Reference: *Biol. Bull.* **227:** 146–160. (October 2014)

Detection and Removal of PCR Duplicates in Population Genomic ddRAD Studies by Addition of a Degenerate Base Region (DBR) in Sequencing Adapters

HANNAH SCHWEYEN, ANDREY ROZENBERG, AND FLORIAN LEESE*

Ruhr University Bochum, Department of Animal Ecology, Evolution and Biodiversity, Universitaetsstrasse 150, D-44801 Bochum, Germany

Abstract. Restriction-site associated DNA sequencing (RAD) has emerged as a powerful marker system for studying genome-wide DNA polymorphisms using next-generation sequencing. A recent technical facilitation of RAD is double-digest RAD (ddRAD), which utilizes two restriction enzymes for library preparation. The more flexible and balanced ddRAD allows analysis of genomic loci in hundreds of individuals. However, in contrast to paired-end sequencing of traditional RAD libraries, PCR duplicates cannot be detected with ddRAD. This is a concern because duplicates can contribute substantially to read coverage data and erroneously inflate the proportion of homozygous loci (allele dropout). Allele dropout can bias population genetic parameter inference and complicate the detection of outlier loci under selection. Here we outline a simple and straightforward approach to detecting PCR duplicates from ddRAD libraries. Our approach introduces a degenerate base region (DBR, 12,288 unique combinations) in the sequencing adapter. We demonstrate the high efficiency and low rate of false positives in simulations. In addition, a pilot study was performed to test this approach on six aquatic invertebrates, sequenced on a HiSeq 2500 sequencer. The reads of the ddRAD libraries consisted of 33.48% PCR duplicates distributed on 19.40% of the loci. A disproportionate number of PCR duplicates were detected in only 4.66% of the loci. While this should not be a concern for general parameter inference, outlier loci detection in particular would be improved by the DBR technique. Given the easy and straightforward application of the technique in other RAD protocols as well, we suggest that DBR regions should generally be included in PCR-based RAD studies.

Received 3 March 2014; accepted 6 August 2014.

* To whom correspondence should be addressed. E-mail: florian.leese@rub.de

Abbreviations: ADO, allele dropout; DBR, degenerate base region; ddRAD, double-digest RAD; RAD, restriction-site associated DNA sequencing.

Introduction

Next-generation sequencing methods have revolutionized molecular ecological and evolutionary research (McCormack *et al.*, 2013). In population genomics, restriction-site associated DNA sequencing (RAD; Miller *et al.*, 2007) has recently become a powerful tool (Narum *et al.*, 2013; Andrews and Luikart, 2014). RAD allows the detection and analysis of hundreds to hundreds of thousands of single nucleotide polymorphisms (SNPs) across a genome. It has recently been applied to a broad spectrum of questions ranging from genetic mapping (*e.g.,* Baird *et al.*, 2008) and tests of selection (*e.g.,* Hohenlohe *et al.*, 2012) to population genomic and phylogeographic studies (*e.g.,* Emerson *et al.*, 2010). The key advantage of RAD is that it is applicable to organisms with genomic resources such as a reference genome (Hohenlohe *et al.*, 2012) but also to organisms previously not used in genomic studies (Emerson *et al.*, 2010; Hess *et al.*, 2013). RAD allows reducing genomic complexity very specifically by selecting a restriction enzyme according to the genome size of the organism and its GC-content (Fig. 1). Recently, several less laborious and more cost-effective modifications of the original RAD protocol were developed, namely 2b-RAD (Wang *et al.*, 2012), ezRAD (Toonen *et al.*, 2013), and double-digest RAD (ddRAD, Peterson *et al.*, 2012). ddRAD is a particularly efficient and popular approach, which uses two restriction

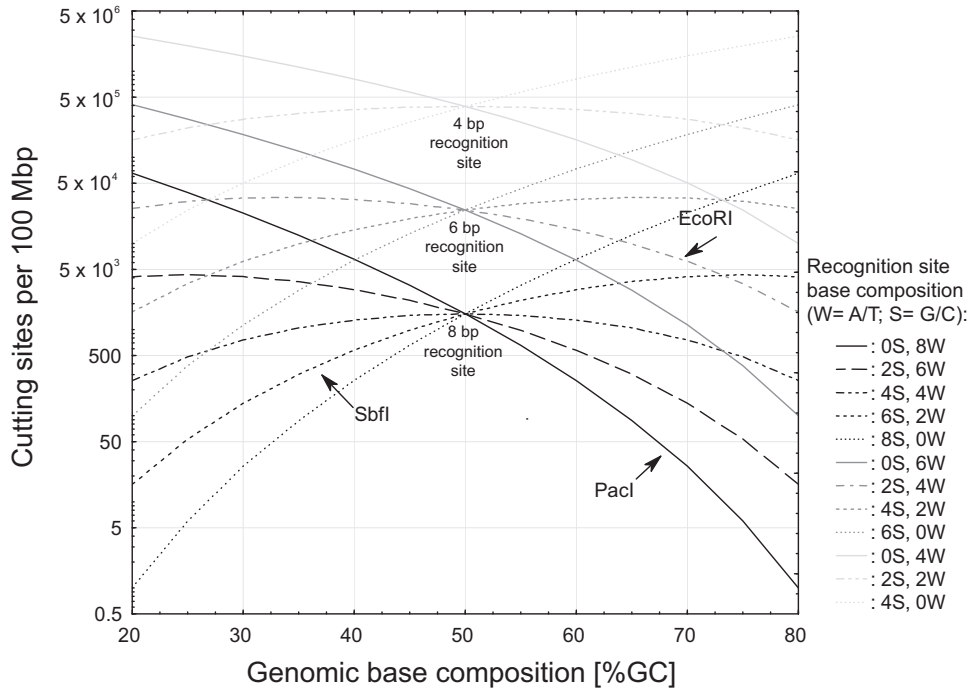

Figure 1. Restriction-site associated DNA (RAD) sequencing reduces the genomic complexity with specific restriction enzymes (REs). Depending on the length of the recognition site of an RE (4, 6, or 8 bp, as indicated in the figure) and the GC/AT content of the recognition site motifs, different amounts of cutting sites are retained after digestion (listed logarithmically per 100 Mbp on the *y*-axis). The number of cutting sites for a certain restriction enzyme strongly depends on the GC content of the genome (*x*-axis). Three examples of restriction enzymes typically used in RAD studies are shown with arrows.

enzymes to reduce genomic complexity and produce a more even distribution of the sequenced DNA fragments across the genome. Furthermore, it uses a two-barcode indexing system (Peterson *et al.*, 2012).

