MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 1163–1171 doi: 10.1111/1755-0998.12378

AftrRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data

MICHAEL G. SOVIC,*† ANTHONY C. FRIES* and H. LISLE GIBBS*†
*Department of Evolution, Ecology, and Organismal Biology, Aronoff Laboratory, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210, USA, †Ohio Biodiversity Conservation Partnership, Aronoff Laboratory, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210, USA

Abstract

An increase in studies using restriction site-associated DNA sequencing (RADseq) methods has led to a need for both the development and assessment of novel bioinformatic tools that aid in the generation and analysis of these data. Here, we report the availability of AftrRAD, a bioinformatic pipeline that efficiently assembles and genotypes RADseq data, and outputs these data in various formats for downstream analyses. We use simulated and experimental data sets to evaluate AftrRAD's ability to perform accurate *de novo* assembly of loci, and we compare its performance with two other commonly used programs, STACKS and PYRAD. We demonstrate that AftrRAD is able to accurately assemble loci, while accounting for indel variation among alleles, in a more computationally efficient manner than currently available programs. AftrRAD run times are not strongly affected by the number of samples in the data set, making this program a useful tool when multicore systems are not available for parallel processing, or when data sets include large numbers of samples.

Keywords: bioinformatics, *de novo* assembly, genotyping, locus identification, RADseq

Received 7 October 2014; revision received 29 December 2014; accepted 15 January 2015

Introduction

Restriction site-associated DNA sequencing (RADseq) methods (Miller *et al.* 2007; Baird *et al.* 2008; Davey & Blaxter 2011; Etter *et al.* 2011) are now widely used in studies of population genetics, phylogeography and conservation biology. By targeting SNP variation adjacent to restriction sites, RADseq provides a means to efficiently sample homologous SNPs across multiple individuals with little or no prior information about the genome. As a result, RADseq has become a popular tool for studying genetic variation in nonmodel organisms (i.e. Reitzel *et al.* 2013; Gamble & Zarkower 2014; Leache *et al.* 2014; Martin & Feinstein 2014; Viricel *et al.* 2014).

An important step in RADseq studies is *de novo* assembly of the short-read sequences into orthologous loci. The large size of these short-read sequence data sets, along with factors such as paralogy, sequencing and PCR error, and multiple mutation types (i.e. indel variation in addition to SNPs), make efficient and accurate assembly challenging. A number of bioinformatic tools have been developed to date for analysing RADseq data sets, including STACKS (Catchen *et al.* 2011), RAINBOW (Chong *et al.* 2012) and PYRAD (Eaton 2014). Of these, STACKS has been the most widely adopted, in part due to is low computational demand. However, this increase in analysis speed comes at the cost of not aligning reads to each other, and as a result, STACKS analyses do not account for indel variation that is likely to be present. Recently, PYRAD (Eaton 2014) was described as an alternative pipeline which effectively accounts for indel variation, but this program is computationally demanding, and can require multicore processing, especially as the number of individuals in the data set increases (see PYRAD documentation).

Here, we describe AftrRAD (align from total reads), a novel pipeline for the *de novo* assembly and genotyping of RADseq data. We use both simulated and experimental data sets to compare the performance of this pipeline to other commonly used programs such as STACKS (Catchen *et al.* 2011) and PYRAD (Eaton 2014) and discuss advantages and disadvantages of AftrRAD in relation to these widely used pipelines. In brief, AftrRAD, like PYRAD, effectively handles indel variation among alleles by aligning reads prior to *de novo* assembly of loci. However, the computational demand of AftrRAD is less dependent on the number of samples in the data set, making it useful for studies using large numbers of individuals, and in most cases, it runs very efficiently

Correspondence: Michael G. Sovic, Fax: 614-292-2030;
E-mail: sovic.1@osu.edu

without the need for parallel analyses. AftrRAD also provides a number of unique options for evaluating the quality of the RADseq data set and can currently output data for analyses in programs such as STRUCTURE (Pritchard *et al.* 2000), GENEPOP (Raymond & Rousset 1995), and SNAPP (Bryant *et al.* 2012) and can also produce unfolded site frequency spectra for analyses with fastsimcoal (Excoffier *et al.* 2013).

Methods

We first describe the conceptual approach used by AftrRAD to assemble reads into loci and score individual genotypes. Further details on these approaches can be found in Fig. 1, Fig. S1 (Supporting information) and in the documentation provided with the program, which is available at u.osu.edu/sovic.1 and at https://github.com/mikesovic/AftrRAD.git. We then describe the methods used in this study to evaluate the performance of AftrRAD using both simulated and experimental data.

