Received: 28 June 2017 | Revised: 29 September 2017 | Accepted: 6 October 2017

DOI: 10.1111/1755-0998.12734

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study

Sarah P. Flanagan[1,2] | Adam G. Jones[1,3]

[1]Biology Department, Texas A&M University, College Station, TX, USA

[2]National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, USA

[3]Department of Biological Sciences, University of Idaho, Moscow, ID, USA

Correspondence
Sarah P. Flanagan, National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, USA.
Email: spflanagan.phd@gmail.com

Funding information
Postdoctoral Fellowship from the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from The University of Tennessee, Knoxville; NSF, Grant/Award Number, Grant/Award Number: #DEB-1119261, DEB-1401688, DGE-1252521

Abstract

The trade-offs of using single-digest vs. double-digest restriction site-associated DNA sequencing (RAD-seq) protocols have been widely discussed. However, no direct empirical comparisons of the two methods have been conducted. Here, we sampled a single population of Gulf pipefish (*Syngnathus scovelli*) and genotyped 444 individuals using RAD-seq. Sixty individuals were subjected to single-digest RAD-seq (sdRAD-seq), and the remaining 384 individuals were genotyped using a double-digest RAD-seq (ddRAD-seq) protocol. We analysed the resulting Illumina sequencing data and compared the two genotyping methods when reads were analysed either together or separately. Coverage statistics, observed heterozygosity, and allele frequencies differed significantly between the two protocols, as did the results of selection components analysis. We also performed an in silico digestion of the Gulf pipefish genome and modelled five major sources of bias: PCR duplicates, polymorphic restriction sites, shearing bias, asymmetric sampling (i.e., genotyping fewer individuals with sdRAD-seq than with ddRAD-seq) and higher major allele frequencies. This combination of approaches allowed us to determine that polymorphic restriction sites, an asymmetric sampling scheme, mean allele frequencies and to some extent PCR duplicates all contribute to different estimates of allele frequencies between samples genotyped using sdRAD-seq versus ddRAD-seq. Our finding that sdRAD-seq and ddRAD-seq can result in different allele frequencies has implications for comparisons across studies and techniques that endeavour to identify genomewide signatures of evolutionary processes in natural populations.

KEYWORDS
allele dropout, next-generation sequencing, PCR duplicates, selection components analysis, *Syngnathus scovelli*

1 | INTRODUCTION

Many questions in modern evolutionary biology require genetic information from individuals. Important aspects of the evolutionary process, including adaptive divergence, phylogenetic relationships, mating system dynamics and the influence of neutral processes on population differentiation, can be estimated only from reliable genotypes of individuals, which often must be compared across populations and species. Consequently, the field of evolutionary biology has eagerly adopted next-generation sequencing technologies and the myriad genotyping techniques that go along with them.

One popular genotyping technique is restriction site-associated DNA sequencing (RAD-seq), a reduced-representation approach that targets DNA sequences near restriction sites. This family of techniques allows individuals to be genotyped at the same randomly sampled regions throughout the genome and yields thousands of single nucleotide polymorphism (SNP) genotypes. Several versions of RAD-seq have been developed, and each one uses slightly different methods to achieve the same purpose. Single-digest RAD-seq (sdRAD-seq) was the original RAD-seq method; it uses one infrequently cutting restriction enzyme plus a sonication step to generate short fragments for sequencing (Baird et al., 2008; Miller, Dunham, Amores, Cresko, & Johnson, 2007). Double-digest RAD-seq (ddRAD-seq) uses two restriction enzymes and omits the sonication step (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). A method called 2bRAD-seq takes advantage of type-IIB restriction enzymes, which cut twice at a specified distance from the restriction site, releasing a short DNA fragment of around 30 base pairs. Hence, a single type-IIB restriction enzyme can be used to generate fragments, but they will be very short by today's sequencing standards (Wang, Meyer, McKay, & Matz, 2012). These methods have been reviewed in detail elsewhere (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Andrews & Luikart, 2014; Puritz et al., 2014), and the utility of the RAD-seq approach in general for studying adaptive variation has recently been called into question (Lowry et al., 2017a,b), although researchers agree that RAD-seq can be a useful tool for molecular ecologists (Catchen et al., 2017; Lowry et al., 2017b; McKinney, Larson, Seeb, & Seeb, 2017). The benefits of sdRAD-seq versus ddRAD-seq have also been debated in the literature (Andrews & Luikart, 2014; Andrews et al., 2014; Puritz et al., 2014). Ultimately, each method has its own sets of biases and technical issues, and the choice will depend on the focus of the study, the study organism and the budget allocated to the project (Andrews et al., 2016).

A major issue in RAD-seq studies is that restriction enzyme cut sites can be polymorphic. Polymorphic restriction sites result in some individuals not being genotyped at particular loci or having a homozygous genotype called when the individual is actually heterozygous (Davey et al., 2013), a phenomenon known as allelic dropout. These false homozygous calls decrease the observed heterozygosity at loci with polymorphic restriction sites (Andrews et al., 2016), resulting in biased summary statistics (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013). False homozygote calls can lead to high error rates even with sufficient depth of coverage (Henning, Lee, Franchini, & Meyer, 2014), because polymorphic sites are simply not genotyped (Davey et al., 2013). Bias due to allelic dropout may be limited unless effective population sizes are large ($N_e > 10^5$; Andrews et al., 2016; Gautier et al., 2013) or when polymorphism is high (Cariou, Duret, & Charlat, 2016). Although Andrews et al. (2016) suggest that loci with null alleles may be identified in a data set by high variance in coverage depth across samples, this type of filtering step is not typically a part of RAD-seq analyses. Nevertheless, typical filtering steps to retain loci with high coverage across individuals and with allele frequencies above a cut-off may remove many loci experiencing allelic dropout (Andrews et al., 2016), and a recent method uses a Bayesian model to identify loci likely suffering from bias due to polymorphic restriction sites and flags them for removal from the analysis (Cooke et al., 2016). Of the various RAD-seq methods, those using multiple restriction enzymes (e.g., ddRAD-seq) are likely to be more greatly affected by allelic dropout than those using a single restriction enzyme (e.g., sdRAD-seq) because there are at least twice as many potentially polymorphic restriction sites involved (Andrews et al., 2016; Arnold et al., 2013).

Other sources of error in RAD-seq studies emerge from the PCR amplification step. The most important problem is the production of PCR duplicates, which stem from a random allele at a given locus being amplified more than the other allele and result in a falsely homozygous genotype (Andrews et al., 2016). These false homozygote calls cause the same biases as allelic dropout and also result in variance in coverage depth at a locus (Andrews et al., 2016), making it difficult to differentiate between allelic dropout and PCR duplicates [unless the study design incorporates a way of identifying PCR duplicates, see Andrews et al. (2016), Casbon, Osborne, Brenner, and Lichtenstein (2011), Davey et al. (2011), Schweyen, Rozenberg, and Leese (2014), Tin, Rheindt, Cros, and Mikheyev (2015)]. Although PCR duplicates should impact ddRAD-seq and sdRAD-seq libraries at a similar rate as long as the number of PCR cycles is the same, sdRAD-seq produces fragments with different random break points, permitting PCR duplicates to be removed during filtering steps (Andrews et al., 2014). A related problem arises from GC bias during the PCR steps (Andrews et al., 2016; Davey et al., 2011), and both GC bias and PCR duplicates are minimized by the use of a high-fidelity polymerase such as Phusion (Puritz et al., 2014). Finally, PCR preferentially amplifies shorter fragments. This issue will affect ddRAD-seq methods more than sdRAD-seq because the two restriction sites determine the fragment length of RAD loci, whereas in sdRAD-seq the length of RAD loci is randomly determined by shearing (Andrews et al., 2016).

A final source of variance in coverage depth among loci is a result of the shearing step in sdRAD-seq. Shorter fragments (<10 kb) shear less efficiently than longer fragments, resulting in loci from shorter fragments having fewer reads (Davey et al., 2013). This phenomenon only affects the sdRAD-seq method, and Andrews et al. (2016) suggest that its effects should be minimal, because most sdRAD-seq studies use restriction enzymes whose recognition sites occur rarely in the genome, resulting in mostly large fragments prior to shearing.

Although the biases in sdRAD-seq and ddRAD-seq have been deduced and evaluated using simulation models, no study has used both methods and evaluated the impact of using two different RAD-seq library preparation approaches to genotype individuals from a single population. Here, we provide a case study, which demonstrates that the two methods produce different allele frequency distributions for individuals from the same wild-caught population of the Gulf pipefish, *Syngnathus scovelli*.

2 | METHODS

2.1 | Collection methods

The Gulf pipefish, *Syngnathus scovelli*, is a sex-role-reversed marine fish in the family Syngnathidae (seahorses, pipefishes and seadragons). The species is found in the Gulf of Mexico and along the Atlantic coast of Florida in shallow seagrass beds. Gulf pipefish were collected from the Gulf of Mexico by seine net in Corpus Christi, TX (27°41′33″N, 97°10′54″W). Each fish was euthanized in MS-222, preserved in ethanol and frozen until DNA could be extracted.

We extracted DNA from tissue from the adult heads and from entire embryos using the PUREGENE DNA extraction kit (QIAGEN). Genomic DNA quality was evaluated by visualizing each sample on an agarose gel, and each sample was quantified using a QUBIT FLUOROMETER 2.0 (Life Technologies).

2.2 | sdRAD-seq library preparation

We prepared a sdRAD-seq library using DNA from 30 pregnant males and 30 females following the protocols described in Baird et al. (2008). From each sample, 1 μg of DNA was digested with 100 Units of *Pst*I-HF (New England Biolabs) at 37°C for 90 min. The digestions were cleaned up using the DNA Clean & Concentrator-5 (Zymo) kit before the first adapter ligation. These P1 adapters were identical to those used by Baird et al. (2008), and each adapter contained a 6-bp barcode and an Illumina sequencing primer. The ligation used 1000U of T4 DNA ligase (New England Biolabs) and NEB Buffer 2 (New England Biolabs) and was incubated at 16°C for 30 min before heat inactivation. We pooled 25 μl from each of 12 samples and sheared these pooled samples using a Bioruptor. Sheared DNA was cleaned up using the DNA Clean & Concentrator-5 kit (Zymo) and eluted in 20 μl. Each set of 12 pooled samples remained separate until the final pooling step. This sdRAD-seq library was electrophoresed on a 1.25% agarose gel stained with SafeView (ABMGood), and fragments in the range of 300–700 bp were excised from the gel with a razor blade. DNA was recovered from the gel slices using a Zymoclean Gel DNA Recover Kit (Zymo). The 5′ and 3′ overhangs were removed using the Quick Blunting Kit (New England Biolabs), and then, an adenosine base was added to the 3′ end of the fragments in a reaction with Klenow enzyme (New England Biolabs) at 37°C for 30 min. The library was subsequently cleaned up using a DNA Clean & Concentrator-5 kit (Zymo), and the P2 adapter, containing only the sequencing adapters and no barcodes, was ligated onto the fragments with T4 DNA ligase (New England Biolabs). After another clean-up step with the DNA Clean & Concentrator-5 kit (Zymo), PCR with Phusion polymerase was run for 18 cycles (cycle conditions: 98°C for 30 s; 18 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 10 s; 72°C for 5 min) in two separate reactions. Those reactions were pooled and purified with DNA Clean & Concentrator-5 (Zymo). The cleaned PCR products for each set of 12 pooled samples were pooled into a final library whose quality was determined using a QUBIT FLUOROMETER 2.0 (Invitrogen), and the library was sent to the University of Oregon for 100 bp single-end Illumina HiSeq 2000 sequencing.

