MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 28–41 doi: 10.1111/1755-0998.12291

Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference

A. MASTRETTA-YANES,* N. ARRIGO,† N. ALVAREZ,† T. H. JORGENSEN,‡ D. PIÑERO§ and B. C. EMERSON*¶

*Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, 14 Norwich NR4 7TJ, UK, †Department of Ecology and Evolution, Biophore Building, University of Lausanne, 1015 Lausanne, Switzerland, ‡Department of Bioscience, Aarhus University, Universitets Parken, 8000 Aarhus C, Aarhus, Denmark, §Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado postal 70-275, Mexico, DF 04510, Mexico, ¶Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), C/Astrofísico Francisco Sánchez 3, La Laguna, Tenerife, Canary Islands 38206, Spain

Abstract

Restriction site-associated DNA sequencing (RADseq) provides researchers with the ability to record genetic polymorphism across thousands of loci for nonmodel organisms, potentially revolutionizing the field of molecular ecology. However, as with other genotyping methods, RADseq is prone to a number of sources of error that may have consequential effects for population genetic inferences, and these have received only limited attention in terms of the estimation and reporting of genotyping error rates. Here we use individual sample replicates, under the expectation of identical genotypes, to quantify genotyping error in the absence of a reference genome. We then use sample replicates to (i) optimize *de novo* assembly parameters within the program *Stacks*, by minimizing error and maximizing the retrieval of informative loci; and (ii) quantify error rates for loci, alleles and single-nucleotide polymorphisms. As an empirical example, we use a double-digest RAD data set of a nonmodel plant species, *Berberis alpina*, collected from high-altitude mountains in Mexico.

Keywords: *de novo* assembly, error rate, optimization, RAD sequencing, replicates, *Stacks*

Received 12 December 2013; revision received 27 May 2014; accepted 4 June 2014

Introduction

Restriction site-associated DNA sequencing (RADseq) is a genotyping method that allows subsampling of a genome at putatively homologous locations across many individuals to identify and type single-nucleotide polymorphisms (SNPs) in short DNA sequences. The method was created by Baird *et al.* (2008) and has been subsequently developed into a family of related approaches (also called genotyping-by-sequencing and reviewed by Davey *et al.* 2011). These approaches can be applied to nonmodel organisms to potentially sequence thousands of loci for hundreds of individuals, rapidly and at low cost, regardless of genome size and previous genomic knowledge. As a result, RADseq is increasingly being used across the spectrum of evolutionary analysis, ranging from phylogenetic relationships within a genus (e.g. Jones *et al.* 2013), to genome-wide association studies to identify regions under selection (e.g. Parchman *et al.* 2012; Richards *et al.* 2013), through to ecological and conservation studies (Narum *et al.* 2013).

Although the validity of RADseq data has been demonstrated, genotyping errors are to be expected. RADseq is prone to both technical and human sources of error (Table 1), similar to those identified for traditional molecular markers (e.g. Bonin *et al.* 2004) and for whole-genome sequencing (Pool *et al.* 2010; Gompert & Buerkle 2011). Wet laboratory procedures, parallel sequencing and species-specific genome properties also contribute to error in several ways (Table 1), leading to variance in: (i) the total number of reads per individual; (ii) the number of loci represented in each individual; (iii) read count per locus; and (iv) the read counts of alternative alleles at polymorphic loci (Hohenlohe *et al.* 2012). For example, differences in

Correspondence: Alicia Mastretta-Yanes, Fax: +34 922 260135; E-mail: A.Yanes@uea.ac.uk

Table 1 Potential causes of genotyping error for restriction site-associated DNA sequencing data

Source	Reason	Reference
Technical and human error		
Technical	Errors related inversely to the quality of reagents and equipment, and to the organization of the laboratory in different rooms to avoid contamination	Bonin *et al.* (2004)
Human	Sample mislabelling, sample contamination, pipetting error and error during DNA concentration measurements	Bonin *et al.* (2004)
Wet laboratory		
Enzyme sensitivity to DNA quality and quantity	Digestion and PCR efficiency may be uneven among samples, which can result in the underrepresentation of some restriction fragments	Bonin *et al.* (2004)
Pooling concentration	Samples with higher concentration can be overrepresented in the sequencing output if they are not pooled in equimolar amounts	Baird *et al.* (2008); Peterson *et al.* (2012)
PCR error	PCR error may get further amplified and can appear in multiple reads resembling an alternative allele at a locus. PCR error may differ among samples depending on reaction conditions and experimental design	Hohenlohe *et al.* (2012)
PCR bias	PCR amplification success may be variable across different alleles or barcodes, biasing their representation. Differences in amplification success lead to variation of coverage among loci and individuals, potentially resulting in allelic dropout, nonrepresentation of some loci or PCR duplicates	Bonin *et al.* (2004); Peterson *et al.* (2012); Hohenlohe *et al.* (2012)
Size selection (double digest)	Different fragments may be selected if more than one excision is performed. Imprecise size selection can include fragments of lengths relatively distant from the size-selection target mean	Peterson *et al.* (2012)
Exposure to UV light	Can produce fragmentation (that could lead to locus/allele dropouts) and mutation of DNA strands (that introduces nonbiological variation)	Grundemann & Schomig (1996)
Next-generation sequencing (NGS)		
Sequencing error	NGS introduces sequencing error (0.1–1.0% per nucleotide) that can vary across samples, restriction site-associated DNA (RAD) sites and positions in the reads for each site	Hohenlohe *et al.* (2012), Meacham *et al.* (2011); Nielsen *et al.* (2011); Loman *et al.* (2012)
Sequencing sampling	The sampling process of a heterogeneous library inherent in NGS introduces sampling variation in the number of reads observed across RAD sites as well as between alleles at a single site	Hohenlohe *et al.* (2012)
Barcode error	PCR or sequencing errors at the DNA-tag of a fragment can reduce the number of reads obtained for it	Hohenlohe *et al.* (2012)
Genome intrinsic		
GC content	At large numbers of PCR cycles, RAD loci with high GC content are sequenced at higher depths compared to RAD loci with low GC content. But at the same time, high GC content loci could be undersequenced if too few PCR cycles are performed. GC bias contributes to PCR duplicates	Davey *et al.* (2013)
Restriction site variation	Variation in the restriction site within a locus will result in allelic dropout	Davey *et al.* (2013); Gautier *et al.* (2013b)
DNA methylation	For some restriction enzymes, digestion is impaired or blocked by methylated DNA. The same gene may or may not be methylated in different individuals or tissues	Roberts *et al.* (2010)
Bioinformatic		
Variation in coverage	Coverage is an important filter to distinguish real variation from sequencing errors, repetitive regions and duplicates. But if there is coverage heterogeneity among samples and alleles, or if the general coverage is low, setting the filters with minimal coverage values too high can lead to allele dropout. Setting it too low, however, can lead to incorrect single-nucleotide polymorphism (SNP) calls	Hohenlohe *et al.* (2012); Davey *et al.* (2013); Catchen *et al.* (2013)
PCR duplicates	PCR duplicates occur when more than one copy of the same original DNA molecule attaches to different beads/cells during sequencing. This can result in high coverage of PCR error variation, or it can produce heterogeneous coverage distribution due to GC and PCR bias	Davey *et al.* (2013)
Fragment length	Alleles will drop out as restriction fragment length decreases because RAD loci from short restriction fragments have low read depths. The efficacy of different bioinformatics tools at dealing with this varies	Davey *et al.* (2013)

Table 1 (Continued)

Source	Reason	Reference
Paralogs and repetitive regions	Paralogous and repetitive regions with similar sequences can be erroneously merged together as a single locus	Hohenlohe *et al.* (2012); Dou *et al.* (2012)
Presence of indels	*Stacks* and *RADtools* are unable to handle indels; therefore, indel-containing loci are not clustered together, while they can be recovered by *RaPiD* and *pyRAD*	Peterson *et al.* (2012); Davey *et al.* (2013)
Mapping using a reference genome	Mapping of alleles that are different from the reference genome is less probable than for a reference-matching allele, causing a bias in allele frequency towards the allele found in the reference sequence. It may additionally reduce the number of SNPs discovered and bias estimates of nucleotide diversity towards smaller values	Pool *et al.* (2010)

amplification success during the PCR step may lead to variation in the depth of coverage among loci and individuals, potentially causing locus or allelic dropout (Appendix S1, Supporting Information).

