Received: 2 April 2019 | Revised: 12 August 2019 | Accepted: 16 August 2019

DOI: 10.1111/1755-0998.13087

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Attack of the PCR clones: Rates of clonality have little effect on RAD-seq genotype calls

Peter T. Euclide[1] | Garrett J. McKinney[2] | Matthew Bootsma[1] | Charlene Tarsa[3] | Mariah H. Meek[3] | Wesley A. Larson[4]

[1]Wisconsin Cooperative Fishery Research Unit, College of Natural Resources, University of Wisconsin-Stevens Point, Stevens Point, WI, USA

[2]School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA

[3]Department of Integrative Biology and AgBio Research, Michigan State University, East Lansing, MI, USA

[4]U.S. Geological Survey, Wisconsin Cooperative Fishery Research Unit, College of Natural Resources, University of Wisconsin-Stevens Point, Stevens Point, WI, USA

Correspondence
Peter T. Euclide, Wisconsin Cooperative Fishery Research Unit, College of Natural Resources, University of Wisconsin-Stevens Point, Stevens Point, WI 54481, USA.
Email: peter.euclide@uwsp.edu

Abstract

Interpretation of high-throughput sequence data requires an understanding of how decisions made during bioinformatic data processing can influence results. One source of bias that is often cited is PCR clones (or PCR duplicates). PCR clones are common in restriction site-associated sequencing (RAD-seq) data sets, which are increasingly being used for molecular ecology. To determine the influence PCR clones and the bioinformatic handling of clones have on genotyping, we evaluate four RAD-seq data sets. Data sets were compared before and after clones were removed to estimate the number of clones present in RAD-seq data, quantify how often the presence of clones in a data set causes genotype calls to change compared to when clones were removed, investigate the mechanisms that lead to genotype call changes and test whether clones bias heterozygosity estimates. Our RAD-seq data sets contained 30%–60% PCR clones, but 95% of RAD-tags had five or fewer clones. Relatively few genotypes changed once clones were removed (5%–10%), and the vast majority of these changes (98%) were associated with genotypes switching from a called to no-call state or vice versa. PCR clones had a larger influence on genotype calls in individuals with low read depth but appeared to influence genotype calls at all loci similarly. Removal of PCR clones reduced the number of called genotypes by 2% but had almost no influence on estimates of heterozygosity. As such, while steps should be taken to limit PCR clones during library preparation, PCR clones are likely not a substantial source of bias for most RAD-seq studies.

1 | INTRODUCTION

Advances in high-throughput sequencing technology have made it possible to collect terabytes of genomic data in systems where significant funding opportunities are not available, such as for non-model organisms and organisms of conservation concern (reviewed in Allendorf, Hohenlohe, & Luikart, 2010; Bernatchez et al., 2017; Hohenlohe, Amish, Catchen, Allendorf, & Luikart, 2011). As a result, an increasing number of researchers, many without substantial bioinformatic experience, are finding themselves with large genomic data sets (Willette et al., 2014). Many previous reviews have stressed the importance of exploring genomic data sets and parameter space to avoid potential sources of bias (Hendricks et al., 2018; Meirmans, 2015; O'Leary, Puritz, Willis, Hollenbeck, & Portnoy, 2018). However, fully exploring every data set is impossible. Therefore, understanding which data processing steps have the largest impact on the final results can help to achieve a balance between efficiency and data quality (Mastretta-Yanes et al., 2015; Paris, Stevens, & Catchen, 2017).

One of the most commonly used methods to generate genomic data in nonmodel organisms is restriction site-associated DNA sequencing (RAD-seq). RAD-seq has become a method of choice because it is capable of targeting a subset of the genome, thereby substantially reducing the overall costs of sequencing

without sacrificing depth of coverage. Analysis methods for RAD-seq have been refined over the last half decade, and numerous studies have explored how varying analysis parameters can influence final data sets (Fountain, Pauli, Reid, Palsbøll, & Peery, 2016; Mastretta-Yanes et al., 2015; Paris et al., 2017). For example, Paris et al. (2017) explored the parameter space associated with the analysis program STACKS (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Rochette & Catchen, 2017) and isolated a subset of variables that had the largest influence on the number of retained loci. However, the potential biases associated with many aspects of RAD-seq have yet to be explored. One potential source of bias in RAD-seq and other genomic methods that incorporate PCR amplification is the influence of PCR clones (also known as PCR duplicates) on genotype calls (Li & Wren, 2014). PCR clones (i.e. any two or more sequences originating from the same template fragment of DNA) are discussed in at least five major reviews of RAD-seq as an important consideration during library preparation and analysis (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Andrews & Luikart, 2014; Davey et al., 2013; Flanagan, Forester, Latch, Aitken, & Hoban, 2018; Puritz et al., 2014). RAD-seq protocols and analysis pipelines have been amended to incorporate steps that reduce, flag or remove clones, but the impact of PCR clones on final data sets has not been quantified.

The primary concerns associated with PCR clones are as follows: (a) an inflation of homozygosity caused by preferential amplification of one allele in a heterozygote (allelic drop out), (b) misidentification of PCR errors as SNPs leading to erroneous heterozygote calls and (c) artificial inflation of sequencing depth caused by retaining PCR clones, resulting in overconfidence in genotype calls (Andrews et al., 2016; Davey et al., 2013). While in theory, all of these possibilities are important, few studies have evaluated how removing PCR clones from actual data sets influences downstream genotype calls and diversity estimates (but see Díaz-Arce & Rodríguez-Ezpeleta, 2019). Ebbert et al. (2016) evaluated whether PCR clone removal improved variant calls in whole-genome sequencing data and found that 99.9% genotype calls were identical between SNP chip data, unfiltered next-generation sequence data and clone-filtered sequence data. Therefore, while genotyping errors associated with PCR clones do occur, Ebbert et al. (2016) concluded the actual cost these errors have on genotype call accuracy is likely low compared with the cost of time and data lost by conducting PCR clone removal. However, the data set analysed in Ebbert et al. (2016) only contained 1%–2% PCR clones, whereas many RAD-seq data sets contain 30% or more (e.g. Schweyen, Rozenberg, & Leese, 2014). Flanagan and Jones (2018) evaluated how PCR clones differed between single- and double-digest RAD protocols and still found limited evidence that PCR clones biased study results but were unable to directly evaluate how PCR clones influenced genotype calls in the same data set. Finally, Díaz-Arce and Rodríguez-Ezpeleta (2019) found that estimates of F_{ST} and variance in the number of loci genotyped per individual differed somewhat between STACKS catalogs built with and without PCR clones in RAD-seq data sets containing between 27% and 58% PCR clones but did not evaluate how genotype calls changed as a result of PCR clones.

Most RAD-seq library preparations require at least some PCR amplification (5 to 15 cycles) and will therefore include some level of PCR clones. To identify PCR clones, it is currently necessary to use paired-end sequencing and a RAD-seq protocol that employs either random shearing or degenerate primers to track unique DNA molecules and the amplicons that arise from these molecules (Andrews et al., 2016; Schweyen et al., 2014). For example, the original RAD-seq protocol uses a single restriction enzyme to cut DNA followed by random shearing to reduce fragment length (Baird et al., 2008). The random shearing step results in variable template molecule length so any molecules of identical length and sequence are presumed to be PCR clones. RAD-seq protocols that use two restriction enzymes to normalize template length, such as ddRAD or ezRAD, make it impossible to identify PCR clones in this way because all fragments are of identical length (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012; Toonen et al., 2013). Schweyen et al. (2014), however, developed an additional step for the ddRAD protocol that uses a degenerate base region to identify clones based on variation in the fragment length of template molecules.