Locus assembly

AftrRAD performs an initial filter for low-quality reads by removing those that contain: (i). Any base with Phred scores below a configurable value (default 20) (ii). Restriction enzyme recognition sequences more than 1 base different than the correct sequence, (iii). Long strings of homopolymers (default length of 15) or (iv). The beginning of the P2 adaptor sequence. Reads passing this initial screen are further filtered based on the nonzero mean read depth (coverage) across all individuals for each remaining unique read in the data set. Specifically, unique reads with mean depths below a configurable threshold are eliminated as likely error reads (Fig. S2, Supporting information). Because we evaluate read depths among, rather than within individuals, using a nonzero mean read depth in place of overall mean depths helps ensure that true rare alleles (i.e. singletons) are retained in the data set.

Following these quality control steps, barcodes and restriction sites are removed from the retained reads, and all pairwise comparisons of unique reads are evaluated to identify potentially allelic pairs of sequences. These comparisons are initially performed by searching for short, exactly matching substrings of sequence between each pair of reads, and pairs with matching substrings are flagged as potentially allelic and subsequently aligned using ACANA (Huang *et al.* 2006). The resulting pairwise alignments must meet two criteria to be considered allelic. First, they cannot contain more indels than a configurable value (default three). Second, the percentage identity across the aligned region (ignoring indels) must exceed a configurable threshold (default 90%). The aligned allelic pairs are then assembled into candidate loci by merging all pairs that share a specific read (Fig. S3, Supporting information). Finally, a secondary, global alignment is performed on the set of alleles at each candidate locus using MAFFT (Katoh & Standley 2013).

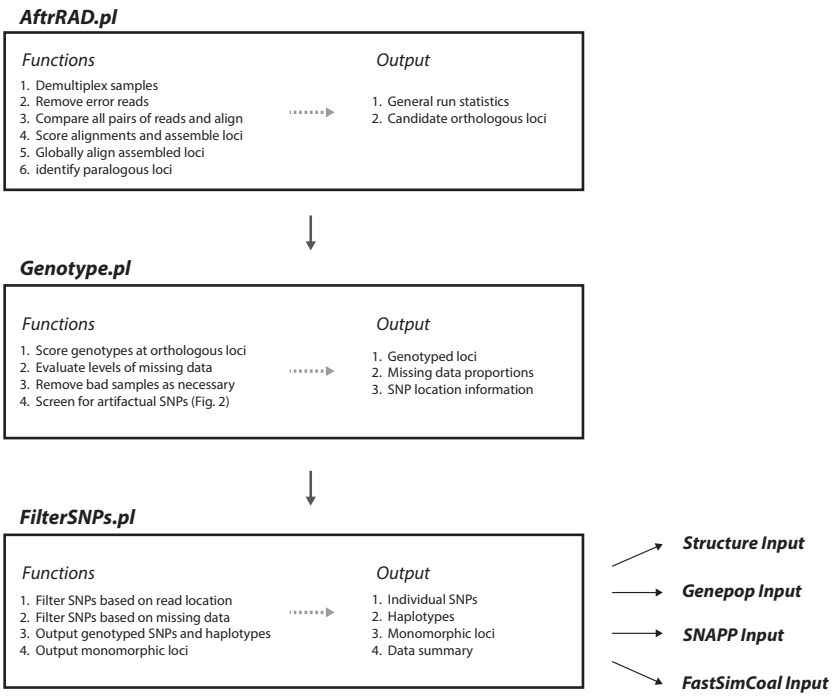

Fig. 1 Brief overview of the AftrRAD pipeline, with important functions and output from each of the component scripts listed. The program provides options for formatting data from the FilterSNPs.pl script into a variety of input files for downstream analyses.

Genotyping

After candidate loci are identified, read counts are obtained for each individual at each locus/allele. For loci identified with >2 alleles in the data set, the 3rd highest read count is obtained for each individual, and loci with values that exceed a configurable threshold (default five) in at least one individual in the data set are flagged as paralogous and removed from further analyses. For loci with two alleles, read counts are screened for evidence of excess heterozygosity, which is another indicator of paralogy (i.e. a duplicated locus that has become fixed for alternative alleles). Loci exhibiting this pattern are also removed as paralogous. Remaining (nonparalogous) loci are genotyped in each individual by applying a binomial test to reduce the chance that any error reads which pass through the upstream quality filtering described above result in an incorrect heterozygote call. Genotyping is only performed at loci for which the sum of read counts exceeds a configurable threshold (default 10). Following the genotyping, polymorphic sites are identified at each locus, and the genotyped SNPs (or haplotypes, in cases where multiple SNPs occur at a single locus) are output in two-dimensional matrices for further formatting and analyses. As part of the genotyping process, AftrRAD plots the frequency of SNPs at each position along the sequence reads and provides the option to omit SNPs after a specified read position. This allows for elimination of artefactual SNPs that accumulate as a result of indel variation towards the ends of reads (Fig. 2).

