Finding the right coverage

Fountain, Emily D.; Pauli, Jonathan N.; Reid, Brendan N.; Palsboll, Per J.; Peery, M. Zachariah

Published in:
Molecular Ecology Resources

DOI:
10.1111/1755-0998.12519

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 966–978 doi: 10.1111/1755-0998.12519

Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates

EMILY D. FOUNTAIN,* JONATHAN N. PAULI,* BRENDAN N. REID,* PER J. PALSBØLL† and M. ZACHARIAH PEERY*

Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, Madison, WI 53706, USA, †Marine Evolution and Conservation, Groningen Institute of Evolutionary Life Sciences, University of Groningen, Groningen, 9747 AG, The Netherlands

Abstract

Restriction-enzyme-based sequencing methods enable the genotyping of thousands of single nucleotide polymorphism (SNP) loci in nonmodel organisms. However, in contrast to traditional genetic markers, genotyping error rates in SNPs derived from restriction-enzyme-based methods remain largely unknown. Here, we estimated genotyping error rates in SNPs genotyped with double digest RAD sequencing from Mendelian incompatibilities in known mother–offspring dyads of Hoffman's two-toed sloth (*Choloepus hoffmanni*) across a range of coverage and sequence quality criteria, for both reference-aligned and *de novo*-assembled data sets. Genotyping error rates were more sensitive to coverage than sequence quality and low coverage yielded high error rates, particularly in *de novo*-assembled data sets. For example, coverage ≥5 yielded median genotyping error rates of ≥0.03 and ≥0.11 in reference-aligned and *de novo*-assembled data sets, respectively. Genotyping error rates declined to ≤0.01 in reference-aligned data sets with a coverage ≥30, but remained ≥0.04 in the *de novo*-assembled data sets. We observed approximately 10- and 13-fold declines in the number of loci sampled in the reference-aligned and *de novo*-assembled data sets when coverage was increased from ≥5 to ≥30 at quality score ≥30, respectively. Finally, we assessed the effects of genotyping coverage on a common population genetic application, parentage assignments, and showed that the proportion of incorrectly assigned maternities was relatively high at low coverage. Overall, our results suggest that the trade-off between sample size and genotyping error rates be considered prior to building sequencing libraries, reporting genotyping error rates become standard practice, and that effects of genotyping errors on inference be evaluated in restriction-enzyme-based SNP studies.

Keywords: ddRAD, genotyping error, Mendelian incompatibility, next-generation sequencing, single nucleotide polymorphism

Received 23 March 2015; revision received 10 February 2016; accepted 11 February 2016

Introduction

The advent of next-generation sequencing (NGS) is revolutionizing the field of molecular ecology by facilitating the discovery and genotyping of thousands of single nucleotide polymorphisms (SNPs) in nonmodel organisms. Moreover, approaches, such as restriction-associated DNA sequencing (RAD-seq) (Baird *et al.* 2008; Peterson *et al.* 2012) and genotyping by sequencing (GBS) (Elshire *et al.* 2011), allow for the genotyping of large numbers of SNP loci across multiple individuals. As such, restriction-enzyme-based methods hold tremendous potential for improving the characterization of demographic history (Pujolar *et al.* 2013; Lozier 2014), quantification of population connectivity (Corrales & Höglund 2012; Funk *et al.* 2012) and detection of loci subject to local adaptation (Lamichhaney *et al.* 2012; Franchini *et al.* 2014). Nevertheless, the application of restriction-enzyme-based sequencing approaches for population-level genotyping is still in its infancy and, in contrast to more traditional genetic markers such as microsatellite loci, genotyping error rates in SNPs remain unknown and factors influencing error rates are largely unexplored (see Larson *et al.* 2014; Palti *et al.* 2015).