2.3 | ddRAD-seq library preparation

We prepared four ddRAD-seq libraries containing a total of 159 pregnant males, eight nonpregnant males, 160 offspring and 57 females (384 individuals total). We followed the ddRAD-seq library preparation method from Peterson et al. (2012) with several modifications described elsewhere (Flanagan & Jones, 2017; Flanagan, Rose, & Jones, 2016). Briefly, 1 μg of genomic DNA from each individual was digested with 100 units of *Pst*I-HF (New England Biolabs) and 25 units of *Mbo*I (New England Biolabs) in a 3-hr, 37°C incubation. Following purification by AMPure XP beads (Agilent), 250 ng of each DNA sample was ligated to barcoded adapters using T4 ligase (Epicentre) in a 23°C incubation lasting 30 min followed by a 10 min 65°C heat shock. These adapters were identical to those used in sdRAD-seq library preparation (see above; Baird et al., 2008). Ninety-six unique barcodes were used for each ddRAD-seq library, so we pooled the adapter-ligated fragments from 96 individuals after an AMPure XP bead (Agilent) purification. We extracted fragments in the range of 300–700 bp from a 1% agarose gel stained with SafeView (ABMGood). Phusion polymerase (New England Biolabs) was used to amplify the size-selected fragments in four separate rounds of PCR, each using twelve cycles (cycle conditions: 98°C for 30 s; 12 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 10 s; 72°C for 5 min). The four PCRs were pooled and cleaned with AMPure XP beads (Agilent). The quality of the final ddRAD-seq library was evaluated by visualizing DNA on a gel and quantifying it with a QUBIT FLUOROMETER 2.0 (Invitrogen). Four libraries, each of which contained 96 barcoded individuals, were sent to the University of Oregon Genomics Core Facility for 100 bp single-end Illumina HiSeq 2000 sequencing.

2.4 | A note on terminology

We conducted several analyses that resulted in 100-bp haplotypes, with each haplotype containing at least one SNP, distributed across the genome. Throughout the rest of the manuscript, we will refer to haplotypes derived from an analysis of both sdRAD-seq and ddRAD-seq sequencing reads together as RAD loci, and will refer to this analysis in general as the "combined RAD-seq data set," with "sdRAD Together" and "ddRAD Together" specifying individuals in this data set. Alternatively, sdRAD loci and ddRAD loci are the haplotypes derived from a separate analysis of the sdRAD-seq or ddRAD-seq sequencing reads, respectively. When referring to individuals analysed separately, they will be labelled as "sdRAD Separate" and "ddRAD Separate." RAD loci, sdRAD loci and ddRAD loci all can contain one or more single nucleotide polymorphisms (SNPs), which we will refer to RAD SNPs, sdSNPs and ddSNPs throughout the manuscript.

2.5 | Aligning raw reads

The raw reads from each sequencing run were separated by barcode using the process_radtags module of Stacks (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), and each individual's reads were aligned to the Gulf pipefish genome (Small et al., 2016) using Bowtie 2.0 (Langmead & Salzberg, 2012) with the –sensitive parameters. The genome contains 22 major linkage groups, corresponding to the

22 chromosomes of *S. scovelli*. Approximately 87% of the genome assembly is arranged on these linkage groups. The remaining genomic data are assembled into 1,574 scaffolds that have not yet been localized to a chromosome (Small et al., 2016).

2.6 | Treating the two data sets as one: genotyping and analysis

Both the ddRAD-seq and sdRAD-seq library preparations used *Pst*I, so the two libraries should share many loci. Therefore, we began by treating them as a single data set. We used the ref_map.pl module in Stacks (Catchen et al., 2011, 2013) to identify RAD loci from the aligned reads. For RAD loci to be assembled, we required a minimum of three raw reads (-m 3), and we allowed two mismatches when generating the catalog of RAD loci (cstacks –n 2). We then ran the populations module in Stacks (Catchen et al., 2013), requiring a minimum allele frequency of 0.05, the presence of each locus in at least 50% of the individuals and that each locus be present in males, females and offspring. We subsequently randomly chose one SNP per RAD locus.

Using the resulting vcf file, we compared the coverage statistics between sdRAD-seq and ddRAD-seq individuals using Wilcoxon rank-sum tests. Specifically, we calculated per-individual means and variances for each SNP in addition to the total number of reads per individual. Both allelic dropout and PCR duplicates result in uneven coverage of the two alleles at a locus (Andrews et al., 2016; Arnold et al., 2013; Davey et al., 2013; Gautier et al., 2013), so we compared the coverage of the reference and alternative alleles for each library preparation method. Specifically, at each SNP, we calculated the proportion of reads that belonged to the reference allele in heterozygotes (focusing on heterozygotes controlled for different allele frequencies across SNPs). Larger values therefore represent bias towards the reference allele and smaller values represent bias towards the alternative allele. To determine which loci might have the most extreme coverage, we calculated the mean and variance in coverage for all loci and used those as the null distribution.

Another way to investigate the influence of restriction site polymorphisms is to estimate the allelic dropout rates per SNP. We used GBSTOOLS (Cooke et al., 2016) to estimate SNPs with high rates of restriction site allelic dropout. After adding *Mbo*I cut sites to the GBSTOOLS scripts, we used GBSTOOLS to digest the *S. scovelli* genome (Small et al., 2016) using both *Pst*I and *Mbo*I (to estimate allelic dropout in the ddRAD data set) and using only *Pst*I (to estimate allelic dropout in the sdRAD data set). We then used GBStools and python v. 2.7 to estimate allelic dropout in the combined RAD-seq data set using normalization factors of 1.0.

To identify whether signatures of population structure emerged between the sdRAD-seq and ddRAD-seq individuals, we conducted a principal components analysis of population structure using PCADAPT (Luu, Bazin, & Blum, 2017). We also calculated F_{ST} values between the sdRAD-seq group and the ddRAD-seq group using functions in GWSCAR (https://github.com/spflanagan/gwscaR). For those calculations, $F_{ST} = (H_T - H_S)/H_T$, where $H_T = 2\bar{p}\bar{q}$ and $H_S = (1/n)\Sigma_{i=1}^{n} 2p_iq_i$. In these formulas, the observed allele frequencies are denoted as p and q, with $\bar{p}$ and $\bar{q}$ representing the mean values across n groups (for this analysis, $n = 2$). Thus, H_T is the expected heterozygosity among populations and H_S is the average expected heterozygosity within populations (Nei, 1986; Wright, 1943). We also evaluated how imposing a coverage filter (loci with average per-individual coverage between 3× and 50×) impacted the analysis. These analyses were performed in R version 3.3.1 (R Core Team 2017).

2.7 | Treating the two data sets separately: generating two stacks catalogs

To better understand each data set on its own, we analysed the ddRAD-seq sequences and the sdRAD-seq sequences separately in two separate runs of ref_map.pl in Stacks (Catchen et al., 2011, 2013) and then ran populations once for each data set to generate vcf files using three populations: males, females and offspring. The same parameter settings were used as above (minimum stack depth of 3, 2 mismatches allowed, minimum allele frequency of 0.05, loci present in 50% of individuals). We then repeated the analyses described above: compared coverage per individual and coverage per locus, calculated our allelic imbalance metric, estimated restriction site polymorphism using GBSTOOLS (Cooke et al., 2016), conducted a principal components analysis with PCADAPT (Luu et al., 2017) and compared allele frequencies using F_{ST}.

In addition to comparing the two different library preparation methods to each other, we compared the assembly methods (whether the ddRAD-seq and sdRAD-seq reads were analysed together or separately in Stacks). These comparisons were made using ANOVA on coverage statistics and F_{ST} values, with the statistic as the response variable and the library preparation and assembly methods as explanatory variables. To further investigate the relationship between variation in coverage and F_{ST} values, we binned loci into six coverage categories (3–5×, 5–10×, 10–20×, 20–30×, 30–50× and >50×) and used the R package lattice (Sarkar, 2008) to visualize the mean F_{ST} values for loci in each category.

The number of individuals sequenced using sdRAD-seq was much smaller than the number sequenced using ddRAD-seq. To ensure the patterns we observed in our F_{ST} results were not due to sample size, we randomly chose 60 individuals from the ddRAD-seq data set to compare to the 60 sdRAD-seq individuals. Additionally, we compared 60 ddRAD-seq individuals to a different set of 60 ddRAD-seq individuals as a control. This analysis was performed using the sdSNPs and ddSNPs generated from the separate analyses, and F_{ST} values were compared using ANOVAs where the sample size or the comparison (sdRAD to ddRAD or ddRAD to ddRAD) and the analysis approach (separate filtered; separate unfiltered; together filtered; or together unfiltered) were explanatory variables.

2.8 | Impact of library preparation on selection components analysis

To fully assess the impact of different RAD-seq library preparation methods on the results of an empirical study, we performed an

F_{ST}-based selection components analysis. Selection components analysis identifies signatures of selection on the genome by comparing allele frequencies in individuals from a single population at different life history stages (Christiansen & Frydenberg, 1973; Flanagan & Jones, 2015; Monnahan, Colicchio, & Kelly, 2015). In a previous analysis of the ddRAD-seq individuals from this study, we demonstrated that signatures of sex-biased viability selection and sexual selection are distributed across the genome in *S. scovelli* and that more loci show signatures of sex-biased viability selection than sexual selection (Flanagan & Jones, 2017). In that analysis, we inferred maternal alleles from the 130 father–offspring combinations and compared the inferred maternal alleles to the females collected in the population to identify putative signatures of sexual selection. Collected adult male and adult female allele frequencies were compared to identify sex-biased viability selection (Flanagan & Jones, 2017).

Because the sdRAD-seq individuals do not include any offspring, we could not repeat the sexual selection component of the analysis for sdRAD-seq individuals. However, we used the combined RAD-seq data set (ddRAD individuals and sdRAD individuals analysed together) and evaluated the results in comparison with an analysis based only on the ddRAD-seq data. Additionally, we compared allele frequencies in males and females in all three data sets (RAD-seq, ddRAD-seq and sdRAD-seq).

To perform the selection components analysis, we used the same procedure as described in Flanagan and Jones (2017), converted into R code, which we have made available in a package called GWSCAR (https://github.com/spflanagan/gwscaR). For the sexual selection analysis, we first inferred maternal alleles by subtracting the paternal allele from each offspring's genotype. We then calculated F_{ST} values between the inferred maternal alleles and the collected females, using the same methods described above. The value of $2NF_{ST}(k-1)$ has a chi-square distribution with $(k-1)(n-1)$ degrees of freedom, where k is the number of alleles, N is the total number of individuals and n is the number of populations sampled (Waples, 1987; Workman & Niswander, 1970). We applied this calculation to every locus in our analysis to calculate p-values and then applied the Benjamini and Hochberg (1995) false discovery rate to identify significant loci at the level of $\alpha = 0.05$.