Validation of AftrRAD

Simulated data—Three independent RADseq data sets that differed from each other by the frequency of indel mutations were simulated using the simRRL.py script in the PYRAD package (Eaton 2014). For each data set, 1000 loci were simulated for five individuals from each of three taxa. One of the three simulations introduced only SNP variation in the data set (no indels), while the other two introduced indels as 5% and 10% of the mutations, respectively. These simulated data sets were each analysed (assembled and genotyped) using AftrRAD, STACKS and PYRAD. To the extent possible, parameter settings were chosen to be equivalent across the three programs, although not all parameters have exact analogues across all three (i.e. a minimum threshold Phred score of 20 was set for all three programs; however, STACKS uses a sliding window average to calculate this value, while AftrRAD and PYRAD compare the actual value at each nucleotide site). Specific parameter values used are given in Table S1 (Supporting information). Performance of the three programs was assessed by evaluating and comparing the identity of loci genotyped in all individuals by each of the programs. For loci that were not genotyped in all individuals by all three programs, attempts were made to determine the causes for the discrepancies. Perl scripts used to aid in these comparisons are available on Dryad.

Experimental data—Data sets representing two taxa (eastern massasauga rattlesnake, *Sistrurus catenatus*, and

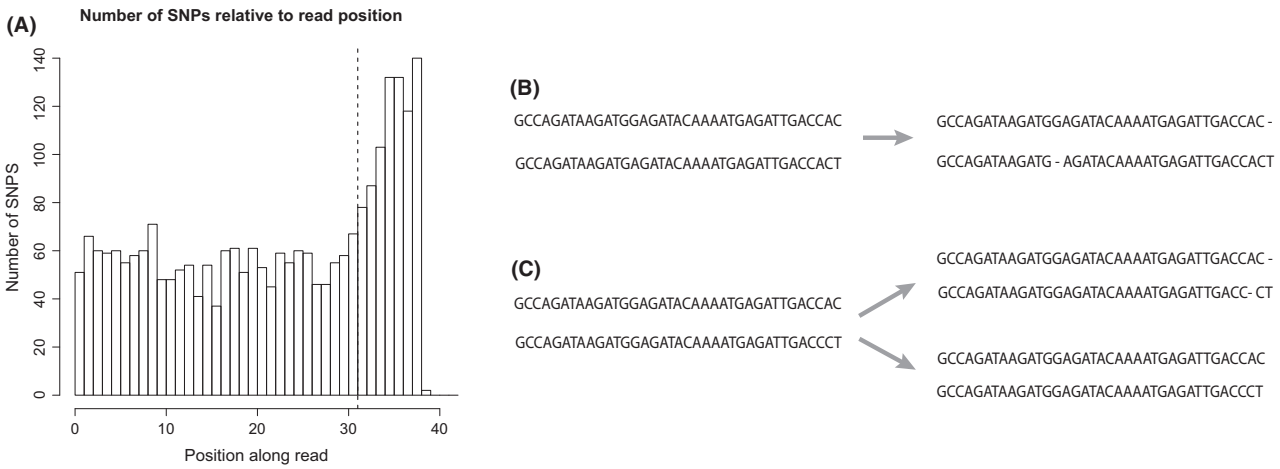

Fig. 2 Example of the build-up of artifactual SNPs toward the ends of reads. (A): Frequency of SNPs relative to read length in the rattlesnake dataset. AftrRAD provides the option to omit SNPs after a chosen position in the read based on this plot. The dotted line represents a position after which SNPs could be omitted due to a build-up of artifacts. (B, C): Example of why artifactual SNPs accumulate towards the ends of reads. In (B), the two alleles (top and bottom sequences left of arrow) are aligned without ambiguity. This is because the indel occurs early in the sequence read, causing a much higher alignment score when accounting for the indel. In (C), the two alleles are the same as in (B), with the exception that the indel occurs near the end of the reads. In this case, the alignment program has two options that may result in similar alignment scores, and will often choose the option associated with the lower arrow, as SNPs are generally penalized less than indels. There is no way to distinguish which of these options is correct without additional sequence information.

mallard, *Anas platyrhynchos*) were generated using a double-digest RADseq method (Peterson *et al.* 2012) and sequenced as 50-bp reads on an Illumina HiSeq 2000. Each of these experimental data sets contained 15 barcoded samples, and analyses of each of these data sets were similar to those described above for the simulated data.

