processing to trim low-quality sections from the beginnings or ends of reads or to eliminate reads entirely. A quality score is also used by several variant callers, including *freebayes* and GATK (Depristo *et al.* 2011; Garrison & Marth 2012), to determine the probability of a SNP call being real or artifactual.

4 Filtering SNP data

Despite attempts to limit the introduction of technical artifacts during library preparation and bioinformatic processing, SNP data sets require rigorous filtering because the inclusion of only a few incorrectly genotyped loci in a data set can create a significant, misleading signal (Davey *et al.* 2013; Li & Wren 2014; Puritz *et al.* 2014b; Meirmans 2015). This is especially important for outlier detection because signal caused by genotyping error is likely to stand out in pattern and magnitude from the signal produced by the background SNP data. Full post-processing exploration of each dataset should include an evaluation of the quality of each locus and individual, the confidence in both SNP calls and genotypes, and the presence of multi-locus contigs. This should involve generating frequency distributions of parameters like missing data per locus and individuals, read depth, and heterozygosity to determine appropriate threshold values for these parameters. In addition, the comparison of multiple filtered data sets generated using different parameter values provides guidance for which combinations of thresholds retain the most loci while minimizing artifacts.

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

To illustrate the effects of various filtering strategies and parameter thresholds, we employed six different filtering schemes (FS) across four different data sets (Portnoy *et al.* 2015; Puritz *et al.* 2016; Hollenbeck *et al.* 2018; O'Leary *et al.* 2018). All data sets were created using the *dDocent* pipeline and differ in terms of the focal organism, type of reference used to map reads, the type of reads and the number of libraries sequenced (**Table 1**). The red snapper data (Puritz *et al.* 2016) set consists of previously published data that has been recalled against a fully

sequenced genome consisting of large contigs (154,064 contigs; N50 = 233,156 bp; total length 1.23 Gb). For all FS, we first filtered genotypes, loci and individuals and as a final step decomposed complex variants, retaining only SNPs. Details of full FS are available in **Table S1** and fully annotated scripts for filtering are available at https://github.com/sjoleary/SNPFILT.

4.1 Low quality loci versus low quality individuals:

Filtering parameters used to identify loci and individuals that did not sequence well include genotype call rate per locus (i.e. proportion of individuals a locus is called in) and missing data per individual, as well as genotype depth and the mean depth per locus, i.e. mean number of reads at a given locus across individuals. For data sets characterized by high levels of missing data (e.g. red snapper, **Figure 2**), applying hard thresholds can result in retaining little to no loci in the filtered data set. For example, for the red snapper data setting hard cut-offs retaining only loci with genotype call rates >95% and individuals with <25% missing data, leads to a final data set of only 10 SNPs on 3 contigs in 262 individuals (raw data set contains 1,106,387 SNPs on 25,168 contigs for 282 individuals, **Table S2**).

As an alternative strategy, starting with low cutoff values for missing data (per locus and individual) and iteratively and alternatively increasing them may result in more high-quality loci and individuals being retained. For example, in the red snapper data set first removing low confidence genotypes by filtering for minimum genotype read depth >5, SNP quality score >20, minor allele count >3, minimum mean read depth per locus >15 and then iteratively increasing the stringency of allowed missing data (final threshold values of a 95% genotype call rate and 25% allowed missing data per individual) results in 9,478 – 12,056 SNPs on 1,626 – 1,680 contigs and 187 – 189 individuals being retained (**Table S2**). This occurs because poor quality individuals tend to deflate genotype call rates in otherwise acceptable loci, and poor-quality loci increase missing data in otherwise acceptable individuals. Applying an iterative filtering strategy consistently results in more loci and individuals being retained, even in data sets consisting of individuals sequenced on the same sequencing run for which the initial distributions of missing data per locus and individuals are more favorable (**Figure 3**). For example, after removing low confidence loci from the flounder data set as described above and then setting a hard cutoff for a genotype call rate of >95% and allowed missing data per individual of <25% results in a data set consisting of 15,682 SNPs on 3,802 contigs over 170 individuals, while iterative filtering results

in data sets consisting of 18,663 – 24,103 SNPs on 4,789 – 5,341 contigs over 164 – 167 individuals (**Table S2**).