The consequences of error, and statistical methods to account for it, have been widely discussed for other molecular makers, from AFLPs and microsatellites (Bonin *et al.* 2004; Pompanon *et al.* 2005; Price & Casler 2012) to whole-genome sequence data (Pool *et al.* 2010; Gompert & Buerkle 2011; Nielsen *et al.* 2011). Error can lead to incorrect biological conclusions, such as an artificial excess of homozygotes (Taberlet *et al.* 1996), false departure from Hardy–Weinberg equilibrium (Xu *et al.* 2002), overestimation of inbreeding (Gomes *et al.* 1999), unreliable inferences about population structure (Miller *et al.* 2002) and incorrectly inferring demographic expansion from the confounding influence of low frequency error-derived SNPs (Pool *et al.* 2010). These potentially inaccurate inferences can be mitigated and accounted for if error rates are reported (Bonin *et al.* 2004; Pompanon *et al.* 2005; Pool *et al.* 2010; Davey *et al.* 2011) or incorporated into data analysis (Gompert & Buerkle 2011; Nielsen *et al.* 2011; Gautier *et al.* 2013a). However, the quantification and reporting of such errors has been largely overlooked by most recent RAD studies.

In addition to errors introduced during wet laboratory and sequencing procedures, errors can arise during the bioinformatic processing of RADseq data (Table 1). For instance, when RAD sequences are assembled into loci and alleles, often using distance-based criteria, genotyping results will vary according to the algorithm used (Davey *et al.* 2013; note that we refer to a *locus* as a short DNA sequence produced by clustering together unique RAD *alleles*; in turn, alleles differ from each other by small number of *SNPs*). Several assembly and genotyping tools for RADseq data have recently been released, such as *RaPiD* (Willing *et al.* 2011), *RADtools* (Baxter *et al.* 2011), graph-based distance clustering approaches (Peterson *et al.* 2012), *Stacks* (Catchen *et al.* 2011, 2013), *Rainbow* (Chong *et al.* 2012) and *pyRAD* (Eaton 2014). Within a given tool, it is to be expected that different parameters and settings will result in different levels of assembly-related error. For instance, *Stacks* relies on a set of core parameters (summarized in Table 2) to first create sets of short-read sequences that match (i.e. stacks) within a given threshold of nucleotide differences, and to then curate and assemble these into genotyped loci within individuals. Catchen *et al.* (2013) have explored how variation in: (i) the minimum number of raw reads required to form a stack (*-m*); (ii) the number of mismatches allowed between stacks (*-M*); (iii) the maximum number of stacks allowed per single locus (*--max_locus_stacks*); and (iv) modulating the assumed rate of sequencing error (using a bounded SNP calling model) affects the recovery of RAD loci. To do so, they ran *Stacks de novo* pipeline using different parameter values and compared results to expectations from a reference genome. They concluded that the optimal values for these parameters will depend upon the polymorphism of the genome being analysed, the amount of sequencing error and the depth of sequencing performed. The authors recommended testing a range of parameter values to optimize the analysis of each RADseq data set. However, their strategy to assess whether true or erroneous loci were assembled involved a reference genome; therefore, alternative criteria are needed for taxa where a reference genome is not available.

Here we show how replicates can be used not only to estimate error rates, but also to optimize the *de novo* assembly of RADseq data. The central premise is that DNA replicates derived from the same DNA should have the same genotype. Thus, after running any *de novo* assembly pipeline with different combinations of parameters, one can evaluate which settings produce both a high number of loci and low differences between replicate pairs (Appendix S1, Supporting Information). Optimizing *de novo* assembly is particularly important for low-coverage data sets, because it facilitates the recovery of more loci than could otherwise be reliably achieved.

To demonstrate how replicates can be used to estimate error rates and optimise *de novo* assembly, we use

Table 2 Role of *Stacks* core parameters in the assembly of loci and potential sources of genotyping error

Parameter	How it affects assembly and genotyping error*
Minimum number of identical, raw reads required to create a stack (*-m*) default 3	Reads with convergent sequencing errors are probably to be erroneously labelled as stacks if $-m$ is too low. True alleles will not be recorded and will drop out if *-m* is too high. *-m* can decrease genotyping error by distinguishing real loci from PCR and sequencing error, but it can increase error by calling a heterozygous locus as homozygous when minimum coverage is set too high and one of the alleles is therefore excluded
Number of mismatches allowed between loci when processing a single individual (*-M*) default 2	If *-M* is too low, some real loci will not be formed, and their alleles will be treated as different loci (undermerging). If *-M* is too large, repetitive sequences and paralogs will form large nonsensical loci (overmerging)
Number of mismatches allowed between loci when building the catalogue (*-n*) default 0	For $n = 0$, there would be loci represented independently across individuals that are actually alleles of the same locus. If $n > 0$, the consensus sequence from each locus is used to attempt to merge loci. This is important for population studies where monomorphic or fixed loci may exist in different individuals. Merging fixed alleles as a single locus can increase the probability of assembling real loci and therefore decrease the allele error rate. However, erroneous loci will be created if *-n* is too high
Maximum number of stacks at a single *de novo* locus (*--max_locus_stacks*) default 3	The expectation for nonrepetitive genomic regions is that a monomorphic locus will produce a single stack because the two sequences on the two homologous chromosomes are identical and thus indistinguishable. In contrast, a polymorphic locus will produce two stacks representing alternative alleles. Confounding cases that may arise from short, sequencing error-based stacks or from repetitive sequences, where hundreds of loci in the genome may collapse to a single putative locus. *--max_locus_stacks* allows for the identification and blacklisting of confounding cases
Single-nucleotide polymorphism (SNP) calling model	In the default SNP calling model, the error parameter is allowed to vary freely, whilst in a bounded-error model, the boundary value is substituted if the maximum-likelihood value of ε exceeds a lower or upper bound. One consequence is that reducing the upper bound increases the chance a homozygous loci being called heterozygous. The SNP calling model allows the tolerance for false positive vs. false negative rates in calling genotypes to be tuned, which in turn influences the genotyping error

*Parameters explanation as in Catchen *et al*. (2013) and *Stacks* documentation, effect on genotyping error as discussed here.

double-digest RADseq (Parchman *et al*. 2012; Peterson *et al*. 2012) data generated from populations of *Berberis alpina*, a nonmodel plant species limited to high-altitude mountains in Mexico. We use the program *Stacks*, an efficient and well-documented software, that is increasingly being used by molecular ecologists, but the principle of comparing replicates could be applied to other assembly and genotyping tools for RADseq data. Optimizing RAD data assembly is important to achieve good results (Davey *et al*. 2013), and accounting for error is essential for the robustness of any individual study or meta-analysis. However, the approach presented here could be particularly useful when focal species lack previous genomic knowledge and when data sets are characterized by low coverage.

Methods

Study system and sampling

The focal species is *Berberis alpina* (Zamudio 2009), a diploid plant with a probable genome size of between 0.5 and 1.83 Gbp, based on values of related species (Rounsaville & Ranney 2010). *Berberis alpina* inhabits the Transmexican Volcanic Belt (TMVB), a biodiversity hot spot for temperate forest plant species (Myers *et al*. 2000) where the species is restricted to a few mountain tops (Fig. 1).