Currently, the primary ways to address PCR clones in RAD-seq data are to (a) minimize the number of PCR cycles used in library preparation and (b) identify and remove all clone reads after sequencing. Once PCR clones are identified, most protocols call for discarding all but a single copy of each unique molecule. In theory, this should eliminate most PCR bias (Davey et al., 2013). However, discarding PCR clones without fully understanding whether they create bias may result in less robust data sets. For example, when the level of clonality is high, removing clones can translate into discarding a large portion of data, and therefore a large and potentially unnecessary reduction in depth of coverage. Schweyen et al. (2014) found that average coverage dropped by approximately a third once PCR clones were removed from their data. Additionally, for methods such as ddRAD where additional steps in the protocol must be added to incorporate degenerative primers that allow for the detection of PCR clones, PCR clone removal may unnecessarily consume time and resources (Schweyen et al., 2014). Finally, there is no consensus on whether removing clones is necessary, and many studies already exist that could not or did not account for PCR clones. Therefore, the future comparability and application studies that address PCR clones differently depends on a formal classification of PCR clone bias.

While PCR clones certainly affect read depth, and potentially magnify the effects of allelic dropout, the impact of PCR clones on downstream genotype calls and analyses is unclear. Here, we compare four RAD-seq data sets generated using the library preparation method outlined in Ali et al., 2016 to (a) evaluate clonality across libraries and individuals, (b) identify how PCR clones influence genotype calls, (c) investigate why genotype calls change after PCR clones are removed and (d) determine whether removing PCR clones influence estimates of heterozygosity. We found that PCR clones represent a relatively large portion of overall reads (20%–60%) but

removing PCR clones changed less than 6% of genotype calls, and primarily resulted in an increased amount of missing genotype calls. We conclude that typical conservation and population genetic RAD-seq studies are unlikely to be biased by PCR clones, but suggest that researchers investigate clonality in their own data sets when possible.

2 | METHODS

We compiled RAD-seq data from four species (brook trout [*Salvelinus fontinalis*], chum salmon [*Oncorhynchus keta*], cisco [*Coregonus artedi*] and walleye [*Sander vitreus*]) generated separately in three different laboratories using the *SbfI* enzyme, methods outlined in Ali et al. (2016), and 11 (brook trout) to 12 (chum salmon, cisco and walleye) PCR cycles. All data sets were collected for ongoing population genomic studies, and only a subset of the data was used for the present analysis. Brook trout libraries were prepared in the Genomic Variation Lab at the University of California—Davis and sequenced on Illumina NextSeq 500 (PE 75 bp reads, 96 samples/lane) at the Cornell Institute of Biotechnology, chum salmon libraries were prepared in the Seeb Laboratory at the University of Washington and sequenced on a HiSeq 4000 (PE 150 bp reads, 96 samples/lane) at the University of Oregon Genomics and Cell Characterization Core Facility (McKinney, McPhee, Pascal, Seeb, & Seeb, 2019), and walleye and cisco libraries were prepared in the Larson Laboratory at the University of Wisconsin-Stevens Point and sequenced on a HiSeq 4000 (PE 150 bp reads, 96 samples/lane for cisco and 192 individuals/lane for walleye) at the Michigan State Genomics Core Facility. Two rounds of sequencing were conducted for chum salmon, and the volume of DNA for each individual was adjusted in the second round of sequencing to reduce variation in sequence reads per individual (Larson et al., 2017). The random shearing step of library preparation was conducted using a sonicator for the chum salmon and brook trout libraries and fragmentase enzyme for the cisco and walleye libraries (NEBNext® dsDNA Fragmentase®, New England BioLabs, M0348L).

2.1 | SNP identification

SNP identification and genotyping were conducted in STACKS 1.46 (brook trout, cisco, walleye) or 1.47 (chum salmon) following similar procedures and flags (Catchen et al., 2011). Samples were demultiplexed with *process_rad-tags* (flags = c, -q, -r, -t 140 [-t 65 for brook trout]), and PCR clones were filtered from individual files by running the *clone_filter* step in STACKS with default settings. The *clone_filter* step works by searching for identical sequences in paired-end reads to identify fragments that are the same size and likely to be PCR clones. Because the Ali et al. (2016) protocol utilizes random shearing, it is unlikely two copies of a particular tag will have identical lengths; therefore, identical tags are presumed to be PCR clones, and all but a single copy are removed. After *clone_filter*, all remaining STACKS steps were conducted on both the non-clone-filtered (hereafter referred to as unfiltered) data set and the data set with clones removed (hereafter referred to as filtered data set). Stacks of similar sequences (loci) for each individual were identified with *ustacks* (flags = -m 3, -M 5, -H --max_locus_stacks 3, --model_type bounded, --bound_high 0.05 for cisco, walleye and brook trout, and 0.01 for chum salmon), and a catalog was created using a subset of individuals (brook trout = 31, cisco = 29, walleye = 60, chum salmon = 36) with *cstacks* (-n of 2 for chum salmon, 3 for walleye and cisco and 4 for brook trout). Putative SNPs within each individual were then matched against the catalog with *sstacks*. Finally, all genotype calls were output as VCF files with *populations* (flags = -r 0.3, --min_maf 0.05) and all samples in a given data set grouped as a single population in the popmap file.

2.2 | Locus selection strategy

To ensure that loci used for the core analysis reflected the type and quality of loci that would generally be used in a RAD-seq study, low quality or otherwise problematic SNPs were removed from both the clone-filtered and clone-unfiltered data sets prior to analysis. Therefore, we removed potential paralogs, low-quality individuals and loci that were not shared between filtered and unfiltered data sets, or could not be genotyped consistently. Paralogs are prevalent in salmon genomes, do not conform to diploid expectations and cannot be genotyped consistently with RAD-seq data (Allendorf et al., 2015; McKinney, Waples, Pascal, Seeb, & Seeb, 2018). To ensure that paralogs did not influence our findings, we ran HDPlot on unfiltered data for all species and removed loci identified as potential paralogs (McKinney, Waples, Seeb, & Seeb, 2017). Parameters for this analysis were set for each species by visually choosing threshold values for read depth ratio and proportion of heterozygotes that identified the loci conforming to theoretical expectations for singletons (McKinney et al., 2017). We then removed any loci that were not genotyped in both filtered and unfiltered data sets and removed individuals that had genotype calls at less than 30% of loci in either data set. Finally, any locus that failed to genotype in >50% of individuals in either data set was removed. Because STACKS determines reference alleles based on the first nucleotide identified during *ustacks*, references sometimes differed between filtered and unfiltered data sets. To correct this, all reference alleles that differed were converted to the unfiltered data set references.

2.3 | Objective 1: clone summary

The per cent of sequences identified as PCR clones in each individual was determined by dividing the total number of clones identified with *clone_filter* by the total number of retained reads. This clonality rate was then used as the response variable in analyses of variance (ANOVAs) to determine whether clonality varied among species or sequencing libraries within species. Any significant differences ($\alpha = 0.05$) were investigated using Tukey's honestly significance test calculated in R. Next, we summarized the level of clonality per RAD-tag by plotting the distribution of clones identified per RAD-tag for

each individual and calculating the per cent of molecules that contained 5 or fewer clones.

