MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 5937–5942

NEWS AND VIEWS

COMMENT

Demystifying the RAD fad

JONATHAN B. PURITZ,* MIKHAIL V. MATZ,† ROBERT J. TOONEN,‡ JESSE N. WEBER,§ DANIEL I. BOLNICK§ and CHRISTOPHER E. BIRD¶

**Marine Genomics Laboratory, Harte Research Institute, Texas A&M University-Corpus Christi, 6300 Ocean Drive, Corpus Christi, TX 78412-5869, USA; †Department of Integrative Biology, University of Texas at Austin, 205 W 24th ST C0990, Austin, TX 78712, USA; ‡Hawai'i Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawai'i at Mānoa, PO Box 1346, Kāne'ohe, HI 96744, USA; §Department of Integrative Biology, Howard Hughes Medical Institute, University of Texas at Austin, Austin, TX 78712, USA; ¶Department of Life Sciences, Texas A&M University-Corpus Christi, 6300 Ocean Drive, Corpus Christi, TX 78412-5869, USA*

We are writing in response to the population and phylogenomics meeting review by Andrews & Luikart (2014) entitled 'Recent novel approaches for population genomics data analysis'. Restriction-site-associated DNA (RAD) sequencing has become a powerful and useful approach in molecular ecology, with several different published methods now available to molecular ecologists, none of which can be considered the best option in all situations. A&L report that the original RAD protocol of Miller *et al.* (2007) and Baird *et al.* (2008) is superior to all other RAD variants because putative PCR duplicates can be identified (see Baxter *et al.* 2011), thereby reducing the impact of PCR artefacts on allele frequency estimates (Andrews & Luikart 2014). In response, we (i) challenge the assertion that the original RAD protocol minimizes the impact of PCR artefacts relative to that of other RAD protocols, (ii) present additional biases in RADseq that are at least as important as PCR artefacts in selecting a RAD protocol and (iii) highlight the strengths and weaknesses of four different approaches to RADseq which are a representative sample of all RAD variants: the original RAD protocol (mbRAD, Miller *et al.* 2007; Baird *et al.* 2008), double digest RAD (ddRAD, Peterson *et al.* 2012), ezRAD (Toonen *et al.* 2013) and 2bRAD (Wang *et al.* 2012). With an understanding of the strengths and weaknesses of different RAD protocols, researchers can make a more informed decision when selecting a RAD protocol.

Correspondence: Jonathan B. Puritz, Fax: (361) 825 2050; E-mail: jpuritz@gmail.com

Keywords: genomics, next-generation sequencing, population, restriction, restriction-site-associated DNA

Received 12 March 2014; revised 30 May 2014; accepted 7 June 2014

Mitigating PCR artefacts in RAD protocols

With a firm understanding of the molecular biology of RAD protocols, PCR artefacts and their impact on allele frequency estimates within loci can be effectively mitigated in mbRAD (Davey *et al.* 2013) as well as ddRAD, ezRAD and 2bRAD See Box 1 for method descriptions. Consequently, one should not be compelled to conduct mbRAD solely for the ability to identify putative PCR duplicates in the sequence data. It is well known that, among loci, high GC content is negatively correlated with read depth which adds complication to RAD analysis, but PCR duplicate identification does not solve this problem (Davey *et al.* 2013). The primary concern of molecular ecologists, in terms of PCR artefacts, is whether or not PCR skews RAD allele frequencies within loci, thereby causing consistent and predictable genotyping errors. In all RAD variants, none of the artefacts introduced by PCR are expected to result in statistical bias within loci because there is little relationship between allelic identity and factors that bias PCR amplification (i.e. inconsistent priming sites, AT/GC content, or fragment length). In all RAD protocols, all priming sites are identical; AT/GC content varies very little within loci, and with the exception of mbRAD, DNA fragments within loci are equal lengths. Ultimately, there is no reason to expect that allele frequencies within loci will be biased by PCR, but this remains untested. The majority of PCR artefacts are introduced during library enrichment on the laboratory bench as opposed to cluster generation or sequencing by synthesis in the Illumina sequencer (Aird *et al.* 2011); therefore, the impact of PCR artefacts on RAD genotyping can most effectively be controlled in library preparation. Effective strategies that reduce the amount of statistical error in allele frequency estimation due to PCR include:

1 *Modify PCR enrichment*: Aird *et al.* (2011) report that PCR artefacts due to AT/GC content can be drastically reduced by simple modifications to the library enrichment protocol, such as reducing temperature ramp rate, extending the initial denaturation step to 180 s and subsequent denaturation steps to 80 s, adding betaine to Phusion HF polymerase reactions can reduce hairpins, or alternative polymerases can be used with better performance than Phusion HF such as AccuPrime Taq HiFi or KAPA HiFi. Oyola *et al.* (2012) found that adding

tetramethylammonium chloride to PCR increases the thermostability of AT base pairs and reduces bias against the amplification of AT-rich regions. PCR enrichment can be modified in all RAD protocols.

2 *Replication and amplicon labelling*. Replication of RAD libraries can be used to estimate the error introduced by library preparation. From empirical estimates of genotyping error, which includes error due to PCR, filters can be created to identify and remove loci prone to bias and to determine the heterozygote discovery rate. These features are currently included in the 2bRAD pipeline. Another replication strategy employed in ddRAD that could also be applied to any library enrichment is to conduct multiple independent PCR amplifications per sample and combine them (Peterson *et al.* 2012). Alternatively, the use of degenerate base regions (Casbon *et al.* 2011) allows direct identification of the original alleles prior to PCR amplification similar to the random shearing advocated for mbRAD by A&L.

3 *In Silico QC*. The latest generation of RAD analysis pipelines for ddRAD, ezRAD and 2bRAD (see links for our pipelines below) offers empirical estimates of genotype call confidence that are valid despite PCR amplification of the library. Likewise, Stacks (Catchen *et al.* 2011), the mbRAD pipeline software, identifies reads with paired-end reads starting at identical positions as PCR duplicates. Identification of PCR duplicates helps satisfy the assumptions of SNP calling, thereby resulting in improved call confidence. Haplotype callers such as GATK (DePristo *et al.* 2011; van der Auwera *et al.* 2013) and FreeBayes (Garrison & Marth 2012) calculate frequencies of DNA fragments rather than just SNPs and more effectively identify erroneous reads caused by base substitutions due to polymerase errors. FreeBayes, employed in the dDocent pipeline for ddRAD and ezRAD Puritz *et al.* 2014, also models the nonindependence of homologous reads due to both mitotic and PCR duplicates, resulting in more robust tests of heterozygosity. Using conservative criteria for calling a novel allele, such as repeatability among individuals and/or replicates, relative lack of allele bias (close to 50% representation in a heterozygote) and, in 2bRAD, comparable representation of both strands (lack of strand bias, Guo *et al.* 2012) can lead to high confidence in called genotypes. False homozygote calls – miscalled heterozygotes due to random allele dropout (as opposed to systematic allele dropout discussed in the next section) – still happen but are rare: approximately 10% of all heterozygotes called *de novo* and about 5% of all heterozygotes called in a reference-based pipeline based on replicate genotyping with 2bRAD (M. V. Matz, personal observations).

4 *Remove PCR enrichment*. As noted by A&L, ezRAD can be conducted with Illumina TruSeq PCR-Free prep kits, thereby negating any PCR-related biases. This is the only way to completely remove PCR artefacts during RAD library preparation.

Potential biases when conducting RADseq

In addition to PCR artefacts, there are other biases to consider when selecting a RAD protocol, but there are strategies to mitigate these sources of bias and error for all RAD protocols. Davey *et al.* (2013) experimentally identified restriction fragment size bias and heterozygous restriction sites (the root cause of allele dropout-ADO) as serious problems in mbRAD genotyping. Guo *et al.* (2012) further identified strand bias, where forward and reverse reads of the same DNA fragment result in different gentoypes, as a potential problem for reliable genotyping. While heterozygous restriction sites and strand biases plague all RAD methodologies discussed here, restriction fragment size bias on within-locus allele frequency estimates is a phenomenon associated with fragment shearing via sonication, a methodology employed only by mbRAD. Thus, although ADO can clearly bias some types of analyses (Arnold *et al.* 2013; Gautier *et al.* 2013), it may be relatively unimportant for others and can often be dealt with by simply excluding problematic loci (Davey *et al.* 2013). One such simple strategy to mitigate the influence of fragment size bias and heterozygous restriction sites in all RAD protocols is to filter from consideration any loci that are not represented in all genotyped individuals. Similarly, any loci exhibiting strand bias can be removed from consideration, but strand bias can, at present, only be identified in 2bRAD where each restriction fragment can be sequenced in either direction.