Seven mountains where *B. alpina* and one where *B. moranensis* (a closely related species with which *B. alpina* potentially hybridizes) occur were sampled in the TMVB and nearby areas of the Sierra Madre Oriental (SMOr) during September–October 2010 and April–May 2011 (Sampling localities: doi:10.5061/dryad.g52m3). The sampling locations for *Berberis alpina* encompass the full range of the species within the TMVB (Fig. 1). Fresh young leaves of 6–25 specimens per mountain (depending upon population sizes) were collected and kept on ice while transported to the molecular ecology laboratory within the Instituto de Ecología, Universidad Nacional Autónoma de México (UNAM). Herbarium specimens were prepared and deposited within the Herbario Nacional in Mexico City. *Berberis pallida* and *B. trifolia* collected in the TMVB in October 2012 were used as outgroups. For each sample, half the tissue was stored at −80°C at UNAM, with the remainder dried in silica gel for transport to the University of East Anglia (UEA), England, where samples were maintained at −20°C until

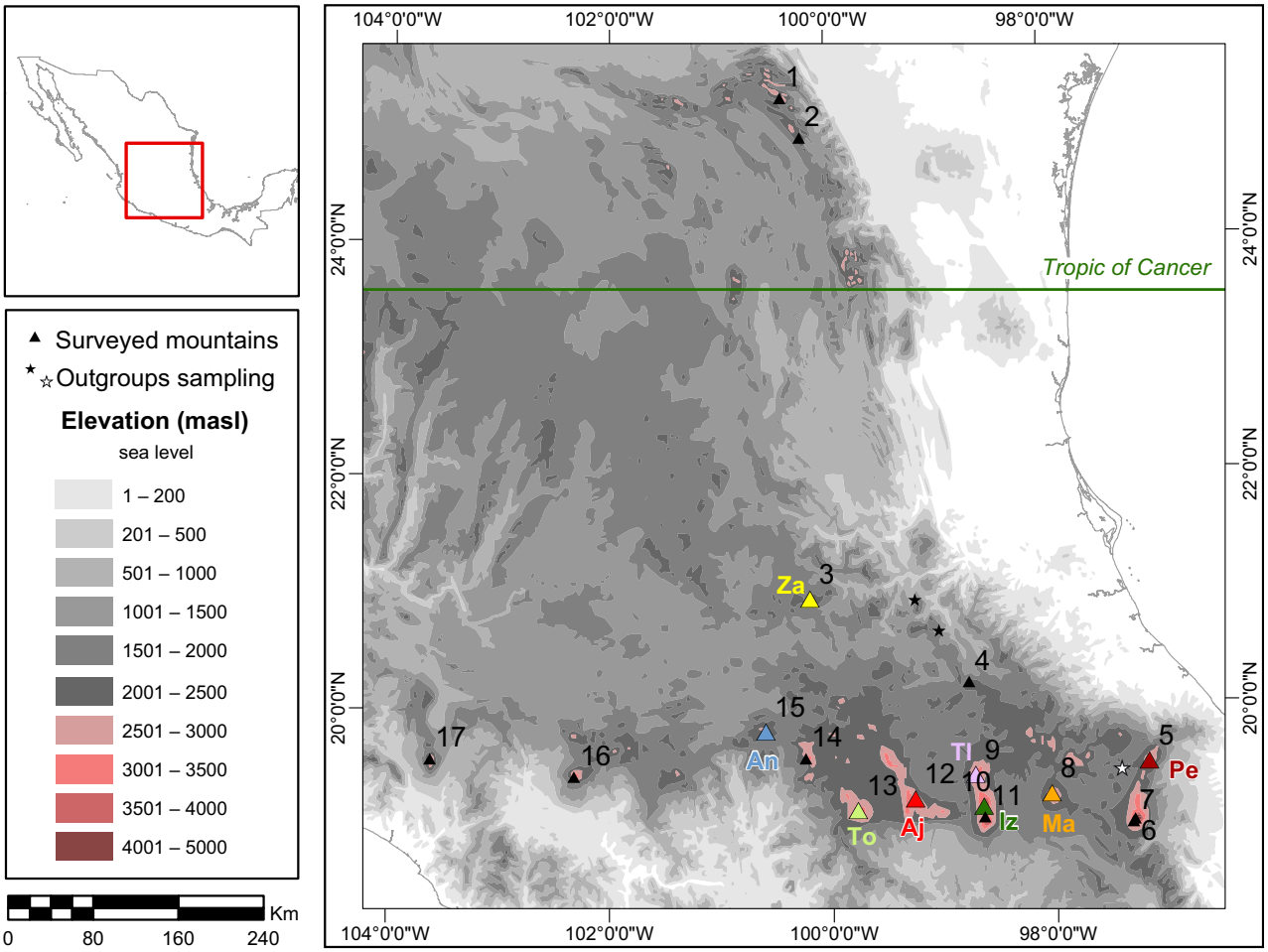

Fig. 1 Mountains surveyed for the presence of *Berberis alpina* within the Sierra Madre Oriental (1–3) and the Transmexican Volcanic Belt (4–17). *Berberis alpina* was found on El Zamorano (Za), Nevado de Toluca (To), Ajusco (Aj), Tlaloc (Tl), Iztaccihuatl (Iz), La Malinche (Ma) and Cofre de Perote (Pe). *Berberis moranensis* was found on Cerro San Andrés (An). *Berberis pallida* (black stars) and *B. trifolia* (white star) were sampled as outgroups.

extraction. Samples were collected with SEMARNAT permit No. SGPA/DGGFS/712/2896/10.

Molecular methods

DNA extractions of *Berberis alpina* and *B. moranensis* were performed at UEA using the Qiagen DNeasy Plant Mini Kit (69106). DNA extractions of outgroup samples were performed at UNAM using a CTAB method (Vázquez-Lobo 1996) with fresh tissue. Seventy-five specimens of *B. alpina* and *B. moranensis* (6–10 per sampling site) plus three samples of *B. trifolia* and three of *B. pallida* (outgroup species) were used to prepare double-digest RAD libraries (Parchman *et al.* 2012; Peterson *et al.* 2012) using the enzymes EcoRI-HF and MseI, T4 DNA Ligase and Phusion Taq from New England Biolabs. Appendix S2 (Supporting Information) contains the complete laboratory protocol, including reaction mixes and sequencing quality details. Individual DNA extracts were randomly divided into three groups (BERL1, BERL2 and BERL3), each corresponding to pools of final libraries sequenced in an independent lane. Each group was comprised of 27 *Berberis* sp. individuals and five replicates for a total of 32 barcoded (sequence tagged) individuals. For each group, the five replicates consisted of four intralibrary (group) replicates and one interlibrary replicate. Replicates had the same DNA source but were processed and barcoded independently. Replicates were chosen randomly but included at least one replicate per outgroup and sampling location. Within each group of 32 barcoded individuals, positions on PCR plates were randomly selected. The digestion, ligation and PCR steps were performed in the same plate for the three groups. Samples of the same group were then pooled together, and size selection for all three groups was performed in the same gel. The three groups were each sequenced with

single-end reads (100 bp long) in a separate lane of an Illumina HiSeq2000, using the Lausanne Genomic Technologies Facility service provider, Switzerland.

Basic quality filtering and general bioinformatics pipeline

All raw reads were trimmed to 84 bp because a considerable drop in quality was identified after position 85 of BERL3. Quality filtering and demultiplexing were performed with a custom Perl script equivalent to the *Stacks* program *process_radtags* (this custom script was developed prior to the release of the update of *process_radtags* that allows processing single-end double-digested data). Demultiplexed data were then *de novo* assembled and genotyped using *Stacks* v. 1.02 (Catchen *et al*. 2013), first with the default settings and all samples as an exploratory run, and then with the settings and subset of samples described below for the following two experiments: (i) *exploratory analysis of Stacks key assembly parameters and SNP calling model using replicates* and (ii) *effect of using different parameters on the output amount of data and on the detection of genetic structuring*. Trimming, demultiplexing and *Stacks de novo* assembly were performed using a computer cluster (Westmere Dual 6 core Intel X5650 2.66 GHz processor systems of 12 cores with 48 GB of RAM).