Users of restriction-enzyme-based sequencing methods are confronted with a trade-off between the number of loci and individuals genotyped and coverage. Limiting genotyping to SNPs with high coverage (i.e. read depth) reduces error rates in heterozygotes by increasing

Correspondence: Emily D. Fountain, Fax: +1 608 262 9922; E-mail: efountain@wisc.edu

the likelihood that both alleles will be detected, but also limits the number of individuals and loci sampled (Li *et al.* 2008; Harismendy *et al.* 2009). Reducing coverage allows for greater sample sizes of individuals and loci, but increases the risk of incorrectly calling a heterozygote a homozygote due to sequencing errors because fewer correct sequences are available per gene copy. Several studies have investigated methods for incorporating sequencing error rates in data subject to low coverage (Bansal 2010; Malhis & Jones 2010; Le & Durbin 2011; Nielsen *et al.* 2012; Mastretta-Yanes *et al.* 2015), but the effects of sequencing error and coverage on genotyping error rates are uncertain. Therefore, subjectively chosen coverage and sequence quality criteria often involve low to medium coverage thresholds in order to maximize the number of loci (Table 1), and these criteria may result in error rates that compromise inference (Nielsen *et al.* 2011).

In addition to the effects sequencing error and coverage, genotyping error rates are likely influenced by the quality of genome reference data available for alignment. When available, aligning raw sequence reads to a reference genome prior to SNP and genotype calling may reduce error rates. A high-quality genome (one that contains no discernable misassemblies and for which some of the gaps in assembly have been resolved) increases alignment accuracy due to the removal of contaminated sequences and a mostly resolved genome (Chain *et al.* 2009). However, researchers must often conduct alignments using draft genomes with low scaffold coverage, incomplete assembly and an unknown amount of error (Chain *et al.* 2009; Gnerre *et al.* 2011). Moreover, even low-quality reference genomes are unavailable for many nonmodel organisms. For nonmodel species, *de novo* assembly presents additional challenges for SNP discovery and genotyping including spurious SNP calls as a result of collapsing paralogs into unique loci (Zhang *et al.* 2011).

In this study, we estimated error rates in SNP genotypes derived from double digest restriction-associated DNA sequencing (ddRAD) as a function of coverage and sequence quality criteria, and compared error rates between reference-aligned and *de novo*-assembled data sets. We inferred genotyping error rates based on Mendelian incompatibilities in mother–offspring pairs of Hoffman's two-toed sloth (*Choloepus hoffmanni*), relationships confirmed through previous field observations of radiocollared individuals and microsatellite-based parentage analyses (Peery & Pauli 2012). Estimating genotyping error rates based on deviations from Mendelian expectations between closely related individuals such as parents and their offspring (Douglas *et al.* 2002; Hao *et al.* 2004; Saunders *et al.* 2007; Haaland & Skaug 2013; Chen *et al.* 2014) provides a cost–effective alternative to sequencing the same individuals multiple times (He *et al.* 2013; Mastretta-Yanes *et al.* 2015). We subsequently conducted maternity assignments for each offspring present in our sample at different coverage thresholds to assess the effect of genotyping error rates on a common application in molecular ecology. Our study is the first to empirically evaluate how SNP genotyping error rates are influenced by the combined effects of coverage and sequence quality in a nonmodel species using ddRAD sequencing. Based upon these results, we provide guidelines for selecting coverage criteria in assays of nonmodel species.

Methods

Field sampling and DNA extraction

We captured 16 mother–offspring pairs (Tables S1–S3, Supporting information) of *Choloepus hoffmanni* 85 km northeast of San José, Costa Rica. Individuals were wild caught from February 2010 to February 2011 following methods described in Peery & Pauli (2012). We collected

Table 1 Examples of coverage and quality score used in previous genetic studies for single nucleotide polymorphism obtained by restriction-enzyme-based sequencing using *de novo* assembly and reference genome aligned data. 'Lowest' represents the lowest coverage minimum used in a study and 'highest' represents the highest coverage minimum used in a study

	Coverage/Quality score					
	Linkage mapping		Population diversity		Phylogenetics	
	Lowest	Highest	Lowest	Highest	Lowest	Highest
Reference aligned	5/30[c]	–	5/30[f]	10/–[g]	–	–
De novo-assembled	8/20[a]	15/–[b]	6/10[d]	10/15[e]	15/20[h]	12/30[i]

–: Information on coverage and/or quality score was lacking.
[a]Zhou *et al.* (2014), [b]Recknagel *et al.* (2013), [c]Kai *et al.* (2014), [d]Ruegg *et al.* (2014), [e]Lozier (2014), [f]Hohenlohe *et al.* (2010), [g]Jezkova *et al.* (2015), [h]Jones *et al.* (2013), [i]Wagner *et al.* (2013).