This analysis was performed in the ddRAD-seq data set with 130 father–offspring combinations and in the combined RAD-seq data set with 153 father–offspring combinations. The comparison of males and females involved 57 females and 159 males in the ddRAD-seq data set, 30 males and 30 females in the sdRAD-seq data set, and 87 females and 189 males in the RAD-seq data set. In each of the comparisons, we used a single SNP from each RAD locus, and each locus was required to be present in at least 50% of the individuals in each group. We also retained only SNPs with a minor allele frequency of at least 0.05 and with an average coverage value between 5× and 20×. For each of the two selection components, we used ANOVA to compare the F_{ST} values using the type of analysis method as the explanatory variable.

2.9 | Comparing the results to samtools

Bias resulting from the Stacks analysis may have been a result not of the sequencing methods per se, but rather an artefact of the Stacks pipeline. Therefore, we used SAMTOOLS (Li, 2011; Li et al., 2009) to create consensus loci using both the 60 sdRAD individuals and the subset of 60 ddRAD individuals. We subsequently used bcftools to call variant SNP sites. In calling SNPs, we required that 50% of the individuals had data for a particular SNP (-d 0.5) and excluded sites where all samples were phased with the phase bit set at 0.05 (-p 0.05 –P full). Using VCFTOOLS (Danecek et al., 2011), we further filtered SNPs to remove indels and retain biallelic SNPs with a minor allele frequency of at least 0.05, and in R (R Core Team 2017), we removed loci with an average per-individual coverage below three reads per locus. To identify whether the quality scores were meaningful in the context of reducing bias, we also filtered the data set to include only loci with a quality score ≥30. We used the same samtools-bcftools-vcftools pipeline on the 60 sdRAD individuals and the subset of 60 ddRAD individuals separately. Shared loci between these two separate assemblies were identified using custom scripts in R (R Core Team 2017).

To identify whether similar patterns emerged using the samtools assembly as in the Stacks assembly, we calculated per-locus coverage statistics for the samtools data sets and the quality-filtered samtools data sets. We also calculated F_{ST} values between ddSNPs and sdSNPs in four different data sets: (i) analysed together by samtools; (ii) analysed together by samtools and filtered based on quality scores; (iiii) analysed separately by samtools; and (iv) analysed separately by samtools and filtered based on quality scores.

2.10 | In silico digestion of the reference genome

To model the impact of the different sources of error, we wrote a C++ program (https://github.com/spflanagan/SCA/tree/master/programs/insilico_radseq) to perform an in silico digestion of the reference genome sequences and model shearing bias, polymorphic restriction sites, PCR bias and uneven coverage. The program performs both a single-digest and a double-digest of the reference. When modelling polymorphic restriction sites, we used an approach similar to that of Gautier et al. (2013). We assumed that 10% of the restriction sites would be constant, and the constant loci were chosen randomly. The loci that were polymorphic had an expected proportion of nucleotides with a segregating mutation of $\theta = 4N_e\mu$, and θ was the same for both the single digestion and double digestion but each locus had its own μ. We altered the value of θ by changing both the effective population size (N_e) to be 5,000, 10,000 and 20,000, and by drawing μ from a uniform distribution of either $[10^{-9},10^{-8}]$ or $[10^{-8},10^{-7}]$.

This program parses the fasta file containing the reference genome and finds the restriction enzyme recognition sites. When conducting a single digest, the length of the fragment determines how many sheared fragments are generated. Because the sdRAD-seq library preparation sheared to an average fragment size of 500 bp, the fragment was sheared s times based on the length of the fragment, l, at s random locations on the fragment, according to the formula: $s = l/500$. The resulting RAD loci from either end of the fragment (any extra sheared parts from the middle of the fragment were discarded) were kept if their length was between 250 and 700 bp. If shearing bias was modelled, as each fragment was generated, it was only kept if adding that fragment to the set of processed fragments maintained an average fragment length above 500 bp, biasing the shearing towards longer fragments. To model the double digestion, only fragments with both restriction enzyme sites that were 250–700 bp in length were kept.

Once the in silico digestion was complete, a simulated population needed to be sampled at both sdRAD and ddRAD loci generated by the in silico digestion. Each locus was given a population-level allele frequency, either a random uniformly distributed number in the range of [0,1) or drawn from a normal distribution centred around 0.8 with a standard deviation of 0.13 ("skewed"). This skewed distribution used the mean and standard deviation of the SNPs from our combined data set. If the locus was shared between the single and double digest, it had the same population-level allele frequency in both. We modelled one biallelic SNP per locus, and individuals were randomly assigned genotypes based on the population-level allele frequencies. We then determined whether polymorphic restriction sites or PCR duplication affected the genotype at each locus for each individual. If the locus could have a polymorphic restriction site (i.e., it was not one of the 10% of sites that were constant), a number was drawn from a Poisson distribution with a mean $\theta_i * l_{RS}$, where θ_i is the proportion of nucleotides with a mutation for locus i and l_{RS} is the length of the restriction site (6 for *Pst*I and 4 for *Mbo*I). If the Poisson distribution returned a 1, then one of the alleles was randomly chosen to be dropped and the genotype for that locus became homozygous for the selected allele. If the Poisson distribution returned 2 or higher, then both alleles were dropped and the locus was missing for that individual. For the individuals sampled at the in silico ddRAD loci, the locus was evaluated twice in this manner as either restriction site could be polymorphic.

To model PCR duplication events, a Poisson distribution was used with a mean representing the percentage of reads per PCR cycle that would be duplicated multiplied by the number of PCR cycles. This approach assumed that more PCR cycles would result in higher duplication rates. If the Poisson distribution returned a 1 or higher, one of the two alleles was randomly chosen as the duplicated allele and replaced the genotype at the second allele. We varied the PCR duplication rate from 0 to 5% per cycle, and the number of cycles was 12 for the ddRAD loci and 20 for the sdRAD loci, mirroring the number of cycles we used in the library preparation steps.

Once the genotypes were assigned and affected (or not) by polymorphic restriction sites and PCR duplications, we calculated F_{ST} at the loci shared by the single and double digests between the individuals sampled by in silico ddRAD and in silico sdRAD. We sampled a total of 400 individuals, and either had a symmetric sampling scheme ($n_{sd} = n_{dd} = 200$) or asymmetric ($n_{sd} = 60$, $n_{dd} = 340$). F_{ST} was calculated as $F_{ST} = (H_T - H_S)/H_T$, where H_S is the weighted average expected heterozygosity in each subpopulation and H_T is the expected heterozygosity in the entire population. We ran the in silico digestion with different restriction site mutation rates, PCR duplication rates and population-level allele frequencies and compared the resulting F_{ST} values to each other and to the observed values from the empirical library preparation.

3 | RESULTS

3.1 | Assembly statistics

The analysis treating sdRAD-seq and ddRAD-seq reads together included data from 444 individuals, 60 of which were prepared using the sdRAD-seq method. After the pruning imposed by the populations module of Stacks (minor allele frequency ≥0.05, SNPs present in 50% of females, males and offspring; Catchen et al., 2013), 84,851 SNPs from 36,007 RAD loci were retained. The analysis of only the 60 sdRAD-seq individuals resulted in 250,425 sdSNPs from 115,708 sdRAD loci. The ddRAD-seq data set contained 69,109 ddSNPs from 31,956 ddRAD loci. The sdRAD and ddRAD data sets shared 49,893 SNPs on 23,396 RAD loci.

3.2 | Differences in coverage when assembled together versus separately

All of the coverage data were skewed, so we used the natural log to transform the following coverage metrics. Total coverage was affected by the interaction of the library preparation method (sdRAD-seq vs ddRAD-seq) and whether the data were analysed together or separately ($F_{1,169977} = 304.2$, $p = 2 \times 10^{-16}$). The sdRAD-seq individuals had more reads than ddRAD-seq individuals when they were analysed separately (Tukey's HSD $p < .0001$) but not together (Tukey's HSD $p = .58$; Figure 1). When the sdRAD-seq reads were analysed together with the ddRAD-seq reads, the average number of reads per individual was lower than coverage for sdRAD individuals when the two library preparation methods were analysed separately (Tukey's HSD $p < .0001$; Figure 1). The ddRAD-seq individuals had the same coverage regardless of whether they were analysed together or separately (Tukey's HSD $p = .16$; Figure 1).

The log-transformed per-SNP, per-individual coverage was higher in ddRAD than sdRAD when analysed separately (Tukey's HSD $p < .0001$) and when analysed together (Tukey's HSD $p < .0001$). When sdRAD-seq reads and ddRAD-seq reads were analysed together, sdSNPs had higher coverage than sdSNPs from the separate analysis (Tukey's HSD $p < .0001$), but ddRAD showed no difference (Tukey's HSD $p = .9996$; Figure 1).

One indication that loci may be biased due to restriction site polymorphism or PCR duplicates is whether one allele has higher coverage than the other allele at a given locus. We found that both library preparation methods and analysis approach impacted the proportion of reference reads in heterozygotes ($F_{1,211838} = 20.78$, $p = 5.16 \times 10^{-6}$). The ddRAD-seq individuals had higher proportions than the sdRAD-seq individuals when analysed together (Tukey's

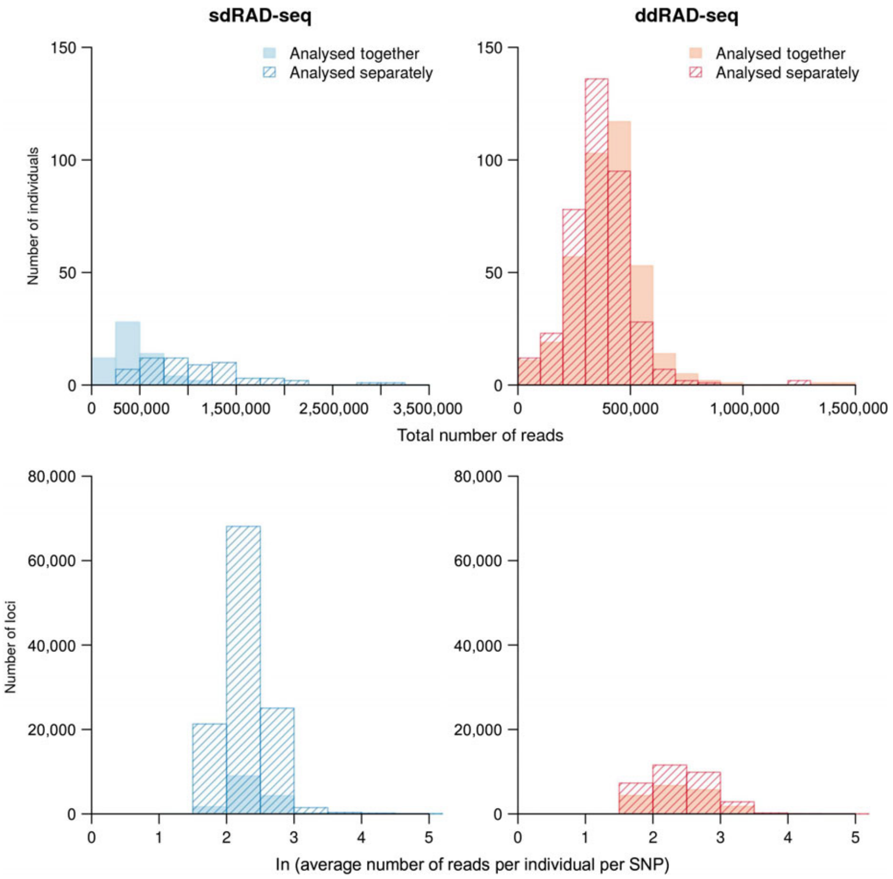

FIGURE 1 Comparison of locus coverage statistics in sdRAD-seq and ddRAD-seq data sets when analysed together and separately. The sdRAD-seq data set had a higher average number of reads per individual when assembled separately from the ddRAD-seq data set, whereas the ddRAD-seq loci had higher per-individual coverage when assembled together with the sdRAD-seq reads (top row). The ddRAD-seq data set had more reads per individual per locus than the sdRAD-seq data set (bottom row)