Experiment 1. Exploratory analysis of Stacks key assembly parameters and SNP calling model using replicates

We explored the effect of using different *de novo* assembly conditions and SNP calling model settings within *Stacks* on error rates and number of loci recovered. To do so, we used the 11 replicates that sequenced successfully (yielding sufficient reads to have >50% of the mean number of loci in a first exploratory analyses of the full data set) to run *Stacks* multiple times with a range of parameter values. For the assembly, the following key parameters were tested with the values specified in parentheses: the minimum number of raw reads required to form a stack (-*m* 2–15), the maximum number of mismatches allowed between stacks when processing an individual (-*M* 2–10), the allowed number of mismatches between loci when building the catalogue (-*n* 0–5) and the maximum number of stacks per locus (--*max_locus_stacks* 2–6). Only one parameter was varied at a time while keeping the other parameters fixed to $m = 3$, $M = 2$, $n = 0$ and *max_locus_stacks* = 3. The value of -*N* was always defined as $M + 2$. For the SNP calling model, we compared the default (where error rate varies freely) and the bounded model, testing different values (0.5, 0.25, 0.15, 0.1, 0.05 and 0.0056) for the upper bound (*sequencing error upper bound*, a parameter used by the bounded model: Catchen *et al*. 2013). Note that values >0.15 represent high and unrealistic levels of sequencing error. The minimum was set to 0.0056 because this was the PhiX estimate of sequencing error for BERL3 (which had the largest sequencing error of all lanes) at cycle 100 (instead of 75, to compensate for a slight quality drop at 80–84 bp). As for the remaining settings, three different minimum coverage values were explored ($m = 3$, 4 and 10) and the other parameters were set to the values considered to perform better in the assembly exploratory analyses ($M = 2$, $N = 4$, $n = 3$, *max_locus_stacks* = 3, see Results).

Outputs were then processed as detailed in *General processing of* Stacks *outputs* (see below), and the results were analysed in R v. 2.15.1 (R Core Team 2012) to estimate: (i) the number of output loci and SNPs; (ii) locus, allele and SNP error rates (as defined in *Error rates*, see below); and (iii) Euclidean distance matrices among individuals to build neighbour-joining (NJ) dendrograms (to examine whether replicate pairs cluster together, as would be expected).

Experiment 2. The effect of parameter values on output amount of data and the detection of genetic structuring

To examine the effect of using different *Stacks* settings on the full data set (78 specimens), we ran *Stacks* with four *de novo* parameter profiles, namely default, optimal, near optimal and high coverage. The default values were $m = 3$, $M = 2$, $N = 4$, $n = 0$, *max_locus_stacks* = 3 and the default SNP calling model. The other parameter profiles were given values that provided the highest number of loci and SNPs at the lowest error rates in the exploratory analysis using the replicate pairs ($M = 2$, $N = 4$, $n = 3$, *max_locus_stacks* = 3 and a SNP calling model with an upper bound of 0.05, see Results) but increasing the minimal coverage: $m = 3$ (optimal), $m = 4$ (near optimal) and $m = 10$ (high coverage). Note that we define optimal as the profile that performed better in *experiment 1* for our data, and thus, optimal parameter values will vary for other RADseq data. Each parameter profile was used to run *Stacks* with all individuals of *B. alpina* and *B. moranensis* (75), the three individuals of the closest outgroup (*B. trifolia*) and the replicates (14).

Outputs were then processed as detailed in *General processing of* Stacks *outputs*, and locus, allele and SNP error rates (as defined in *Error rates*) were estimated for each profile. After error rate estimation, subsequent analyses were run with only one of the replicates of each replicate pair. This data set was used to: (i) estimate an Euclidean distance matrix based on SNPs; (ii) perform a principal coordinates analysis (PCoA) based on the distance matrix to summarize data into the four-first

eigenaxes that account for 90% of the total variance; (iii) normalize the distance matrix and extract the distances between individuals of the same sampling location; and (iv) run the population program of *Stacks* to estimate F_{ST} between population pairs using only samples from *B. alpina* and *B. moranensis*.

General processing of Stacks outputs

Stacks outputs from experiments 1 and 2 were imported to a desktop computer, where data were visualized and exported as allele and coverage matrices. These matrices were then analysed with R to: (i) estimate the number of reads and coverage per locus, per individual and per lane; (ii) filter data to keep only those individuals having more than 50% of the mean number of loci per individual, and only those loci present in at least 80% of the barcoded individuals; and (iii) output loci and individuals that passed the previous filter as plink format. Further analyses were performed as described above for each experiment.

Error rates

Replicate pairs were used to estimate three error rates using R: (i) locus error rate, corresponding to missing data at the locus level and measured as the number of loci present in only one of the samples of a replicate pair, divided by the total number of loci found; (ii) allele error rate, calculated as the number of allele mismatches between replicate pairs, divided by the number of loci being compared; and (iii) SNP error rate, measured as the proportion of SNP mismatches between replicate pairs.

Note that we refer to a locus as a short DNA sequence produced by clustering together unique RAD alleles; in turn, alleles differ from each other by a small number of SNPs. We define a missing locus as absent in at least one sample of a replicate pair, but present in any other individual of the data set. In addition to the locus error rate, we further examined the distribution of missing data within replicate pairs by estimating: (i) the number of missing loci per replicate pair; (ii) the proportion of missing loci (number of missing loci per replicate pair over the total); and (iii) the percentage of missing loci of a given replicate that were not the same missing loci in the other replicates (proportion of missing loci different within a replicate pair). Appendix S1 (Supporting Information) provides a diagram detailing the differences between replicates estimated here.

The R scripts utilized here used the following packages: adegenet_1.3-7 (Jombart 2008), ape_3.0-8 (Paradis *et al.* 2004), gtools_2.7.1 (Warnes *et al.* 2013), multicore_0.1-7 (Urbanek 2011) and stringr_0.6.2 (Wickham 2012).

Results and discussion

RAD sequencing output and coverage

An average of 1 632 914 reads per tagged individual were obtained after demultiplexing, with no major differences between lanes or sampling localities. Full details of sequencing output are provided in Appendix S2 (Supporting Information). In a first exploratory analysis (using *Stacks* default settings and postfiltering the data with the >50% and 80% criteria described in the basic quality exploration section), fifteen of the 96 samples had too few reads and therefore did not pass the filter of sharing >50% of the mean number of loci with the rest of the individuals. Among these were the interlibrary replicate sequenced in lane BERL1 (PeB01_ir1) and one sample of a replicate pair (MaB21). Also, a strong lane effect associated with lane BERL3 was found. Samples sequenced from this lane clustered together within a NJ dendrogram, while the samples from BERL1 and BERL2 were intermixed, clustering typically by geography. The source of the lane effect was determined to be a single SNP found in position 70 of many reads, which was then identified as an artefact by the sequencing service provider. Deleting position 70 in all the demultiplexed reads removed the lane effect.

In general, mean coverage per locus was low (increasing the min. coverage *-m* from 3 to 10 produced a substantially lower number of loci, Fig. 2 and Table 3). For *Stacks*, coverage is the main filter to distinguish sequencing error from real variation. However, if coverage is generally low, a high-filter threshold for coverage can lead to allele dropout, which in turn becomes genotyping error. Assembling and genotyping a low-coverage RADseq data set like that of *Berberis* is thus challenging, and may lead researchers to keep only a small fraction of the loci and alleles that have high coverage for all individuals which, as shown below, may not be the most reliable data. Many RADseq data sets may have low coverage, particularly for species for which genome size is unknown, or if a study design aims for more individuals or loci to increase the accuracy of population genetic parameters (Buerkle & Gompert 2013).