Like any marker system, RAD and its modifications have some conceptual and methodological limitations that can bias the inferred population parameters (Arnold *et al.*, 2013; Davey *et al.*, 2013; Gautier *et al.*, 2013). One particular concern is null alleles, that is, the inability of RAD to correctly assess all allelic variants, primarily because of mutations in the recognition sites (see Arnold *et al.*, 2013). Additionally, other factors, for example, unequal PCR success, can also cause underrepresentation of one of the two alleles (Casbon *et al.*, 2011). All these effects can be summarized under the term "allele dropout" (ADO). ADO inflates homozygosity and thus deflates estimates of genetic diversity. Therefore, inferred population genetic parameters such as heterozygosity, θ and F_{ST}, and the detection of outlier loci under selection can be biased (Arnold *et al.*, 2013; Gautier *et al.*, 2013; McCormack *et al.*, 2013). ADO is especially an issue in studies on populations with high effective population sizes (Gautier *et al.*, 2013). As a possible solution to the ADO bias, both Arnold *et al.* (2013) and Gautier *et al.* (2013) recommend comparison of read coverage data: loci affected by ADO are expected to show a lower coverage. Nevertheless, read coverage data can be strongly distorted by PCR artifacts, which are a common issue in Illumina sequencing data (Kozarewa *et al.*, 2009; Aird *et al.*, 2011; Skelly *et al.*, 2011). PCR bias for a locus can be random or caused by conditions such as differences in GC content, whereby a higher GC content can lead to an increased PCR amplification (Davey *et al.*, 2013). The resulting PCR duplicates can contribute disproportionally to read coverage data. This is a concern in both genomic and transcriptomic studies, in particular if only low amounts of template DNA/RNA are used (Casbon *et al.*, 2011), and PCR-free methods of library preparation have been suggested as a solution (*e.g.,* Kozarewa *et al.*, 2009; Mamanova *et al.*, 2010; Toonen *et al.*, 2013). However, except for a variant of ezRAD, all RAD protocols include a PCR amplification step. Therefore, PCR duplicates can inflate coverage data and possibly mask null alleles. In contrast to more recent PCR-based RAD approaches, paired-end sequencing data generated from traditional sonication-based RAD libraries can identify PCR duplicates. The sonication and thus the random shearing step prior to the PCR generate fragments of a homologous RAD locus that vary in length. Therefore, the starting points of the second sequence reads at a locus should be different (Fig. 2), whereas reads originating from PCR duplicates have the same starting point

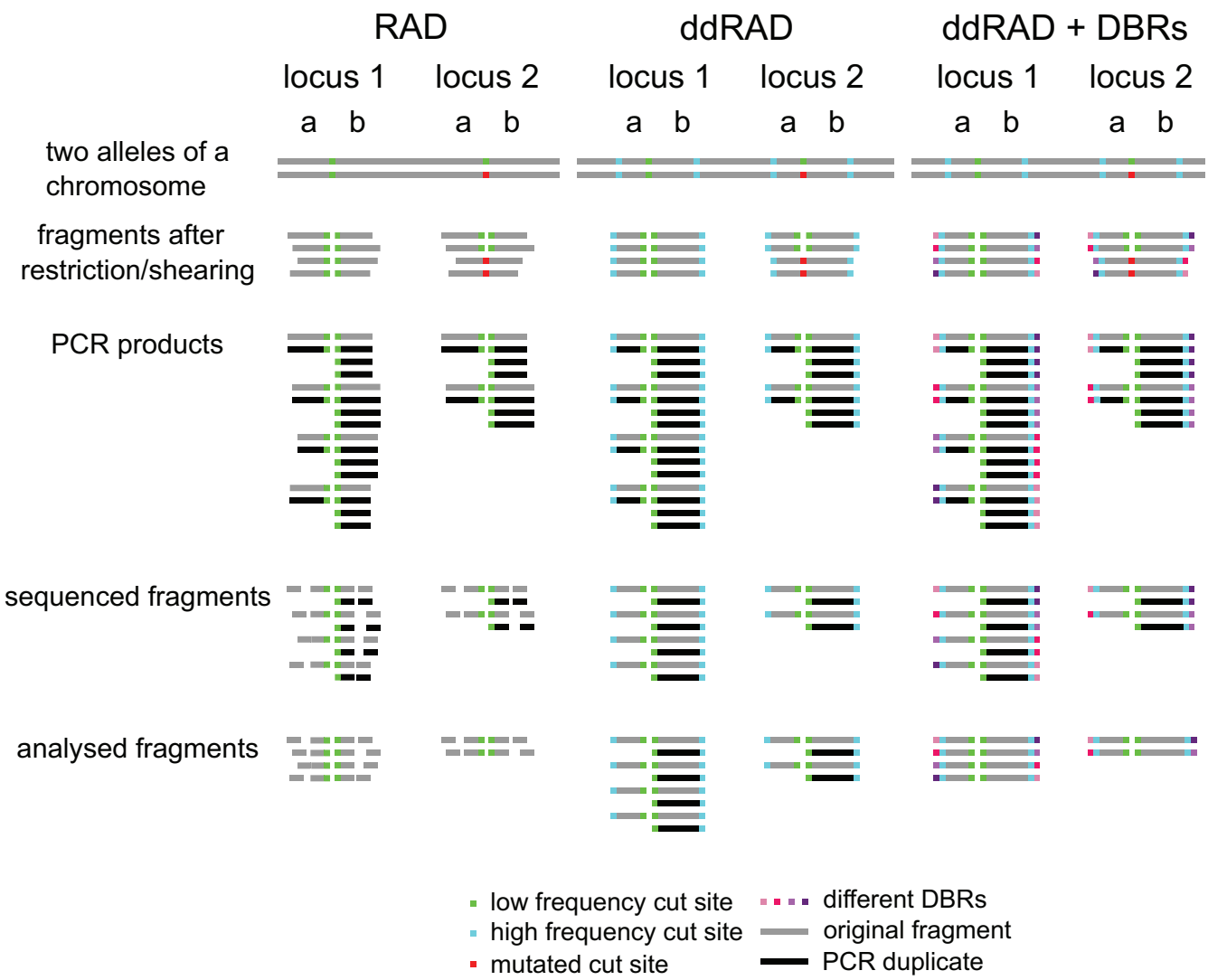

Figure 2. The possibility of detecting PCR duplicates with three RAD approaches. Two genomic regions are compared: locus 1 having the functional restriction site (green = low-frequency cutter, blue = high-frequency cutter) and locus 2 having one of the two alleles of a chromosome with a mutation in the low-frequency cutter recognition site (red). All methods start with the digestion of DNA with either one (RAD) or two (ddRAD and ddRAD+DBRs) restriction enzymes. Due to a mutation in the restriction site in the second allele of locus 2, a restriction is inhibited. In RAD, a shearing step is performed, leading to different fragment lengths. Sequencing adapters are then ligated. In the case of ddRAD+DBRs, sequencing adapters and DBRs are ligated to the fragments. During a PCR, duplicates of the DNA fragments arise (black). Due to PCR biases, some fragments will be amplified more often than others (compare frequency for b and a, respectively). A fraction of the amplified PCR products is subsequently sequenced on an Illumina sequencer. If the PCR is biased, more than one sequence of a unique DNA fragment will be obtained. These duplicates can be identified in paired-end RAD data sets due to their identical starting point for the second read, and coverage values of biased and unbiased PCR are similar after excluding such PCR duplicates (loci 1a and b). However, for ddRAD, these PCR duplicates cannot be identified by their length because the second restriction enzyme makes all fragments of similar length. Thus, all sequenced fragments are counted as individual fragments regardless of whether they are generated by a PCR duplication step or not. Therefore, PCR-biased loci show a higher coverage value than unbiased loci (loci 1a and b). PCR-biased loci with a mutation in the restriction site (locus 2b) can have coverage values similar to those of PCR unbiased fragments without a mutation in the restriction site (locus 1a), if PCR bias is substantial. Using the ddRAD+DBR method, PCR duplicates can be reliably identified. Many different DBRs are ligated to the DNA fragments prior to PCR, leading to the association of different DBRs to different DNA fragments, especially for low-coverage data. Therefore, PCR duplicates can be identified as having identical DBRs. After their exclusion, loci affected by alleles with a mutation in the restriction site can be distinguished by differences in their read coverage (loci 1 and 2) comparable to paired-end RAD.

(Davey *et al.*, 2013; Hohenlohe *et al.*, 2013). Thus, with a reliable filtering step after sequencing, most or all PCR duplicates can be excluded to obtain an unbiased sequencing coverage. In this filtered data set, sequences with too low a coverage can afterward be excluded, assuming ADO as the cause. Parameter inferences based on this data set should lead to adequate results (Fig. 2) (Arnold *et al.*, 2013; Gautier *et al.*, 2013). With this method, Hohenlohe *et al.* (2013) found PCR duplicates at a relative frequency of 22% in fish-species libraries.