Advantages and disadvantages of different RAD protocols

Each alternative RAD method has advantages and drawbacks (See Table 1, Miller *et al.* 2007; Wang *et al.* 2012; Peterson *et al.* 2012; Toonen *et al.* 2013; Elshire *et al.* 2011; Sonah *et al.* 2013; Poland *et al.* 2012). In our estimation, all RAD protocol variants are effective and each has varying utility, bias and technical challenge. We demonstrate our premise by summarizing four RAD variants that have broad applicability to taxa without reference genomes (>99.9% of all species), the mbRAD protocol (Miller *et al.* 2007 & Baird *et al.* 2008), ddRAD (Peterson *et al.* 2012), ezRAD (Toonen *et al.* 2013) and 2bRAD (Wang *et al.* 2012), and outlining two advantages and disadvantages per method.

Advantages of mbRAD

1 The random shearing of the 3′ end of each RAD locus helps with the identification of putative PCR duplicates, the assumption being that any read pairs with identical starting position of the paired-end read resulted are duplicates.

2 Random shearing, combined with larger insert size ranges (determined by library size selection), also makes

Box 1. Four Different RAD Protocols

The original RAD protocol (Miller *et al.* 2007 and Baird *et al.* 2008) involves six steps. Genomic DNA is first digested with a single restriction enzyme (usually a low-frequency cutter). Barcode containing adapters are then ligated onto digested 5′ ends. Ligated genomic DNA is then sonicated, and a 3′ adapter is ligated to the randomly sheared end. After ligation, the library is size-selected. Finally, RAD fragments with both adapters properly ligated are enriched with PCR.

The double digest RAD protocol (Peterson *et al.* 2012) uses two enzymes to digest genomic DNA in a four-step protocol. Genomic DNA is simultaneously digested with two restriction enzymes (usually a low-frequency cutter combined with a high-frequency cutter). Barcoded P1 adapters (with an overhang matching the first restriction site) and P2 adapters (with an overhang matching the second enzyme restriction site) are ligated onto digested fragments in a single sticky-end ligation. Samples are then pooled and size-selected. Lastly, PCR is used to enrich the library and also to introduce a second barcode in the form of an Illumina index, increasing multiplexing potential. It should be noted that GBS (Poland *et al.* 2012) is extremely similar to ddRAD and can be considered a specific ddRAD protocol. Most of the pros and cons associated with ddRAD are also relevant to RESTseq (Stolle & Moritz 2013).

The ezRAD protocol (Toonen *et al.* 2013) uses two high-frequency cutter isochimozer enzymes (for the same cut site) to digest genomic DNA. Subsequently, digested DNA is inserted directly into a commercially available Illumina TruSeq library preparation kit. Using the Illumina kit, DNA is end-repaired and adapters using either single or dual indexing are ligated onto genomic fragments. Samples are then pooled and size-selected. Depending on the Illumina kit, libraries can either be enriched via PCR or using the non-PCR kits are finished after size selection.

The 2bRAD protocol (Wang *et al.* 2012) relies on a IIb-type restriction endonuclease to excise 36-bp fragments containing the 6-base recognition site and adjacent 5′ and 3′ base pairs. To these fragments, adapters with dual barcodes are ligated, and the target band is excised out of an agarose gel after PCR enrichment. There are no intermediate purification stages and no fragment size selection. The procedure can be customized to represent less loci in the genome by the use of base-selective adapters. Current laboratory and bioinformatic protocols implement dual indexing and the use of replicates to derive empirical quality filtering criteria.

it more likely to *de novo* assemble RAD loci (contigs) of greater length than other RAD methods. Longer contigs increase the likelihood of clustering/aligning loci to existing genomic resources of other organisms, critical for identifying function and gene ontology.