Exploratory analysis of Stacks assembly parameters and SNP calling model using replicates

We ran *Stacks* with 11 replicate pairs (22 samples). After filtering the output so that all individuals shared >50% of the mean number of loci per individuals, most assembly parameter profiles recovered 19–20 samples and only runs with $n \geq 3$ recovered all 22. The samples that were not recovered for some of the parameter profiles explored for *Stacks* either had a small number of reads

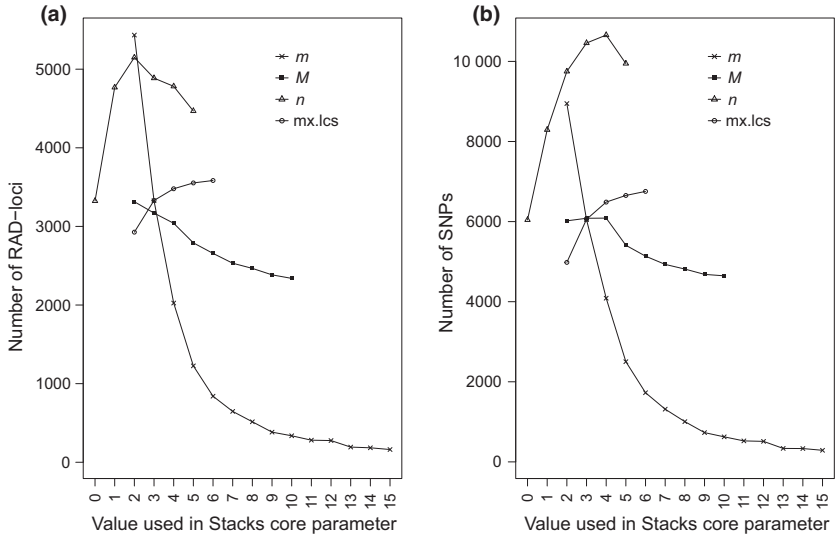

Fig. 2 Total number of (a) restriction site-associated DNA loci and (b) single-nucleotide polymorphisms obtained using different *Stacks* core parameter settings. For each run, only one parameter varied, with the remaining set to $m = 3$, $M = 2$, $n = 0$ and *max_locus_stacks* (mx.lcs) = 3 and $N = M + 2$.

Table 3 Information content, error rates and efficacy to detect structuring of genetic variation for the full data set processed with different *Stacks* parameter settings

	Optimal	Near optimal	High coverage	Default
Number of restriction site-associated DNA loci	6292	2449	292	4554
Total number of single-nucleotide polymorphisms (SNPs)	11057	4353	502	7736
Mean read coverage per sample	10.32 (SD 4.16)	15.30 (SD 5.9)	58.92 (SD 21.9)	11.50 (SD 4.65)
Mean locus error rate	0.1738 (SD 0.103)	0.1657 (SD 0.100)	0.0882 (SD 0.088)	0.1590 (SD 0.094)
Mean allele error rate	0.0592 (SD 0.013)	0.0599 (SD 0.010)	0.0879 (SD 0.023)	0.0841 (SD 0.017)
Mean SNP error rate	0.0243 (SD 0.006)	0.0321 (SD 0.006)	0.0578 (SD 0.019)	0.0423 (SD 0.010)
Variation explained by first two axes of principal coordinates analysis*	80 (39)%	82 (34)%	47 (22)%	57 (32)%
Mean of F_{ST} pairwise matrix*	0.19 (0.07)	0.15 (0.04)	0.03 (0.01)	0.07 (0.04)

*Results outside parenthesis were obtained using all the samples of the data set, and the value inside parenthesis corresponds to the results if excluding the samples from El Zamorano and the outgroup. El Zamorano (*Berberis alpina* population from SMOr) was excluded because it explained as much variation as the *B. trifolia* outgroup (Appendix S5, Supporting Information).

relative to other individuals, or belonged to the more distant outgroup (*B. pallida*, *OutBs*). These samples shared <50% of the mean number of loci with the remainder of the data set and thus were excluded by the filtering step. When both samples of a replicate pair passed filtering, they clustered together in the NJ dendrogram (Appendix S3, Supporting Information), with two exceptions: (i) the interlibrary replicate (PeB01) pair clustered together in only 18 of 36 parameter profiles tested, and in the remaining analyses it formed a paraphyletic group with other samples from the same sampling location and (ii) one replicate pair (AjB21) did not cluster together in nine occasions, with each replicate clustering instead with samples from another locality. Importantly, the parameter profiles at which incorrect clustering occurred were high values for minimal coverage (*-m*) and the number of mismatches between loci when processing an individual (*-M*). This suggests that setting *-m* too high can lead to locus/allele dropout large enough to cause incorrect inferences of individual differentiation. It is less evident why setting *-M* to high values causes differences between replicates, but it is probably related to over-merging (e.g. merging paralogs as a single locus), leading to the formation of nonsensical loci (Catchen *et al.* 2013). The fact that not all replicate pairs clustered together indicates that differentiation among individuals should be interpreted with care. However, this only occurred with some parameter values, indicating that assembly settings can be tuned to minimize differences between replicates.

Across all explored parameter profiles, the number of loci recovered ranged from approximately 200 to >5000 (Fig. 2a), the number of SNPs ranged from approximately 200 to >8000 (Fig. 2b) and the total number of

missing loci ranged from 50 to >500 (Fig. 3a). In general, the parameters that control the minimal coverage (*-m*) and number of mismatches allowed between loci when building the catalogue of loci (*-n*) contributed most to the variance of the amount of data (Fig. 2a) and missing loci (Fig. 3a,b).

A key source of variation between replicate pairs is that the identity of most (>70%) of the missing loci in a given replicate is not the same in the corresponding replicate (Fig. 3c), which leads to a locus error rate typically >10% (Fig. 3d) regardless of the parameter values used. As these differences are between samples from the same DNA source that were processed together, it seems that stochastic PCR/sequencing sampling events and imprecise size selection are the main sources of heterogeneous coverage among loci.

Allele error rates ranged from approximately 5% to >15%, depending on the parameter profile used to execute *Stacks* (Fig. 4a). Allele mismatches between replicates can be caused by allelic dropout, or by the acceptance of error-based variation (probably enhanced by PCR duplicates) during assembly. Similarly, the SNP error rate ranged from approximately 2% to 12% (Fig. 4b). Again, the most important differences were related to changes in *-m* and *-n*. Increased values of *-m* decreased the allele error rate, but not to a level below 10%, and at a cost of yielding fewer loci. Similarly, the SNP error rate was reduced from approximately 7% at $n = 0$ to approximately 2.5% at $n = 3$.

The parameter *-m* controls the total number of raw reads per individual to create a stack, so the higher it is set, the lower is the probability that there will be enough reads per locus to assemble an allele. Setting *-m* to a higher value could also result in genuine alleles being considered as secondary reads (reads that are not used to assemble reference alleles and that are set aside), and as a consequence treated as sequencing errors (see *Stacks* documentation for further details). For the *Berberis* data set, the danger of labelling stacks with concurrent sequencing errors is reduced by the fact that the data were run in three different lanes with a randomized sample design.