Unfortunately, this approach to detect PCR duplicates cannot be applied for the more advanced RAD protocols like 2b-RAD, ddRAD, and ezRAD (if a PCR step is included) because starting points of the second sequencing run do not vary (see Fig. 2 for ddRAD) (Narum *et al.*, 2013). This leads to a greater uncertainty if coverage values are used to exclude ADO since PCR duplicates that erroneously inflate coverage cannot be discriminated.

To distinguish PCR duplicates in high-throughput sequencing data, Casbon *et al.* (2011) have suggested the use of degenerate base regions (DBRs) to tag individual template molecules prior to amplification. We utilized this idea to develop a straightforward approach to identify and correct loci with a strong PCR bias in ddRAD. Our key modification of the original ddRAD protocol is that a short DBR is included in the P7 adapter and serves as an individual label for different DNA molecules belonging to the same locus. In the current study we provide a statistical justification of the success of this approach and quantify the effects of PCR duplicates in a pilot experiment using ddRAD libraries obtained from three aquatic invertebrate species.

Materials and Methods

Simulating DBR success

In a first step, we evaluated the general potential of a variable region (DBR) in the P7 adapter for the identification of PCR biases. We simulated the proportion of recovered DBRs for various coverage values with custom R scripts (R Core Team, 2013): DBR_simulation.R available at the Evoeco radtools Github repository (2014). With the script *C* (coverage), different DBRs were randomly drawn out of a pool of *n* different DBRs from a uniform distribution (assuming 12,288 different DBRs according to the herein described fragment, Appendix Fig. A1). Afterward, the number of different recovered DBRs was counted. We then performed 20,000 permutations (*perm*) to obtain probability distributions. In addition, we calculated the cumulative probabilities for the different *C* values. To infer threshold values for unbiased PCR reactions, we determined the probability of recovering identical DBRs using a 5% and 10% type-I error threshold limit (false positively detected PCR duplicates). We performed calculations for two typical population genomic cases: (1) individual-based ddRAD studies with a low average target coverage value of 20×, and (2) population pool-based ddRAD studies with a coverage value of 200×.

Adapter design

In a second step, we tested the conceptual approach in the laboratory. As a starting point for our modified P5 and P7 adapters, we took the sequences used by Peterson *et al.* (2012) (Fig. 3). Both adapters were complemented with Ins, that is, inserts of 0 to 3 bases that increase the sequence variability to preserve the lasers of the sequencer from blinding by the identical bases in the restriction sites (Appendix Fig. A1). In our design, the P7 adapter also directly included the second barcode in contrast to the protocol of Peterson *et al.* (2012), in which it was introduced by the PCR using specific primers. Additionally, the DBR sequence was added to the P7 adapter between the lns and the overhang. Our DBR construct was designed in such a way that the bottom strand (in 3′→5′ direction) consists of two fixed G-positions and 8 degenerate bases (two M, one H, and five N). On the top strand five Ns are located next to three 2′-deoxyinosine (1), which can pair with any of the four DNA bases, preferring, in decreasing order, C, A, and T (Watkins and SantaLucia, 2005), followed by two C-residues: 5′ NNNNNIIICC 3′. The design of this fragment is expected to minimize mispairings of the top and the bottom strands during hybridization. Hybridization of both strands was performed according to Peterson *et al.* (2012), with the exception of the temperature profile (97 °C for 5 min followed by a decrease in temperature of 2 °C per min in a 1.5-ml Eppendorf tube in a thermal block).

Library preparation

DNA was extracted from six individuals of three freshwater invertebrate species using a modified salt-extraction protocol (Sunnucks and Hales, 1996): one specimen of *Ancylus fluviatilis* O. F. Müller, 1774 (A2), two of *Gammarus fossarum* Koch, 1835 (G1 and G2), and three of *Sericosotoma personatum* (Spence in Kirby & Spence, 1826) (Appendix Table A1). The extracted DNA was incubated with 1 μl of RNase A (10 mg/ml; Thermo Scientific) for 30 min at 37 °C. Next, the samples were purified using the MinElute Reaction cleanup kit (Qiagen) and eluted in 30 μl of molecular grade H_2O. DNA concentrations were measured using the Qubit dsDNA BR assay kit and the Qubit fluorometer (Life Technologies) with 1 μl of DNA. Between 350 and 600 ng of DNA were used in the double digestion of the samples. One point five microliters of NsiI and 1.5 μl of Csp6I restriction enzymes (both FastDigest, Thermo Scientific) were added to the 3 μl FastDigest buffer and appropriate amount of DNA. The volume was made up with water to 30 μl. The reaction was conducted at 37 °C

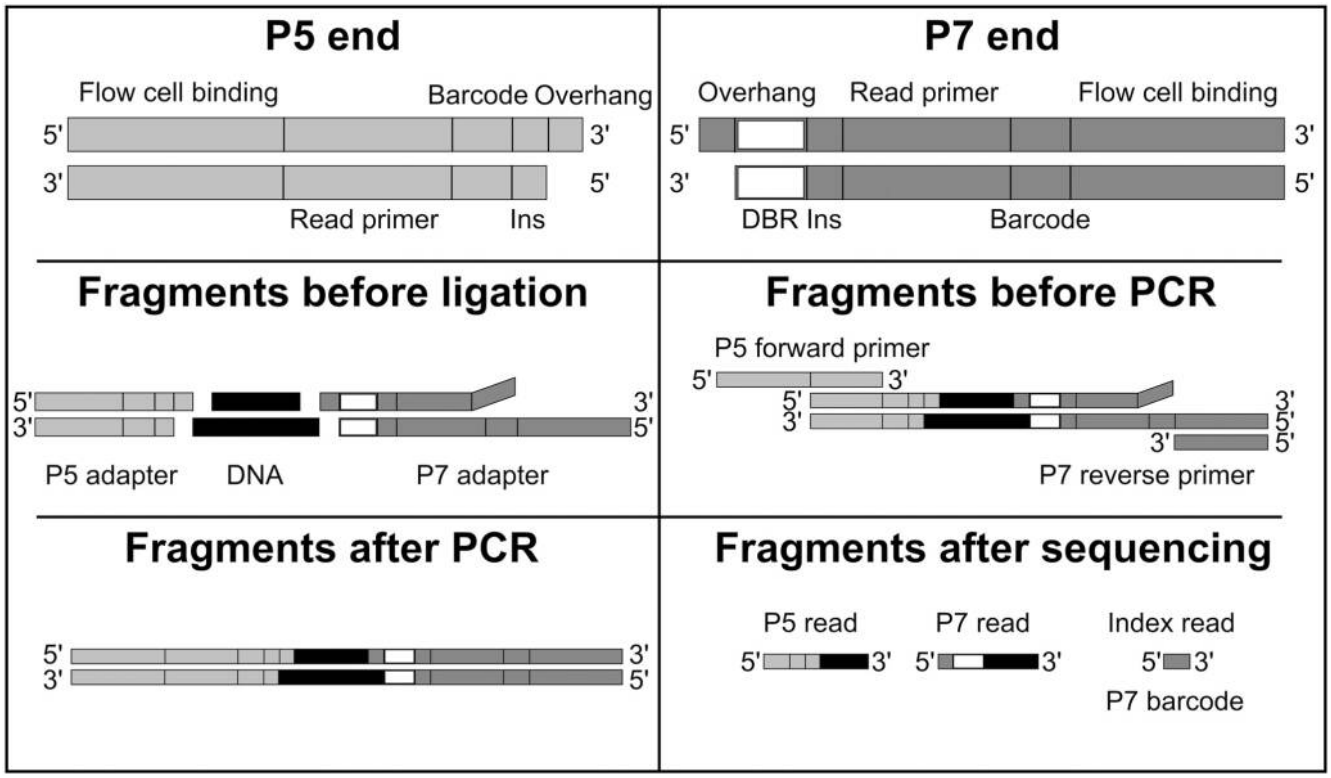

Figure 3. Construction principle of the two (P5 and P7) sequencing adapters (top row) and schematic overview of the laboratory steps. Both adapters have a Barcode and an insert (Ins) of 0–3 bp that help generating a high diversity of nucleotides. The P7 adapter also includes a degenerate base region (DBR) that allows the detection of PCR duplicates.

