MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 5943–5946

NEWS AND VIEWS

REPLY

Trade-offs and utility of alternative RADseq methods: Reply to Puritz *et al.* 2014

KIMBERLY R. ANDREWS,[*] PAUL A. HOHENLOHE,[†] MICHAEL R. MILLER,[‡] BRIAN K. HAND,[§] JAMES E. SEEB[¶] and GORDON LUIKART[§]

[*]School of Biological & Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, UK; [†]Department of Biological Sciences, Institute of Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844-3051, USA; [‡]Department of Animal Science, University of California, One Shields Avenue, Davis, CA 95616, USA; [§]Flathead Lake Biological Station, Fish and Wildlife Genomics Group, University of Montana, Polson, MT 59860, USA; [¶]School of Aquatic and Fishery Sciences, 1122 NE Boat Street Box 355020, Seattle, WA 98195-5020, USA

Puritz *et al.* provide a review of several RADseq methodological approaches in response to our 'Population Genomic Data Analysis' workshop (Sept 2013) review (Andrews & Luikart 2014). We agree with Puritz *et al.* on the importance for researchers to thoroughly understand RADseq library preparation and data analysis when choosing an approach for answering their research questions. Some of us are currently using multiple RADseq protocols, and we agree that the different methods may offer advantages in different cases. Our workshop review did not intend to provide a thorough review of RADseq because the workshop covered a broad range of topics within the field of population genomics. Similarly, neither the response of Puritz *et al.* nor our comments here provide sufficient space to thoroughly review RADseq. Nonetheless, here we address some key points that we find unclear or potentially misleading in their evaluation of techniques.

Keywords: allele dropout, genotyping error, next generation sequencing, PCR duplicates, pooling, population genomics

Received 18 July 2014; revised 01 October 2014; accepted 08 October 2014

Puritz *et al.* (2014) focus their discussion on RADseq PCR artefacts that have the potential to cause problems for population genomics analyses by producing genotyping errors, skewing allele frequency estimates and causing false positive alleles (Pompanon *et al.* 2005). One such PCR artefact that was described in our meeting review is PCR duplicates (Fig. 1). PCR duplication rates can vary greatly across RADseq projects and samples, and can occur at high frequencies (e.g. >20% of reads, Hohenlohe *et al.* 2013; up to 60%, B. K. Hand & G. Luikart, unpublished data). The impact of PCR duplicates on population genomics analyses has not been quantified in the literature, but high frequencies of duplicates are expected to impact analyses by falsely increasing homozygosity and by making PCR errors appear to be true alleles (false alleles, Pompanon *et al.* 2005). Moreover, failure to remove PCR duplicates can spuriously inflate confidence in genotype calls because most genotyping approaches heavily rely upon read coverage to inform genotype likelihoods under the assumption that each read represents an independent observation (Nielsen *et al.* 2011). Puritz *et al.* downplay the importance of PCR duplicates by describing several methods aimed at avoiding, detecting or correcting for them. Unfortunately, the efficacy of these methods has not been tested through empirical or modelling studies. In some cases, the suggested methods are unlikely to improve genotyping. For example, Puritz *et al.* suggest using a very conservative criterion of nearly 50% representation of each allele before calling a heterozygote to account for PCR duplicates; however, implementing this criterion would lead to a severe homozygote bias for data generated using any RADseq protocol, and especially data containing PCR duplicates.

Puritz *et al.* also suggest the use of PCR-free methods to avoid PCR duplicates. These methods have strong potential (Andrews & Luikart 2014). However, these methods are rarely feasible with present-day technology due to high per-sample cost and high per-sample DNA quantity requirements (i.e. ezRAD using PCR-free Illumina kits, approximately $30/sample and 1–2 μg of DNA/sample, Toonen *et al.* 2013), except for research questions that can be answered using pooled samples (Futschik & Schlöetterer 2010).

As described in our meeting review, the most straightforward method currently developed for identifying RADseq PCR duplicates can only be used for data generated using methods that have a random-shearing step and also generate paired-end sequences (PE-RADseq). For these methods, PCR duplicates can be identified as fragments that are identical in insert length and sequence composition, because random shearing ensures that fragments at a given locus are unlikely to be of equal length unless they are duplicates (Fig. 1; Davey *et al.* 2011; Hohenlohe *et al.* 2013). In contrast, for methods without a random-shearing step, all fragments at a given locus are expected to be of equal length whether or not they are PCR duplicates, and therefore fragment length and sequence composition cannot be used to identify duplicates. Currently, the only

Correspondence: Kimberly R. Andrews, Fax: +44 191 334 1201; E-mail: kimandrews@gmail.com