Computational efficiency and sensitivity to locus read depths

To compare the computational time required for the programs, all analyses were performed on a PowerEdge T410 server with a Linux operating system, with PYRAD set to use up to 6 processors in each run. AftrRAD runs were repeated on a MacBook Pro with a 2.9 GHz Intel Core I7 processor to demonstrate differences in computational efficiency across platforms for this program. We also analysed a third data set with each of the programs in which 90 additional samples were added to the 15 *S. catenatus* samples above. This allowed for an evaluation of the effect of sample size on run times for the programs.

Finally, we evaluated the effect of diminishing levels of read depths in data sets evaluated with AftrRAD by repeating analyses with progressively smaller subsets of the original rattlesnake data. For each analysis, we recorded the number of alignments performed in the analysis (generally the most computational-demanding part of the analysis), the number of polymorphic, monomorphic and total loci identified and scored in all samples and the estimated heterozygosity values for both polymorphic loci and total loci. For each of the four data sets (the full data set and the three progressively smaller subsets), we performed multiple (10) AftrRAD runs. In each of these 10 runs, the minDepth parameter was set to a value between one and ten to assess the sensitivity of the program to this parameter.

Results/Discussion

Simulated data sets

For the simplest simulated data set, which included no indel variation, 995 of the 1000 loci were scored similarly by all three programs (Fig. 3A). The proportion of loci shared by all three decreased with increasing frequency of indels, down to 919 of the 1000 loci in the data set containing the greatest number of indels (Fig. 3). In most cases, loci not scored similarly by all three programs were shared between AftrRAD and PYRAD and contained indel variation. Differences between PYRAD and AftrRAD were attributable, at least in part, to either large indels that exceeded the configurable 'numIndels' parameter value set for the AftrRAD runs or to infrequent oversplitting of loci that likely results from the heuristic search methods used during the assembly process in AftrRAD. In general, the programs performed well with the simulated data, as all correctly assembled >95% of the loci in each data set, with the single exception that STACKS

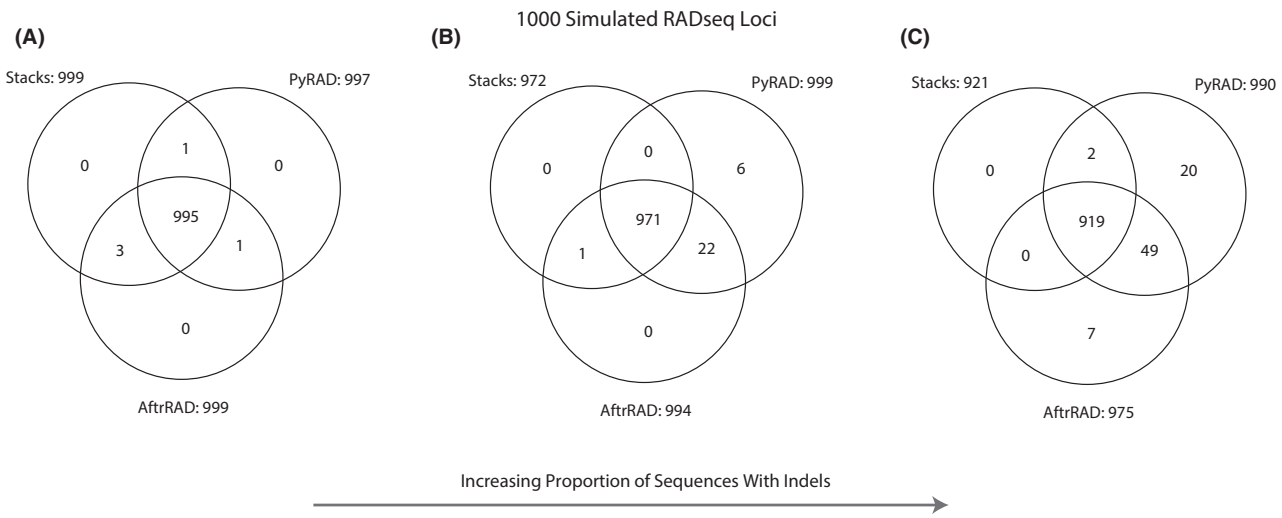

Fig. 3 Venn diagrams representing 1000 RADseq loci simulated under each of three simulation conditions which differed by the proportion of indel mutations (proportion of total mutations 0%, 5%, and 10% for A–C, respectively). Each simulated dataset was analyzed with STACKS, PYRAD, and AftrRAD, and the loci identified and genotyped by each of the three programs was compared. The proportion of loci identified similarly across all three programs ranged from 91.9% in the dataset with the greatest number of indels (C) to 99.5% in the dataset with no indel variation (A). Numbers outside the circles represent the total number of loci identified and genotyped in all samples in each of the programs.

correctly identified ~92% of the loci in the data set with the greatest level of indel variation. However, some of the significant challenges regarding assembly of orthologous loci from RADseq data arise from factors such as sequence error, variation in read counts among loci and also alleles within loci and paralogy, which can all be difficult to simulate in a way that reflects real data. As a result, we further assessed performance on experimental data sets to better evaluate AftrRAD under more realistic scenarios.