HSD $p < .0001$) but not when analysed separately (Tukey's HSD $p = .0685$). The proportion of reference reads in heterozygotes was lower in ddRAD when analysed separately from sdRAD (Tukey's HSD $p < .0001$), whereas sdRAD proportions were consistent regardless of analysis method (Tukey's HSD $p = .8934$; Table 1).

The GBSTOOLS analysis estimated the number of dropped alleles per individual per SNP. The estimated dropped allele count was influenced by both the library preparation (ddRAD-seq vs sdRAD-seq) and the analysis method ($F_{1,101116} = 85.93$, $p < 2 \times 10^{-16}$; Table 1). The sdRAD-seq library displayed more evidence for dropped alleles than ddRAD-seq library when they were analysed alone (Tukey's HSD, $p < .0001$) or together (Tukey's HSD, $p < .0001$). The ddRAD-seq showed more dropouts when analysed together than alone (Tukey's HSD, $p < .0001$), and we saw the same pattern for sdRAD-seq (Tukey's HSD, $p < .0001$).

PCR duplicates are expected to result in variance in coverage depth at a locus (Andrews et al., 2016). Increasing the number of PCR cycles will increase the number of PCR duplicates in the sequencing library, so we expected that the sdRAD-seq data set would suffer more from the problem of PCR duplicates. We compared the average variance in coverage across all SNPs between the two library preparation methods and between the two analysis approaches. We log-transformed the variance in coverage for each SNP and found that the library preparation method and the analysis method interact to influence variance in coverage ($F_{1,169977} = 474.5$, $p < 2 \times 10^{-16}$; Table 1). The sdSNPs had higher variance in coverage than ddSNPs both when analysed together (Tukey's HSD $p < .0001$) and separately (Tukey's HSD $p < .0001$). Within the sdSNPs, the variance in coverage was higher when assembled together than alone (Tukey's HSD $p < .0001$), whereas ddSNPs had higher variance in coverage when assembled alone (Tukey's HSD $p < .0001$).

3.3 | Observed heterozygosity

Observed heterozygosity is expected to be affected by polymorphic restriction sites, leading to decreased observed heterozygosity at some loci. In contrast to our expectations, the ddRAD-seq data set had higher proportions of heterozygotes than sdRAD-seq regardless of whether they were analysed together (Tukey's HSD $p < .0001$) or separately (Tukey's HSD $p < .0001$), although the interaction term was significant ($F_{1,169977} = 128.10$, $p < 2 \times 10^{-16}$; Table 1), indicating that the library preparation method impacted heterozygosity in different ways depending on the way the data were analysed. The subset of 60 ddRAD individuals used in the F_{ST} analysis also had a higher mean proportion of heterozygotes than the sdRAD individuals (Wilcoxon signed-rank test $W = 1033200000$, $p < 2.2 \times 10^{-16}$).

3.4 | Analysis of population structure using principal components analysis

When we applied the principal components approach of PCADAPT (Luu et al., 2017) to identify population structure in the data set with

TABLE 1 The impact of analysing ddRAD-seq and sdRAD-seq reads together or separately on variables related to allelic dropout and PCR duplications

	sdRAD separately	ddRAD separately	sdRAD together	ddRAD together
Average Coverage Per SNP	11.55 (78.26)	13.26 (29.43)	15.12 (173.16)	13.44 (31.08)
Mean (*SD*) [min–max]	[4.06–20075.63]	[3.76–2440.11]	[4.2–20075.63]	[4.43–2222.23]
Coverage Variability (in *SD*)	6.76 (39.14)	6.69 (21.14)	8.59 (83.67)	6.52 (19.27)
Mean (*SD*) [min–max]	[1.43–9542.12]	[1.68–2405.49]	[1.35–9542.12]	[1.68–1491.7]
Proportion Heterozygotes	0.2 (0.14)	0.24 (0.16)	0.22 (0.14)	0.24 (0.15)
Mean (*SD*) [min–max]	[0–1]	[0–1]	[0–1]	[0–1]
Proportion of Reference Reads in Heterozygotes	0.4998 (0.0609)	0.5008 (0.0543)	0.4995 (0.0669)	0.5034 (0.0697)
Mean (*SD*) [min–max]	[0.0231–0.9741]	[0.0067–0.9094]	[0.0073–0.9741]	[0.0067–0.9745]
Number of Dropped Alleles	0.15 (0.23)	0.07 (0.13)	0.23 (0.35)	0.11 (0.19)
Mean (*SD*) [min–max]	[0–1.01]	[0–0.95]	[0–1.92]	[0–1.4]

These are the results of the analysis using Stacks. Shown for each variable are the mean, variance (in parentheses) and range (in square brackets: [min. – max.]). When analysed separately, ddRAD-seq has higher per-SNP coverage, but the pattern is reversed when the data sets are analysed together. Regardless of analysis approach, the ddRAD-seq data set has slightly higher heterozygosity compared to the sdRAD-seq data set. A decrease in heterozygosity may reflect allelic dropout as a consequence of polymorphic restriction sites. The sdRAD-seq data set has higher mean variance in coverage than the ddRAD-seq data set, possibly reflecting the impact of PCR duplicates (note that the values presented in the table for the coverage variability are the summary statistics for the coverage standard deviations, rather than the variances). The proportion of reference reads in heterozygotes indicates whether one allele was overrepresented at a given locus. The ddRAD-seq library had higher proportions of reference reads in heterozygotes than the sdRAD-seq library, and the proportion was higher in the ddRAD-seq library when it was analysed together with the sdRAD-seq library than when it was analysed separately. The number of dropped alleles are estimated by GBStools. Analysing the data set together results in more dropout alleles being maintained in the data set, and sdRAD-seq has a higher number of dropout alleles than ddRAD-seq, regardless of analysis method.

ddRAD-seq and sdRAD-seq individuals combined, we found that 24.6% of the variation was explained by library preparation method (Figure 2). Alternatively, in the analysis of the separate ddRAD-seq and sdRAD-seq data sets, individuals did not sort based on library preparation method, and the first axis of variation only explained 6.5% of the variation, suggesting that differences between the two sets of individuals were small.

3.5 | Comparison of allele frequencies in ddRAD-seq and sdRAD-seq

We compared allele frequencies between the ddRAD-seq and sdRAD-seq individuals in two ways: when they were analysed separately (as if comparing results from separate studies) and when they were analysed together (treating them as if they were one data set). When the individuals were analysed together, many more loci were fixed for one allele or the other in both the ddRAD-seq and the sdRAD-seq data sets, but the overall distributions of allele frequencies were similar between ddRAD individuals and sdRAD individuals (Figure S1). The assembly method significantly impacted F_{ST} values between ddRAD and sdRAD individuals ($F_{1,74171} = 228.3$, $p < 2 \times 10^{-16}$; Figures 3, S2).

When the ddRAD and sdRAD individuals were analysed separately (14,324 shared SNPs), the mean major allele frequency was significantly higher in sdSNPs ($\mu = 0.7901$) than in ddSNPs ($\mu = 0.7888$; one-sided paired Wilcoxon signed-rank test $V = 49279000$, $p = .00015$). This pattern was reflected in the F_{ST} values, which had a mean of 0.00307 and ranged from 0 to 0.4883 (Figures 3, S2). When the two sets of individuals were analysed together (27,334 SNPs), the mean major allele frequency was higher in ddRAD individuals than sdRAD individuals ($\mu_{ddRAD} = 0.8211$, $\mu_{sdRAD} = 0.7038$; one-sided paired Wilcoxon signed-rank test $V = 379090000$, $p < 2.2 \times 10^{-16}$), and mean F_{ST} was 0.00604 and ranged from −1.6309 to 0.9576. This significantly higher mean (Tukey's HSD $p < .0001$) did not include the 4,440 SNPs that were fixed for different alleles in the combined analysis. Any similar sites in the combined analysis would not have been polymorphic in at least one of the separate data sets, and so, those loci were not retained.

One approach to ameliorate the bias of RAD-seq, particularly the bias due to PCR duplicates, is to remove SNPs from the analysis with high per-SNP, per-individual coverage (Schweyen et al., 2014). Therefore, we imposed a filter to retain SNPs with an average coverage between 3× and 50× on both data sets, and this filter significantly impacted the F_{ST} values ($F_{1,74171} = 292.2$, $p < 2 \times 10^{-16}$; Figures 3, S2), although its effect interacted with the analysis approach ($F_{1,74171} = 133.7$, $p < 2 \times 10^{-16}$). In the comparison of allele frequencies at sdSNPs and ddSNPs generated from separate analyses, the filter removed 4,363 SNPs (9,961 SNPs were retained) and changed the mean F_{ST} to 0.00266, although this was not a significant change (Tukey's HSD $p = .94$). The filter did significantly decrease the mean F_{ST} to −0.0033 for the analysis of the SNPs together (Tukey's HSD $p < .0001$). Filtering for coverage removed 16,778 loci from the combined analysis, with 10,556 SNPs retained. The filter removed most of the SNPs with fixed alleles between sdRAD and ddRAD, with only 301 such SNPs remaining. Keeping loci with coverage between 3× and 50× in both the separate and combined analyses removed any difference in the mean between the two sets of F_{ST} values (Tukey's HSD $p = .637$). The importance of