tissue by ear punch from captured individuals. Genomic DNA was extracted from the 32 *C. hoffmanni* using a Qiagen DNeasy Blood and Tissue Kit™ (Qiagen, catalog #59506) following the manufacturer's protocol for tissue samples with a modifications to the last step: two separate, final elutions of each 65 μL each were conducted.

ddRAD-Seq library preparation and sequencing

We followed the ddRAD protocol by Peterson *et al.* (2012) for preparing double digest libraries for sequencing using the restriction enzymes *Eco*RI (HF) and *Msp*I from New England Biolabs, Beverly, MA, USA. Appendix S1 (Supporting information) contains complete laboratory protocol details. Each library was comprised of 50 individuals per library with one library sequenced per lane. The 32 individuals in this study were not sequenced in more than one library. The ddRAD libraries consisted of 50 high-quality DNA digests (run 1), 50 low-quality DNA digests (run 2) or a mix of both high- and low-quality DNA extracts to total 50 (run 3). We defined a DNA extract as 'high quality' if the majority of DNA fragments were >500 bp based on analysis on Agilent 2100 Bioanalyzer with a DNA 1000 Chip (Agilent Technologies, Santa Clara, CA, USA), the DNA extraction had ~1.8 260/280 ratio, DNA extraction concentration prior to digestion was ≥60 ng/μL, and DNA concentration after digestion was >5 ng/μL; and 'low quality' if the majority of DNA fragments were <500 bp, the DNA extraction had <1.7 260/280 ratio, DNA concentration prior to digestion was <60 ng/μL, and DNA concentration after digestion was below 5 ng/μL (Tables S1–S3, Supporting information). We pair-end sequenced samples to 101 bp with the Illumina HiSeq2000 high throughput in three lanes, which contained a single PhiX control lane. Sequencing was conducted at University of Wisconsin-Madison Biotechnology Center, USA.

Single nucleotide polymorphism and genotype calling

We conducted all data cleaning, demultiplexing, SNP and genotype calling in STACKS 1.20 (Catchen *et al.* 2013). Raw reads were demultiplexed by barcode and index, and were trimmed to 90 bp using the `process_radtags` command. We cleaned the raw reads at different average quality score values: ≥30, ≥20 and ≥10 by setting the `–s` command in `process_radtags` to the desired average quality score. The sliding window (`-w`) was kept at the default value of 0.15. The sliding window calculates the average quality score within a fraction of the read length and discards any reads that drop below the set `-s` value within the sliding window. Individual genotypes were called using coverage thresholds: ≥30, ≥20, ≥15, ≥10 and ≥5 for each quality score, leading to 18 combinations of coverage and sequence quality criteria for each data set (reference-aligned *reduced*, reference-aligned *full* and *de novo*) for assessing how the parameters impacted genotyping error rates. We retained SNPs that were shared between a minimum two individuals and with a minor allele frequency (MAF) ≥0.02.

Reference alignment. After demultiplexing and cleaning, we pair-end aligned the remaining R1 and R2 reads to the *C. hoffmanni* reference genome version 1.0 (GCA_000164785.1) using SOAP2 (Li *et al.* 2009). We chose SOAP2 as this aligner does not soft mask; STACKS 1.20 converts soft-masked regions to Ns which may make calling some polymorphisms or mapping haplotypes impossible on soft-masked reads. Parameters for the ungapped alignment included a maximum of two mismatches permitted in the seed and the insert size constrained between 400 and 600 bps.

We created two reference-aligned data sets in STACKS 1.20. The first data set (hereafter referred to as 'reference-aligned *full*') included the pair-end reads that aligned to the reference genome without the possibility of aligning to more than one contig (unflagged reads) and all reads that were flagged as possibly residing on different contigs. The reference-aligned *full* data set allowed for a more direct comparison of genotyping error rates with the *de novo*-assembled data set, given that the latter included all demultiplexed reads. The second data set (from here on referred to as 'reference-aligned *reduced*') was constructed using only unflagged reads. We generated the reference-aligned *reduced* data set to determine whether including secondary alignments in the analyses influenced genotyping error rates. Reference-aligned reads were processed using the `ref_map` pipeline, which creates stacks of loci for each sample. All individuals were subsequently loaded into the catalog, and matches were made against the catalog. The number of identically aligned sequences required to build a stack (`-m` parameter) was evaluated using the values 3 and 5 given that using higher `-m` values tends to split true loci into several different loci (Catchen *et al.* 2011; Mastretta-Yanes *et al.* 2015). The two values resulted in similar numbers of loci and therefore we elected to use setting of `-m` = 3 for all subsequent analyses.