Experimental data sets

As with the simulated data, we used STACKS, PYRAD and AftrRAD to analyse RADseq data sets from massasauga rattlesnakes and mallards. Of the 3719 loci scored in all rattlesnake samples by at least one of the three programs, 3085 (83.0%) were identified similarly across all three programs (Fig. 4A). Likewise, in the mallard data set, of the 3596 loci identified and scored in all samples by least one of the three programs, 2920 (81.2%) were identified similarly across all three programs (Fig. 4B).

Some general patterns emerged upon evaluating the loci not scored in all three programs. First, as expected, loci shared between PYRAD and AftrRAD, but not identified by STACKS, almost always contained at least one indel, causing STACKS to incorrectly split the alleles into multiple loci. Of the loci shared by PYRAD and STACKS, but not scored in AftrRAD, a large proportion were identified as paralogous by AftrRAD (i.e. 63/78 in the rattlesnake data set and 35/74 in the mallard data set). In general, it appears that the methods used by AftrRAD to identify paralogous loci are more conservative than those in the other two programs (AftrRAD tends to identify a larger number of loci as paralogous than the other two). There was a notable difference between the two data sets in the number of loci shared by STACKS and AftrRAD, to the exclusion of PYRAD. Specifically, for the rattlesnake data set, 1.6% of the total loci (60/3719) fit this pattern, in contrast to 7.2% (259/3596) of the mallard loci. A small number of these were identified as paralogous by PYRAD, but in many cases, the reason for their absence in PYRAD was not readily apparent. One possibility may be related to the type of scenario demonstrated in Fig. S3 (Supporting information), in which two alleles at a locus exhibit a level of identity that exceeds the minimum threshold for assigning those alleles to the same locus, while a third allele achieves this threshold level of identity with just one of those other two alleles. PYRAD may have oversplit some loci with this type of pattern because it uses a single 'reference' sequence to represent the locus during assembly, and any alleles that do not meet the threshold level of identity in comparison to this reference allele may get incorrectly assigned to a separate locus. This explanation is supported by the fact that the proportion of loci shared by STACKS and AftrRAD was higher in the mallard data set, which has a greater level of polymorphism than the rattlesnake data set, and is therefore more likely to have a greater number of loci that exhibit this pattern of variability (mean heterozygosity values across all loci were 0.186 and 0.049 for the two data sets, respectively).

In both experimental data sets, STACKS identified the majority of loci scored by only one of the three programs. Two main factors account for this. First, a number of these loci were identified as paralogous in at least one of

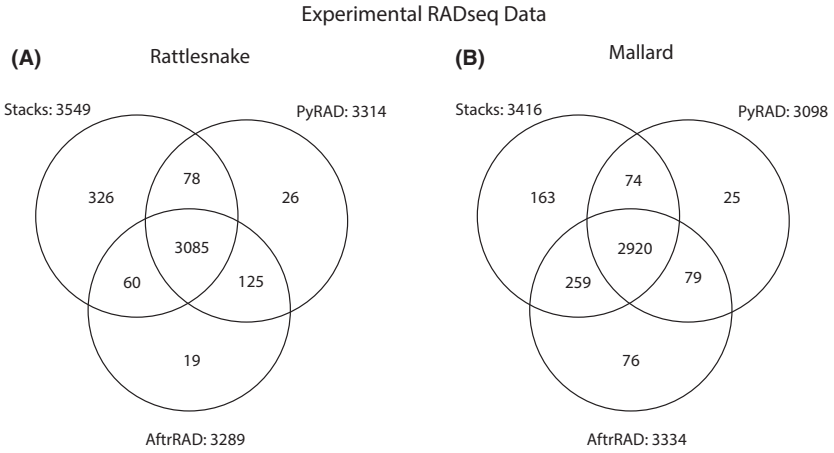

Fig. 4 Venn diagrams representing RADseq loci identified and genotyped by STACKS, PYRAD, and AftrRAD from eastern massasauga rattlesnake (A) and mallard (B). Factors that accounted for differences among the programs include indel variation, differential identification and removal of paralogous loci, and also minor genotyping differences, in which the locus was assembled in a similar way across all three programs, but was not genotyped in all samples by at least one program. Numbers outside the circles represent the total number of loci identified and genotyped in all samples in each of the programs.

the other two programs (27.9% and 42.9% of the unique STACKS loci for the rattlesnake and mallard data sets, respectively), suggesting that PYRAD and AftrRAD identify and remove paralogous loci more efficiently than STACKS. However, it is important to note that additional downstream analyses, such as tests of Hardy–Weinberg equilibrium, can allow for identification and removal of some of these paralogous loci from the STACKS data sets. Of the remaining loci unique to STACKS, most were assembled in a similar way across all three programs and resulted from minor genotyping differences that arose from slight differences in the configurable parameter values set for each run. In most of these cases, STACKS genotyped a locus in all 15 individuals, while that locus was genotyped in most (i.e. 14/15), but not all of the samples in the other two programs. As a result, these loci were not identified in the set of those shared among all three programs.