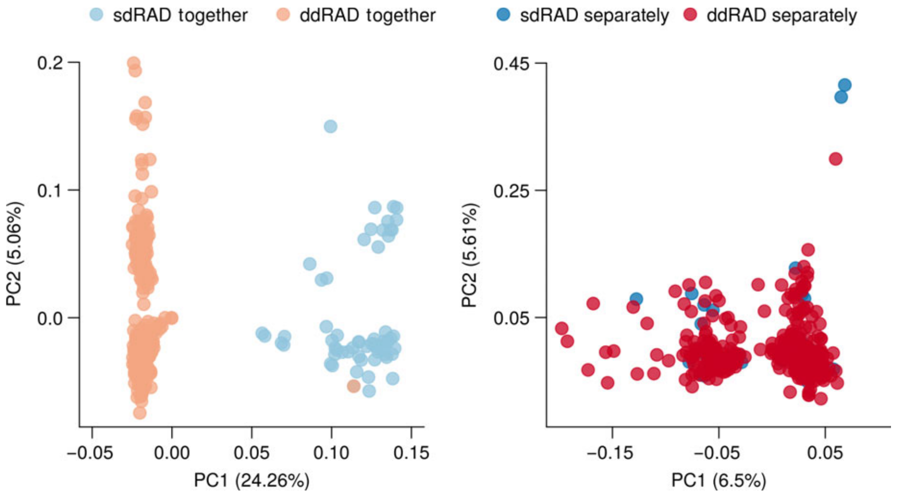

FIGURE 2 Principal components analysis of the sdRAD-seq and ddRAD-seq data sets when analysed together (left; light blue and orange) or separately (right; dark blue and dark red). When analysed together, the primary axis of variation in the data set separates the individuals sequenced by sdRAD and ddRAD, whereas when the data were analysed separately the sdRAD and ddRAD individuals overlap

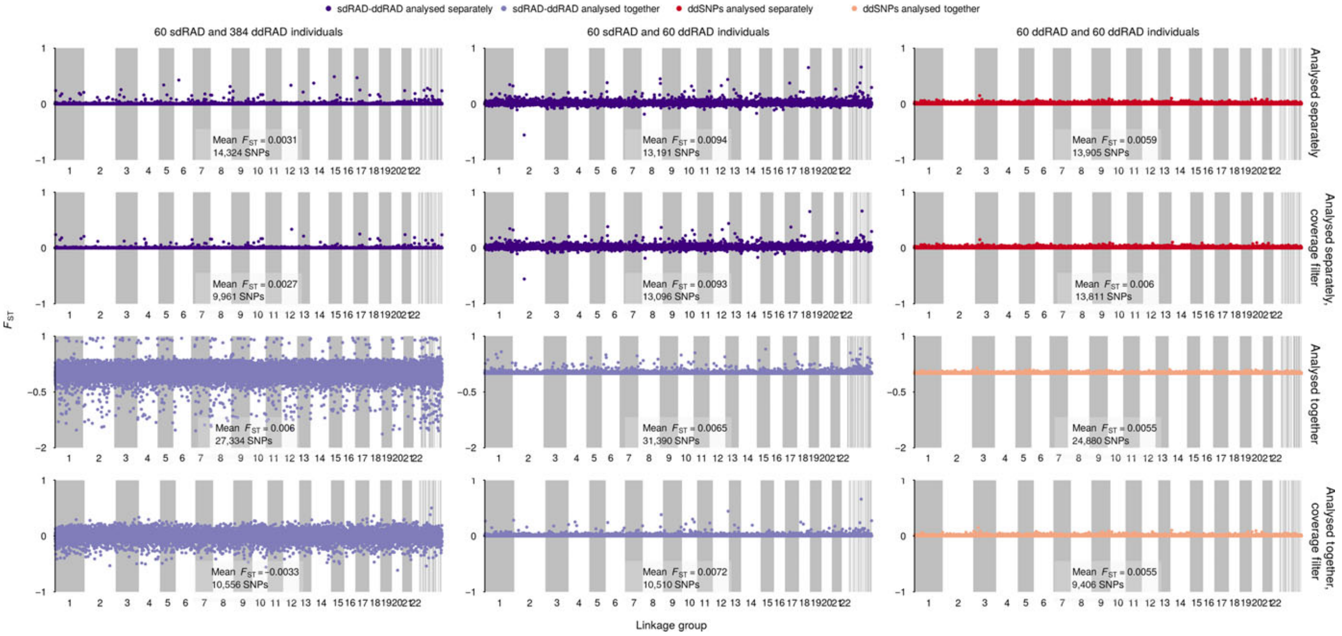

FIGURE 3 Comparison of sdRAD-seq and ddRAD-seq allele frequencies using F_{ST}. We analysed the complete data sets (left column), the complete sdRAD data set and a subset of the ddRAD data set (middle column) and two subsets of the ddRAD data set (right column) to evaluate the impact of sample size on the allele frequencies. For each of the comparisons, we analysed the sdRAD-seq and ddRAD-seq data in two separate runs of Stacks (Catchen et al., 2011, 2013), which are presented in the top two rows. We also pooled the sdRAD-seq and ddRAD-seq data and analysed them in a single analysis in Stacks (Catchen et al., 2013; bottom two rows; Catchen et al., 2011). For each of the analyses, we also investigated the effect of imposing a coverage filter to remove SNPs with average per-SNP coverage ≤3× or >50×. The number of SNPs and the mean F_{ST} values for the analyses are presented in each panel

coverage filters is emphasized when looking at a heatmap of mean F_{ST} values for SNPs binned by coverage (Figure 4). This figure shows that the most extreme F_{ST} values emerge when one group has mid-range coverage (e.g., 10–20×), but the other group has exceedingly high coverage (>50×).

To test for the effects of the number of sampled individuals, we also compared allele frequencies between all 60 sdRAD individuals and 60 randomly selected ddRAD individuals and between 60 ddRAD individuals and a separate randomly chosen set of 60 ddRAD individuals. For each of these comparisons, we performed the same four F_{ST} comparisons as above (analysed together, analysed together with a coverage filter, analysed separately and analysed separately with a coverage filter). We found that sampling fewer individuals in the ddRAD-seq data set resulted in higher F_{ST} values ($F_{1,121180}$ = 4.997, p = .0254; Figures 3, S2), although the type of analysis interacted with the effect of sample size ($F_{3,121180}$ = 6.298,

p = .0003; Figures 3, S2). The comparison of 60 ddRAD-seq individuals to 60 other ddRAD-seq individuals resulted in overall lower F_{ST} values than the corresponding comparisons of sdRAD-seq and ddRAD-seq individuals ($F_{1,179742}$ = 13.04, p = .0003; Figures 3, S2), although the analysis type interacted with the effect of whether ddRAD individuals were being compared to sdRAD or other ddRAD individuals ($F_{3,179742}$ = 15.72, $p = 3.21 \times 10^{-10}$; Figures 3, S2).

3.6 | Differences in the outcome of selection components analysis

For the selection components analysis, we filtered loci to retain only those with coverage between 5× and 20×. After the filtering step, we compared allele frequencies in males and females using 28,230 SNPs, 48,743 ddSNPs and 92,710 sdSNPs. The comparison of maternal alleles to collected females could only be conducted in the combined RAD data set and the ddRAD data set, as no offspring were genotyped using sdRAD-seq. The inference of maternal alleles reduced the number of loci used to 16,099 SNPs and 35,666 ddSNPs, because not all father–offspring genotype combinations facilitate inference of the maternal allele (Flanagan & Jones, 2017). However, it is worth noting that the inference of maternal alleles does not bias allele frequencies unless error rates are high (Flanagan & Jones, 2017).

The comparison of allele frequencies between males and females to test for sex-biased viability selection yielded dramatically different results depending on which sequencing and analysis methods were used. The male–female analysis using sdRAD-seq had higher F_{ST} values than the equivalent comparison in the ddRAD-seq analysis (ANOVA $F_{2,169680}$ = 4,567, $p < 2 \times 10^{-16}$; Figures 5, S3) and higher values than the analysis of both ddRAD-seq and sdRAD-seq individuals together (Tukey's HSD $p < 2 \times 10^{-16}$), although the combined analysis had higher F_{ST} values than the ddRAD comparison. The combined analysis also identified 4,315 significant SNPs after correcting for multiple comparisons whereas the ddRAD analysis identified only 58 significant ddSNPs (Figures 5, S3). Note that this number is different from the number identified in Flanagan and Jones (2017) because our previous analysis used slightly more stringent filtering methods than those used here.

To ensure that the differences between the ddRAD-seq and sdRAD-seq selection components analysis comparing males and females were not simply driven by the different sample sizes, we randomly sampled 30 males and 30 females from the ddRAD-seq data set and compared their allele frequencies. The smaller sample size yielded significantly higher F_{ST} values than the full ddRAD-seq data set (one-sided Wilcoxon signed-rank test, W = 1,435,300,000, $p < 2.2 \times 10^{-16}$), but the F_{ST} values were still significantly lower than the sdRAD-seq data set (one-sided Wilcoxon signed-rank test, W = 1,967,400,000, $p < 2.2 \times 10^{-16}$).

The test for sexual selection yielded similar results to that of males versus females. Again, the combined analysis had many more significant SNPs. The selection components analysis of the combined data set identified 125 significant SNPs after Benjamini and Hochberg (1995) false discovery rate correction, whereas only 16 ddSNPs were significant in the ddRAD selection components analysis. However, the ddRAD-seq analysis had a higher mean F_{ST} value than the combined analysis in this case (ANOVA $F_{1,51763}$ = 8.668, p = .00324; Figure 5), although the range of F_{ST} values was larger in the combined analysis (0.0000–0.2915 in combined, 0.0000–0.1966 in ddRAD).

3.7 | Comparison to samtools analysis

To verify that the results were due primarily to differences in the underlying data rather than to artefacts in the analysis methods, we re-analysed the sdRAD individuals and the subset of 60 ddRAD individuals with samtools, bcftools and vcftools. The samtools analysis of sdRAD and ddRAD individuals together resulted in 133,946

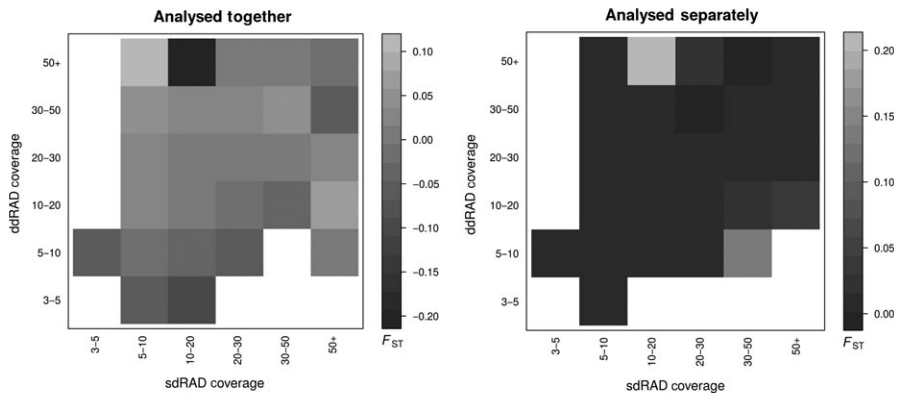

FIGURE 4 The impact of sdRAD and ddRAD coverage on F_{ST} values when the data sets were analysed separately (left) and together (right). The values presented here are mean F_{ST} values for each coverage category without any coverage filters imposed. When analysed separately, the highest F_{ST} values (light grey) occurred when one data set had medium coverage (5–20×) and the other had high coverage (30+). When analysed together, many of the F_{ST} values were negative, which is indicative of major issues with the data set. These extreme values occurred when ddRAD had high coverage (≥50×) and sdRAD had medium coverage (10–20×). Note that the two panels are on different scales