2.4 | Objective 2: effect of PCR clones on genotype calls

The influence that PCR clones have on genotype calls was evaluated by comparing genotypes between unfiltered and filtered data sets for each species. First, genotype call changes were identified by looking for differences in all genotype calls between unfiltered and filtered data sets and changes were grouped by type and direction of change (e.g. no-call-to-homozygote and homozygote-to-heterozygote). Second, genotype call changes were summarized by type, species and library to determine whether certain types of genotype changes were more common than others and whether the pattern was consistent across all data sets. Third, all genotype call changes were analysed by individual to estimate the percentage of genotypes that changed per individual (individual change rate) and this value was used as the response variable to test for potentially predictive variables of genotype changes among individuals (overall genotyping rate and average read depth per individual). Support for predictive variables was estimated using linear regression conducted on natural log-transformed individual change rates to correct for non-normal distributions identified using Q-Q plots. Because genotyping rate and mean read depth were autocorrelated (adjusted R^2 = 0.56, p < .001), we limit our discussion to comparisons of mean read depth but include summary statistics for genotyping rate in the supplemental materials. Finally, all genotype call changes were pooled by SNP to estimate the per cent of genotypes changed per SNP (SNP change rate) and this value was used as the response variable to test for potential predictive variables of genotype changes among SNPs (SNP genotyping rate, average read depth, minor allele frequency and position). Support for predictive variables was estimated using linear regressions. Linear regressions for both individual- and SNP-specific models were run first using all genotype changes and again using only call-to-call changes (i.e. homozygote-to-heterozygote, heterozygote-to-homozygote or homozygote-to-homozygote) because call-to-call changes likely impact data sets more significantly.

2.5 | Objective 3: causes of genotype change

We used a modified GTScore pipeline (see https://github.com/gjmckinney/GTscore; G. McKinney, in review) to evaluate genotype call changes after clones were removed for the SNPs with the highest change rate in the walleye data set. Target SNPs were chosen by identifying the 1,000 SNPs with the highest change rate; however, because multiple SNPs had the same change rate, this resulted in a total of 1,228 SNPs that changed genotype call in 15 or more individuals. We chose to use the walleye data for this analysis because when compared to the other three data sets used in the study, the walleye data contained an moderate level of per cent clonality, genotype call change rate and number of loci. GTScore was used because STACKS does not currently report read counts for the alternate allele of homozygote calls or when no genotype is called. To implement the modified GTScore pipeline, we first created probes that represented the full sequence of each allele at a locus. Next, reads for each allele were counted as the number of sequences that matched each allele-specific sequence. Although genotyping algorithms are similar between STACKS and GTScore, GTScore generally produced slightly lower read counts. This is likely because GTScore does not allow mismatches in read counting (so-called secondary reads in STACKS).

Allele count data generated from GTScore were used to test three hypotheses associated with specific genotype change types: (a) call-to-no-call changes are caused primarily by a loss in reads that results in read depths dropping below the minimum threshold set in *ustacks*, (b) no-call-to-call changes occur only when the original call is ambiguous prior to clone filtering (i.e. edge cases that just meet call criteria) and (c) call-to-call changes primarily occur when SNP read counts are skewed in favour of one allele or the other (i.e. PCR bias). Hypotheses 1 and 2 were tested by comparing the median read depth of unfiltered and filtered data sets for all call-to-no-call and no-call-to-call changes. Median read depths below the STACKS minimum depth threshold of three reads were taken as evidence that genotypes changed call due to a loss of read depth. As a control group, we also report median read depths for observations where no change between unfiltered and filtered data sets. We used the median rather than mean for these analyses because we found that the mean was highly skewed by a few SNPs with very high read counts. Hypothesis 3 was tested by first calculating the proportion of total reads identified as clones at each allele for a particular genotype and then calculating the difference in proportion of clones between the allele with the greatest number of reads and least number of reads. In this situation, evidence for PCR related bias would be supported by one allele having a higher proportion of clones than the other while a one-to-one relationship would indicate both alleles are amplified similarly. We calculated the difference in proportion of clones between alleles for genotypes that changed as well as genotypes where no change was observed to determine whether bias in the number of clones at each allele was higher for changed genotypes compared with genotypes that did not change.

2.6 | Objective 4: effect of PCR clones on heterozygosity

To determine how much clones influence estimates of genetic diversity, average heterozygosity before and after clone filtering was taken from the *sumstats* file compiled by STACKS. Statistically significant differences in observed heterozygosity between unfiltered and filtered data sets were evaluated using a Welch two-sample *t* test run separately for each species, and the effect size of differences was estimated simply as the per cent change in heterozygosity between unfiltered and filtered data sets.

TABLE 1 Summary of the four RAD-seq data sets used for all analyses

Species	*N*	Number of SNPs	Total retained reads	Total clones	Mean retained reads	*SD* retained reads	Mean clones	*SD* clones
Brook trout	152	17,456	609.29	348.55	4.01	2.02	2.29	1.19
Chum salmon	178	42,292	1,231.77	360.14	6.42	2.05	1.88	1.08
Cisco	129	11,185	444.73	177.77	3.45	1.42	1.38	0.65
Walleye	152	14,102	157.24	65.57	1.03	0.65	0.43	0.30

Note: Libraries for each species were prepared using the Ali et al. (2016) RAD protocol and retained reads and PCR clones were identified using STACKS. The last four columns are averages for each individual (i.e. mean clones is mean number of reads identified as clones for each individual). All read counts are in millions of reads. *N* = number of individuals that passed quality checks; *SD* = standard deviation.

3 | RESULTS

3.1 | Objective 1: clone summary

As expected, the absolute number of clones per data set was positively correlated to the number of retained reads for all individuals (Table 1). Per-individual clonality ranged from 1% to 68% and averaged 40% (*N* = 625). Individual clonality was relatively consistent within library preparations, except for chum salmon. However, average clonality differed substantially among data sets, from ~25% clonality in chum salmon to ~55% in brook trout (see Figure 1 for ANOVA results; Table S1). Clonality also differed significantly between library preparations for brook trout and to a lesser degree for cisco, and not at all for walleye and chum salmon (Figure 1). The number of clones per sequenced tag was generally low, and the majority of tags had no clones at all (Figure 2; brook trout = 67.6%, cisco = 85.0%, chum salmon = 70.3%, walleye = 77.5%). For RAD-tags that did have clones, most (~95%) had fewer than five (brook trout = 96.3%, cisco = 94.0%, chum salmon = 94.7%, walleye = 98.6%).