for 20 min. Next, the samples were purified with the MinElute Reaction cleanup kit and eluted in 25 μl of H_2O. Subsequently, the adapters were ligated to the DNA fragments. For the ligation step the amount of adapters used was adjusted to the amount of expected cut-sites and the DNA concentration using the Mol Calculator of Peterson *et al.* (2012) with adapter excess set to 5× for P5 and 10× for P7. Finally, 3 μl of T4 ligation buffer and 0.5 μl of T4 Ligase (2,000,000 units/μl; both New England Biolabs) were added and made up to 40.5 μl with H_2O.

For a preliminary size selection, a gel extraction was conducted with a 1.5% TAE agarose gel at a size range of about 250 to 600 bp. The DNA was purified using the MinElute gel extraction kit (Qiagen) and eluted in 25 μl of H_2O. Next, the fragments were amplified in a PCR using 10 μl of Q5 buffer, 0.5 μl of Q5 polymerase (both New England Biolabs), 5 μl of dNTP (2 mmol l^{-1}), and 5 μl of each 10 μmol l^{-1} primer. Twelve to 25 μl of DNA was added and made up to a total reaction volume of 50.5 μl with water. An initial denaturation step of 30 s at 98 °C was followed by either 21 or 22 amplification cycles of 10 s at 98 °C, 30 s at 65 °C, and 30 s at 72 °C. The final elongation step was 5 min at 72 °C. The samples were purified with the MinElute PCR purification kit (Qiagen) and eluted in 13 μl of H_2O. The DNA concentration was measured using 1 μl of DNA with the Qubit dsDNA BR assay kit (Life Technologies). Equimolar pools of DNA were size-selected with the LabChip XTe and the DNA 750 assay kit, using a size range of 308 to 462 bp. All samples were pooled together and the DNA was concentrated using the MinElute Reaction cleanup kit to about 12 μl of 20 ng/μl of DNA. The library was sequenced on one lane of an HiSeq 2500 sequencer (Illumina) by GATC (Konstanz, Germany) to produce paired-end sequences 101 bases long.

The procedure was repeated for four specimens (two of *Ancylus fluviatilis* (A1 & A2), one of *Gammarus fossarum* (G2), and one of *Sericostoma personatum* (S2)) with a P7 adapter similar to the ones described above but without the DBR region, to test for the general performance of the adapters with and without DBRs. All work was identical to the workflow outlined above except that only 12 PCR cycles were run.

Data analysis

The raw reads were first subject to filtering with trimmomatic (Bolger *et al.*, 2014): sequences containing fragments similar to adapters and primers were discarded (ILLUMI

NACLIP:illumina.fa:2:30:10) and mild quality-trimming (SLIDINGWINDOW:7:10) was applied afterward. The minimal length of the reads was set to 95 bases.

As the currently available software products for RAD-assembly have no options to accommodate non-standard adapter designs, we implemented our own preprocessing script, which uses the paired reads retained after filtering as input (preprocess_ddradtags.pl, available at the Evoco radtools Github repository, 2014). The script (i) searches for the presence of the barcode in the P5-read, (ii) detects the DBR present in the P7-read, (iii) discards putative duplicates, (iv) removes the Ins, and (v) trims the resulting sequences to make them of equal length. These steps were performed as follows for the six libraries. The barcodes were provided as a predefined list containing the respective sequences and the expected length of Ins in the P7 reads. We tolerated one mismatch in the respective region. In the script, the DBR pattern was specified with the standard International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes (NNNNNHMMGG in our case) and matching was performed allowing a one-base shift or a one-base mismatch. Since the assembly of the reads was not the purpose of the script, we used a straightforward strategy to define groups of similar reads, referred to as "loci": for every P5 read, the first 30 bases after the cut-site were divided into three blocks (10 bases each), and individual "loci" incorporated reads that had at least two identical blocks. The reads assigned to the same "locus" and having similar DBRs (a maximum of one mismatch allowed) were considered PCR duplicates. To speed up the processing of the reads and reduce memory footprint, this comparison was performed not exhaustively but sequentially. The Ins and the DBRs were excised from the final sequences, while the barcodes were left intact to facilitate assignment of reads to individuals later on (*e.g.,* with the aid of the process_radtags.pl script from the Stacks package, Catchen *et al*., (2013)). The script produced a detailed report, a list of all DBRs found, and a file with all encountered barcode-"locus"-DBR combinations together with respective counts. This last output was used in the current paper to produce summary statistics for different per-"locus" coverage ranges. These ranges were not corrected for the calculated false positive estimates originating from the simulations due to a non-uniform distribution of DBR recovery. Instead, we used four coverage ranges to infer differences in PCR duplicates associated with the coverage: 11–20, 21–30, 31–40, and 41–50 reads per "locus."

The information content of an individual DBR given the observed frequency distribution was measured in terms of Shannon entropy according to the classical formula (Shannon, 1948).

The ddRAD data have been deposited in the NCBI SRA database (BioProjectPRJNA264244).

Results

Simulation of degenerate base region efficiency

We tested the efficiency of DBRs in distinguishing biased from unbiased PCR reactions in individual and population-pool ddRAD studies with our 12,288 unique DBRs using our custom R script. In the simulated data the number of recovered DBRs increases with coverage. However, with increasing coverage the probability increases that a specific DBR will be obtained multiple times by chance and not due to biased PCR (false positive), leading to saturation. For a target coverage of 20× we expect to recover 20 unique DBRs with a high probability ($P = 0.9843$, Appendix Fig. A2A). For a target coverage of 200×, the probability function is bell-shaped with the modal value being 199 unique DBRs ($P = 0.3257$, Appendix Fig. A2B).

Our simulation results show that unique DBRs can be expected for coverage values up to 36× when applying a 5% error threshold (50× coverage for a 10% threshold; Appendix Table A2).

Data analysis of double-digest restriction-site associated DNA screening

We obtained 17,868,677 reads from the four libraries produced without DBRs after quality filtering (Table 1). Reads were assigned to 2,592,053 "loci" with an average coverage of 6.89. The six samples with DBRs yielded 41,331,110 reads assigned to 3,721,011 "loci" with an average coverage of 11.11.

Overall, the sequencing success of the different samples was unevenly distributed. While the sample with the highest number of reads contributed 19,140,118 reads assigned to 1,303,473 "loci" (G1 with DBRs), the sample with the lowest number of reads contributed only 246,841 reads to 140,766 "loci" (G2 with DBRs).

To account for PCR biases, all reads belonging to the same "locus" (group of similar reads, likely corresponding to the same locus or at least allele) and having similar DBRs were interpreted as PCR duplicates. They contributed to 33.48% of the reads. After exclusion of PCR duplicates, the average coverage dropped to 7.39 for the data set with DBRs.