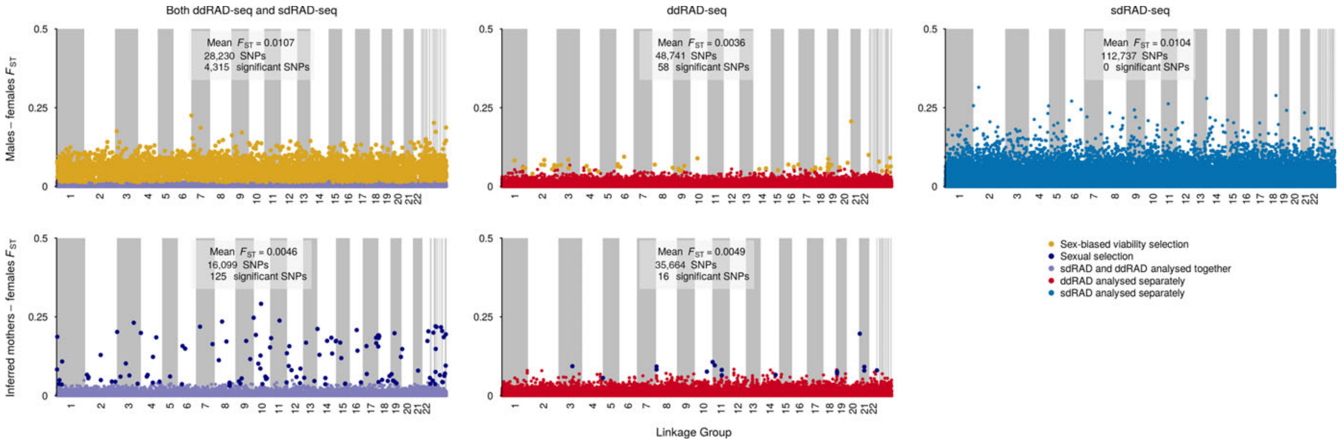

FIGURE 5 The type of sequencing method and the data analysis method greatly impacted the results of the selection components analysis. Analysing the ddRAD-seq and sdRAD-seq data sets together (left column) yielded higher F_{ST} values than when using only ddRAD-seq (middle column). Analysing only the sdRAD-seq data set (right column) resulted in higher F_{ST} values between males and females than both the combined analysis and the ddRAD analysis. The colours represent the different types of analyses (sdRAD and ddRAD analysed together in purple, ddRAD analysed separately in red and sdRAD analysed separately in blue) and the different types of inferred selection (sexual selection in dark blue and sex-biased viability selection in yellow)

variant SNPs, after filtering for coverage. The samtools analysis of sdRAD individuals separately resulted in 232,101 variant SNPs, and the samtools analysis of ddRAD individuals separately yielded 63,934 variant SNPs, with 53,057 shared variant SNPs between the two analyses.

These samtools results share some similarities with the Stacks analysis of the sdRAD-seq individuals and a subset of 60 of the ddRAD-seq individuals. Average coverage depth per locus per individual (total depth at the locus/the number of genotyped individuals) was determined by an interaction between library preparation method (ddRAD-seq vs sdRAD-seq) and analysis approach (analysed together or separately; $F_{1,563923} = 1526.5$, $p < 2 \times 10^{-16}$), with both sdSNPs and ddSNPs analysed together having lower average coverage than when analysed separately (sdRAD Tukey's HSD $p < .0001$; ddRAD Tukey's HSD $p < .0001$). When analysed separately, the sdRAD-seq data set had higher coverage than the ddRAD-seq data set (Tukey's HSD $p < .0001$; Table 2). Similar to Stacks, when analysed separately, the ddRAD-seq data set had higher coverage. However, in samtools the average coverage was lower when the data sets were analysed together, whereas coverage increased for sdRAD but not ddRAD in the Stacks analysis. Note that samtools outputed locus-wide coverage statistics instead of per-locus, per-individual coverage statistics, so we were unable to detect such trade-offs in the samtools data set.

The proportion of heterozygotes in the samtools analysis was determined by the interaction between library preparation method and whether the data were analysed together or separately ($F_{1,563923} = 24.667$, $p < 2 \times 10^{-16}$), similar to the Stacks analysis. When analysed separately, the ddRAD-seq data set had a higher proportion of heterozygotes than the sdRAD-seq data set (Tukey's HSD $p < .0001$). The sdRAD-seq data set had a higher proportion of heterozygotes than when it was analysed with the ddRAD-seq data (Tukey's HSD $p < .0001$), whereas the ddRAD-seq data set had a higher proportion of heterozygotes when it was analysed without the sdRAD individuals (Tukey's HSD $p < .0001$). These patterns are consistent with the overall analysis with Stacks (Table 1) and with the analysis of the Stacks data set with a subset of 60 ddRAD individuals (Table 2).

The samtools analysis also resulted in large F_{ST} values between the two data sets. We compared F_{ST} values between ddSNPs and sdSNPs with a basic coverage filter (equivalent to the baseline filtering performed in the Stacks analysis), but we also used the samtools quality scores to impose an additional filter. F_{ST} values between data sets were dependent on both the analysis method and whether the additional quality filter was imposed ($F_{1,360635} = 211.4$, $p < 2 \times 10^{-16}$). Without the filter, the F_{ST} values were not significantly different when the data sets were analysed together or separately (Tukey's HSD $p = .6053$), but after including the quality filter, the separate analysis had lower F_{ST} values than the analysis of the data together (Tukey's HSD $p < .0001$). For both analysis approaches, including the quality filter reduced the F_{ST} values (separate Tukey's HSD $p < .0001$; together Tukey's HSD, $p < .0001$).

3.8 | Analysis of the in silico digestion

The in silico digestions of the reference genome demonstrated the importance of PCR duplications, restriction site polymorphism, shearing bias, mean allele frequency skew and asymmetric sampling schemes for allele frequency estimation in studies employing ddRAD-seq or sdRAD-seq. We began by isolating one or two sources of bias at a time. Both the percentage of reads resulting from PCR duplicates ($F_{1,84148} = 38.4$, $p < 5.82 \times 10^{-10}$) and whether the allele frequencies were skewed towards large major allele frequencies affected F_{ST} values, although the effect of a skewed allele frequency spectrum had a more dramatic effect ($F_{1,84148} = 2056.9$, $p = 2 \times 10^{-16}$) and the interaction was not

TABLE 2 Analysing the data sets with Stacks and samtools results in similar differences in coverage, proportion of heterozygotes and extreme F_{ST} values

	Stacks				Samtools			
	sdRAD separately	**ddRAD separately**	**sdRAD together**	**ddRAD together**	**sdRAD separately**	**ddRAD separately**	**sdRAD together**	**ddRAD together**
Average Coverage Per SNP; mean (variance) [range]	11.55 (6,124.26) [4.06–20,075.63]	13.8 (1,364.11) [4–2,870.55]	15.12 (29,985.22) [4.2–20,075.63]	13.05 (775.57) [3.68–2,145.45]	8.95 (35.41) [3–253.95]	10.55 (63.27) [3–251.53]	8.13 (55.69) [3–252.74]	8.13 (55.69) [3–252.74]
Proportion Heterozygotes; mean (variance) [range]	0.2 (0.02) [0–1]	0.25 (0.02) [0–1]	0.22 (0.02) [0–1]	0.21 (0.03) [0–1]	0.25 (0.02) [0–1]	0.27 (0.02) [0–1]	0.26 (0.02) [0–1]	0.26 (0.07) [0–1]
F_{ST}; mean (variance) [range]	0.01 (0.0007) [−0.56 to 0.66]		0.01 (0.0005) [0–0.79]		0.01 (0.0009) [0–0.77]		0.01 (0.0016) [0–0.81]	
Filtered F_{ST}; mean (variance) [range]	0.01 (0.0006) [−0.56 to 0.66]		0.01 (0.0003) [0–0.66]		0 (0.0005) [0–0.56]		0.02 (0.0013) [0–0.71]	

This table presents the results of analysing the 60 sdRAD-seq individuals and a random subset of 60 ddRAD-seq individuals either together or separately. In both cases, the reads were aligned to the *S. scovelli* genome prior to analysis. For each variable displayed above, we report the mean, variance (in parentheses) and range (in square brackets: [min. – max.]). In both analyses, coverage differed between sdRAD SNPs and ddRAD SNPs. The samtools data set had less extreme coverage values and variances, likely because samtools removes indels, which Stacks does not. This additional filter likely removed the most problematic loci, which in Stacks were those with extremely high coverage. In both the Stacks and samtools analyses, the ddRAD-seq data set had a higher proportion of heterozygotes than the sdRAD-seq data set when they were analysed separately. Both Stacks and samtools resulted in some extreme F_{ST} values between ddRAD-seq and sdRAD-seq data sets, even after filtering for coverage thresholds (Stacks analysis) or using quality filters (samtools).

significant ($F_{1,84148}$ = 0.019, p = .89; Table 3). Unsurprisingly, higher rates of polymorphic restriction sites ($\theta = 4N_e\mu$) impacted the differentiation between sdRAD and ddRAD loci ($F_{1,42015}$ = 18.5, $p = 1.7 \times 10^{-5}$), although this effect was primarily driven by altering the N_e value ($F_{1,42013}$ = 40.963, $p = 1.57 \times 10^{-10}$) rather than the mutational distribution ($F_{1,42013}$ = 1.096, p = .295; Table 3).

One source of bias that could create differences between sdRAD loci and ddRAD loci is shearing bias. However, shearing bias did not significantly affect the distribution of F_{ST} values ($F_{1,32658}$ = 0.179, p = .671898). More important factors were the symmetry of the sampling scheme ($F_{1,32658}$ = 212.879, $p < 2 \times 10^{-16}$) and whether the average allele frequency was set at 0.8 instead of 0.5 ($F_{1,32658}$ = 836.636, $p < 2 \times 10^{-16}$), two factors which had a significant interaction ($F_{1,32658}$ = 11.528, p = .000686; Table 3).

In an actual study, all of these sources of bias are expected to be present. Therefore, we ran the in silico digestion with all of the biases and changed one bias at a time to see how different parameters interacted in a complex system of bias. The variables that increased the F_{ST} values the most were skewed allele frequencies (t = 11.756, $p < 2 \times 10^{-16}$) and the symmetry of the sampling scheme (t = 5.679, $p = 1.37 \times 10^{-8}$; Table 3). Although shearing bias alone did not inflate F_{ST} values, when combined with other factors, shearing bias had a significant effect on F_{ST} (t = 2.037, p = .0417; Table 3).

4 | DISCUSSION

Here, we present a case study, which involves sampling a single population using both sdRAD-seq and ddRAD-seq. This analysis provides a unique empirical opportunity to investigate sources of bias in two different RAD-seq methods. Pairing our empirical analysis with an in silico digestion of the Gulf pipefish genome allows us to assess the importance of factors such as shearing bias, polymorphic restriction sites, PCR duplicates, allele frequencies, and sampling schemes on differentiation between individuals genotyped using sdRAD-seq and those genotyped by ddRAD-seq.