Computational efficiency

For both data sets, AftrRAD runs took longer to complete than STACKS runs, but were faster than PYRAD runs (Table 1). The difference in speed compared to STACKS is due to the alignment process, which is computationally demanding, and is not a part of STACKS analyses. The difference in speed between AftrRAD and PYRAD, which both perform alignments, is substantial and becomes more pronounced as the number of individuals in the data set increases. This property of AftrRAD may be especially relevant for large-scale population genomics studies, which are often likely to include sample sizes on the order of hundreds of individuals. It is important to note that PYRAD takes advantage of parallel computing to decrease run time. PYRAD analyses in this study were performed using 6 processors, and the use of larger numbers of processors would probably reduce the differences in run times between PYRAD and AftrRAD. On the other hand, the differences between AftrRAD and PYRAD would be exaggerated if PYRAD were run with a single processor, similar to AftrRAD. This suggests that when multicore computing resources are not available (i.e. such as on a personal laptop), AftrRAD provides a good alternative to currently available programs.

Table 1 Run times (min) required for STACKS, PYRAD and AftrRAD to complete analysis of rattlesnake data sets containing 15 and 105 samples, respectively. All analyses were performed on a Linux operating system, as described in the text, with the exception that AftrRAD runs were repeated on a MacBook Pro laptop to demonstrate differences in speed for this program across platforms

	15 Samples	105 Samples
Stacks	6	37
AftrRAD (Mac)	25	110
AftrRAD	80	335
PyRAD	308	1465

The efficiency of AftrRAD is due in large part to efficient removal of error reads prior to alignment. This is achieved in two steps. The first is similar to methods in STACKS and PYRAD and removes error that is represented by low-quality (Phred) scores indicative of low-confidence sequencing calls. However, a second important source of error that occurs in short-read sequencing data sets is due to PCR artefacts that arise during library construction. These artefacts are likely to be read with high confidence by the sequencer and as a result are not removed by the Phred score filter. Due to the exponential nature of PCR amplification, and because these artefactual sequences do not exist at the beginning of the PCR, the relative frequency of these PCR artefacts is generally very low compared to true reads. In most cases, the artefacts occur as a single sequencing read in a single individual in the data set (see Fig S2, Supporting information, and Fig. 5). These reads are effectively eliminated by STACKS and PYRAD in the process of genotyping, as the relative frequency of true reads to artefacts at the locus will be highly skewed towards the true reads. However, these reads are included in the assembly process by these programs. AftrRAD eliminates these reads prior to assembly by applying a threshold for the mean nonzero read count for each read across all samples in the data set (configurable minDepth parameter). The effect of this threshold is demonstrated in Fig. 5, and is most apparent in Fig. 5A, which represents analysis of the full rattlesnake data set. In this case, application of increasing threshold values significantly reduces the number of alignments necessary, but has little effect on the number of loci identified, or the observed heterozygosity levels estimated with the data.

Sensitivity to locus read depths and sequencing quality

The data sets evaluated in this study had relatively high sequence coverage and quality scores. For the rattlesnake data set, the mean read depth per locus was 80.3 and the median was 43, and for the mallard data set, the mean depth was 73.6 and the median was 46. Because these values are higher than those in many short-read sequencing data sets currently being generated, we evaluated subsets of the rattlesnake data to evaluate the effects of diminishing read depths on the analyses.

Using estimates of heterozygosity as a metric, there were few differences between the full data set (mean depths per locus ranged from 32.2 to 101.3, depending on the minDepth of reads threshold used) and the second-largest data set, for which mean depths ranged from