4.2 Confidence in SNP identification

The ability to filter loci depends on the pipeline used to reconstruct and genotype loci and the set of parameters reported. As discussed previously variant callers such as *freebayes* report PHRED-like quality scores for variants (SNPs) indicating the confidence in the SNP call being correct. Further, users often choose to set a minor allele count to remove potentially artifactual SNP calls. For example, a minor allele count of three requires an allele to be observed in at least two individuals (homozygote and heterozygote). Unfortunately, this strategy will also remove true rare alleles from the data set. For the paired-end data sets examined here, the number of SNPs in the final data set increases from 22 – 27% when no filter for minor allele count is applied, relative to a minor allele count cutoff of three. For the single-end data the difference is considerably greater, as 58% more SNPs are present in the final data set when no filter for minor allele count is applied. The number of SNP-containing loci in the final paired-end data sets is slightly larger when no filter for minor alleles is applied (2 – 3% increase), but considerably larger (46% increase) for the single-end data. The result is due to the number of SNPs per contig increasing from 3.3 – 6.6 to 5.8 – 7.2 for the paired end data, while remaining consistent (1.8 vs 1.9) for single-end data, when no minor allele filter was applied. It is important to note that this could also be related to the focal organism, as the single-end data comes from a shark and the paired-end data come from bony fishes and the taxa likely vary in number of expected polymorphisms within 300 bp fragments (**Table 2**).

4.3 Confidence is genotypes: allele dropout/coverage effects

While artifactual SNPs as described above will result in genotyping error (individuals called heterozygous for alleles that do not exist), genotyping error at real SNPs may also occur. Allele dropout and coverage effects can lead to unsampled alleles and individuals incorrectly genotyped as homozygotes. Whereas coverage effects can be technically mitigated by ensuring that each individual receives adequate reads, allele dropout is an unavoidable artifact of using restriction enzymes and size selection during library preparation. For targeted fragments to be amplified and sequenced, adapters must be correctly ligated to the “sticky” ends left by the enzymes, but polymorphisms may occur in the enzyme recognition site (cut-site polymorphisms) resulting in alleles that are not cut by the restriction enzymes. Similarly, length polymorphisms

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

Allele dropout cannot be avoided by optimizing standard laboratory procedures, but can be accounted for during filtering by removing genotypes below a certain threshold of minimum reads, and by identifying loci with high variance in read depth among individuals (Davey *et al.* 2013; Cooke *et al.* 2016). Low coverage can result in false homozygotes because the number of reads may not high be enough to successfully call both alleles. Loci can be filtered based on a threshold of minimum mean depth per locus and users can code individuals' genotypes at specific loci as missing if they fall below a minimum depth threshold that reflects the number of reads required to confidently call homozygotes. This increases the confidence in individual genotypes, and results in the removal of loci that consistently have genotypes not called with high confidence across individuals. Lastly, it is important to consider the statistical model being used for variant calling, and how the model relates to read depth. For example, *freebayes* and GATK (Depristo *et al.* 2011; Garrison & Marth 2012) are Bayesian callers that integrate data across all samples when determining genotypes, meaning lower read-depth genotypes can be called with greater accuracy. This is in contrast to genotyping models implemented in STACKS or PyRAD (Catchen *et al.* 2011; Eaton 2014) which genotype individuals one at a time without the ability to integrate data across samples until genotyping is completed. In addition, when deviations from Hardy-Weinberg proportions are not expected, χ^2 tests of Hardy-Weinberg expectations for individual loci within demes can also indicate heterozygote deficits that may indicate allele dropout.

4.4 Identification of multi-locus contigs

Multi-locus contigs can be identified by assessing distributions of read depth, excess heterozygosity, and the number of haplotypes observed per each individual at each marker (Ilut *et al.* 2014; Li & Wren 2014; Willis *et al.* 2017). In general, total or mean read depth per locus should be approximately normally distributed. Loci with coverage falling well above this distribution may be reads clustered or mapped from multiple loci. Loci with excess coverage are best identified by generating a frequency distribution of coverage and choosing thresholds, for example, two times the mode (Willis *et al.* 2017) or the 90th quantile (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters; **Figure 4**). Appropriate thresholds will vary between datasets and species. Because fixed or near-fixed differences may

exist between non-orthologous loci, multi-locus contigs often have an excess number of heterozygotes (Hohenlohe *et al.* 2011; Willis *et al.* 2017). *VCFtools* (Danecek *et al.* 2011) provides a statistical framework for assessing heterozygote-excess via a χ^2 test of Hardy-Weinberg expectations for VCF files. Finally, reads in multi-locus contigs often exhibit more than two haplotypes per individual, and therefore markers can be filtered based on a threshold for the number of individuals with excess haplotypes (Ilut et al. 2014, Willis et al. 2017). While each of these filters applied alone may catch many or even the majority of multi-locus contigs, the most effective strategy to remove multi-locus contigs appears to be applying each filter in parallel and removing markers flagged by any of the three filters (Willis et al. 2017).