3.2 | Objective 2: effect of PCR clones on genotype calls

Overall, 94% of genotypes did not change after PCR clones were removed. The per cent of genotypes per library that did change calls

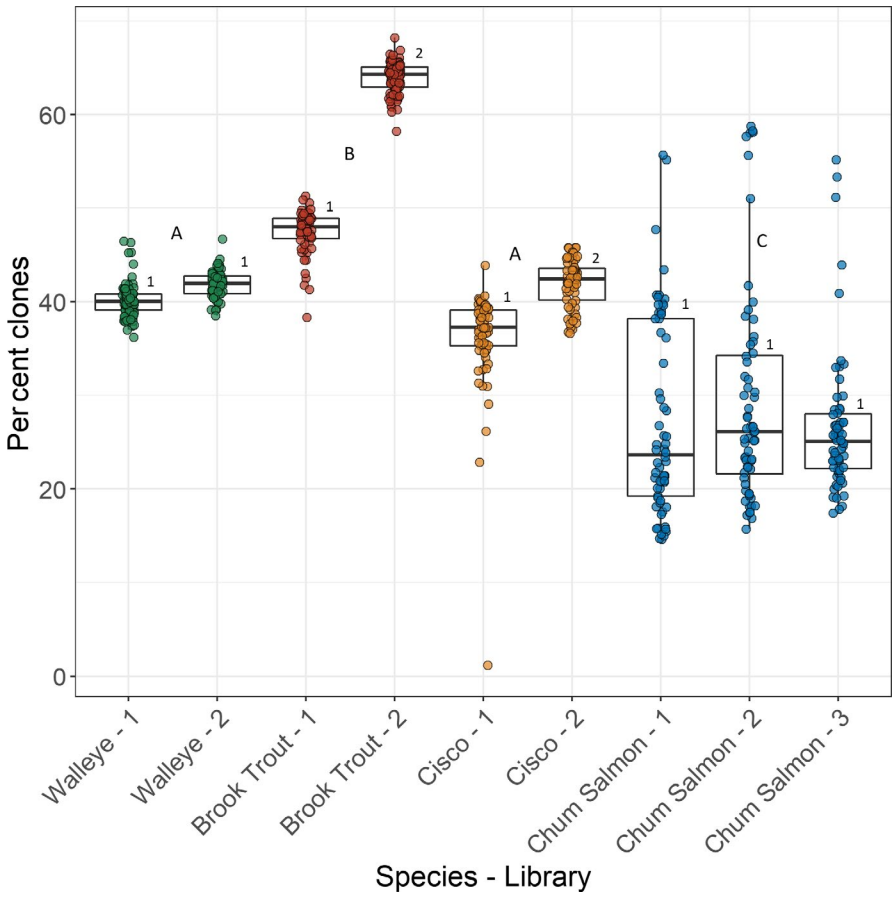

FIGURE 1 Per cent of RAD-tags identified as PCR clones for each individual, library and species. Each dot is an individual, and box plots denote the median, and 25th and 75th percentile. Species are denoted with different colours. Letters denote Tukey HSD significant differences between species and number denote significant differences between libraries within species ($p < .05$) [Colour figure can be viewed at wileyonlinelibrary.com]

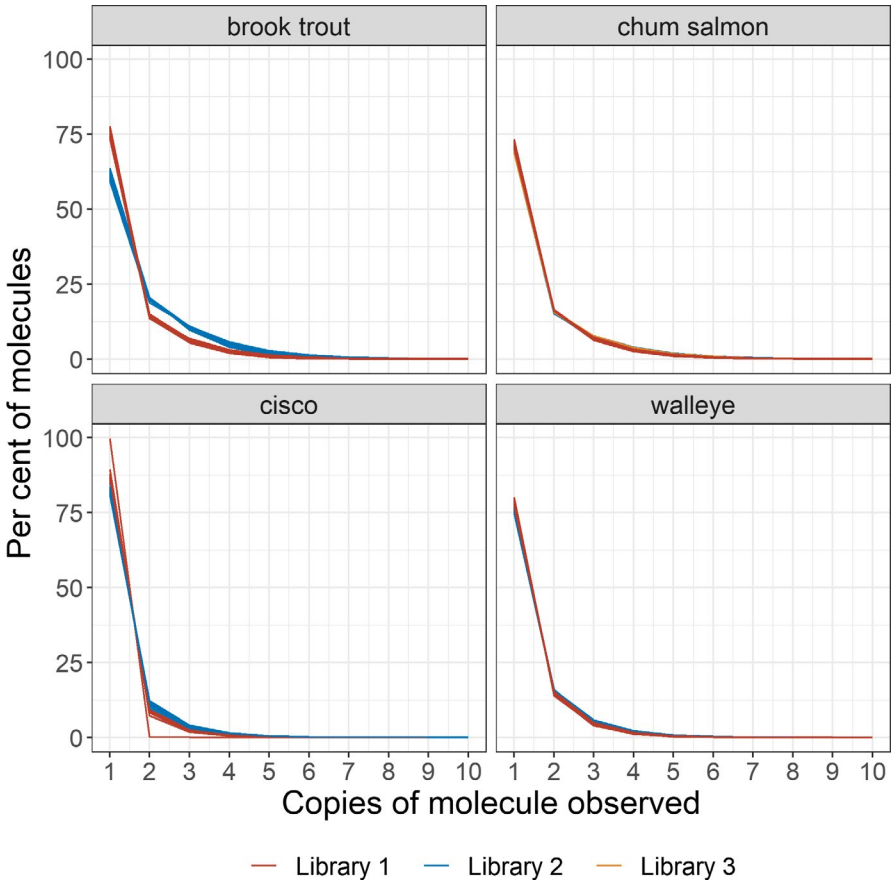

FIGURE 2 The percentage of RAD-tags (molecules) within each individual that were observed between one and ten times. The number of clones for each molecule is the number of copies of molecule minus 1; therefore, RAD-tags with a value of 1 have 0 clones. Each plot represents a different species, each individual is denoted as a separate line, and individuals are colour-coded by library. RAD-tags that contained more than 10 clones were uncommon (~0.2% of reads) and therefore not shown for easier interpretation of the plot [Colour figure can be viewed at wileyonlinelibrary.com]

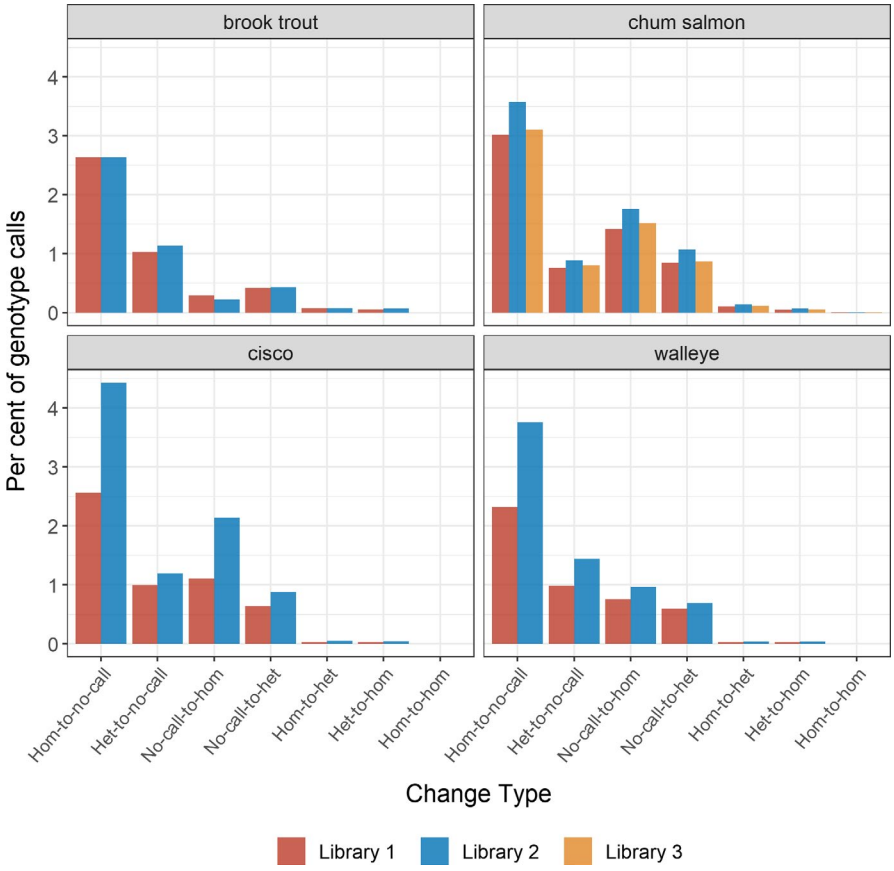

FIGURE 3 Per cent of calls changed between unfiltered and filtered data sets by type (i.e. het-to-no call is heterozygote in the unfiltered data set that changed to a no call in the filtered data set). Each plot represents a different species, and colour denotes unique library preparations. Heterozygote = het, homozygote = hom [Colour figure can be viewed at wileyonlinelibrary.com]