To investigate the impacts of PCR duplicates for different coverage ranges, summary statistics were calculated for different per-"locus" coverage intervals (1–10, 11–20, 21–30, 31–40, 41–50, >50) (Appendix Table A3). Most PCR duplicates were found in the coverage range of above 50 (52.51% of all reads in this coverage range were PCR duplicates). In the coverage ranges between 11× and 50×, which are typical for ddRAD studies using individually barcoded specimens, much fewer PCR duplicates were found (*e.g.,* 6.29% PCR duplicates for 11–20× coverage). For the total coverage range, those samples with the highest

Table 1

Overview of the sequencing success from 10 invertebrate ddRAD libraries with and without degenerate base regions (DBRs) for the whole coverage range

DBR	Sample	Reads	Unique reads	"Loci"	PCR duplicates	% PCR duplicates	Coverage without PCR duplicates	Coverage with PCR duplicates
YES	A2	3813519	3131053	432212	682466	18	7.24	8.82
	G1	26293578	14755762	1574450	11537816	44	9.37	16.70
	G2	277549	245207	156394	32342	12	1.57	1.77
	S1	5638120	3884820	731231	1753300	31	5.31	7.71
	S2	14538968	10093641	1119741	4445327	31	9.01	12.98
	S3	334283	293713	153254	40570	12	1.92	2.18
NO	A1	2871551	*	475330	*	*	*	6.04
	A2	4783036	*	615891	*	*	*	7.77
	G2	5819036	*	1053336	*	*	*	5.52
	S2	9385069	*	990601	*	*	*	9.47

* No PCR duplicates could be identified due to the use of adapters without DBRs.

number of reads also had the highest proportion of PCR duplicates; this was not found in distinct coverage ranges.

In a next step, we tested how equally the PCR duplicates were distributed at the different "loci." Overall, 19.40% of the "loci" had PCR duplicates (Table 2). In 4.66% of the cases, PCR duplicates contributed to more than 30% of the total coverage of the "locus." For the four coverage intervals between 11× and 50×, the number of duplicates was much higher (Table 2). Here, higher coverage values showed increased numbers of "loci" with PCR duplicates (58.30% for the range 11–20× to 99.21% for 41–50×). In contrast, strong PCR bias (>30% PCR duplicates per "locus") was less prominent in comparison to the total coverage range, with the highest amounts for the range of 41–50× (0.68%). It is noticeable that the strength of this bias differs among samples (Appendix Table A4). While for library A2 only 0.04% of the "loci" had more than 30% PCR duplicates in the 11–20× coverage range, in library G2 1.76% of the "loci" were affected.

To compare the frequency distribution of sequenced DBRs with unbiased *in silico* data, all recovered DBR sequences were sorted by their frequency (after exclusion of the PCR duplicates; Fig. 4). Our results showed that the sequenced DBRs were exponentially distributed, indicating that many of them were encountered only a few times, while a small number were overrepresented. Due to this bias the actual entropy (as a measure of available variation) was slightly decreased to 12.95 bits (equivalent to 7916 equally frequent tags) in comparison to 13.58 bits for the ideal case of the even distribution of all 12,288 DBRs.

Discussion

The aim of population genomic studies is to infer estimates on genetic diversity across populations and individuals as well as to detect genomic regions under selection. The recently developed double-digest restriction-site associated DNA sequencing (ddRAD) technique (Peterson *et al.,* 2012) is increasingly being applied in this context. However, technical artifacts such as PCR biases cannot be

Table 2

Overview of ddRAD "loci" unaffected by PCR duplicates and those affected by different proportions of PCR duplicates (in 10% steps from 0% to 100%)

Coverage	"Loci" without duplicates	"Loci" with duplicates (frequencies of PCR duplicates at "loci" given in 10% bins)									
		>0–10%	>10–20%	>20–30%	>30–40%	>40–50%	>50–60%	>60–70%	>70–80%	>80–90%	>90–100%
11–20	108,066	96,091	44,577	7,615	905	166	18	6	3	7	5
21–30	17,921	65,804	28,910	3,499	357	31	4	1	5	1	4
31–40	2,603	33,922	20,882	2,836	247	22	4	1	1	0	3
41–50	266	15,626	15,209	2,395	207	14	4	1	1	1	1
Total	2,978,438	236,006	230,719	78,082	53,448	107,222	1,730	7,950	1,329	313	145

The data are shown for the complete range of coverage values as well as for typical coverage values between 11 and 50. Most "loci" recovered have coverage values <11×.

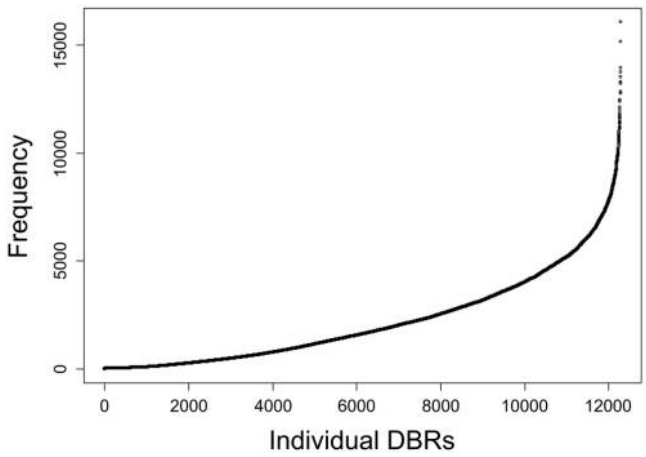

Figure 4. Frequency of the 12,288 unique degenerate base regions (DBRs) in the ddRAD libraries.

distinguished with this approach. This is a concern as PCR biases may lead to incorrect parameter estimates due to allele dropout (ADO). Comparing contemporary population genomic methods based on the general RAD methodology (Miller *et al.*, 2007; Baird *et al.*, 2008), only paired-end sequencing of RAD libraries has the possibility of identifying PCR duplicates and thus discriminating loci affected by PCR bias (Narum *et al.*, 2013). With paired-end RAD, however, the variation in fragment length due to nonrandom shearing of the fragments by sonication can still introduce some biases, as shown by Davey *et al.* (2013). The authors found that short fragment lengths have in general a lower read coverage, thereby erroneously leading to assumed null alleles at those loci. Conversely, long fragments have in general higher values of read coverage. Therefore, longer fragments affected by ADO could be misinterpreted as alleles without a mutation in the restriction site. This problem can be avoided with ddRAD, for which two restriction enzymes are utilized (Davey *et al.*, 2013). Here, the average fragment length depends on the cleavage frequency of the frequent cutter. However, this design does not allow distinguishing among PCR duplicates.

The modification of the ddRAD protocol outlined in this study solves this problem by the addition of a highly degenerate base region (DBR, Casbon *et al.*, 2011) in the P7 adapter.

Probability of false positives

Using a simulation approach, we have assessed the frequency of false positives, that is, reads with the same DBR that are actually not PCR duplicates but bear the same DBR by chance. This is of critical importance for identifying coverage thresholds above which the technique is limited. The simulation results showed that within typical coverage intervals used for the analysis of individually barcoded samples, the number of false positives is negligible (36× coverage for a 5% type-I error level). For the higher coverage values required for analyzing pooled samples (about 200×), on average only one such coincidence is expected.

These *in silico* data demonstrate the great potential of the outlined DBR concept in identifying PCR duplicates in comparison to paired-end sequencing of traditional RAD libraries, which is the only RAD method capable of dealing with this issue (Hohenlohe *et al.,* 2013). Paired-end RAD distinguishes unique fragments on the basis of differences in their length. Nevertheless, the available variability of the fragment lengths is naturally very restricted due to (i) the limited distance between two restriction sites, and (ii) the usually very stringent size-selection step during library preparation. For example, Hohenlohe *et al.* (2013) used a size-selection step with a length variation of 70 bp, meaning 70 potential starting positions for the second read. For 20× coverage values, this would still lead to the exclusion of at least one false positive per locus with 95% probability. Using the outlined DBR approach, the probability for a false positive under the same scenario is only 1.6%. Therefore, our outlined DBR approach allows for a much greater resolution than using the length variability information of the second read.