The analysis approach played an important role in the outcomes of the data analysis at every step, from the coverage of loci, to restriction site dropout, to differences between ddRAD-seq and sdRAD-seq allele frequencies. We analysed the data sets using both Stacks and samtools, and identified similar trends with a few small differences. In both analyses, the sdRAD-seq data set had higher per-individual, per-SNP coverage (Figure 1, Table 2), an unsurprising result as the 60 sdRAD individuals were sequenced in one lane, whereas the ddRAD individuals were pooled 96 per lane. Analysing the data with both Stacks and samtools resulted in a higher proportion of heterozygotes in the ddRAD-seq data sets than the sdRAD-seq data sets, especially when the data were treated separately. Finally, both the Stacks and samtools analyses resulted in large F_{ST} values between the sdRAD-seq and ddRAD-seq individuals, suggesting that the patterns we observed are due to underlying differences in the data set, not due to bias arising from differences between the analysis pipelines. In both cases, however, it is clear that more bias is introduced to the analysis when ddRAD-seq and sdRAD-seq are analysed together as one data set. For this, we can offer a clear suggestion to researchers: if dealing with data generated by different methods, analyse them separately and then identify overlap between the data sets.

Despite the fact that all sampled Gulf pipefish came from a single population, we observed significantly different allele frequencies between individuals that were genotyped using sdRAD-seq and those genotyped using ddRAD-seq. If coverage filters were not in place, the F_{ST} values ranged up to 0.9788 when the two types of libraries were analysed together and 0.4883 when analysed separately. These extreme values were much larger than the maximum values observed between geographically distinct populations of *S. scovelli* (Flanagan et al., 2016). The difference between genotyping methods was not simply due to the fact that we sampled 60 individuals using sdRAD-seq and 384 using ddRAD-seq, because a comparison of 60 randomly selected ddRAD individuals to the 60 sdRAD individuals yielded F_{ST} values that were substantially higher than a comparison of 60 ddRAD individuals to another 60 ddRAD individuals (Figure 3). However, skewed sample sizes did exaggerate differences between the two data sets in our in silico digestion of the reference genome (Table 3).

The differences between the data sets generated by different RAD-seq methods were partly a result of the differences in coverage we observed between the data sets. Our sdRAD-seq data set had higher per-individual coverage than the ddRAD-seq data set primarily because the sdRAD-seq data set contained only 60 individuals sequenced in one Illumina lane, whereas the ddRAD-seq data set comprised data from four lanes of Illumina sequencing, each with 96 individuals. Similarly, the ddRAD-seq data set had fewer SNPs (31,956 ddSNPs compared to 113,166 sdSNPs used in the analyses; Figure 1). Variance in coverage and allelic dropout can result from pooling of individuals right after the ligation rather than immediately before sequencing (daCosta & Sorenson, 2014), but as both the sdRAD-seq and ddRAD-seq individuals were pooled at the same step in the present analysis, this factor is unlikely to explain the differences between the two data sets. Regardless of the source of the variance in coverage, differences in coverage between the two data sets appear to have driven extreme F_{ST} values (Figures 3 and 4). Additionally, the results from our in silico digestion suggest that loci with major allele frequencies skewed towards a value of one result in elevated F_{ST} (Table 3). As F_{ST} is calculated from allele frequencies, it is unsurprising that we found increased F_{ST} values between methodologies when the mean allele frequency was elevated in our in silico digestion (see Jakobsson, Edge, & Rosenberg, 2013; Jost, 2008). However, our results are a reminder to researchers that inflated allele frequencies due to bias in RAD-seq can yield inflated estimates of F_{ST}.

Polymorphic restriction sites were found to play a major role in bias in this RAD-seq study. In our in silico digestion, increasing the rate of mutations at restriction sites significantly increased F_{ST} values between sdRAD-seq and ddRAD-seq samples, although this effect

TABLE 3 Results of the in silico digestion

	PCR Dup.	n_{sd}	n_{dd}	Mean AF	Shearing Bias	Restriction Site Mut.	N_e	Mean ± SE F_{ST}	Max. F_{ST}
PCR Duplication and Skewed AF	0	200	200	0.5	NotBiased	0	10,000	0.00056 ± 0.000013	0.01679
	1	200	200	0.5	NotBiased	0	10,000	0.00054 ± 0.000012	0.01309
	2	200	200	0.5	NotBiased	0	10,000	0.00058 ± 0.000013	0.01529
	3	200	200	0.5	NotBiased	0	10,000	0.00059 ± 0.000012	0.01268
	4	200	200	0.5	NotBiased	0	10,000	0.00061 ± 0.000013	0.01438
	5	200	200	0.5	NotBiased	0	10,000	0.00063 ± 0.000013	0.01349
	1	200	200	0.8	NotBiased	0	10,000	0.00097 ± 0.000017	0.01569
	2	200	200	0.8	NotBiased	0	10,000	0.00096 ± 0.000017	0.01730
	3	200	200	0.8	NotBiased	0	10,000	0.00095 ± 0.000017	0.01603
	4	200	200	0.8	NotBiased	0	10,000	0.00103 ± 0.000018	0.01708
	5	200	200	0.8	NotBiased	0	10,000	0.00100 ± 0.000017	0.01963
Restriction site polymorphisms	0	200	200	0.5	NotBiased	10^{-7} to 10^{-8}	5,000	0.00054 ± 0.000012	0.01698
	0	200	200	0.5	NotBiased	10^{-7} to 10^{-8}	10,000	0.00056 ± 0.000012	0.01398
	0	200	200	0.5	NotBiased	10^{-7} to 10^{-8}	20,000	0.00063 ± 0.000014	0.01809
	0	200	200	0.5	NotBiased	10^{-8} to 10^{-9}	5,000	0.00053 ± 0.000011	0.01072
	0	200	200	0.5	NotBiased	10^{-8} to 10^{-9}	10,000	0.00056 ± 0.000013	0.01861
	0	200	200	0.5	NotBiased	10^{-8} to 10^{-9}	20,000	0.00060 ± 0.000013	0.01426
Skewed AF, shearing bias and asymmetric sampling	0	200	200	0.5	NotBiased	0	10,000	0.00056 ± 0.000013	0.01679
	0	200	200	0.5	Biased	0	10,000	0.00051 ± 0.000031	0.01386
	0	200	200	0.8	Biased	0	10,000	0.00085 ± 0.000037	0.01338
	0	200	200	0.8	NotBiased	0	10,000	0.00093 ± 0.000016	0.01625
	0	340	60	0.5	Biased	0	10,000	0.00040 ± 0.000022	0.00649
	0	340	60	0.5	NotBiased	0	10,000	0.00040 ± 0.000009	0.01042
	0	340	60	0.8	Biased	0	10,000	0.00082 ± 0.000034	0.00863
	0	340	60	0.8	NotBiased	0	10,000	0.00071 ± 0.000012	0.01137
Multiple sources of bias	1	200	200	0.8	Biased	10^{-8} to 10^{-9}	10,000	0.00098 ± 0.000040	0.00934
	1	340	60	0.5	Biased	10^{-8} to 10^{-9}	10,000	0.00041 ± 0.000022	0.00804
	1	340	60	0.8	Biased	10^{-7} to 10^{-8}	10,000	0.00072 ± 0.000033	0.01037
	1	340	60	0.8	Biased	10^{-8} to 10^{-9}	10,000	0.00076 ± 0.000033	0.00765
	1	340	60	0.8	NotBiased	10^{-8} to 10^{-9}	10,000	0.00072 ± 0.000012	0.01084
	2	200	200	0.8	Biased	10^{-8} to 10^{-9}	10,000	0.00107 ± 0.000048	0.01244
	2	340	60	0.5	Biased	10^{-8} to 10^{-9}	10,000	0.00048 ± 0.000027	0.00911
	2	340	60	0.8	Biased	10^{-7} to 10^{-8}	10,000	0.00075 ± 0.000032	0.00898
	2	340	60	0.8	Biased	10^{-8} to 10^{-9}	10,000	0.00077 ± 0.000034	0.01008
	2	340	60	0.8	NotBiased	10^{-8} to 10^{-9}	10,000	0.00074 ± 0.000013	0.01122
	3	200	200	0.8	Biased	10^{-8} to 10^{-9}	10,000	0.00106 ± 0.000044	0.01189
	3	340	60	0.5	Biased	10^{-8} to 10^{-9}	10,000	0.00049 ± 0.000025	0.00688
	3	340	60	0.8	Biased	10^{-7} to 10^{-8}	10,000	0.00072 ± 0.000029	0.00823
	3	340	60	0.8	Biased	10^{-8} to 10^{-9}	10,000	0.00080 ± 0.000036	0.00940
	3	340	60	0.8	NotBiased	10^{-8} to 10^{-9}	10,000	0.00076 ± 0.000013	0.01384

The table is divided into four sections: (i) the effects of PCR duplication rates ("PCR Dup.") and a skewed allele frequency spectrum on F_{ST} values ("Mean AF"); (ii) influence of restriction site mutation rates on F_{ST} by changing the effective population size (N_e) and by changing the average per-nucleotide mutation rate ("Restriction Site Mut."); (iii) the effects of skewed allele frequencies (mean p = .5 vs. mean p = .8), shearing bias and asymmetric sampling (n_{sd} = 60, n_{dd} = 340) on F_{ST}; and (iv) how F_{ST} is affected by multiple sources of bias. F_{ST} is significantly affected by asymmetric sampling, which when combined with PCR bias actually elevated F_{ST} values. Skewed allele frequencies also resulted in greater differentiation between simulated sdRAD and ddRAD loci. We present the mean and standard error for F_{ST} as well as the maximum F_{ST} value ("Max. F_{ST}") for each parameter combination.

was less pronounced when combined with other sources of bias (Table 3). Skewed coverage of reference and alternative alleles is expected because of allelic dropout, primarily due to polymorphic restriction sites (Andrews et al., 2016; Gautier et al., 2013). In our sequencing data, when the data sets were analysed separately, we found that the difference in skew in coverage towards the reference allele in ddRAD-seq SNPs and sdRAD-seq SNPs was less pronounced (Table 1). More important than the means is the range and the variance of the skew, as the skew should occur in either direction. Although the ddRAD-seq individuals had different mean proportions of reference reads in heterozygotes than sdRAD-seq when analysed together, the two did not differ when analysed separately. In fact, the sdRAD-seq data set analysed separately had a larger range of values (0.951) compared to the ddRAD-seq (0.903; Table 1), indicating that sdRAD-seq might be experiencing more allelic dropout. In addition, the GBStools analysis indicates that the sdRAD-seq data set has more dropout alleles than the ddRAD-seq data set (Table 1). This result is surprising because the ddRAD-seq data set includes more restriction sites and is expected to be more impacted by polymorphic restriction sites than sdRAD-seq (Andrews et al., 2016; Arnold et al., 2013). Additionally, polymorphic restriction sites are expected to falsely inflate homozygosity (Davey et al., 2013), and our ddRAD-seq data set had higher proportions of heterozygous individuals than the sdRAD-seq data set (Table 1), a pattern which was not due to the larger sample size and supports the finding that our sdRAD-seq data set seems to be more greatly affected by polymorphic restriction sites than our ddRAD-seq data set.