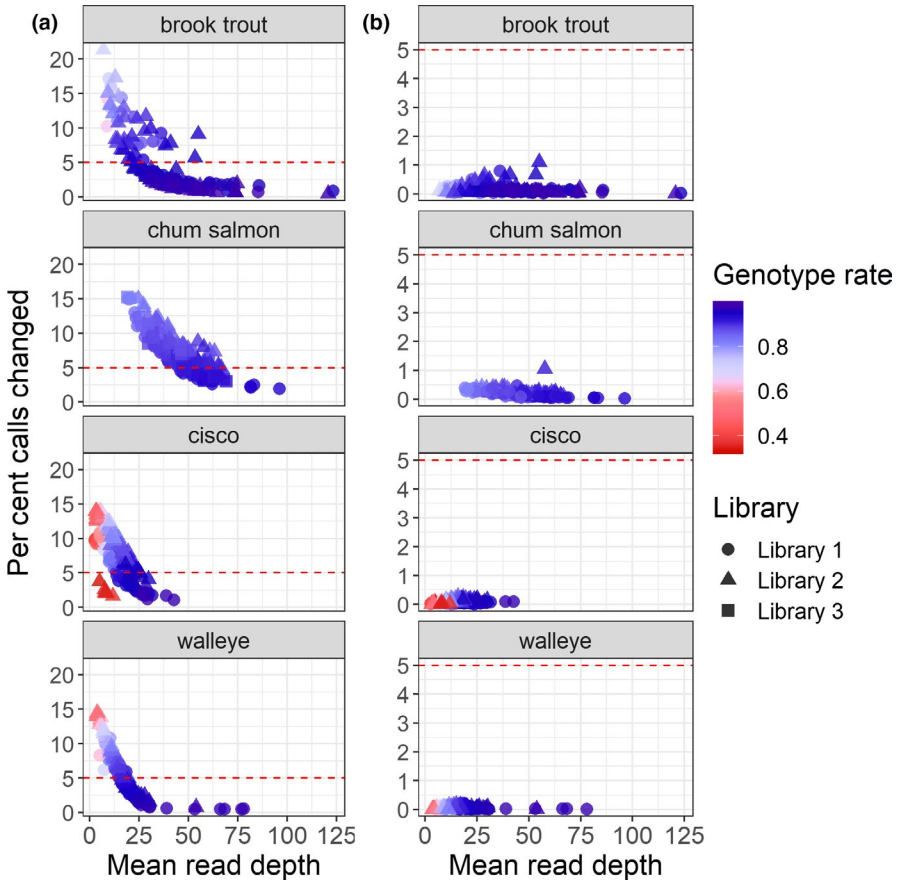

FIGURE 4 Relationship between per cent of calls changed, read depth and genotype rate (colour) across all change types including no calls (a) and only calls that changed between a heterozygote and a homozygote (i.e. call-to-call change) (b). Species are plotted separately, and red line added for context and denotes 5% calls changed [Colour figure can be viewed at wileyonlinelibrary.com]

postclone filter varied from 5% to 9% depending on species and library preparation (Table S2). The majority of genotypes that changed were call-to-no-call or no-call-to-call changes (Figure 3). Homozygote-to-no-call changes were found most often (2%–4% of genotype observations), followed by heterozygote-to-no-call, no-call-to-homozygote and no-call-to-heterozygote changes (0.2%–2% of genotype observations each), changes between heterozygotes and homozygotes (0.03%–0.15%), and finally homozygote-to-homozygote changes (0.003%) (Figure 3). Call-to-no-call changes resulted in an average decline in per-individual genotype rate of 2% (standard deviation = 0.8%).

Between 0.5% and 21.4% of genotype calls changed per individual once clones were removed. Individual change rate was negatively correlated with average read depth and genotyping rate (Figure 4; Table S3). Because genotype rate and mean read depth are autocorrelated (i.e. higher depth and higher genotyping rate), we will focus only on mean read depth here. The relationship between individual change rate and mean read depth was strongest for walleye (adjusted R^2 = 0.80) and weakest for cisco (adjusted R^2 = 0.57). However, the relationship was nonlinear, and more closely followed a pattern of exponential decay whereby once mean read depth reached about 20, individual change rate levelled off to between 0% and 5% for most individuals (Figure 4a [% change by depth]). When only call-to-call changes were assessed, the adjusted R^2 between individual change rate and mean read depth dropped substantially for brook trout, walleye and cisco to 0.01, 0.02 and 0.07, respectively, and dropped less substantially for chum salmon to 0.35 (Figure 4b).

We also investigated whether certain SNPs were more sensitive to PCR clones than others. Almost all SNPs (95%–100% per data set) changed genotype in at least one individual. However, call-to-call changes only occurred at 14.7% of SNPs on average but ranged from just 8% in walleye and cisco to 17% in brook trout and 26% in chum salmon. Most SNP change rates were fairly low and only changed in 1%–10% of individuals (average = 6.1%). However, some SNPs changed more frequently (max = 46.6%). If only call-to-call changes included, the average SNP change rate dropped to 0.8% (max = 11.2%; Figure S1). Unlike individual change rate, there was no clear relationship between the SNP change rate and locus-specific read depth, genotyping rate, minor allele frequency or SNP position (Table S3).