Experimental application

To test the applicability of our approach and quantify PCR duplicates, we constructed adapters with a short DBR included (Fig. 3) and developed six ddRAD libraries for aquatic invertebrates from three distinct evolutionary groups as a test case. Analysis of the resulting sequences showed that all 12,288 possible DBR sequences were recovered. Their occurrence, however, was not evenly distributed. This deviation can be explained by unequal hybridization or ligation success of the adapters for different DBR sequences, which can be minimized by adding more nondegenerate positions between the overhang and the DBR region of the P7 adapter. In addition, there was a 1-bp mismatch between our adapter and the primer used by the sequencing company, which could have also affected the results. The loss of DBR variability due to these biases, however, was less than 1 bit, which still allowed us to detect PCR duplicates with very high confidence. Because of the discrepancies between the simulated and found variability, we did not use the simulated cut-off values for false positive ratios on the experimental data but rather analyzed different coverage ranges.

Evidence for PCR duplicates

A key finding of our study is the large number of PCR duplicates in ddRAD libraries. Of all reads, 33.48% were identified as PCR duplicates, and 19.40% of the "loci" contained at least one duplicate. This shows that PCR biases

play a major role in ddRAD sequencing, as was found for paired-end RAD by Hohenlohe *et al.* (2013). To understand how these numbers of PCR duplicates could inflate population genomic studies, we analyzed to what degree they contributed to the coverage of the different "loci." For most of the "loci," only a small proportion of the coverage originated from the PCR duplicates. Therefore, most "loci" affected by ADO can be detected by coverage thresholds, as suggested by Gautier *et al.* (2013), regardless of the use of DBRs. Nevertheless, few "loci" had very high proportions of PCR duplicates. For example, in library G2, 1.33% of the loci in a coverage range usual for individual-based ddRAD studies (11–50×) had more than 30% of the reads being PCR duplicates. These are the cases in which PCR duplicates can inflate homozygosity. Because of their comparatively low frequency, they will probably not inflate most population genomic parameters. However, in the case of genome-wide outlier detection analyses, undetected PCR duplicates in combination with ADO could lead to false positive results for purifying selection as well as false negative results for balancing selection. Thus, it should be best practice to discard reads originating from PCR duplicates from the data sets as well as loci that show too-low coverage values after the removal of the duplicates.

Transferability of the degenerate-base-region approach to other protocols for restriction-site associated DNA sequencing

The DBR based approach described here can in general be applied to other RAD protocols beside ddRAD. First, a short DBR could be included in the adapter used for single-end RAD sequencing such as, for example, in Baird *et al.* (2008). The same holds true for the 2b-RAD protocol, which is based on type IIB restriction enzymes (Wang *et al.*, 2012), cutting the DNA both up- and downstream of the recognition site. Here, the variable fragments could be added to the selective adapter. In contrast, we do not see how the approach could be included in the ezRAD procedure (Toonen *et al.*, 2013) without losing the central advantage of using the Illumina TruSeq kits. For several applications this should not be a concern because ezRAD can be used with the PCR-free library preparation kit (Toonen *et al.*, 2013). The convenient option to circumvent PCR duplicates by not performing a PCR in ezRAD will not, however, work for specimens that are small and thus low in DNA. Furthermore, as the genomic complexity is less reduced in ezRAD, this technique is probably not applicable to directly generate population genomic data for many specimens; instead single nucleotide polymorphisms are identified for their further use with other scoring approaches (*e.g.,* microarrays). While ezRAD has these highlighted advantages, it cannot be seen as alternative to the use of DBRs in population genomic screens on many individuals.

Conclusion

The incorporation of highly variable, degenerate base regions into the P7 ddRAD sequencing adapter substantially improves this technique because it reliably detects PCR duplicates without introducing new laboratory steps or substantially increasing initial costs.

Acknowledgments

We thank Martina Weiss, Johannes Köster, Shobhit Agrawal, and Jennifer Jackson, as well as two anonymous reviewers for helpful comments on this manuscript. HS and FL are supported by the *GeneStream* Junior Research Group funded by the Kurt Eberhard Bode Foundation within the Deutsches Stiftungszentrum.

Literature Cited

Aird, D., M. Ross, W. Chen, M. Danielsson, T. Fennell, C. Russ, D. Jaffe, C. Nusbaum, and A. Gnirke. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12:** R18.

Andrews, K., and G. Luikart. 2014. Recent novel approaches for population genomics data analysis. *Mol. Ecol.* **23:** 1661–1667.

Arnold, B., R. Corbett-Detig, D. Hartl, and K. Bomblies. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22:** 3179–3190.

Baird, N., P. Etter, T. Atwood, M. Currey, A. Shiver, Z. Lewis, E. Selker, W. Cresko, and E. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3:** e3376.

Bolger, A., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* btu170.

Casbon, J., R. Osborne, S. Brenner, and C. Lichtenstein. 2011. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39:** e81.

Catchen, J., P. Hohenlohe, S. Bassham, A. Amores, and W. Cresko. 2013. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22:** 3124–3140.

Davey, J., T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. Blaxter. 2013. Special features of RAD sequencing data: implications for genotyping. *Mol. Ecol.* **22:** 3151–3164.

Emerson, K., C. Merz, J. Catchen, P. Hohenlohe, W. Cresko, W. Bradshaw, and C. Holzapfel. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci. USA* **107:** 16196–16200.

Evoeco radtools Github repository. 2014. [Online]. Available https://github.com/evoeco/radtools [2014, 17 Oct].

Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J. Cornuet, and A. Estoup. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* **22:** 3165–3178.

Hess, J., N. Campbell, D. Close, M. Docker, and S. Narum. 2013. Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Mol. Ecol.* **22:** 2898–2916.

Hohenlohe, P. A., S. Bassham, M. Currey, and W. A. Cresko. 2012. Extensive linkage disequilibrium and parallel adaptive divergence across three spine stickleback genomes. *Philos. Trans. R. Soc. B* **367:** 395–408.

Hohenlohe, P. A., M. D. Day, S. J. Amish, M. R. Miller, N. Kamps-Hughes, M. C. Boyer, C. C. Muhlfeld, F. W. Allendorf, E. A.

Johnson, and G. Luikart. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol. Ecol.* **22:** 3002–3013.

Kozarewa, I., Z. Ning, M. Quail, M. Sanders, M. Berriman, and D. Turner. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6:** 291–295.

Mamanova, L., R. Andrews, K. James, E. Sheridan, P. Ellis, C. Langford, T. Ost, J. Collins, and D. Turner. 2010. FRT-seq: Amplification-free, strand-specific, transcriptome sequencing. *Nat. Methods* **7:** 130–132.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* **66:** 526–538.

Miller, M., J. Dunham, A. Amores, W. Cresko, and E. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17:** 240–248.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller, and P. A. Hohenlohe. 2013. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* **22:** 2841–2847.

Peterson, B., J. Weber, E. Kay, H. Fisher, and H. Hoekstra. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7:** e37135.

R Core Team. 2013. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Shannon, C. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* **27:** 379–423, 623–656.

Skelly, D., M. Johansson, J. Madeoy, J. Wakefield, and J. Akey. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21:** 1728–1737.

Sunnucks, P., and D. Hales. 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). *Mol. Biol. Evol.* **13:** 510–524.