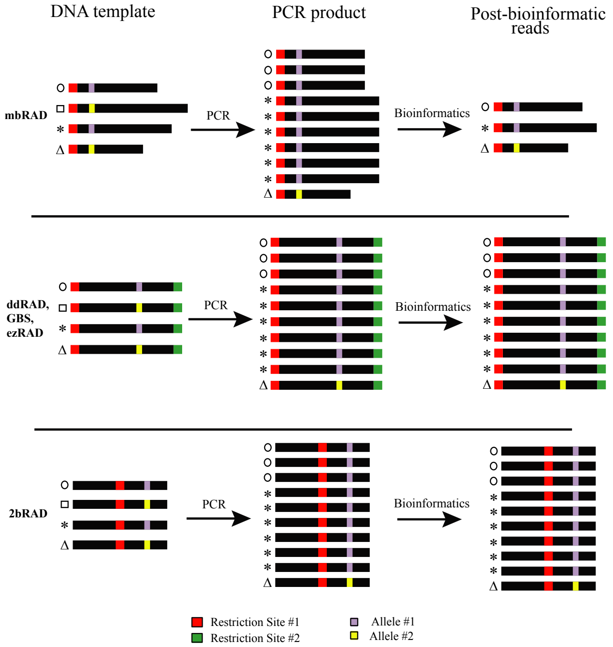

Fig. 1 Example of fragments produced after PCR for one heterozygous locus for different RADseq protocols, and the reads retained after bioinformatic analyses. PCR duplicates are shown with the same symbol (circle, square, asterisk or triangle) as the parent fragment from the original template DNA. By chance, some alleles will amplify more than others during PCR. For all protocols, PCR duplicates will be identical in sequence composition and length to the original template molecule. For mbRAD (the original RADseq, Miller *et al.* 2007; Baird *et al.* 2008), this feature can be used to identify and remove PCR duplicates bioinformatically, because original template molecules for a given locus will not be identical in length. For alternative methods, this feature cannot be used to identify PCR duplicates, because all original template molecules for a given locus are identical in length. High frequencies of PCR duplicates can cause heterozygotes to appear as homozygotes or can cause PCR errors to appear as true diversity.

method employing a random-shearing step is the original RADseq (Miller *et al.* 2007; Baird *et al.* 2008; called 'mbRAD' by Puritz *et al.*). However, future work may reveal effective methods for addressing the issue of PCR duplicates in other RADseq methods. For example, one promising approach for identifying PCR duplicates is the incorporation of degenerate bases into adapter sequences, which enables counting of the number of template molecules (Casbon *et al.* 2011; Tin *et al.* 2014).

Puritz *et al.* describe two additional sources of bias that can be introduced during PCR: preferential amplification of loci based on GC content and fragment size. These biases should not directly impact genotyping when using standard methods for removing loci with low coverage unless indels cause certain alleles to amplify with higher efficiency than others (Davey *et al.* 2013). However, these biases may greatly increase variance in coverage among loci, meaning that higher mean coverage must be attained to produce sufficient depth across all loci. Fragment size bias among loci during PCR is expected to be much higher for methods that rely exclusively on restriction enzymes to fragment genomic DNA, because each locus has a single characteristic fragment size (ezRAD, ddRAD, GBS; excluding 2b-RAD, for which all fragments are equal size, Elshire *et al.* 2011; Peterson *et al.* 2012; Wang *et al.* 2012; Toonen *et al.* 2013). In contrast, random-shearing-based approaches produce a range of fragment sizes for each locus, and therefore these methods should not be affected by fragment size bias during PCR (mbRAD) (Davey *et al.* 2011).

As Puritz *et al.* point out, Davey *et al.* (2013) identified a different type of fragment size bias that results from sonication in mbRAD. After mbRAD restriction digests, all fragments are much larger than the required size range for sequencing, and sonication is used to reduce fragment sizes. Davey *et al.* (2013) showed that incomplete shearing can occur for shorter restriction fragments, resulting in sheared fragments that are too large to be recovered during the size selection step before PCR. Ultimately, this effect may increase variance in coverage across loci. As with fragment size bias introduced during PCR, this variance in coverage should not influence genotyping, but would increase the number of sequence reads required to attain sufficient depth across all loci.