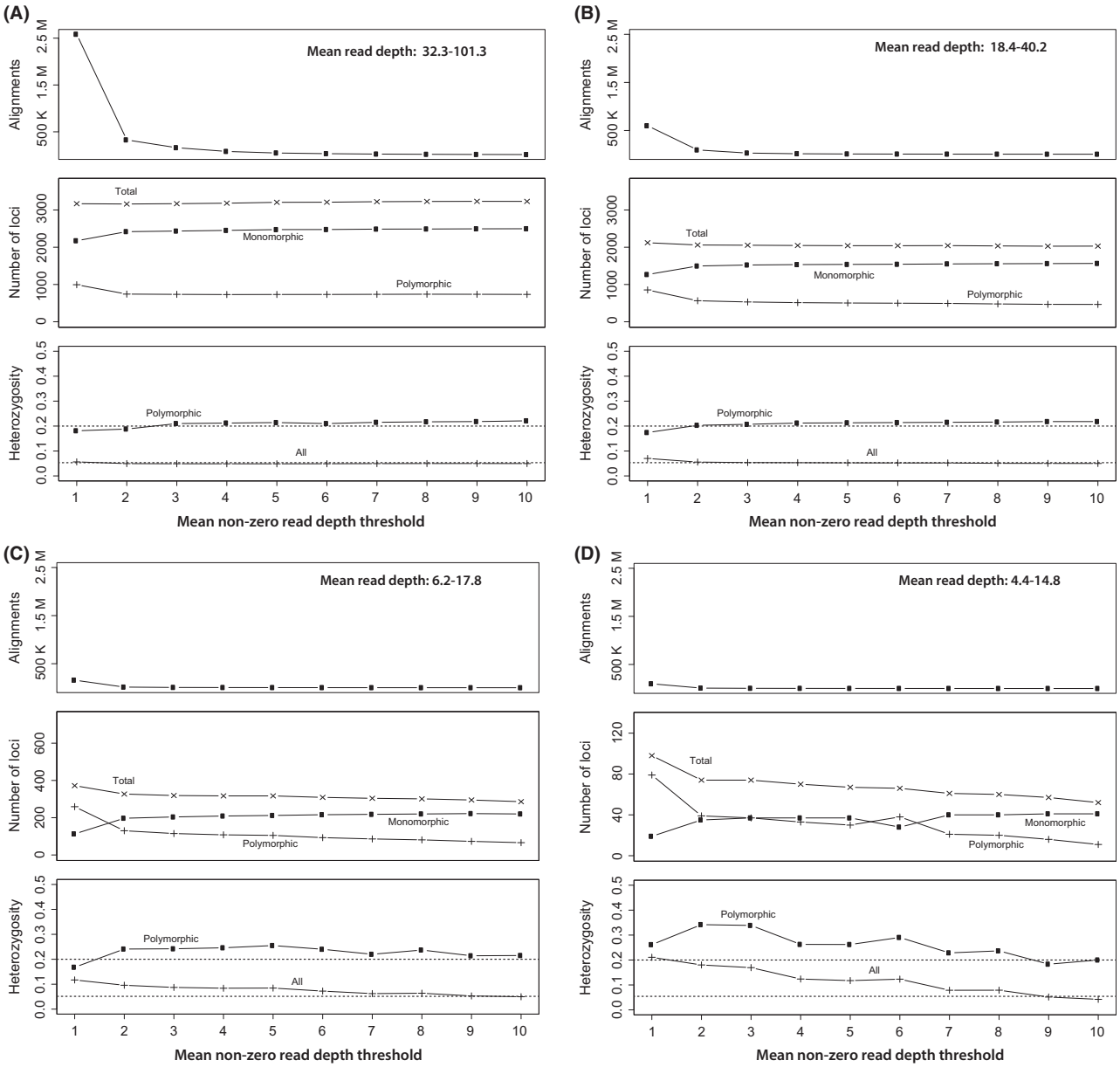

Fig. 5 Effect of threshold for the mean non-zero read depth parameter (minDepth) in AftrRAD in the context of overall read depth for the dataset. The four sets of analyses (A–D) represent progressively smaller subsets of the original dataset (A is based on the entire dataset, while the analysis in D includes only approximately 5% of the total reads in A). The mean read depth values in the top right represent the range for the mean number of reads at each genotyped locus across the minDepth parameter values analyzed (1–10). Applying the minDepth parameter to remove error reads, which generally occur as a single read in a single sample, significantly reduces the number of alignments that must be performed, (see especially 5A, top). These reads are traditionally included in the assembly process of other programs and later removed during the genotype calling stage of analyses. The number of loci (middle plots, note change in scale) and average observed heterozygosity (lower plots; dotted lines represent values of 0.05 and 0.2, respectively) are robust to tested values of the minDepth parameter when read depths are relatively high (A and B). Two different measures of heterozygosity are reported, one of which is based only on polymorphic loci, while the other is calculated from all loci in the dataset.