3.3 | Objective 3: causes of genotype change

Not all genotypes changed calls for the same reason. Changes from a heterozygote or homozygote call in the unfiltered data to a no call in the filtered data appeared to primarily result from the reduction in read depth associated with clone removal (Table 2). About 70% of homozygote-to-no-call changes and 83% of heterozygote-to-no-call changes had at least one allele with fewer than three reads in the postclone filter data set and therefore would not have met the minimum read count threshold in *ustacks*. Additionally, the median read depth of changed genotypes preclone filter was lower (5–7 reads) than the median read depth per allele of unchanged

TABLE 2 Summary of GTScore allele counts for all changes between unfiltered and filtered data sets (NC = no call, AA = homozygote allele A, AB = heterozygote, BB = homozygote allele B) at the 1,228 SNPs with the most genotype changes in the walleye data set

Change group	Change type	*N*	Number of SNPs	Median A1 depth pre-CF	Median A2 depth pre-CF	Median A1 depth post-CF	Median A2 depth post-CF	Mean prop A1 clones	Mean prop A2 clones	Median prop A1 clones	Median prop A2 clones
No-call-to-call	NC->AA	2,022	915	8	0	6	0	0.21	0.37	0.20	0.50
	NC->AB	2,538	939	5	5	4	4	0.21	0.21	0.22	0.23
	NC->BB	413	310	0	8	0	7	0.42	0.23	0.50	0.25
Call-to-no-call	AA->NC	8,401	1,218	5	0	3	0	0.39	0.03	0.33	0.00
	AB->NC	4,415	1,169	3	3	2	2	0.33	0.33	0.33	0.33
	BB->NC	2,294	893	0	6	0	3	0.02	0.40	0.00	0.35
Call-to-call	AA->AB	135	89	16	2	9	2	0.33	0.06	0.35	0.00
	AA->BB	3	3	4	0	2	0	0.39	0.00	0.50	0.00
	AB->AA	112	81	11.5	3	8.5	1	0.26	0.49	0.26	0.50
	AB->BB	40	34	4	11.5	1	9	0.55	0.25	0.60	0.27
	BB->AA	2	2	0	3	0	2.5	0.00	0.17	0.00	0.17
	BB->AB	57	54	2	12	2	6	0.13	0.38	0.00	0.41
No change	NC->NC	34,100	1,221	1	1	1	1	0.12	0.13	0.00	0.00
	AA->AA	78,498	1,220	12	0	9	0	0.26	0.06	0.26	0.00
	AB->AB	38,617	1,221	7	7	5	5	0.24	0.24	0.25	0.25
	BB->BB	14,112	1,163	0	11	0	8	0.06	0.26	0.00	0.27

Note: Number of observations (*N*) indicates the total number of observations of a given change type, while number of SNPs indicates the number of unique SNPs where a given change type occurred. Proportion of clones was calculated as the difference in allele read counts between unfiltered and filtered data divided by the number of unfiltered read counts for an allele.

TABLE 3 Observed heterozygosity (H_O) for each species before and after all PCR clones were removed and summary statistics for Welch two-sample *t* tests

Species	H_O preclone filter	H_O post-clone filter	Degrees of freedom	Test statistic	*p*-value
Brook trout	0.249	0.249	21,662	−0.027	0.979
Cisco	0.209	0.214	22,332	2.926	0.003
Chum salmon	0.262	0.268	84,569	7.522	<0.000
Walleye	0.259	0.261	28,194	1.789	0.073

Note: *p*-values <0.05 indicate a significant difference in H_O between the two data sets.

genotypes preclone filter (11–14 reads). Therefore, the majority of the genotype calls that changed once clones were removed were already closer to the STACKS minimum read threshold than genotype calls that did not change, indicating they would have been more sensitive to changes in read depth (i.e. edge cases). Genotypes that were no calls in the unfiltered data but changed to a homozygote or heterozygote in the filtered data, however, did not appear to be related to changes in coverage or the minimum read count threshold in *ustacks*. While the median number of preclone filter reads for no-call-to-call changes was slightly lower (8–12 reads) than the unchanged calls (11–14 reads), the read depth for most no-call-to-call changes would have met STACKS minimum read threshold. Therefore, no-call-to-call changes may be associated with factors other than PCR clones, such as small differences in the way loci were formed during *ustacks* (e.g. a locus had sufficient reads to be split in one data set but was merged in the other).

Call-to-call changes appeared to largely result from preferential amplification of one allele. When genotypes changed from a homozygote in the unfiltered data set to a heterozygote in the filtered data set, the proportion of clones was larger for the allele with the greatest number of reads 88% of the time (i.e. PCR bias towards the most prevalent allele). Contrastingly, when genotypes changed from a heterozygote to a homozygote, the allele with the greatest number of reads only had a higher proportion of clones 23% of the time (i.e. no PCR bias or bias towards the less prevalent allele). When genotypes were the same before and after clone filtering (i.e. no change in genotype), the proportion of clones was either equal for both alleles (heterozygotes) or highest for the major allele (homozygotes) and there were often zero reads (and therefore no clones) of the alternate allele (Table 2).

3.4 | Objective 4: effect of PCR clones on heterozygosity

We found that average heterozygosity was slightly lower in the unfiltered data set than in the filtered data set for cisco, walleye and chum salmon, and almost identical for brook trout. The size of the difference, however, was universally small (highest change = 0.006) and only significant for cisco and chum salmon (Table 3). The change in observed heterozygosity was similar for chum salmon and cisco, increasing by about 2.3% once clones were removed.

4 | DISCUSSION

We evaluated how PCR clones influence the genotype calls of SNPs identified through RAD-seq for four different species. Even when greater than 60% of sequenced reads were clones, genotype calls were robust to clone removal and about 95% of calls remained unchanged. Call-to-call changes were rare, and because the majority of genotype changes were call-to-no-call, removal of PCR clones resulted in a 2% loss in genotype rate per individual. We found limited evidence that misscalled genotype bias heterozygosity estimates. Therefore, including PCR clones in analysis likely does not have a major influence on estimates of diversity (Andrews et al., 2016). Additionally, the low percentage of calls that changed per locus indicate that PCR clones are also unlikely to bias studies attempting to estimate relatedness or identify outlier loci and search for signatures of selection. Our results suggest that the concern that PCR clones influence analyses conducted using data from RAD-seq may be overstated. We suggest that if steps are taken to ensure that all individuals have sufficient coverage, the effects of PCR clones should be minimal and will be unlikely to significantly bias most genomic studies. Below, we discuss the results of each objective and highlight potential steps and analysis that can be conducted to identify and mitigate the influence of PCR clones on downstream analysis.

4.1 | Objective 1: clone summary

The amount of PCR clones in RAD-seq data sets can be substantial (Andrews et al., 2014; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Hohenlohe et al., 2013; Schweyen et al., 2014). All four of our data sets had a considerable amount of PCR clones, but the patterns of clonality within and among library preparations suggest that good laboratory practices do help mitigate variability in PCR clonality (O'Leary et al., 2018; Van Dijk, Jaszczyszyn, & Thermes, 2014). Higher variability was found in brook trout, where libraries had very different levels of clonality, and chum salmon, where individual variation within libraries was high. Variation among library preparations could result from differences in DNA template quality or number of PCR cycles between libraries. In the present study, both brook trout library preparations used the same number of PCR cycles; therefore, template quality may have caused the difference between libraries we saw. The interindividual variation in chum salmon data could have been the result of the iterative sequencing technique used in the Seeb

Laboratory, whereby DNA for each individual was adjusted in the second round of sequencing. While this process reduces variation in sequence reads per individual, it may boost the amount of PCR clones sequenced for individuals with low-quality template DNA because a larger number of the reads sequenced for these individuals will be PCR clones. Notably, cisco library prep-2 had high variation in DNA quality (some samples appeared highly degraded when run on a gel) and returned high variation in reads per individual but had remarkably consistent levels of clonality across individuals. The results from cisco and chum salmon suggest that individuals with low DNA quality do not inherently produce high levels of clonality but that allocating more sequencing effort to these individuals will increase clonality.

Within each individual, PCR clones were spread over many loci, rather than restricted to a few highly cloned sequences. This means that even when the overall level of clonality is high, the number of clones per molecule stays low for the vast majority (~95%) of molecules. The spread of PCR clones across many loci should help limit the number of miscalled genotypes—when read depth is high (greater than 20), the addition of one or two clones per molecule will have a minimal effect on a genotype call.