4.5 INFO-flag filtering of vcf files

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

Allele balance (AB) compares the number of reads for the reference allele to the number of reads for the alternate allele across heterozygotes. The expected allele balance is 0.5; large deviations may indicate false heterozygotes due to coverage effects, multi-locus contigs, etc. **Figure 5** shows AB for a raw data set, and for data sets that have been filtered for low quality genotypes, loci and individuals. In both unfiltered and filtered data sets, loci with high/low AB are present, indicating that problematic loci will remain unless AB is explicitly filtered for.

Reads supporting either allele in a heterozygote should have similar mapping quality values, the ratio of mapping quality between alleles, therefore, should be approximately one. Systematically large discrepancies between the mapping quality for the reference and alternate alleles at a SNP may indicate read-mapping errors or multi-locus contigs. Standard filtering steps do not remove all loci with biased mapping quality ratios (**Figure 6**). Assessing mapping quality

ratios has the added benefit that it can help to identify minor alleles that are not true alleles (**Figure 6B**), allowing researchers to retain true minor alleles that may contain an important biological signal.

For paired-end libraries, artifacts can also be identified by examining the properly paired status of reads and potential strand bias. The forward and reverse reads of a known pair should always map to the same contig; improper read paring, in which forward and reverse reads of a known pair map to different contigs, indicates mapping anomalies such as multi-copy or improperly assembled loci. Strand bias describes the relationship between forward and reverse reads and SNP-calls at a given locus. For most paired-end RADseq libraries, the forward and reverse reads do not overlap because the actual RAD fragments will be too long. For example a 350 bp RAD fragment characterized with 125 bp pair-end reads will have 100 bp of uncharacterized, intervening sequence. Therefore, a given SNP should only be apparent on either the forward or reverse read. Calls of the same SNP in in both forward and reverse reads often indicate mapping anomalies. However, the implications of this criterion depend on read length and fragment length, and therefore the expected overlap of paired reads in a given data set.

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

5. Physical linkage

After filtering, most RADseq data sets will generally contain sets of SNPs located on the same contig. SNPs located within a few hundred base pairs of each other are generally physically linked (Miyashita & Langley 1988; Hohenlohe *et al.* 2012), whereas most commonly used analyses assume that all genetic markers are independent. Treating physically linked SNPs as independent markers provides biased results, including false signals of population structure. A common method to remove this bias is to retain only one SNP from each contig ("thinning").

This is an appropriate strategy but one that reduces the information content of a given marker if multiple SNPs are contained on a single contig. Another way to deal with physical linkage is to infer haplotypes for each contig based on the combination of filtered SNPs within paired reads (Willis et al. 2017). This strategy will produce the same number of markers as thinning, but many markers will be multi-allelic, therefore, haplotyping manages physical linkage while preserving the total information content of the data set.

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

With the shift from data sets consisting of markers for tens to hundreds of microsatellite loci to several thousand SNP-containing loci, bioinformatic processing has become the only viable means of ensuring data quality. Many studies currently report very few details pertaining to quality control methods applied to the of output from SNP calling pipelines beyond very basic filtering, i.e. removal of markers and/or individuals with low coverage or high levels of missing data. Nevertheless, it is incumbent upon the author to document data preparation and quality control steps and make these available to the scientific community to ensure that data analyses are transparent and fully reproducible (Peng 2014; Leek & Peng 2015). A related problem that enables this under-reporting, however, is a lack of clear quality control standards.

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

Acknowledgements

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

References

Andrews KR, Hohenlohe PA, Miller MR *et al.* (2014) Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. *Molecular Ecology*, **23**, 5943–5946.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, 1–7.

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

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks : Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes|Genomes|Genetics*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: An analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Cooke TF, Yee MC, Muzzio M *et al.* (2016) GBStools: A statistical method for estimating allelic dropout in reduced representation sequencing data (WA Cresko, Ed,). *PLoS Genetics*, **12**, e1005631.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Davey JL, Blaxter MW (2010) RADseq: Next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Depristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–501.

Eaton DAR (2014) PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *Plos One*, **11**, e0151651.

Gautier M, Gharbi K, Cezard T *et al.* (2012) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Graham CF, Glenn TC, Mcarthur AG *et al.* (2015) Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources*, **15**, 1304–1315.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 395–408.

Hollenbeck CM, Portnoy DS, Puritz JB, Samollow P, Gold JR (2018) Fine-scale population structure and genomic islands of divergence in a coastal marine fish, red drum (Sciaenops ocellatus). *Molecular Ecology*.

Ilut DC, Nydam ML, Hare MP (2014) Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering. *BioMed Research International*, **2014**.

Leek JT, Peng RD (2015) Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, **112**, 1645–1646.

Li H, Wren J (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology*, **24**, 3223–3231.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Miyashita N, Langley CH (1988) Molecular and phenotypic variation of the white locus region in Drosophila melanogaster. *Genetics*, **120**, 199–212.