The parameter *-n* modulates the maximum number of mismatches allowed between loci when building the cat-

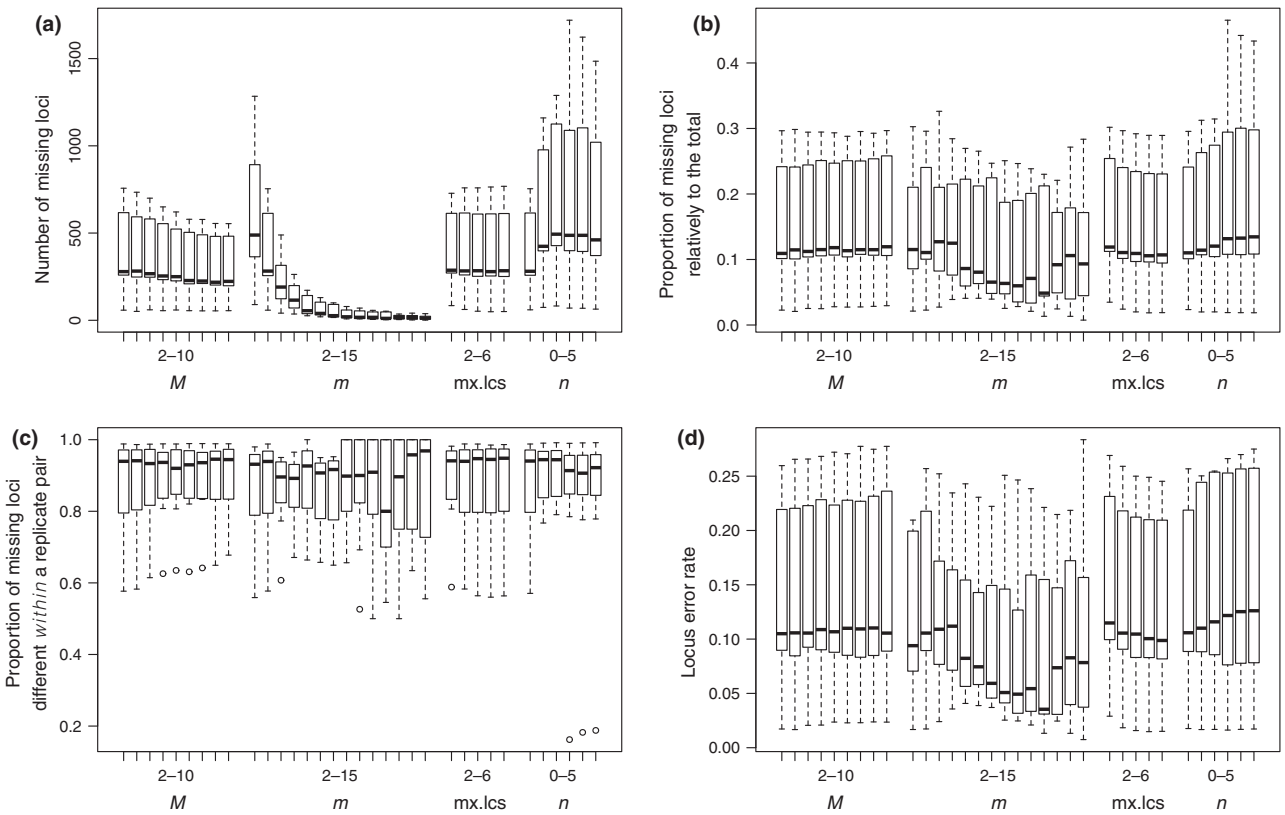

Fig. 3 Effect of different values for *Stacks* core parameters on missing data. In each run, only one parameter varied (shown on the x axis), with settings for the remainder as explained in Fig. 2. (a) total number of missing loci, (b) proportion of missing loci relative to the total, (c) proportion of missing loci different within a replicate pair and (d) locus error rate. See Appendix S1 (Supporting Information) for a diagram detailing the meaning of these estimates.

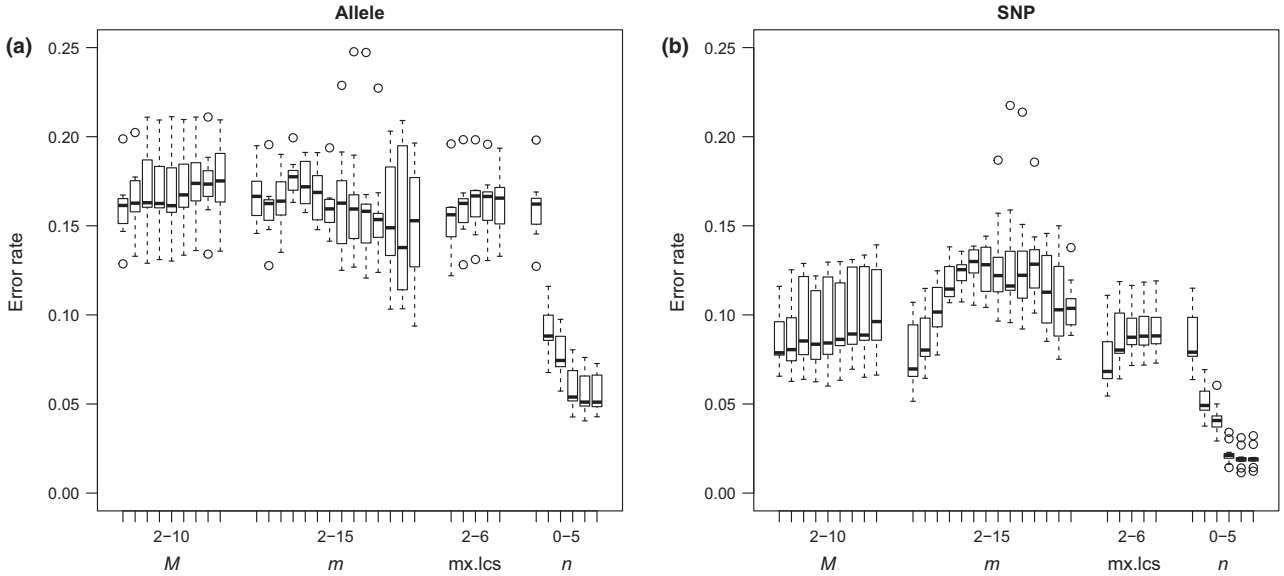

Fig. 4 Effect of different values for *Stacks* core parameters on (a) the allele error rate and (b) the single-nucleotide polymorphism error rate. In each run, only one parameter varied (shown on the x axis) with settings for the remainder as explained in Fig. 2.

alogue (this is a list of all loci and alleles in the population). If $n = 0$, there would be loci represented independently across individuals that are in reality homologous alleles of the same locus. When $n > 0$, *Stacks* uses the consensus sequence from each locus to attempt to merge loci (*Stacks* documentation). Increasing *-n* may have resulted in significant error reduction for the *Berberis* data set because replicates involved samples from geographically isolated localities and outgroups, conditions that would be expected to result in loci that exhibit fixed differences among populations. By merging fixed alleles into a single locus, the allele error rate decreased, probably because the chances of assembling the same true alleles in both replicates increased. A potential negative consequence of a high value of $-n$ is the creation of erroneous loci, which can be assembled for reasons such as the acceptance of sequencing/PCR error-based stacks, and the clustering of repetitive sequence regions or paralogs (Catchen *et al.* 2013). However, the locus error rate did not vary significantly when $-n$ varied from 0 to 5 (Fig. 3d), so it seems that the erroneous loci that were potentially created have less weight than the error reduction benefits gained from increasing the value of $-n$.

Regarding the SNP calling model, reducing the upper bound increases the chance of calling true heterozygous loci instead of wrongly labelling them as homozygous loci with sequencing error (Catchen *et al.* 2013). For the *Berberis* data, differences in genotyping errors were found only after decreasing the upper bound down to 0.0056 in the runs of $m = 3$ and $m = 4$ (Appendix S4, Supporting Information), such that the allele error rate decreased from >5% down to approximately 2.5%. However, this increased the SNP error rate from approximately 2.5% to 7%. Thus, for the *Berberis* data set, it seems better to leave the upper bound of the SNP calling model to a relatively high value. Finally, there were no differences in loci error rate between the SNP calling models (Appendix S4b, Supporting Information).

In summary, for the *Berberis* data set, the parameter values that seemed to both increase the number of loci and reduce the SNP and allele error rates were $m = 3$, $M = 2$, $N = 4$, $n = 3$, *max_locus_stacks* = 3 and a SNP calling model with an upper bound of 0.05.