Our finding that the number of clones per molecule is low is logical given the characteristics of high-throughput sequencing. When a large amount of DNA template material is added to a flow cell, the probability that any one molecule will attach and be sequenced is very small and the probability of a molecule and its clone both attaching and being sequenced is exponentially smaller. Therefore, as long as the amount of PCR product sequenced is relatively high, and the amount amplification per molecule is consistent, the number of clones per molecule should remain small, even if the overall number of copies of each target in the sample is high. Therefore, the number of clones per molecule is likely a better predictor of potential PCR clone bias than the overall amount of PCR clones because most genotype calls are robust to a small change in read counts as long as depth is adequate.

4.2 | Objective 2: effect of PCR clones on genotype calls

The amount of PCR clones varied among data sets and library preparations, but the per cent of genotype changes did not. Andrews et al. (2014) and Andrews et al. (2016) both suggest that when the frequency of PCR clones is high, PCR errors could appear as true alleles, and therefore, the removal of PCR clones should increase genotype accuracy. While we have no way of confirming the correct genotypes to directly evaluate the influence of PCR clones on genotype accuracy, we found that the amount of PCR clones has little to no influence on how frequently genotypes change calls. For example, the library prep-2 of brook trout had an average number of PCR clones that was 20% higher than library prep-1, but almost identical percentage of changed genotypes. Also, while brook trout had some of the highest rates of PCR clonality, it had some of the lowest individual rates of genotype call change, likely due to the relatively high mean read depth achieved in the brook trout data set (>25).

The majority of genotype call changes impacted only the level of missing data, not actual genotype calls (call-to-call changes). Call-to-call changes are the type of change most often speculated to occur from PCR clone bias (Andrews et al., 2016; Davey et al., 2013). However, in practice, call-to-call changes occurred in less than 1% of genotype calls, suggesting that any potential bias associated with PCR clones would likely be due to changes to or from a no-call state. Call-to-no-call genotype changes were the most frequent, but their prevalence decreased rapidly as mean read depth and genotype rate per individual increased, suggesting that they can be at least partially mitigated through sequencing effort. Standard procedures for genotype calling already suggest a target read depth of 5 or greater and the removal of individuals with less than a 70% genotype rate (Fountain et al., 2016; Nielsen, Paul, Albrechtsen, & Song, 2011). Therefore, if researchers follow existing protocols, genotype change rate should be limited to less than 5% per individual. However, if sequencing effort is insufficient, PCR duplicates could have a larger influence on genotype calls and removing PCR clones could result in a substantial decrease in genotyping rate.

We found that essentially all SNPs were at risk of changing genotype calls. Additionally, when call-to-no-call and no-call-to-call changes were removed, the change rate dropped substantially, suggesting again that filtering out PCR clones mostly influences genotyping rate rather than genotype calls at a given SNP. Some SNPs showed higher rates of change than others, but we did not find any factor that could be used to predict SNP-specific change rates.

4.3 | Objective 3: causes of genotype change

About 50%–60% of the call-to-no-call changes we evaluated for Objective 3 appeared to be due to a reduction in coverage, which translated to an increase in missing genotype calls by about 2% in the filtered data set. Therefore, removing PCR clones for low coverage individuals should increase confidence in each genotype by eliminating PCR artefacts but will also decrease the overall number of genotype calls. Calling genotypes from low coverage data can falsely inflate homozygosity when both alleles are not detected (DaCosta & Sorenson, 2014; O'Leary et al., 2018). However, because both alleles of a heterozygote should have equal probability of being sequenced, many of these low coverage genotype calls are likely accurate. It is not possible to directly test genotype accuracy without genotyping specific loci that experienced a call change in our analysis; however, Ebbert et al. (2016) found that concordance between next-generation sequencing data including PCR clones and SNP chip data was above 99%. Therefore, filtering out PCR clones will result in the loss of correct genotype calls. The question then becomes is the increase in confidence for some genotype calls mitigated by higher levels of missing data.

The small fraction of call-to-call genotype changes that we observed are likely a result of PCR bias but our observations do not exactly match theoretical expectations, which postulate that PCR bias will largely result in undercalling true heterozygotes due to

asymmetric amplification of alleles (Andrews et al., 2016, 2014). Our observation that the proportion of clones was higher for the allele with the higher number of reads in homozygote-to-heterozygote changes fits this theoretical expectation. However, homozygote-to-heterozygote changes and heterozygote-to-homozygote changes occurred in similar frequencies, suggesting that PCR errors may result in incorrect genotype calls. Specifically, the higher prevalence of clones at the allele with fewer reads in heterozygote-to-homozygote changes suggests that PCR errors could have been amplified by chance, leading to false heterozygote calls. Unfortunately, removing PCR clones will not completely remove this type of bias. After clones were removed, allele read depth for most call-to-call changes was still skewed. This is because even if the clones from a specific molecule are removed, all unique molecules of a preferentially amplified allele are still more likely to be sequenced than the alternate allele. Therefore, the only way to truly manage the influence of this type of allelic drop out is to start with high-quality DNA and minimize PCR steps during library preparation to avoid over-representation of one allele in the final library.

Finally, our analysis of Objective 3 suggests that many of the genotype changes evaluated in our study may not have been the result of PCR clones. About half of call-to-no-call changes and almost all of no-call-to-call changes had no obvious loss of read depth or PCR bias that could explain the call change between unfiltered and filtered data sets. Therefore, some of our genotyping changes are likely the result of bioinformatic processing, and our estimates of genotype change rates attributed to PCR clones are likely high.

4.4 | Objective 4: effect of PCR clones on heterozygosity

The most common concern associated with PCR clones is that erroneous genotype calls resulting from PCR bias will lead to excess homozygosity (Andrews et al., 2016). Our results indicate that this is not a major concern. Additionally, the sheer number of loci used in most RAD-seq studies means that estimates of diversity and population structure (e.g. pairwise F_{ST}) are fairly robust to a small number of misidentified genotypes. However, more caution may be warranted when high confidence in each genotype is important, such as for studies investigating selection or relatedness (Andrews et al., 2018; Lowry et al., 2017). Nonetheless, we found that call-to-call changes were rare and appeared to be evenly distributed across loci. It is therefore highly unlikely that genotyping error due to PCR clones would be sufficient to substantially influence the results of most genomic studies, as long as best practices are followed.

4.5 | Conclusions and recommendations

PCR clones can lead to the misidentification of genotypes in RAD-seq data sets; however, the prevalence of genotype changes is low and can be mitigated by ensuring that sequencing coverage is adequate, and samples are of high quality. Therefore, we suggest prioritizing these other potential sources of error over PCR clone removal. Best practices, such as collecting fresh samples, checking sample quality prior to sequencing, reducing PCR steps and increasing sequencing effort to reach an average of 20X coverage per individual will reduce the impact PCR clones have on downstream analysis (Davey et al., 2011; Van Dijk et al., 2014). Call-to-call changes made up a minority of call changes compared call-to-no-call genotype changes. Therefore, the primary impact of PCR clone removal may simply be a reduction in overall genotyping rate, without the benefit of a significant increase in genotype confidence. It is important to note that genotypes in our study were called using relatively high coverage and that probabilistic genotyping approaches utilizing low coverage data, which are becoming common (Korneliussen, Albrechtsen, & Nielsen, 2014; Nielsen et al., 2011), may be more impacted by PCR clones. Still, the fact that PCR clones appeared to be relatively randomly distributed across loci and consistent among individuals within libraries suggests that the probability that clones will strongly bias the interpretation of a well-constructed genomic study of any kind is low.