18.4 to 40.2 (Fig. 5A,B). Heterozygosity estimates were similar between these two and, in both cases, were generally unaffected by the minDepth parameter, with the exception of minDepth values of one. When the minDepth was set to one, heterozygosity values based only on polymorphic loci decreased slightly, while those based on all loci (polymorphic and monomorphic) increased. These patterns are expected if error reads are incorrectly incorporated as true alleles in the data set. Heterozygosity estimates vary much more across min-

Depth values in the two data sets with lower coverages (Fig. 5C,D; mean depths per locus ranging from 6.2 to 17.8 and 4.4 to 14.8, respectively). In these data sets, the estimated heterozygosity values approach those in the two data sets with deeper coverage as the minDepth value increases. These results suggest that when average read depths per locus are moderate to high (i.e. >10), results from AftrRAD are robust to variation in the minDepth parameter. In cases of low average read depths per locus (i.e. <5–10), the data presented here suggest that analyses may benefit from the use of higher thresholds of the minDepth of reads parameter, combined with limiting any analyses to loci scored in all individuals. This will reduce the total number of loci available, but will decrease the probability that error reads are being confounded with true alleles at those loci. However, we generally do not recommend analysing extremely low-coverage data with AftrRAD, and any results based on low-coverage data should be interpreted with great caution.

Finally, mean Phred scores for reads in these data sets were generally >35. Because AftrRAD takes a very conservative approach to removing error reads, data sets with low-quality sequencing data may be better analysed with other methods such as STACKS or PYRAD, as these programs allow more flexibility in attempting to utilize information from reads that contain low-quality base calls.

In summary, AftrRAD is a novel pipeline that performs accurate and efficient assembly and genotyping of RADseq data. For most data sets, this program can be run efficiently without the need for muticore analyses. Because AftrRAD effectively accounts for indel variation during locus assembly, it can be applied to data sets representing a wide range of diversity (i.e. population-scale diversity through phylogeographic and potentially phylogenetic-scale diversity) and provides a useful alternative to current bioinformatic tools for analysing RADseq data sets. AftrRAD performs optimally on Mac operating systems, but can also be run (with reduced efficiency associated with the ACANA aligner program) on Linux operating systems.

Acknowledgements

We thank Steven Hara and Crisley de Camargo for assistance in development and evaluation of AftrRAD, and Jordan Satler, Jason Macrander, Kuan-Yu (Alex) Chen, Diana Amazonas, Oleksandr Zinenko, Jenn Hellmann, David Salazar-Valenzuela and Maria Pham for testing and providing feedback on early versions of the program. We also thank Bryan Carstens for helpful suggestions regarding the study and manuscript. This research was funded by the Ohio Biodiversity Conservation Partnership, a collaboration between Ohio State University and the Ohio Division of Wildlife and Ohio State University.

References

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

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, **29**, 1917–1932.

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

Chong Z, Ruan J, Wu C-I (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. *Bioinformatics*, **28**, 2732–2737.

Davey JL, Blaxter MW (2011) RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Eaton DAR (2014) PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogozo V, Rockman MV), pp. 157–178. Humana Press, New York.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLOS Genetics*, **9**, e1003905.

Gamble T, Zarkower D (2014) Identification of sex-specific molecular markers using restriction site-associated DNA sequencing. *Molecular Ecology Resources*, **14**, 902–913.

Huang W, Umbach DM, Li L (2006) Accurate anchoring alignment of divergent sequences. *Bioinformatics*, **22**, 29–34.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Leache AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using Genome-wide SNP data. *Systematic Biology*, **63**, 534–542.

Martin CH, Feinstein LC (2014) Novel trophic niches drive variable progress towards ecological speciation within an adaptive radiation of pupfishes. *Molecular Ecology*, **23**, 1846–1862.

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.

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.

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

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Reitzel AM, Herrera S, Layden MJ, Marindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.

Viricel A, Pante E, Dabin W, Simon-Bouhet B (2014) Applicability of RAD-tag genotyping for interfamilial comparison: empirical data from two cetaceans. *Molecular Ecology Resources*, **14**, 597–605.

M.G.S. and A.C.F. developed AftrRAD and performed analyses of simulated and empirical datasets. M.G.S. wrote the manuscript. H.L.G. and A.C.F. provided samples for the study and assisted in preparation of the manuscript.

Data accessibility

Short-read sequence data sets used in the comparisons of experimental data (mallard and eastern massasauga rattlesnake), barcode information for these data sets and perl scripts used as part of the analyses are available on Dryad (doi:10.5061/dryad.sn034).

AftrRAD is available for download at u.osu.edu/sovic.1/downloads and at https://github.com/mikesovic/AftrRAD.git.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Schematic representation of the steps taken by the AftrRAD pipeline to assemble loci and genotype individuals, and the associated files created during the run.

Figure S2 Representative data from the *S. catenatus* dataset analyzed in this study demonstrating the method used in AftrRAD to remove error reads from the data prior to assembly of reads into loci.

Figure S3 Example sequence reads demonstrating the method used to assemble loci in AftrRAD.

Table S1 Parameters set in each program for comparisons of AftrRAD with Stacks and PyRAD.