Disadvantages of mbRAD

1 The mbRAD protocol is the most technically challenging and complex laboratory protocol of the four RAD methods and requires nonstandard laboratory equipment, such as a sonicator.
2 As reported by Davey *et al.* (2013), the largest source of bias in mbRAD libraries is restriction fragment length bias. This bias, particular to mbRAD, is introduced by the shearing of genomic DNA after restriction digest to random, variable lengths, causing fragments to be sequenced at different depths.

Advantages of ddRAD

1 ddRAD offers the greatest degree of customization. Depending on the enzymes chosen (a single set of uniquely barcoded 'flex-adapters' works with at least five enzyme pairings), and range of fragment sizes selected, a researcher can obtain hundreds of SNPs per individual at very low cost (e.g. sufficient for basic population structure analyses), thousands of SNPs for QTL mapping experiments at moderate cost, or tens of thousands of SNPs for more precise association mapping. Thus, studies that require fewer fragments to obtain robust inferences, or investigators wanting to optimize the number of fragments and/or individuals covered at reasonable depth with a limited number of sequencing reads, can economize.
2 As with any protocol that avoids shearing and tunes fragment numbers with size selection (which includes ezRAD), examining histograms of digested samples early in a project enables researchers to identify and then exclude excessively frequent fragments (e.g. transposons) from libraries. This procedure can be very valuable when studying organisms with large, unsequenced genomes.

Disadvantages of ddRAD

1 Using fragment size selection to tune the quantity of loci sampled can lead to variable representation of some loci

Table 1 Comparison of the utility, technical complexity and sources of bias for different RAD methods

	mbRAD	ddRAD	ezRAD	2bRAD
Restriction cut sites per 10 kb*	~0.2–2.4	~3.7×10^{-5}–39	~39	~2.4
Postdigest fragment reduction	Size selection	Size selection	Size selection	Selective adapters
Contigs > 200 bp†	Yes	No	Some	No
Ability to blast/annotate *de novo* contigs	High	Mid	Mid	Low
Protocol complexity (# Steps)‡	6	4	4–6	3
Level of technical difficulty	High	Mid	Low	Low
Level of technical support	Low	Low	Mid-high	Low
Insert complexity (first × bases)	Low	Low	Very low	High
PCR AT/GC content, copy number Bias among loci	Yes	Yes	Yes, No§	Yes
ID of PCR duplicates	Yes	No	No§	No¶
Uniform locus length	No	No	No	Yes
Oligos required to uniquely identify and build 96 libraries	196**	31	20–22	37
Target insert size range	200–600 bp	Customizable	Customizable	33–36 bp

*These numbers represent only theoretical calculations for one enzyme (or enzyme combination). The number of fragments sampled will depend on size selection, genome composition, the number of enzymes used and the use of restrictive adapters (see 2bRAD).
†When performing 100 bp reads such as on a HiSeq platform.
‡Not counting clean-up steps.
§ezRAD can be used with a PCR-free library preparation kit, thus removing the need to detect PCR duplicates.
¶2bRAD can detect PCR errors by mismatch among forward and reverse reads on individual strands.
**With some effort, the indexing for mbRAD can be modified to reduce the oligo counts to 22–37.

among libraries. This can be minimized using precise size selection tools such as a Pippin Prep (Sage Science). Double digest methods may also be particularly susceptible to ADO (Arnold *et al.* 2013), and this should be considered when performing sensitive population genetic analyses.

2 ddRAD arguably requires the highest quality genomic DNA of all the RAD methods. Proper fragment ligation relies not only on the complete digestion of two enzymes but also on completely intact 5′ and 3′ overhangs.

Advantages of ezRAD

1 ezRAD is the only protocol that relies on Illumina TruSeq kits, which come with an extensive manual, customer support and a guarantee. This means that the work can be sent out to any commercial laboratory that provides Illumina library preparation services and also means that adapter oligos and PCR primers do not need to be custom ordered. For small laboratories without the experience, equipment or resources to develop in-house RAD capability, this approach is probably the simplest path to obtain RAD data.

2 Combined with an Illumina PCR-Free TruSeq kit, ezRAD is the only RAD protocol that can bypass all potential PCR bias.

Disadvantages of ezRAD

1 Illumina TruSeq kits add simplicity and uniformity to the RAD library preparation, but they are also relatively expensive. However, kits have been successfully used with 1/2 and 1/3 reaction volumes, substantially reducing per sample costs.