Toonen, R., J. Puritz, Z. Forsman, J. Whitney, I. Fernandez-Silva, K. Andrews, and C. Bird. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* **1:** e203.

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

Watkins, N., and J. SantaLucia. 2005. Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.* **33:** 6258–6267.

Appendix

Table A1

Specimens used for library preparation

Sample	Species	Coordinates*	DBR†
A1	*Ancylus fluviatilis*	51.2277°N 8.4075°E	−
A2	*Ancylus fluviatilis*	51.2277°N 8.4075°E	+/−
G1	*Gammarus fossarum*	51.2710°N 8.4459°E	+
G2	*Gammarus fossarum*	51.4085°N 7.6527°E	+/−
S1	*Sericosotoma personatum*	51.4436°N 8.2485°E	+
S2	*Sericosotoma personatum*	51.4565°N 8.4363°E	+/−
S3	*Sericosotoma personatum*	51.2036°N 7.7207°E	+

* All samples originate from the Ruhr basin (Germany, North Rhine-Westphalia) and exact coordinates are given.

† DBR indicates if the library was generated with adapters including (+) / not including (−) a degenerate base region.

Table A2

Threshold values for unbiased PCR reactions

Coverage [×-fold]	Number of different DBRs recovered (5% error threshold)	(10% error threshold)
36	36	36
37	36	37
50	49	50
51	50	50
100	98	99
150	147	148
200	196	197
300	293	294

The probability of recovering identical DBRs (degenerate base regions) using a Type-I error limit of 5% and 10% was simulated for different coverage values typical for different population genomic studies. The given number of recovered DBRs (assuming a 5% and 10% error-I threshold, respectively) represents the lowest number of DBRs at which a PCR would still be considered as unbiased. A uniform distribution of 12,288 different DBRs was used for the simulation.

Table A3

Overview of the sequencing success for 10 invertebrate ddRAD libraries (sample column) with or without degenerate base regions (DBR); PCR duplicate statistics are shown for libraries with DBR; statistics are shown for different coverage ranges

DBR	Sample	Reads	Unique reads	"Loci"	PCR duplicates	% PCR duplicates	Coverage without PCR duplicates	Coverage with PCR duplicates
				1–10-fold coverage				
Yes	A2	2116042	1989503	1124428	126539	5.98	1.77	1.88
	G1	210549	202759	142448	7790	3.70	1.42	1.48
	G2	1878852	1792054	811449	86798	4.62	2.21	2.32
	S1	794276	772869	317880	21407	2.70	2.43	2.50
	S2	186540	175587	139450	10953	5.87	1.26	1.34
	S3	1332147	1274504	587392	57643	4.33	2.17	2.27
No	A1	1073858	*	351900	*	*	*	3.05
	A2	1183952	*	417465	*	*	*	2.84
	G2	3205635	*	1091911	*	*	*	2.94
	S2	1773592	*	586464	*	*	*	3.02
				11–20-fold coverage				
Yes	A2	727372	668994	49090	58378	8.03	13.63	14.82
	G1	10048	9462	701	586	5.83	13.50	14.33
	G2	1800237	1681158	120718	119079	6.61	13.93	14.91
	S1	619006	589205	41852	29801	4.81	14.08	14.79
	S2	9999	9138	698	861	8.61	13.09	14.33
	S3	713587	678216	48947	35371	4.96	13.86	14.58
No	A1	535580	*	37734	*	*	*	14.19
	A2	1571783	*	107657	*	*	*	14.60
	G2	907099	*	66417	*	*	*	13.66
	S2	1752836	*	120194	*	*	*	14.58
				21–30-fold coverage				
Yes	A2	669474	598685	26629	70789	10.57	22.48	25.14
	G1	7133	6567	285	566	7.93	23.04	25.03
	G2	1471102	1351050	59027	120052	8.16	22.89	24.92
	S1	456935	427610	18371	29325	6.42	23.28	24.87
	S2	5063	4568	202	495	9.78	22.61	25.06
	S3	410908	385976	16626	24932	6.07	23.22	24.71
No	A1	196281	*	8048	*	*	*	24.39
	A2	550971	*	22817	*	*	*	24.15
	G2	232933	*	9514	*	*	*	24.48
	S2	961523	*	39200	*	*	*	24.53
				31–40-fold coverage				
Yes	A2	640249	559221	18168	81028	12.66	30.78	35.24
	G1	5794	5299	166	495	8.54	31.92	34.90
	G2	1071133	968658	30618	102475	9.57	31.64	34.98
	S1	295663	272163	8478	23500	7.95	32.10	34.87
	S2	4025	3616	114	409	10.16	31.72	35.31
	S3	223434	207449	6435	15985	7.15	32.24	34.72
No	A1	86150	*	2488	*	*	*	34.63
	A2	198037	*	5733	*	*	*	34.54
	G2	123822	*	3556	*	*	*	34.82
	S2	325631	*	9512	*	*	*	34.23
				41–50-fold coverage				
Yes	A2	611502	522220	13490	89282	14.60	38.71	45.33
	G1	4752	4255	105	497	10.46	40.52	45.26
	G2	737605	658475	16384	79130	10.73	40.19	45.02
	S1	167856	151888	3738	15968	9.51	40.63	44.91
	S2	3357	2951	74	406	12.09	39.88	45.36
	S3	117096	107234	2613	9862	8.42	41.04	44.81

(continued)

Table A3 *(Continued)*

DBR	Sample	Reads	Unique reads	"Loci"	PCR duplicates	% PCR duplicates	Coverage without PCR duplicates	Coverage with PCR duplicates
				41–50-fold coverage *(continued)*				
No	A1	42909	*	952	*	*	*	45.07
	A2	85965	*	1920	*	*	*	44.77
	G2	90407	*	2007	*	*	*	45.05
	S2	110906	*	2476	*	*	*	44.79
				>50-fold coverage				
Yes	A2	14375479	7088450	71668	7287029	50.69	98.91	200.58
	G1	67809	43679	397	24130	35.59	110.02	170.80
	G2	6613711	3043150	31124	3570561	53.99	97.78	212.50
	S1	909620	483638	5547	425982	46.83	87.19	163.98
	S2	37857	25990	228	11867	31.35	113.99	166.04
	S3	2024898	725425	5471	1299473	64.17	132.59	370.11
No	A1	425838	*	2034	*	*	*	209.36
	A2	857247	*	3728	*	*	*	229.95
	G2	2396407	*	7535	*	*	*	318.04
	S2	1923010	*	6594	*	*	*	291.63

* No PCR duplicates could be identified due to the use of adapters without DBRs.