De novo assembly. STACKS 1.20 outputs four files after demultiplexing: P1 reads that have a matching P2 read, P2 reads that have a matching P1 read, P1 reads that do not have a matching P2 read and P2 reads that do not have a matching P1 read. Reads are considered to not have a pair-end match when the other read was discarded due to low quality or missing the restriction

enzyme cut site. We processed individuals using the `denovo_pl` pipeline under several scenarios to determine the best parameters to use for our data (Appendix S1, Supporting information). To determine genotyping error rates, we used the paired R1 and R2 reads with the default SNP calling algorithm and the following parameters: `-m 3, -M 3, -n 2`.

Correction algorithm. Version 1.20 of STACKS includes a correction algorithm, `rxstacks`, which may improve individual genotype calls by making corrections based the population-level data in four ways: (i) loci are assigned a log-likelihood score during the stack assembly step of the pipeline, and then using the `rxstacks` command, loci can be removed from the data set if they fall below the user set log-likelihood value; (ii) likely sequencing errors at the individual level are filtered out based on the alleles for a particular nucleotide position in the population and then the nucleotide positions are recalled using the bounded SNP model (bounded by known error rate); (iii) confounded loci (for which an individual has multiple loci that match a single catalog locus) are filtered; and (iv) excess haplotypes are removed.

We tested the `rxstacks` correction algorithm for quality score ≥20 for minimum coverage criteria of 5, 10, 15 and 20. We did not consider a minimum coverage ≥30 because few individuals remained after the correction algorithm was applied to the reference-aligned *full* data set. For the reference-aligned *full* data set, after the `ref_map` pipeline, we ran `rxstacks` and implemented a log-likelihood cut-off of −15.0 under a bounded SNP model with a max bound of 0.01. We chose the log-likelihood threshold on the criteria that at least 3000 loci must be genotyped at a log value in order to retain enough loci for downstream analyses. After removing spurious loci based on the algorithm as described above, the `cstacks` and `sstacks` commands were rerun on the corrected data using the default values which allow for no mismatches when generating the catalog. For the corrected, reference-aligned *full* data set, we removed dyads with <100 shared loci and then we estimated the genotyping error rate. For the *de novo*-assembled SNP genotypes, we employed the same parameter values except the log-likelihood cut-off (−12.0) which was based on the same criteria used for the reference-aligned *full* data set. We conducted a Wilcoxon signed-rank test in R 3.0.1 (R Core Team 2013) between the uncorrected and corrected data sets to test for differences in genotyping error rates and number of loci.

Estimating genotyping error rates. We identified Mendelian errors (MEs; i.e. cases where offspring and their mothers did not share at least one allele at a given locus) in the 16 mother–offspring dyads for all quality score and coverage criteria described above in SAS® 9.4 (SAS Institute Inc. 2013). Dyads that shared fewer than 100 shared loci (Santure *et al.* 2010) for a given quality score and coverage criteria combination were removed from analyses. Although not all genotyping errors result in a ME, there is a linear relationship between the number of genotyping error and ME rates (Saunders *et al.* 2007). The genotyping error rate for each dyad was estimated from the observed ME rate following Saunders *et al.* (2007). We calculated expected MEs as the sum of all values of $P^{(1)}_{\mathrm{ME}}(p_A, m)$ over all SNPs where both the mother and the offspring had genotypes for a given locus, $P^{(1)}_{\mathrm{ME}}(p_A, m) = p_A p_B (2-(1-1/2p_B))^m - (1-1/2p_A)^m + 1/2m)$ p_A and p_B were the population allele frequencies, and m was the number of offspring. In our data, m always equalled one as each mother had only one offspring. In R 3.0.1 (R Core Team 2013), we calculated the median, 25th and 75th percentiles, and 1.5 × interquartile range for genotyping error rates and the number of loci at each quality score and coverage criteria. Mother–offspring dyads represented the sampling unit for all summary statistical estimates. We estimated genotyping error rates at minimum coverage and quality score values (e.g. coverage ≥10 and quality score ≥20). Also, we investigated how loci within a particular coverage interval (e.g. coverage = 10–20) impacted genotyping error rates (see Fig. 3B,C) to determine whether the loci with only low coverage leveraged the genotyping error rates. We referred to the former as 'threshold' analyses and the latter as 'interval' analysis.