O'Leary SJ, Hollenbeck Christopher M. Vega RR, Gold JR, Portnoy DS (2018) Comparative genomics as a tool for restoration enhancement and culture of southern flounder, Paralichthys lethostigma. *BMC Genomics*.

Peng RD (2014) Reproducible Research in Computational Science. *Science*, **1226**.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**.

Portnoy DS, Puritz JB, Hollenbeck CM *et al.* (2015) Selection and sex-biased dispersal: the influence of philopatry on adaptive variation. *PeerJ*, 1–20.

Puritz JB, Gold JR, Portnoy DS (2016) Fine-scale partitioning of genomic variation among recruits in an exploited fishery: causes and consequences. *Scientific Reports*, **6**, 36095.

Puritz JB, Hollenbeck CM, Gold JR (2014a) dDocent : a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **2**, e431.

Puritz JB, Hollenbeck CM, Gold JR (2015) Fishing for Selection, but Only Catching Bias: Examining Library Effects in Double-Digest RAD Data in a Non-Model Marine Species. In: *Plant and Animal Genome XXIII Conference*

Puritz JB, Matz M V., Toonen RJ *et al.* (2014b) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

Schmid S, Genevest R, Gobet E *et al.* (2017) HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution*.

Sovic MG, Fries AC, Gibbs HL (2015) AftrRAD: a pipeline for accurate and efficient de novo assembly of RADseq data. *Molecular Ecology Resources*, **15**, 1163–1171.

Suchan T, Pitteloud C, Gerasimova NS *et al.* (2016) Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS ONE*, **11**, e0151651.

Toonen RJ, Puritz JB, Forsman ZH *et al.* (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.

Wang S, Meyer E, McKay JK, Matz M V (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, **9**, 808–810.

Willis SC, Hollenbeck CM, Puritz JB, Gold JR, Portnoy DS (2017) Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources*.

Data Accessibility

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

Figures and Tables

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

Figure 2: Missing data per locus/individual (indv) for unfiltered red snapper data set (A, B) and after coding genotypes with <5 reads as missing, and removing low quality loci with SNP quality

score <20 and minimum mean depth <15 reads (C, D). Red dashed line indicates mean proportion of missing data.

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

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

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

Figure 6: Ratio of mean mapping quality scores for the reference and alternate allele for southern flounder data set. (A) Genotypes with <5 reads have been coded as missing and loci with SNP quality score <20, mean read depth <15 reads, >30% missing data and/or and minor allele count of <3 removed; (B) same data set without applying minor allele count filter. Red dashed line indicates loci with mapping quality ratio of 1, i.e. the further away the larger the discrepancy between the mapping quality of the reference and alternate allele. Blue dashed lines indicate cut-off values for ratio of mean mapping quality score of 0.25 and 1.75 (alternate to reference allele) as implemented in dDocent_filters (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) to remove loci with high

discrepancy of mapping quality for the alleles of a given locus (indicated in red below the dashed line).

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

Table 1: Comparison of four published ddRAD data sets compiled using the dDocent pipeline. (A) Comparison of sequencing type used to create reference and call genotypes, the number of combined libraries, approximate genome size, and enzymes used to fragment DNA. All data sets were run on the Illumina platform to obtain either paired end (PE) or single end (SE) reads.

Table 2: Comparison of the number of SNPs per reference contig for four ddRAD data sets after applying/not applying a minor allele count (mac) filter (filtering schemes 5/6) and the percent more SNPs and contigs remaining in the data set when not explicitly filtering for minor alleles. All libraries were sequenced on the Illumina platform to obtain either paired-end (PE) or single-end (SE) reads.

Supplementary Information

Table S1: Detailed description of six different filtering schemes applied to example data sets. Applied filters are designed to remove loci with low confidence SNP calls (minimum genotype read depth (minDP), SNP quality score (qual), mean read depth per locus across all individuals (meanDP), minor allele count (mac), missing data (allowed missing data per individual (imiss), genotype call rate (number of individuals that have been called for a given locus (geno)) and INFO-filters as described in the manuscript.

Table S2: Comparison of the number of SNPs, contigs (cont) and individuals (indv) in the raw data sets and retained in each data set for six different filtering schemes (FS) as described in Table S1.