Table A4

Overview of ddRAD "loci" unaffected by PCR duplicates and those affected by different proportions of PCR duplicates (in 10% steps from 0% to 100%) for six invertebrate samples (A1–S3)

Sample	Coverage	"Loci" without duplicates	"Loci" with duplicates (frequencies of PCR duplicates at "loci" given in 10% bins)									
			>0–10%	>10–20%	>20–30%	>30–40%	>40–50%	>50–60%	>60–70%	>70–80%	>80–90%	>90–100%
A2	11–20	20,597	15,110	5,072	439	13	2	0	0	0	0	0
	21–30	3,598	11,067	3,053	107	1	0	0	0	0	0	0
	31–40	509	5,524	1,998	54	1	0	0	0	0	0	0
	41–50	36	2,161	1,275	29	0	0	0	0	0	0	0
	Total	322,800	36,992	23,637	3,629	2,304	3,991	57	90	27	16	5
G1	11–20	1017,716	16,485	9,979	3,094	633	140	17	6	3	2	1
	21–30	17,811	11,844	8,154	2,111	323	30	3	1	5	0	0
	31–40	2,881	7,077	7,269	2,042	233	22	4	1	1	0	0
	41–50	442	3,780	6,313	1,990	199	13	4	1	1	1	0
	Total	1038,904	44,868	74,244	47,271	29,448	50,271	1,073	5,313	910	144	79
G2	11–20	236	236	152	45	5	4	0	0	0	3	0
	21–30	20	97	60	12	1	0	0	0	0	0	0
	31–40	2	63	38	3	1	0	0	0	0	0	0
	41–50	0	29	37	5	0	0	0	0	0	0	0
	Total	130,219	496	1,050	712	1,693	5,925	55	482	67	8	0
S1	11–20	23,820	17,713	6,143	626	45	9	1	0	0	0	0
	21–30	3,625	10,033	2,408	91	4	1	0	0	0	0	0
	31–40	526	4,445	1,160	38	5	0	0	0	0	0	0
	41–50	48	1,796	633	9	2	0	0	0	0	0	0
	Total	563,353	37,825	29,833	7,794	8,180	17,216	204	976	142	35	11
S2	11–20	45,295	46,288	23,124	3,394	208	11	0	0	0	2	4
	21–30	7,751	32,612	15,172	1,171	28	0	1	0	0	1	4
	31–40	1,113	16,710	10,378	696	7	0	0	0	0	0	3
	41–50	130	7,807	6,911	361	6	1	0	0	0	0	1
	Total	787,822	115,135	100,752	17,817	10,094	25,736	330	987	180	109	50
S3	11–20	307	259	107	17	1	0	0	0	0	0	0
	21–30	46	151	63	7	0	0	0	0	0	0	0
	31–40	11	103	39	3	0	0	0	0	0	0	0
	41–50	1	53	40	1	0	0	0	0	0	0	0
	Total	135,340	690	1,203	859	1,729	4,083	11	102	3	1	0

Data are shown for the complete coverage range as well as for coverage values between 11 and 50, which are typically aimed for in population genomic studies.

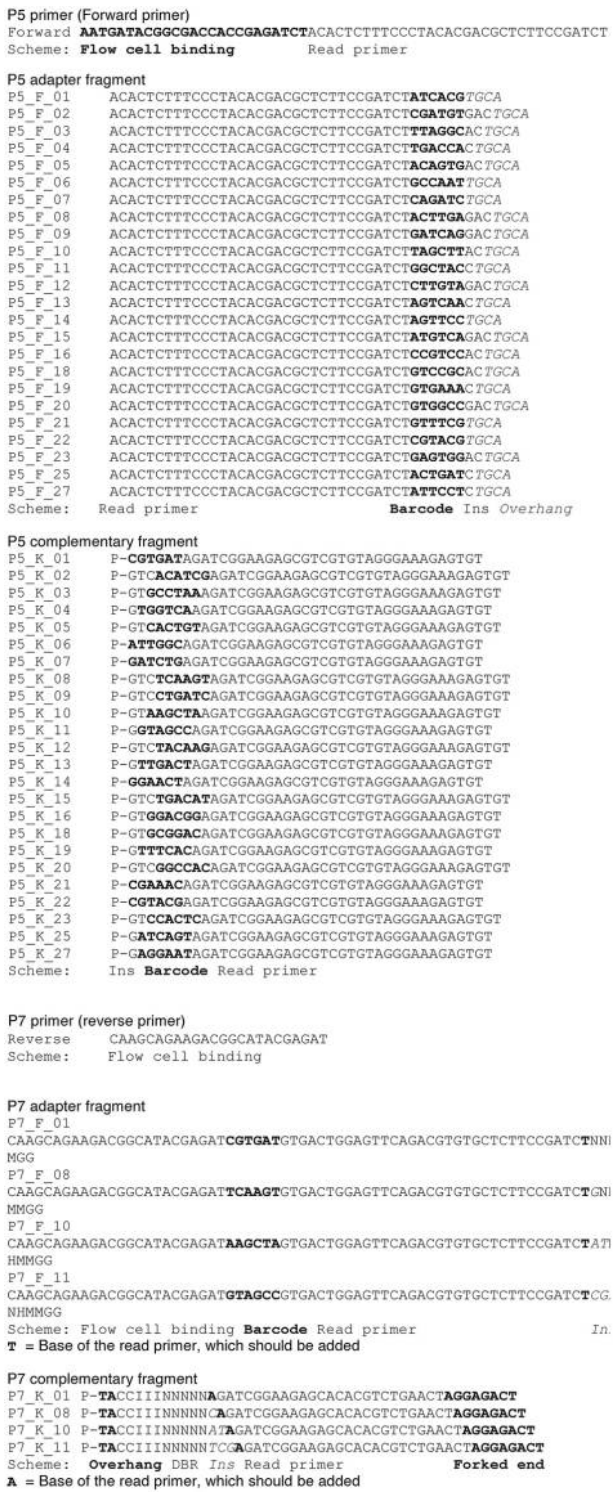

Figure A1. List of oligonucleotides in 5′→3′ direction. Grey-shaded schemes, indicating different parts of the oligonucleotides, are explained below the sequences. Sequences with a P- in front of the sequence are 5′ phosphorylated. Nucleotides are according to the IUPAC Code; I indicates desoxy-Inosine bases.

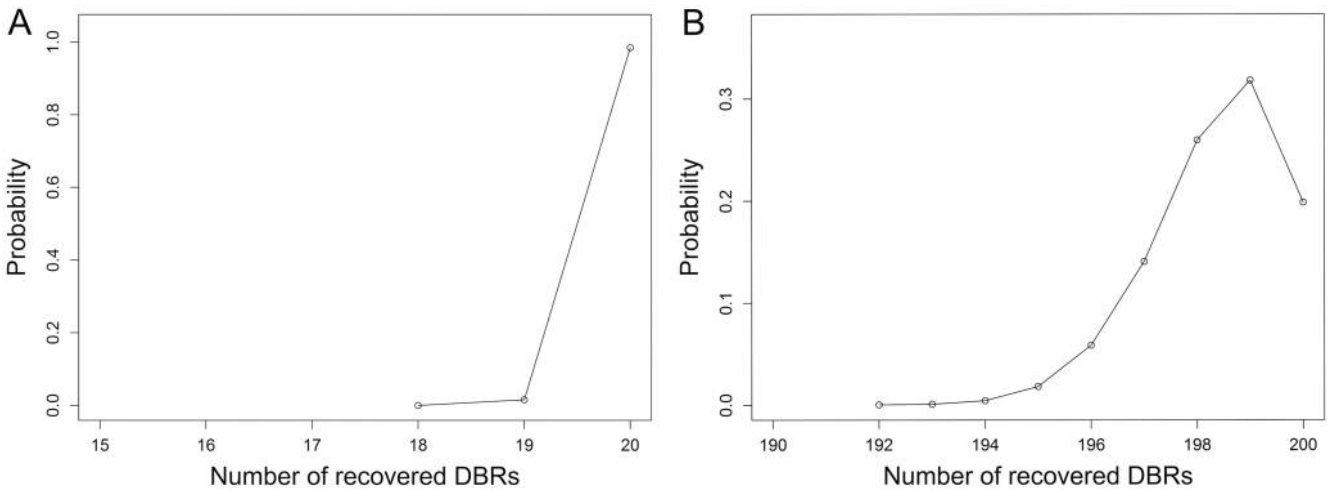

Figure A2. Probability to recover DBRs for two coverage values. A simulation approach was used to generate the recovery probability for different amounts of DBRs for (A) 20× and (B) 200× coverage. A uniform distribution of 12,288 different DBR variants was used in the simulation.