Parentage analysis

To assess the effects of genotyping error rates on a population genetic application, we conducted parentage analyses under different coverage thresholds in CERVUS 3.0 (Marshall *et al.* 1998; Kalinowski *et al.* 2007). Specifically,

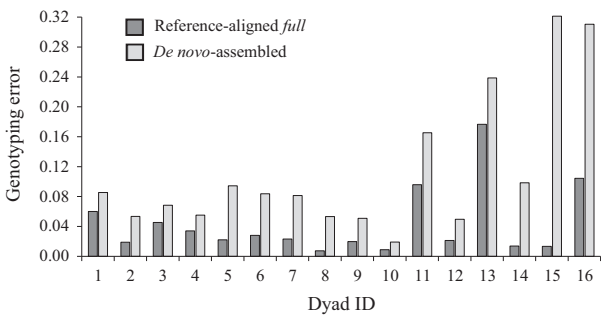

Fig. 1 Dyad-specific genotyping error rates for the reference-aligned *full* and *de novo*-assembled data sets at a quality score ≥20 and coverage ≥10.

we conducted 'single-parent' assignments and treated the maternity of juvenile sloths as unknown using a quality score ≥20 for the five coverage thresholds (5, 10, 15, 20 and 30) for uncorrected *de novo*-assembled and reference-aligned *reduced* data sets and three different coverage cut-offs (5, 10 and 20) for the `rxstacks` corrected reference-aligned *reduced* data sets (further details in Appendix S1, Supporting information). We determined the statistical significance of parentage assignments using critical 'delta' values at the 95% confidence level, assumed complete sampling of mothers and assumed a 0.01 and a 0.05 genotyping error rate in separate analyses. We recorded the number of offspring that were correctly and incorrectly assigned, and unassigned to a mother for each coverage criteria.

Results

After quality control, the average number of pair-end reads retained for the 32 individuals was as follows: 3 990 770 for quality score ≥30, 5 159 585 for quality score ≥20 and 5 401 076 for quality score ≥10. The fewest unmatched P1 and P2 reads were produced at quality score ≥10 because fewer reads were discarded as low quality. Full details of sequencing output are provided in Tables S1–S3 (Supporting information).

Variation in genotyping error rates among dyads

Genotyping error rates varied considerably among the 16 mother–offspring dyads for all quality and coverage criteria in both reference-aligned *full* and *de novo*-assembled data sets (Fig. 1). Some variation in dyad-specific genotyping error rates could have been the result of differences in DNA quality among extractions; however, 'low-quality' DNA extractions occurred in several dyads with low genotyping error rates (e.g. dyads 4, 8 and 10; Fig. 1, Tables S1–S3, Supporting information). Moreover, several 'high-quality' DNA extractions yielded high genotyping error rates (e. g., dyads 11, 14 and 15; Fig. 1, Tables S1–S3, Supporting information). Finally, we did not detect a correlation between dyad-specific genotyping error rates and the shared number of loci or the average number of reads when coverage was ≥5 or ≥30.

Reference alignment

A relatively low percentage of reads (28–38%) aligned with high likelihood to one place in the reference genome, even when relaxed alignment parameter values were used that allowed for lower sequence identity or larger gaps. An additional 35–48% of reads aligned to more than one place in the reference genome. The reference-aligned *reduced* data set (which only included reads that aligned to one place) had slightly greater median genotyping error rates compared to the reference-aligned *full* data set (which included secondary alignments; Fig. 2A). The reference-aligned *reduced* data set had fewer loci than the reference-aligned *full* data set; for example, the median number of SNP loci was 11 099 and 19 016 at coverage ≥5, respectively (Fig. 2B).