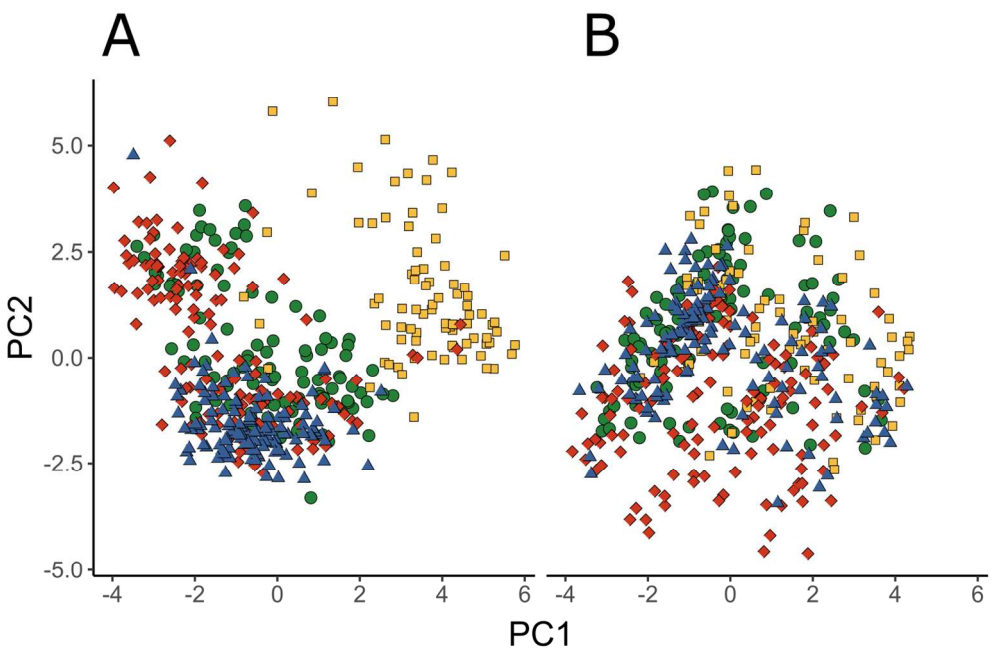

Figure 1: Library effects (adapted from Puritz et al. 2015). PCA of RAD data set combining four libraries (yellow squares, red diamonds, blue triangles, green circles) before (A) and after (B) correcting for library effects by removing affected markers.

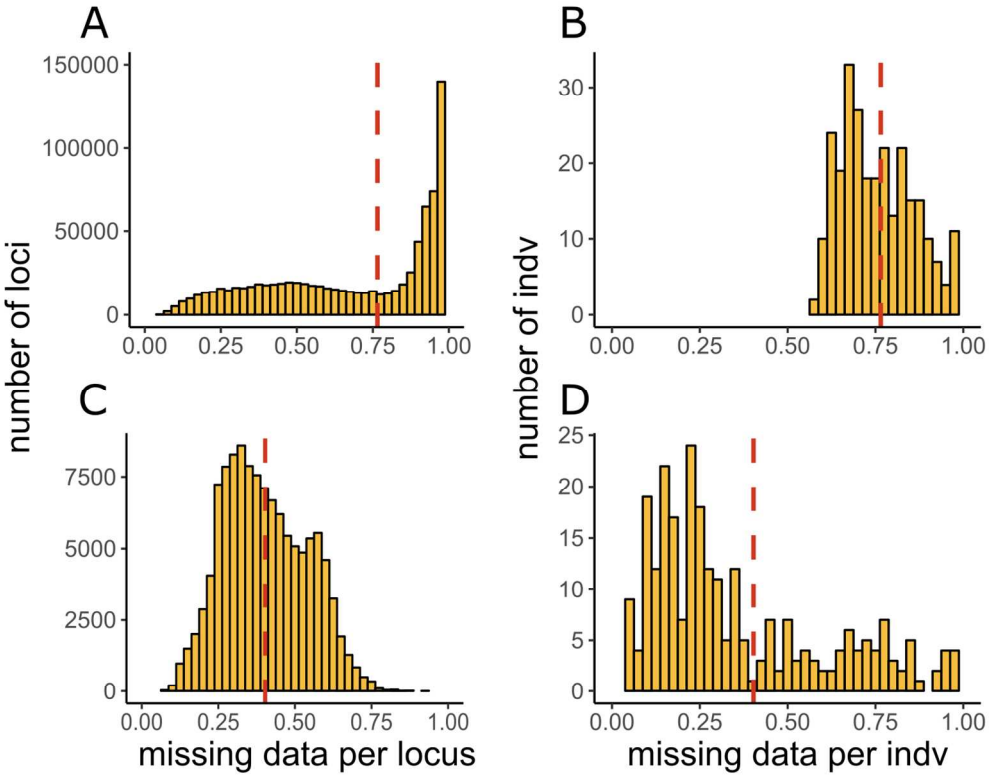

Figure 2: Missing data per locus/individual (indv) for unfiltered red snapper data set (A, B) and after coding genotypes with <5 reads as missing, and removing low quality loci with SNP quality score <20 and minimum mean depth <15 reads (C, D). Red dashed line indicates mean proportion of missing data.