2 All ezRAD reads start with the same four bases (GATC) which can result in poor read quality on Illumina sequencers if not properly addressed. Each cluster on the flow cell of Illumina sequencer produces a sequence read, and the first 4–5 nucleotides of Read 1 are used to discriminate among adjacent clusters. If the first 4–5 nucleotides are all the same on every cluster, the sequencer can mistakenly classify two clusters composed of two different DNA fragments as one, resulting in poor sequence quality. Most sequencing facilities can easily accommodate ezRAD libraries, however, by dark cycling, PhiX spiking, mixing high- and low-diversity libraries or employing custom sequencing primers that are the standard TruSeq sequencing primers with the addition of a 5′-GATC-3′ on the 3′ end.

Advantages of 2bRAD

1 Extreme protocol simplicity and cost-efficiency. The library preparation procedure literally involves sequential addition of reagents to the same 96-well plate and excising a well-defined band out of a gel in the end. There are no intermediate purification stages or need for special instrumentation beyond a PCR instrument and a standard agarose gel. Additionally, 2bRAD requires only 50-bp single-end Illumina sequencing, and restriction fragments are sequenced on either strand allowing the use of strand bias as a quality filtering criteria.

2 Lack of biases due to fragment size selection. In 2bRAD, essentially all endonuclease recognition sites in the genome can be sampled for sequencing. However, 2bRAD can be customized (by the use of selective adapters) to sequence less loci for applications such as population genetics or QTL mapping. Lastly, Current protocol and bioinformatic pipelines implement dual indexing and empirical quality filtering based on replicates.

Disadvantages of 2bRAD

1 Although 36-bp tags are long enough to be highly unique in a genome as large as human, in genomes with many duplications, they would have considerably less chance of being mapped unambiguously. In moderately duplicated genomes such as *Arabidopsis*, however, 2bRAD works well (Wang *et al*. 2012).
2 2bRAD fragments cannot be used to build genome contigs and are less likely to be cross-mappable across large genetic distances, such as across different species.

Conclusions

A&L state 'next-generation sequencing data analysis should be approached with a keen understanding of the theoretical models underlying the analyses and with analyses tailored to each research project'. We agree entirely and advocate that the same diligence be employed when designing a project and choosing from the diverse array of available RAD laboratory protocols. The most important considerations when selecting a particular RAD protocol are the facilities and molecular experience of the researcher applying the approach, as well as the biology of the organisms and the hypotheses being tested. All RAD protocols are powerful tools for SNP discovery and genotyping of nonmodel species, and it is difficult to make a wrong choice. It is important, however, to learn about the potential pitfalls inherent to each method and how to address them. Each approach has inherent strengths and weaknesses, and at present, there is no reason to broad-brush paint any method as the superior or default protocol.

Our RAD analysis pipelines

2bRAD: http://www.bio.utexas.edu/research/matz_lab/matzlab/Methods.html
dDocent: https://github.com/jpuritz/dDocent

References

Aird D, Ross MG, Chen WS *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**, R18.

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 *et al.* (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics.* 11:11.10:11.10.1–11.10.33.

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

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One*, **6**, e19315.

Casbon JA, Osborne RJ, Brenner S *et al.* (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, **39**, e81.

Catchen JM, Amores A, Hohenlohe P *et al.* (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

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

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

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.

Garrison E, Marth G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN].*

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

Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.

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

Oyola SO, Otto TD, Gu Y *et al.* (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics*, **13**, 1.

Peterson BK, Weber JN, Kay EH *et al.* (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, **7**, e37135.

Poland JA, Brown PJ, Sorrells ME *et al.* (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, **7**, e32253.

Puritz JB, Hollenbeck CM, Gold JR (2012) *dDocent:* a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. . *PeerJ*, **2**: e431 doi: 10.7717/peerj.431.

Sonah H, Bastien M, Iquira E *et al.* (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One*, **8**, e54603.

Stolle E, Moritz RFA (2013) RESTseq – efficient benchtop population genomics with RESTriction Fragment SEQuencing. *PLoS One*, **8**, e63960.

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

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

J.B.P. and C.E.B. conceived this response. J.B.P., M.V.M., R.J.T., J.N.W., D.I.B. and C.E.B. contributed to the writing of this manuscript.

doi: 10.1111/mec.12965