Effect of using different parameters on the output amount of data and on detection of genetic structuring

The four combinations of *Stacks* settings (optimal, near optimal, default and high coverage) used to process the full data set differed in the number of recovered loci, number of SNPs and error rates (Table 3). Among the four combinations, the optimal profile generated the highest number of RAD loci (6292) and SNPs (11 057), and had the lowest allele (5.9%) and SNP (2.4%) error rates, although the locus error rate (17%) was high (Table 3). The smallest locus error rate was found with the high coverage setting (8.8%, Table 3), but this parameter profile produced the highest allele and SNP error rates (8.7% and 5.7%, respectively) and the smallest number of loci and SNPs (292 and 502, Table 3). Therefore, without the replicates, meaningful biological variation would have been discarded, and the data would have been assembled with settings that did not minimize error rate.

The SNP error rate is important for population genetic and phylogeographic analyses. As SNP error increases within a given data set, so does the contribution of noise to the genetic distance between individuals. From a drift–mutation–migration equilibrium perspective, individuals collected from the same geographic region should be expected to be genetically more similar in data sets with smaller SNP error rates. As a simple way to test this, the genetic distances between individuals from the same sampling locality were compared among the four combinations of *Stacks* settings explored here. As expected, the data with the smallest SNP error rate (optimal profile) systematically produced shorter genetic distances between individuals of the same sampling locality, when compared to the other three parameter profiles (Fig. 5).

To be of relevance for population genetics and phylogeographic analyses, molecular markers must not only have minimal noise, but also provide meaningful variation (Price & Casler 2012; Zhang & Hare 2012). The *Berberis* data produced by the optimal parameter profile resulted in substantial genetic variation, 80% of which was explained by the first two axes of the PCoA, which clustered samples by sampling locality (Appendix S5, Supporting Information). The same axes of the PCoAs produced with the data from the high coverage and default parameter profiles explained only 47% and 57%, respectively (Table 3, Appendix S5, Supporting Information). Also, the mean value of the pairwise F_{ST} matrix was higher (0.19) for the data with the smallest SNP error rate and larger number of loci (optimal parameter profile) compared to the default (0.07) or any of the other *Stacks* settings examined (Table 3). This is congruent with simulations that show that low-coverage data sets with a larger sample of sites in the genome yield more accurate and precise population genetics parameter estimates (Buerkle & Gompert 2013).

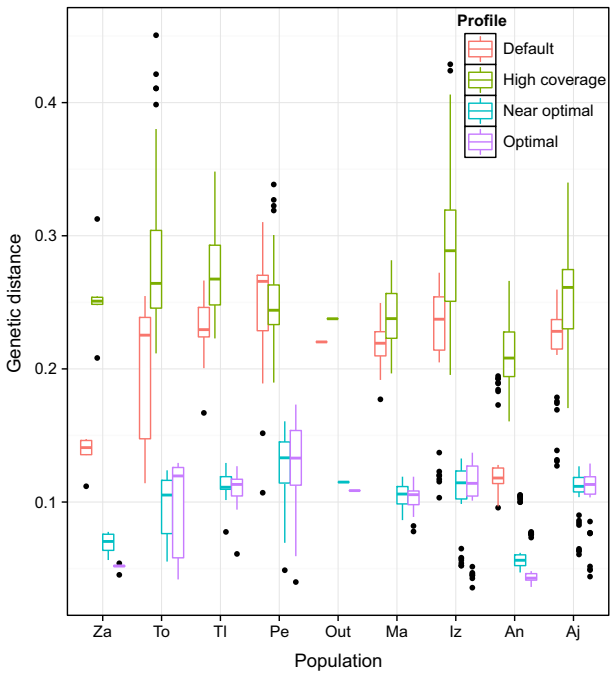

Fig. 5 Effect of different *Stacks* profiles on the genetic distance between individuals of the same sampling location using default values and settings that were considered to perform better in the exploratory parameter analyses, but varying the minimum number of raw reads required to form a stack to: $m = 3$ (optimal), $m = 4$ (near optimal) and $m = 10$ (high coverage).

Assembling *Berberis* data *de novo*, with the optimal parameter profile, maximized the number of informative SNPs and minimized the error that increases intrapopulation variation (Fig. 5). Regardless of the *de novo* assembly tool of choice, we advise researchers (particularly those working with previously unexplored genomes) to include replicates and follow the principles presented here (explore a range of parameter values and choose those that both increase the number of output loci and reduce the SNP and allele error rates). In the case of RADseq data sets already produced without DNA replicates, we recommend the exploration of a range of parameter values to maximize the amount of SNPs recovered and minimize the genetic dissimilarity between individuals from the same sampling locality. This recommendation should be used as a starting point and with care, as locality may be the wrong metric to use when minimizing genetic dissimilarity in some cases (e.g. hybrid zones, breeding areas).

De novo assembly tools and replicates

Restriction site-associated DNA sequencing is ideal to generate genomic data sets for species for which no reference genome is available, making *de novo* assembly a crucial step of data processing. Comparative analyses of some of the available bioinformatic tools show that RAD data are reliable but that it presents special issues that are not fully addressed by existing genotyping tools (Dou *et al.* 2012; Davey *et al.* 2013; Eaton 2014). By comparing *de novo* assembly outputs against a reference genome, Catchen *et al.* (2013) found that there may be substantial variance in the amount and quality of data recovery using different settings within *Stacks*. Using replicates in lieu of a reference genome, we also observed this variance (Figs 2 and 4) and were able to optimise parameter values. We focused on *Stacks*, but the principle of comparing replicates can be applied to evaluate, and reduce, the amount of error produced by different assembly tools in the absence of a reference genome. However, it should be pointed out that there is no single best bioinformatic method to handle RAD data (Davey

et al. 2013). A useful alternative to current tools would be the further development of approaches that use probabilistic base calling (e.g. Li *et al.* 2009; McKenna *et al.* 2010), that would allow uncertainty to be incorporated into the assembly process.

Error rate implications and recommendations for RADseq analyses

Next-generation sequencing methods applied to population genetic inference need to account not only for sequencing error, but also for assembly error and missing data (Pool *et al.* 2010; Davey *et al.* 2011). Including DNA replicates in the preparation of RADseq libraries (see below for some recommendations) improves the characterization of error derived from different sources (Table 1) and provides the ability to partition error into locus, allele and SNP rates. High locus error rates, such as the >10% error for all combinations of parameters evaluated for *B. alpina* (Fig. 3d, Table 3), can be accommodated as missing data and mitigated by appropriate statistical corrections (Pool *et al.* 2010; Davey *et al.* 2011), as is possible with principal components analysis, PCoA and STRUCTURE (Pritchard *et al.* 2000). However, incorrect SNP calling and allelic dropout are more problematic if data analyses are to be performed under the assumption that genotypes are known with complete certainty. Allele error can affect both allele frequency estimates and the accurate discrimination of different genotypes (Bonin *et al.* 2004), with the concomitant inflation of nucleotide diversity and skewing of the SNP frequency spectrum towards rare SNPs (Johnson & Slatkin 2008; Pool *et al.* 2010), thus affecting the meaningful biological interpretation of data. Excitingly, as population genomics and next-generation sequencing technology and analytical tools further develop, genotype uncertainty could be incorporated into the data analysis itself (Nielsen *et al.* 2011; Buerkle & Gompert 2013), using Bayesian hierarchical models and genotype probabilities rather than genotypes *per se* (Gompert & Buerkle 2011; Nielsen *et al.* 2011; Buerkle & Gompert 2013; Gautier *et al.* 2013a). If DNA replicates are included for error rate estimation, genotype uncertainty could account not only for sequencing error, but also for the full range of sources that may affect RADseq (Table 1).