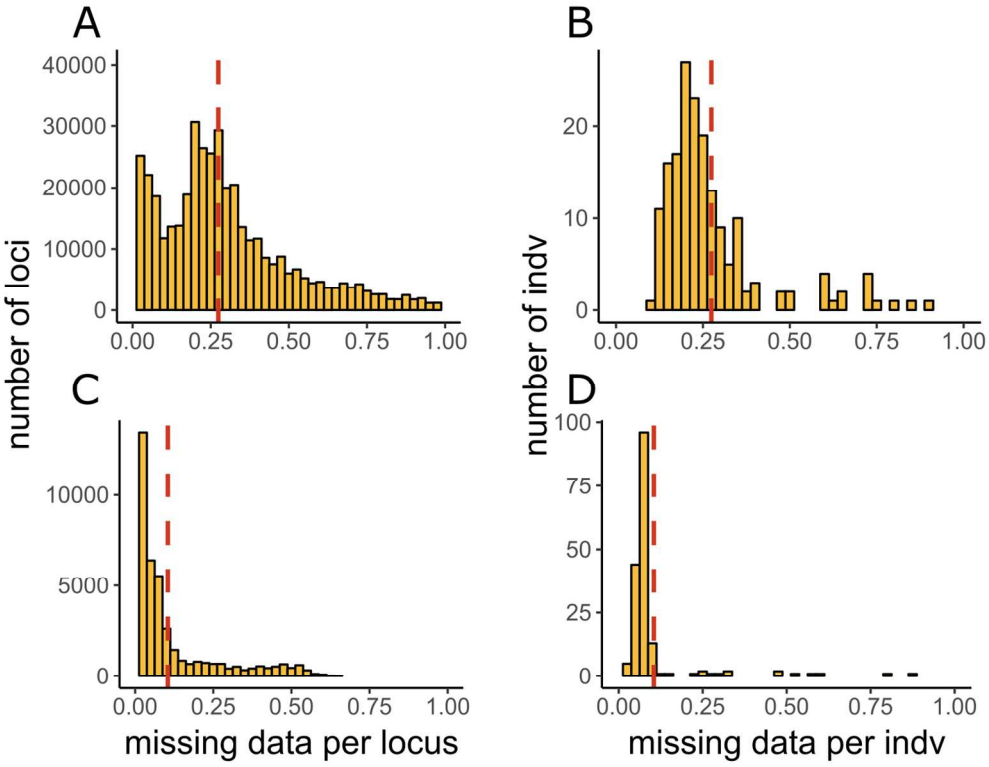

Figure 3: Missing data per locus and individual for unfiltered southern flounder data set (A, B) and after coding genotypes with <5 reads as missing, and removing low quality loci with SNP quality score <20 and minimum mean depth <15 reads (C, D). Red dashed line indicates mean proportion of missing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

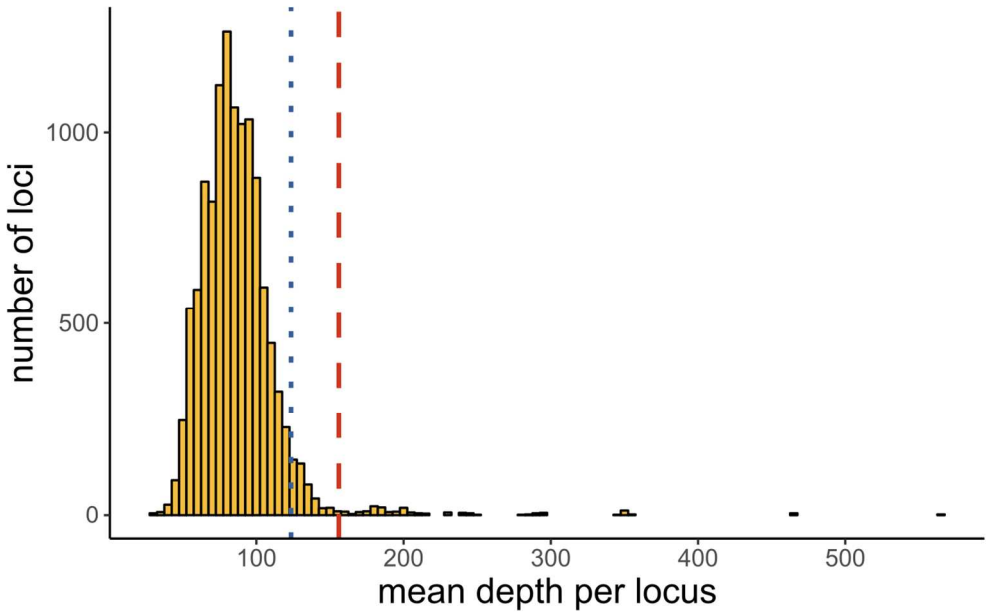

Figure 4: Distribution of mean depth per locus across all loci for red snapper data set after removing low confidence/quality loci (minimum genotype depth >3, SNP quality score >20, minor allele count >3, mean minimum depth across all individuals >15), and iterative filtering of missing data to final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