We expected PCR duplicates to be an important source of bias in our RAD-seq studies (Andrews et al., 2016; Davey et al., 2011; Hoffberg et al., 2016), especially as several recent methods have emerged to reduce their impact (Ali et al., 2016; Hoffberg et al., 2016), including the use of paired-end sequencing to identify and remove PCR duplicates from data sets (Hohenlohe et al., 2013; Schweyen et al., 2014; Smith et al., 2014; Tin et al., 2015). Our in silico digest results suggested that PCR duplicates do affect F_{ST} values, but do not have a marked effect when combined with other sources of bias (Table 3). In our empirical data, we found that our sdRAD-seq SNPs had higher variance in coverage than our ddRAD-seq SNPs, which we expected because we used more PCR cycles in the preparation of the sdRAD-seq library. One suggestion for ameliorating the bias due to PCR duplication is to remove loci with incredibly high coverage (i.e., >50×; Schweyen et al., 2014). When we performed this filtering step, the F_{ST} values shifted towards zero, but only significantly so for the data set containing sdRAD-seq and ddRAD-seq reads analysed together. Therefore, filtering to remove high-coverage loci can remove some of the bias between sdRAD-seq and ddRAD-seq data sets but possibly only if the data are pooled. Our results are therefore mixed with regard to PCR duplicates, because the in silico digestion showed that they may not play a major role in causing differences between sdRAD-seq and ddRAD-seq, but we found some evidence of PCR duplicates in our data set.

Andrews et al. (2016) suggested that shearing bias would not greatly impact sdRAD-seq data sets, but our in silico digestion suggests that shearing bias may be an important factor. We found that shearing bias alone did not substantially alter allele frequencies, but it interacted with other sources of bias, including asymmetric sampling schemes and skewed allele frequencies (Table 3). These results highlight the importance of considering all sources of bias when evaluating RAD-seq data sets, as the effects of various sources of bias may interact to alter allele frequencies.

The bias caused by genotyping some individuals with sdRAD-seq and some with ddRAD-seq strongly affected the results of selection components analysis. In the comparison of males and females, the F_{ST} values were significantly larger when the ddRAD-seq and the sdRAD-seq individuals were analysed together, and a greatly inflated number of loci had significant *p*-values after the false discovery rate correction in both the comparison of males and females and the comparison of inferred maternal alleles and females (Figure 5). While the sexual selection analysis could not be conducted in the sdRAD-seq data set alone, the comparison of males and females using only sdRAD-seq data resulted in much larger F_{ST} values than other analyses. In addition, no loci were identified as significant targets of selection in this analysis. These comparisons imply that selection components analysis may produce different results depending on the choice of ddRAD-seq or sdRAD-seq.

Our comparison of ddRAD-seq and sdRAD-seq was based on different individuals that were genotyped by the two methods. This data set was originally collected for a study of selection components analysis in this population of pipefish, and it was due to logistical constraints (particularly the shearing step) that we switched from sdRAD-seq to ddRAD-seq after genotyping only 60 of the 444 collected individuals (see Flanagan & Jones, 2017 for the analysis of the ddRAD-seq individuals). Our comparison of two subsamples of 60 ddRAD-seq individuals indicates that the differences between sdRAD-seq and ddRAD-seq are too large to be explained by sampling error, and consequently must stem from differences in the methods. A more rigorous comparison of the bias emerging from sdRAD-seq and ddRAD-seq would involve genotyping the same individuals using both methods, but we believe that our analysis reveals several noteworthy patterns and points of caution, especially as syntheses and meta-analyses of RAD-seq studies are conducted.

In conclusion, we have shown that simply using different genotyping methods can result in different allele frequencies, some of which are at the scale of differentiation measured between populations. The major sources of this bias are polymorphic restriction sites, small and asymmetric sampling schemes and to some extent PCR duplicates. These results suggest that bias could jeopardize comparisons of different data sets by inflating observed differentiation, potentially obscuring true evolutionary processes. Encouragingly, researchers are identifying ways to minimize bias (Ali et al., 2016; Hoffberg et al., 2016), and it may be possible to incorporate sources of bias into genotyping methods using Bayesian statistics. In the meantime, it is important for researchers to be cognizant of the factors causing different allele frequencies between data sets and to recognize that not all of these differences are the result of evolutionary processes.

ACKNOWLEDGEMENTS

This work was conducted in part while SPF was a Postdoctoral Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award #DBI-1300426, with additional support from The University of Tennessee, Knoxville. This work was supported by the National Science Foundation, grant number DEB-1119261 to AGJ, grant number DEB-1401688 to AGJ and SPF and the National Science Foundation Graduate Research Fellowship under grant number DGE-1252521 to SPF.

CONFLICT OF INTEREST

The authors have no conflict of interests to declare.

DATA ACCESSIBILITY

Sequence data are archived on the NCBI sequence read archive (SRP096542). Data are archived on Dryad at https://doi.org/10.5061/dryad.qf916. The GWSCAR code is available from GITHUB (https://github.com/spflanagan/gwscaR). All other programs and scripts used in the analyses can be downloaded from GITHUB (https://github.com/spflanagan/https://github.com/spflanagan/SCA) or can be obtained by contacting the authors.

AUTHOR CONTRIBUTIONS

S.P.F. and A.G.J. designed the project and wrote the paper. S.P.F. collected and analysed the data.

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*, 389–400. https://doi.org/10.1534/genetics.115.183665

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*, 1661–1667. https://doi.org/10.1111/mec.12686

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

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*, e3376. https://doi.org/10.1371/journal.pone.0003376

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300.

Cariou, M., Duret, L., & Charlat, S. (2016). How and how much does RAD-seq bias genetic diversity estimates? *BMC Evolutionary Biology*, *16*, 240. https://doi.org/10.1186/s12862-016-0791-0

Casbon, J. A., Osborne, R. J., Brenner, S., & Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, *39*, e81. https://doi.org/10.1093/nar/gkr217

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*, 171–182.

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

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Moleular Ecology Resources*, *17*, 362–365. https://doi.org/10.1111/1755-0998.12669

Christiansen, F. B., & Frydenberg, O. (1973). Selection component analysis of natural polymorphisms using population samples including mother-offspring combinations. *Theoretical Population Biology*, *4*, 425–445. https://doi.org/10.1016/0040-5809(73)90019-1

Cooke, T. F., Yee, M. C., Muzzio, M., Sockell, A., Bell, R., Cornejo, O. E., . . . Kenny, E. E. (2016). GBStools: A statistical method for estimating allelic dropout in reduced representation sequencing data. *PLoS Genetics*, *12*, e1005631. https://doi.org/10.1371/journal.pgen.1005631

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

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

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*, 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*, 499–510. https://doi.org/10.1038/nrg3012

Flanagan, S. P., & Jones, A. G. (2015). Identifying signatures of sexual selection using genomewide selection components analysis. *Ecology and Evolution*, *5*, 2722–2744. https://doi.org/10.1002/ece3.1546

Flanagan, S. P., & Jones, A. G. (2017). Genome-wide selection components analysis in a fish with male pregnancy. *Evolution*, *71*, 1096–1105. https://doi.org/10.1111/evo.13173

Flanagan, S. P., Rose, E., & Jones, A. G. (2016). Population genomics reveals multiple drivers of population differentiation in a sex-role-reversed pipefish. *Molecular Ecology*, *25*, 5043–5072. https://doi.org/10.1111/mec.13794

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

Henning, F., Lee, H. J., Franchini, P., & Meyer, A. (2014). Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: Benefits and pitfalls of using RAD markers for dense linkage mapping. *Molecular Ecology*, *23*, 5224–5240. https://doi.org/10.1111/mec.12860

Hoffberg, S. L., Kieran, T. J., Catchen, J. M., Devault, A., Faircloth, B. C., Mauricio, R., & Glenn, T. C. (2016). RADcap: Sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, *16*, 1264–1278. https://doi.org/10.1111/1755-0998.12566

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*, 3002–3013. https://doi.org/10.1111/mec.12239

Jakobsson, M., Edge, M. D., & Rosenberg, N. A. (2013). The relationship between FST and the frequency of the most frequent allele. *Genetics*, *193*, 515–528. https://doi.org/10.1534/genetics.112.144758

Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology*, *17*, 4015–4026. https://doi.org/10.1111/j.1365-294X.2008.03887.x

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

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

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017b). Responsible RAD: Striving for best practices in population genomic studies of adaptation. *Molecular Ecology Resources*, *17*, 366–369. https://doi.org/10.1111/1755-0998.12677

Luu, K., Bazin, E., & Blum, M. G. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*, 67–77.

McKinney, G. J., Larson, W. A., Seeb, L. W., & Seeb, J. E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: Comment on Breaking RAD by Lowry *et al.* (2016). *Molecular Ecology Resources*, *17*, 356–361. https://doi.org/10.1111/1755-0998.12649

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*, 240–248. https://doi.org/10.1101/gr.5681207

Monnahan, P. J., Colicchio, J., & Kelly, J. K. (2015). A genomic selection component analysis characterizes migration-selection balance. *Evolution*, *69*, 1713–1727. https://doi.org/10.1111/evo.12698

Nei, M. (1986). Definition and estimation of fixation indices. *Evolution*, *40*, 643–645. https://doi.org/10.1111/j.1558-5646.1986.tb00516.x

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*, 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*, 5937–5942. https://doi.org/10.1111/mec.12965

R Core Team (2017) *R: A language and environment for statistical computing* (ed. Computing RFfS), Vienna, Austria: R Core Team.

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. New York: Springer. https://doi.org/10.1007/978-0-387-75969-2

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. *The Biological Bulletin*, *227*, 146–160. https://doi.org/10.1086/BBLv227n2p146

Small, C. M., Bassham, S., Catchen, J., Amores, A., Fuiten, A. M., Brown, R. S., ... Cresko, W. A. (2016). The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biology*, *17*, 258. https://doi.org/10.1186/s13059-016-1126-6

Smith, E. N., Jepsen, K., Khosroheidari, M., Rassenti, L. Z., D'Antonio, M., Ghia, E. M., ... Frazer, K. A. (2014). Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome Biology*, *15*, 1–10.

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*, 329–336. https://doi.org/10.1111/1755-0998.12314

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

Waples, R. S. (1987). A multispecies approach to the analysis of gene flow in marine shore fishes. *Evolution*, *41*, 385–400. https://doi.org/10.1111/j.1558-5646.1987.tb05805.x

Workman, P. L., & Niswander, J. D. (1970). Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. *American Journal of Human Genetics*, *22*, 24–49.

Wright, S. (1943). Isolation by distance. *Genetics*, *28*, 114–138.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Flanagan SP, Jones AG. Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. *Mol Ecol Resour*. 2017;00:1–17. https://doi.org/10.1111/1755-0998.12734