The estimation of genotyping error is affected by sample size, as exemplified by the variance of error rate estimation across replicates for the *Berberis* data (Figs 3 and 4 and Table 3). Including multiple replicates is thus useful, but there is no minimum number for RADseq studies. For *B. alpina,* we aimed to replicate approximately 15% of samples, but as some failed we achieved 11%. The number of replicates for a given study will be a function of the final use of the data, the targeted coverage depth and the precision in error rate estimation needed. Replicates should be randomly chosen while also broadly representing important data features such as geography and taxonomy. In the case of geographic sampling, we would recommend the inclusion of at least one replicate per sampling location. In addition to including replicates in the final data set, replicates could be particularly useful during trial stages, as a way of evaluating the success of a given bench protocol.

Regarding recommendations to reduce error rate, as has been suggested for traditional molecular markers (e.g. Bonin *et al.* 2004; Pompanon *et al.* 2005), good laboratory practice and experimental design will help to minimize error rate. In the case of RADseq data, locus and allele recovery depend on the level of coverage of reads for each allele, locus and individual, but as shown here large numbers of markers can be recovered reliably from relatively low-coverage data sets (down to approximately 7x, as the mean for BERL1 here). Thus, given budget limitations, coverage depth may be traded-off for increased sampling for the number of individuals or sites in the genome, both of which can provide better estimates of population genetic parameters (Buerkle & Gompert 2013). However, studies that require very low error rates should consider increasing the coverage up to 60x (Davey *et al.* 2011). Using automated size selection methods (e.g. Pippen Prep) reduces variance among size-selected libraries, thus decreasing the amount of missing data at the wet laboratory stage. As we have shown here, error rates can also be reduced during *de novo* assembly using an optimal combination of parameter values. Other recommendations have been provided elsewhere (Davey *et al.* 2013).

The acceptable error rate for RADseq studies will be case specific. In the case of *Berberis alpina*, the quantification of allele and SNP error rates found for the optimised *Stacks* settings (5.9% and 2.4%, respectively, Table 3) provides reassurance that the geographic structuring of genetic variation is biologically meaningful, but would warn against more fine-scale analyses of individual relatedness if differences between individuals fell within the error rate threshold.

Estimating error rates for a low-coverage data set allows for the recovery of more loci than could otherwise be reliably achieved, and comparing replicates can be used to aid *de novo* assembling and to validate variation. Thus, including replicates can prove particularly useful for low-coverage data sets and for species lacking a reference genome. We suggest that the use of replicates for *de novo* RADseq studies should be encouraged and we consider it pertinent to extend Crawford *et al.*'s (2012) call for more transparent reporting of genotyping error to RADseq data.

Acknowledgements

We thank Subject Editor Alex Buerkle, Brant Faircloth and three anonymous referees for their constructive comments on an earlier version of the manuscript; L. Figueroa, C. Berney, T. Wyss and A. Brelsford for laboratory work assistance; O. Trejo for assistance with sampling permits and SMG, JRPPK, JJRL, AOM, ROF, SSF, RAF, TSA, JAA, FDRG, FQB y MJLF for fieldwork assistance. Part of the analyses was carried out on the high-performance computing (HPC) cluster supported by the Research and Specialist Computing Support service at UEA. This work has been supported by a CONACYT doctorate scholarship to AMY (213538), by a CONACYT Grant to DP (178245), by a SSE Rosemary Grant Student Research Award to AMY and by an SNSF Grant (PP00P3_144870) to N. Alvarez.

References

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.

Bonin A, Bellemain E, Bronken Eidesen P *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

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

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

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

Crawford LA, Koscinski D, Keyghobadi N (2012) A call for more transparent reporting of error rates: the quality of AFLP data in ecological and evolutionary research. *Molecular Ecology*, **21**, 5911–5917.

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.

Dou J, Zhao X, Fu X *et al.* (2012) Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct*, **7**, 17.

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, doi:10.1093/bioinformatics/btu121.

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.

Gomes I, Collins A, Lonjou C *et al.* (1999) Hardy–Weinberg quality control. *Annals of Human Genetics*, **63**, 535–538.

Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation population genomics. *Genetics*, **187**, 903–917.

Grundemann D, Schomig E (1996) Protection of DNA during preparative agarose gel electrophoresis against damage induced by ultraviolet light. *BioTechniques*, **21**, 898–903.

Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. *Methods in Molecular Biology (Clifton, N.J.)*, **888**, 235–260.

Johnson PLF, Slatkin M (2008) Accounting for bias from sequencing error in population genetic estimates. *Molecular Biology and Evolution*, **25**, 199–206.

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of Xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Loman NJ, Misra RV, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meacham F, Boffelli D, Dhahbi J *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.

Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, **160**, 357–366.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

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

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Parchman TL, Gompert Z, Mudge J *et al.* (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

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.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Price DL, Casler MD (2012) Simple regression models as a threshold for selecting AFLP loci with reduced error rates. *BMC Bioinformatics*, **13**, 268.

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

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards PM, Liu MM, Lowe N *et al.* (2013) RAD-Seq derived markers flank the shell colour and banding loci of the *Cepaea nemoralis* supergene. *Molecular Ecology*, **22**, 3077–3089.

Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, **38**, D234–D236.

Rounsaville TJ, Ranney TG (2010) Ploidy levels and genome sizes of Berberis L. and Mahonia Nutt. species, hybrids, and cultivars. *HortScience*, **45**, 1029–1033.

Taberlet P, Griffin S, Goossens B *et al*. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Urbanek S (2011) *Multicore: Parallel Processing of R Code on Machines with Multiple Cores or CPUs*. R package version 0.1-7. Available from http://CRAN.R-roject.org/package=multicore.

Vázquez-Lobo A (1996) *Filogenia de hongos endófitos del género Pinus: Implementación de técnicas moleculares y resultados preliminares*. Sc. Bach. Dissertation, Facultad de Ciencias, Universidad Nacional Autónoma de México, México.

Warnes GR, Bolker B, Lumley T (2013) *gtools: Various R Programming Tools*. R package version 2.7.1. Available from http://CRAN.R-project.org/package=gtools.

Wickham H (2012) *stringr: Make It Easier to Work with Strings*. R package version 0.6.2. Available from http://CRAN.R-project.org/package=stringr.

Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

Xu J, Turner A, Little J, Bleecker ER, Meyers DA (2002) Positive results in association studies are associated with departure from Hardy–Weinberg equilibrium: hint for genotyping error? *Human Genetics*, **111**, 573–574.

Zamudio S (2009) Notas sobre el género Berberis (Berberidaceae) en México. *Acta Botánica Mexicana*, **87**, 31–70.

Zhang H, Hare MP (2012) Identifying and reducing AFLP genotyping error: an example of tradeoffs when comparing population structure in broadcast spawning versus brooding oysters. *Heredity*, **108**, 616–625.

A.M.Y., B.C.E., D.P., N. Alvarez and T.H.J. designed the research. A.M.Y. and D.P. provided the samples. A.M.Y. performed laboratory work and data analyses. A.M.Y. and N. Arrigo contributed analytical tools. A.M.Y. and B.C.E. wrote the manuscript. All authors contributed to results and discussion, and manuscript edition.

Data and code availability

Raw RADseq data Sequence Read Archive (SRA) accession SRP035472. Sampling information, custom R & Perl scripts and jobs with settings used to run *Stacks*, output data to compare error rates and population differentiation: doi:10.5061/dryad.g52m3. Error rate R functions updated and versioned: https://github.com/AliciaMstt/RAD-error-rates.

Supporting Information

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

Appendix S1 Schematic diagram of RAD data genotyping and differences between replicates

Appendix S2 Summary of ddRAD lab work reaction mixes used and the characteristics of the resulting libraries

Appendix S3 Dendrograms obtained from the analyses of replicates analyses with different *Stacks* parameters

Appendix S4 Allele and single-nucleotide polymorphism (SNP) error rates for the SNP calling model analyses

Appendix S5 PCoA for each of the four *Stacks* parameter profiles tested