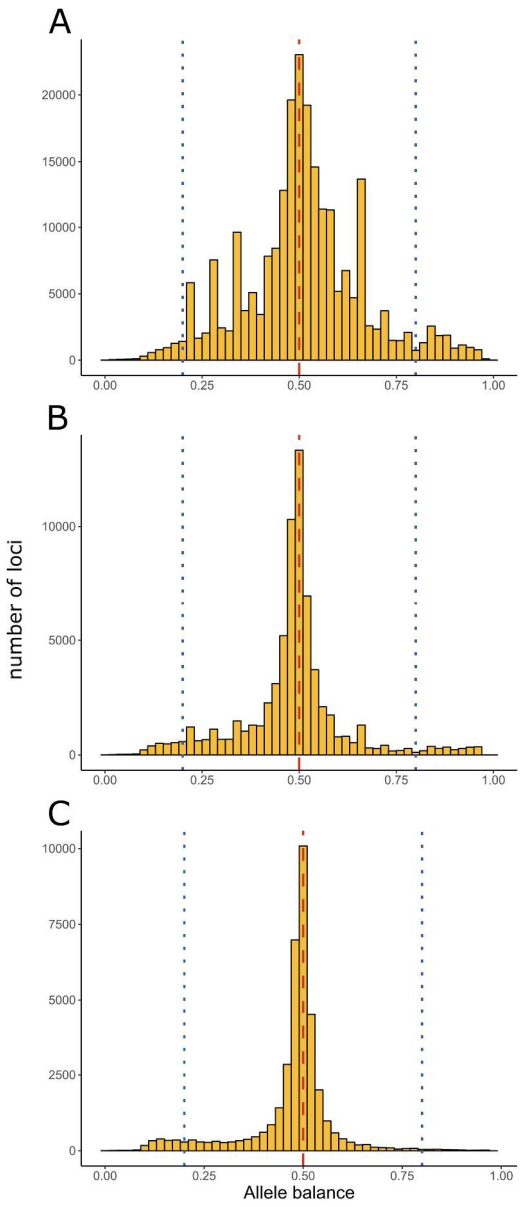

Figure 5: Allele balance in heterozygous genotypes (proportion of reads corresponding to the reference allele) for (A) unfiltered red drum data set, (B) data set with genotype read depths <3 reads coded as missing and loci with SNP quality score <20, mean depth <15 reads and/or >30% missing data removed,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