Even if PCR clones are not a substantial source of bias in most cases, understanding the level and influence PCR clones have on your own data set is still important. For studies using random shearing protocols, such as the Ali et al. (2016) protocol, we suggest clone filtering at least a subset of data to estimate the level of clonality in the data set, which can be a useful diagnostic tool to determine the success of library preparation in capturing the diversity present in the original samples. Further, because clones can lead to some genotype changes, we suggest that in studies requiring high confidence in genotypes, researchers run a subset of their data through the STACKS pipeline with, and without the *clone_filter* step (as in Objective 2) to ensure genotype calls are robust. To aid in this, we have provided a series of scripts that can be used to recreate analyses presented here (see Supplement). Finally, we believe our study is particularly informative for RAD sequencing techniques that cannot remove PCR duplicates, such as ddRAD (Schweyen et al., 2014; Tin, Rheindt, Crow, & Mikheyev, 2015). Rather than spending time and resources on degenerative primers, we suggest that investment in sequencing and library preparation will improve overall data quality while minimizing the influence of clones. In conclusion, while we still strongly encourage researchers to maximize DNA quality and minimize PCR cycles when conducting library preparations, our results suggest that robust data can be obtained without filtering out PCR clones, even when the data sets contain a high level of clonality.

ACKNOWLEDGEMENTS

We thank Keith Turnquist and Carita Pascal for their assistance preparing RAD libraries. We also thank Christian Smith and three anonymous reviewers for providing valuable feedback on the manuscript. Any use of trade, product or company name is for descriptive purposes only and does not imply endorsement by the U.S. Government.

AUTHOR CONTRIBUTIONS

PE and WL designed the study with input from GM and MM. Data analyses were conducted by PE, GM, MB and CT. All authors contributed to the writing of the manuscript.

DATA AVAILABILITY STATEMENT

Raw data for the brook trout, cisco and walleye used in this study were deposited to the NCBI sequence read archive (BioProject: PRJNA557717), and VCF files of genotypes are available on DRYAD (https://doi.org/10.5061/dryad.3mq4631). Data for chum salmon were taken from McKinney et al. (2019) (https://doi.org/10.1101/729574). Python and R scripts used to generate data are available at https://github.com/WesLarson-Lab/Attack_of_the_clones.

ORCID

Peter T. Euclide https://orcid.org/0000-0002-1212-0435
Garrett J. McKinney https://orcid.org/0000-0002-6267-2203
Mariah H. Meek https://orcid.org/0000-0002-3219-4888
Wesley A. Larson https://orcid.org/0000-0003-4473-3401

REFERENCES

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

Allendorf, F. W., Bassham, S., Cresko, W. A., Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived Salmonid fishes. *Journal of Heredity*, *106*(3), 217–227. https://doi.org/10.1093/jhered/esv015

Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. *Nature Reviews Genetics*, *11*(10), 697–709. https://doi.org/10.1038/nrg2844

Andrews, K. R., Adams, J. R., Cassirer, E. F., Plowright, R. K., Gardner, C., Dwire, M., … Waits, L. P. (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RADseq data. *Molecular Ecology Resources*, *18*(6), 1263–1281. https://doi.org/10.1111/1755-0998.12910

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

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

Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis. *Molecular Ecology*, *23*(7), 1661–1667. https://doi.org/10.1111/mec.12686

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

Bernatchez, L., Wellenreuther, M., Araneda, C., Ashton, D. T., Barth, J. M. I., Beacham, T. D., … Withler, R. E. (2017). Harnessing the power of genomics to secure the future of seafood. *Trends in Ecology and Evolution*, *32*(9), 665–680. https://doi.org/10.1016/j.tree.2017.06.010

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. *G3:Genes|Genomes|Genetics*, *1*(3), 171–182. https://doi.org/10.1534/g3.111.000240

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

DaCosta, J. M., & Sorenson, M. D. (2014). Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE*, *9*(9), e106713. https://doi.org/10.1371/journal.pone.0106713

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

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq data analysis parameters for population genetics: The more the better? *Frontiers in Genetics*, *10*, 1–10. https://doi.org/10.3389/fgene.2019.00533

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., … Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17*(S7), 491-500. https://doi.org/10.1186/s12859-016-1097-3

Flanagan, S. P., Forester, B. R., Latch, E. K., Aitken, S. N., & Hoban, S. (2018). Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. *Evolutionary Applications*, *11*, 1035–1052. https://doi.org/10.1111/eva.12569

Flanagan, S. P., & Jones, A. G. (2018). Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. *Molecular Ecology Resources*, *18*(2), 264–280. https://doi.org/10.1111/1755-0998.12734

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Molecular Ecology Resources*, *16*(4), 966–978. https://doi.org/10.1111/1755-0998.12519

Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., … Luikart, G. (2018). Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, *11*(8), 1197–1211. https://doi.org/10.1111/eva.12659

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, *11*, 117–122. https://doi.org/10.1111/j.1755-0998.2010.02967.x

Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., … Luikart, G. (2013). Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, *22*(11), 3002–3013. https://doi.org/10.1111/mec.12239

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), https://doi.org/10.1186/s12859-014-0356-4

Larson, W. A., Limborg, M. T., McKinney, G. J., Schindler, D. E., Seeb, J. E., & Seeb, L. W. (2017). Genomic islands of divergence linked to ecotypic variation in sockeye salmon. *Molecular Ecology*, *26*(2), 554–570. https://doi.org/10.1111/mec.13933

Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*(20), 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

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

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

McKinney, G., McPhee, M. V., Pascal, C., Seeb, J. E., & Seeb, L. W. (2019). Patterns of linkage disequilibrium reveal genome architecture in chum salmon. *bioRxiv*, 729574. https://doi.org/10.1101/729574

McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources*, *18*(3), 570–579. https://doi.org/10.1111/1755-0998.12763

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. https://doi.org/10.1111/1755-0998.12613

Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology*, *24*(13), 3223–3231. https://doi.org/10.1111/mec.13243

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*, 3193–3206. https://doi.org/10.1111/mec.14792

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

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

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

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

Schweyen, H., Rozenberg, A., & Leese, F. (2014). Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biological Bulletin*, *227*(2), 146–160. https://doi.org/10.1086/BBLv227n2p146

Tin, M. M. Y., Rheindt, F. E., Cros, E., & Mikheyev, A. S. (2015). Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, *15*(2), 329–336. https://doi.org/10.1111/1755-0998.12314

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

Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, *322*(1), 12–20. https://doi.org/10.1016/j.yexcr.2014.01.008

Willette, D. A., Allendorf, F. W., Barber, P. H., Barshis, D. J., Carpenter, K. E., Crandall, E. D., … Seeb, J. E. (2014). So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. *Bulletin of Marine Science*, *90*(1), 79–122. https://doi.org/10.5343/bms.2013.1008

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Euclide PT, McKinney GJ, Bootsma M, Tarsa C, Meek MH, Larson WA. Attack of the PCR clones: Rates of clonality have little effect on RAD-seq genotype calls. *Mol Ecol Resour*. 2020;20:66–78. https://doi.org/10.1111/1755-0998.13087