Another source of bias described by Puritz *et al.* is allele dropout, which has the potential to affect genotyping by causing heterozygotes to be scored as homozygotes or causing some individuals to produce no sequence data at affected loci (Taberlet *et al.* 1999; Davey *et al.* 2013; Gautier *et al.* 2013b). Simulation studies predicted that allele dropout would have a greater impact on data generated by ddRAD than mbRAD as a result of nucleotide variation causing gains and losses of restriction cut sites at either end of each locus for ddRAD versus just one end for mbRAD (Arnold *et al.* 2013). Over a range of reasonable conditions, the effect of allele dropout may be slight and possible to account for in estimates of summary statistics (Arnold *et al.* 2013; Gautier *et al.* 2013b).

Another important comparison between RADseq methods is the cost and technical complexity of a method, given the needs of a particular study. For instance, if sequence data from a relatively large number of individual samples are required (rather than pools of individual samples, Futschik & Schlöetterer 2010), scaling of costs and protocol complexity differs widely among RADseq methods. In most RADseq methods, individually barcoded samples can be multiplexed relatively early in library preparation, and thus subsequent steps are conducted on a much smaller number of pools. In contrast, library preparation steps and cost increase linearly with the total number of individual samples in the ezRAD method (Toonen *et al.* 2013).

Given that some RADseq methods will usually only be time-efficient and cost-effective when using pooled samples (i.e. samples without barcoded individuals), another important consideration when designing a RADseq study is whether the research question can be answered using pooled samples. Pooling individuals without barcoding reduces library preparation time and cost for any RADseq method, with the most dramatic time and cost reduction for ezRAD. Pooling shows some promise for producing accurate estimates of allele frequencies (Futschik & Schlöetterer 2010; Ferretti *et al.* 2013; Gautier *et al.* 2013a; Lynch *et al.* 2014; but see Venter 2010; Anderson *et al.* 2014). However, data from pooled samples cannot be used for many widely used population genetics statistics including tests that assign individuals to populations, such as STRUCTURE (Pritchard *et al.* 2000), parentage tests, and tests that rely on estimates of linkage disequilibrium, such as some tests for selection (e.g. Kayser *et al.* 2003; Nielsen *et al.* 2005). Another disadvantage to using pooled samples for population genomics data analysis is that any cryptic population structure that has not been identified *a priori* will go undetected in pooled samples. Researchers should also be aware that errors in allele frequency estimates caused by sequencing error, mapping error and paralogous loci are more difficult to identify for pooled data (Gautier *et al.* 2013a).

There are numerous other considerations in choosing among RADseq methods, including the research questions and goals of the study, genome size and complexity, and quantity and quality of available DNA per sample. Each method may be most appropriate for a particular situation. However, further empirical study and modelling of the extent and consequences of various sources of error and bias are needed to ensure reliable RADseq data production and interpretation. Given the recent exponential growth in use of RADseq in the fields of population genomics, molecular ecology, and conservation genetics, studies of RADseq reliability would strongly contribute to the advancement of these fields.

Acknowledgements

We thank Jeff Good and three anonymous reviewers for helpful comments regarding the manuscript. We thank S. Amish for helpful feedback on PCR duplication rates. KA was supported by a European Union Marie Curie International Incoming Fellowship. PH was supported by NIH grant P30GM103324. BH was supported by the Department of the Interior Northwest Climate Science Center. GL was supported in part by funding from US National Science Foundation Grants DEB-1258203 and DEB-1067613.

References

Anderson EC, Skaug HJ, Barshis DJ (2014) Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, **23**, 502–512.

Andrews KR, Luikart G (2014) Recent novel approaches for population genomics data analysis. *Molecular Ecology*, **23**, 1661–1667.

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

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, **39**, 1–8.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

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

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Ferretti L, Ramos-Onsins SE, Perez-Enciso M (2013) Population genomics from pool sequencing. *Molecular Ecology*, **22**, 5561–5576.

Futschik A, Schlöetterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Gautier M, Foucaud J, Gharbi K *et al.* (2013a) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, **22**, 3766–3779.

Gautier M, Gharbi K, Cezard T *et al.* (2013b) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Hohenlohe PA, Day MD, Amish SJ *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, **22**, 3002–3013.

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, **20**, 893–900.

Lynch M, Bost D, Wilson S, Maruki T, Harrison S (2014) Population-genetic inference from pooled-sequencing data. *Genome Biology and Evolution*, **6**, 1210–1218.

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

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014) Demystifying the RAD fad. *Molecular Ecology* (this issue).

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution*, **14**, 323–327.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2014) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12314.

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

Venter JC (2010) Multiple personal genomes await. *Nature*, **464**, 676–677.

Wang S, Meyer E, McKay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, **9**, 808–812.

K.A. led the writing of the paper, and all authors contributed to the writing. B.H. conducted data analysis. P.H., M.M., B.H., J.S., and G.L. were instructors at the ConGen workshop.

doi: 10.1111/mec.12964