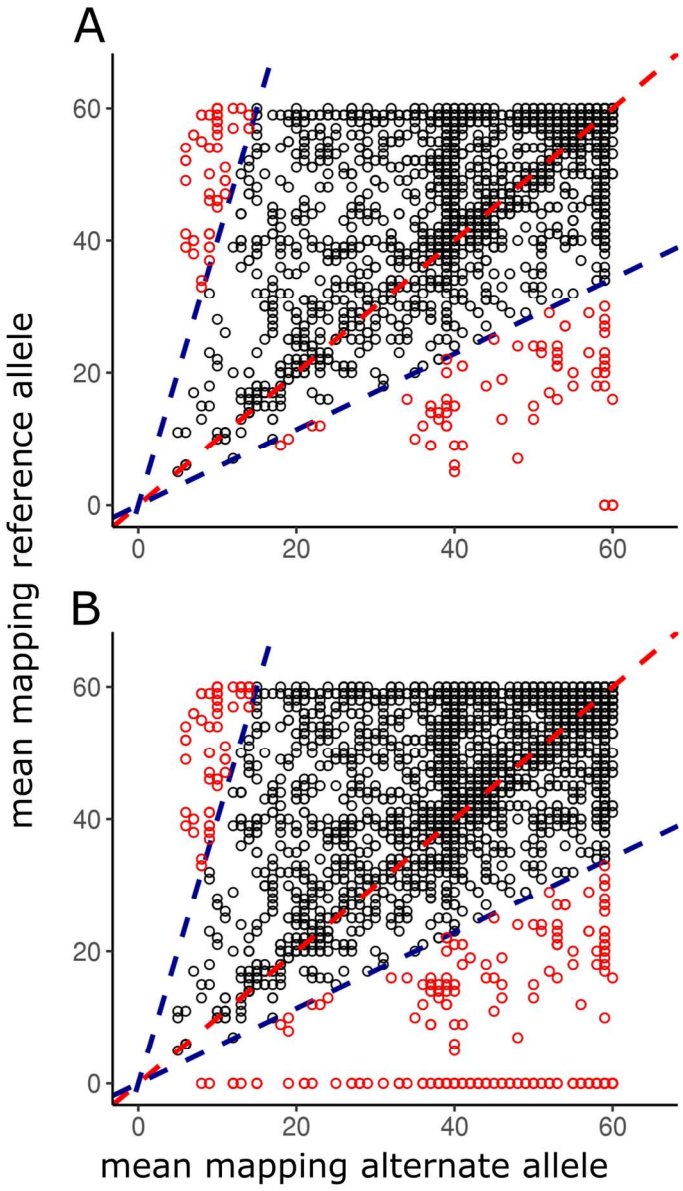

Figure 6: Ratio of mean mapping quality scores for the reference and alternate allele for southern flounder data set. (A) Genotypes with <5 reads have been coded as missing and loci with SNP quality score <20, mean read depth <15 reads, >30% missing data and/or and minor allele count of <3 removed; (B) same data set without applying minor allele count filter. Red dashed line indicates loci with mapping quality ratio of 1, i.e. the further away the larger the discrepancy between the mapping quality of the reference and alternate allele. Blue dashed lines indicate cut-off values for ratio of mean mapping quality score of 0.25 and 1.75 (alternate to reference allele) as implemented in dDocent_filters (https://github.com/jpuritz/dDocent/blob/master/scripts/dDocent_filters) to remove loci with high discrepancy of mapping quality for the alleles of a given locus (indicated in red below the dashed line).

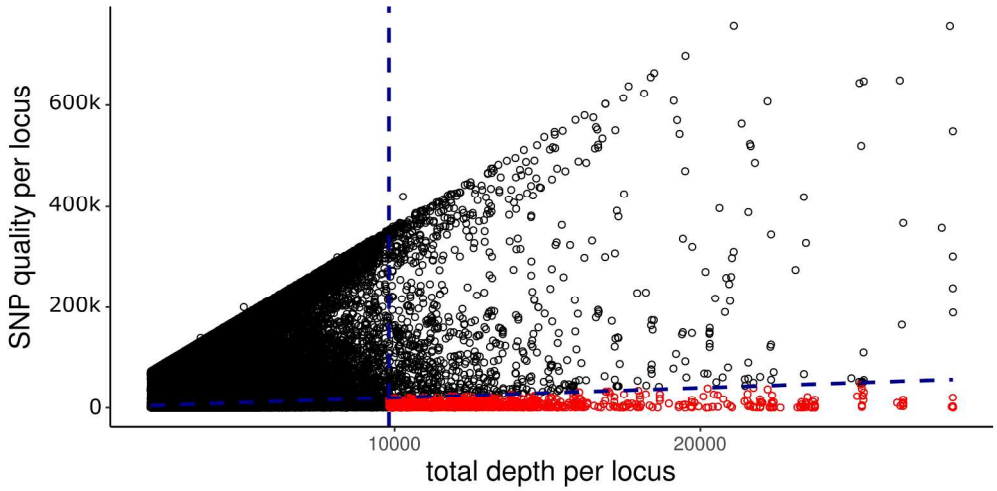

Figure 7: Comparison of SNP quality score and total depth per locus for the bonnethead shark data set. Vertical blue dashed line identifies loci with high depth (mean + 1 standard deviation). Loci with a quality score <2x the depth at that locus are below the diagonal blue dashed line (indicated in red).

	Red drum	**Red snapper**	**Southern flounder**	**Bonnethead sharks**
Reference	Illumina MiSeq PE	genome (scaffold)	Illumina MiSeq PE	Illumina HiSeq PE
Sequencing	Illumina HiSeq PE	Illumina HiSeq PE	Illumina HiSeq PE	Illumina HiSeq SE
Number of libraries	4	4	1	2
Genome size	~ 1 Gb	~ 1.4 Gb	~ 0.6 Gb	~ 3.5 Gb
Enzymes	EcoRI/MspI	EcoRI/MspI	EcoRI/MspI	EcoRI/MspI

data set	reference contigs	% more SNPs	% more contigs	SNPs per contig (mac > 0/3)
Red drum	PE MiSeq (300 bp*)	22.9	3.4	6.6/3.3
Southern flounder	PE MiSeq (300 bp*)	25.7	9.7	4.5/3.9
Red snapper	Genome assembly	27.2	2.2	7.2/5.8
Bonnethead shark	PE HiSeq (300 bp**)	58.9	46.5	1.9/1.8

*300 bp forward and reverse reads overlap

**300 bp forward and reverse reads do not overlap.