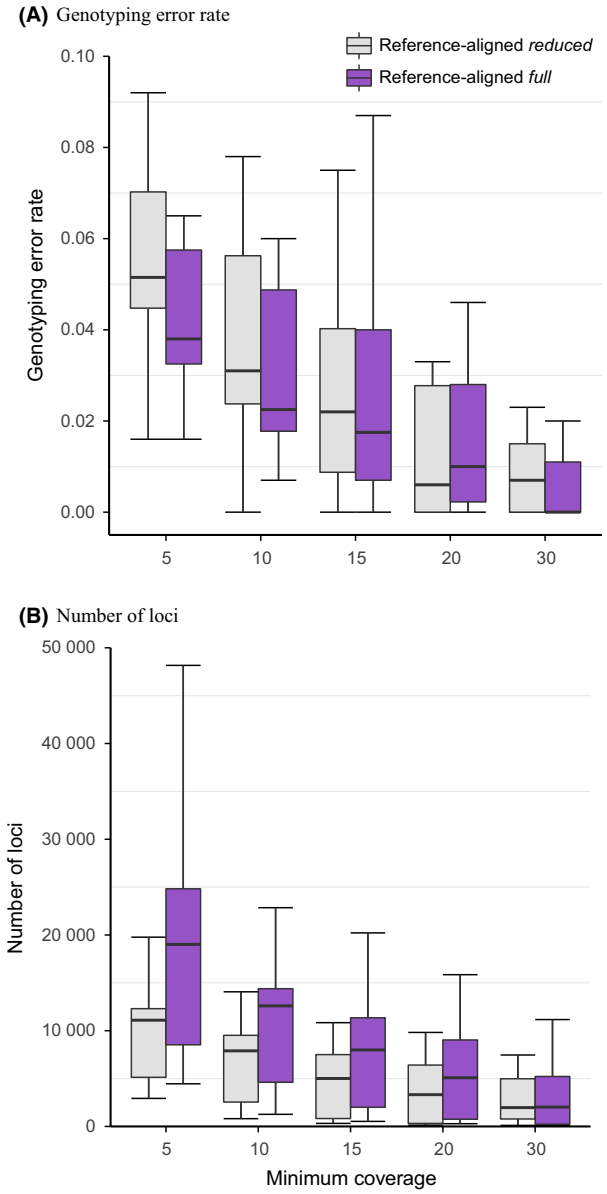

Fig. 2 Genotyping error rates (A) and number of loci shared by dyad (B) as a function of coverage in the reference-aligned *reduced* and *full* data sets at a quality score ≥20. Box plots represent the median, 25th and 75th percentiles, with whiskers extending to a maximum of 1.5 × IQR. (A) genotyping error rates, (B) number of shared loci.

In the reference-aligned *full* data set, three dyads had high (>0.08) genotyping error rates (dyads 11, 13 and 16; Fig. 1, Tables S1–S3, Supporting information). Removing these dyads reduced the median and among-dyad variability in genotyping error rate (Fig. S1, Supporting information). However, we retained these dyads when estimating population-level genotyping error rates because, in the absence of an association between genotyping error rates and DNA extraction quality, SNP-based studies will typically have no *a priori* basis for censoring samples.

Genotyping error rates decreased with increasing coverage for all quality scores (Figs. 3A, top panel, and S2, Supporting information). At a coverage ≥5, genotyping errors exceeded 0.03, but declined to ≤0.01 at coverage ≥30. The median number of loci declined substantially as the coverage increased; for example, sampled loci declined from 15 770 to 1 598 at coverage ≥5 to coverage ≥30 when quality score was ≥30 (Fig. 3A, bottom panel). The median genotyping error rates were greater in the coverage interval analysis (where loci were restricted to those within specific coverage intervals) than in the coverage threshold analysis (Fig. 3A,B). Median genotyping error rates were greatest when coverage was restricted to between ≥5 and ≤10 (>0.10 for all minimum quality scores). The majority of loci were at coverage ≥30 for each quality score (Fig. 3B, bottom panel), but the inclusion of low coverage SNPs nevertheless greatly increased median genotyping error rates. Quality score criteria did not greatly impact mean genotyping error rate except at lower coverage (≥5 and ≥10) (Fig. 3C, to panel). The majority of reads were above a quality score ≥30 and, therefore, changes in number of loci were relatively small across different quality score intervals (Fig. 3C; bottom panel).

De novo-assembled

We observed the same decline in the median genotyping error rate as a function of decreasing coverage in the *de novo*-assembled data set as occurred in the reference-aligned *full* data set. The median genotyping error rate was ≥0.11 in the *de novo*-assembled data set at a coverage ≥5 (as opposed to ≥0.03 in the reference-aligned *full*) and

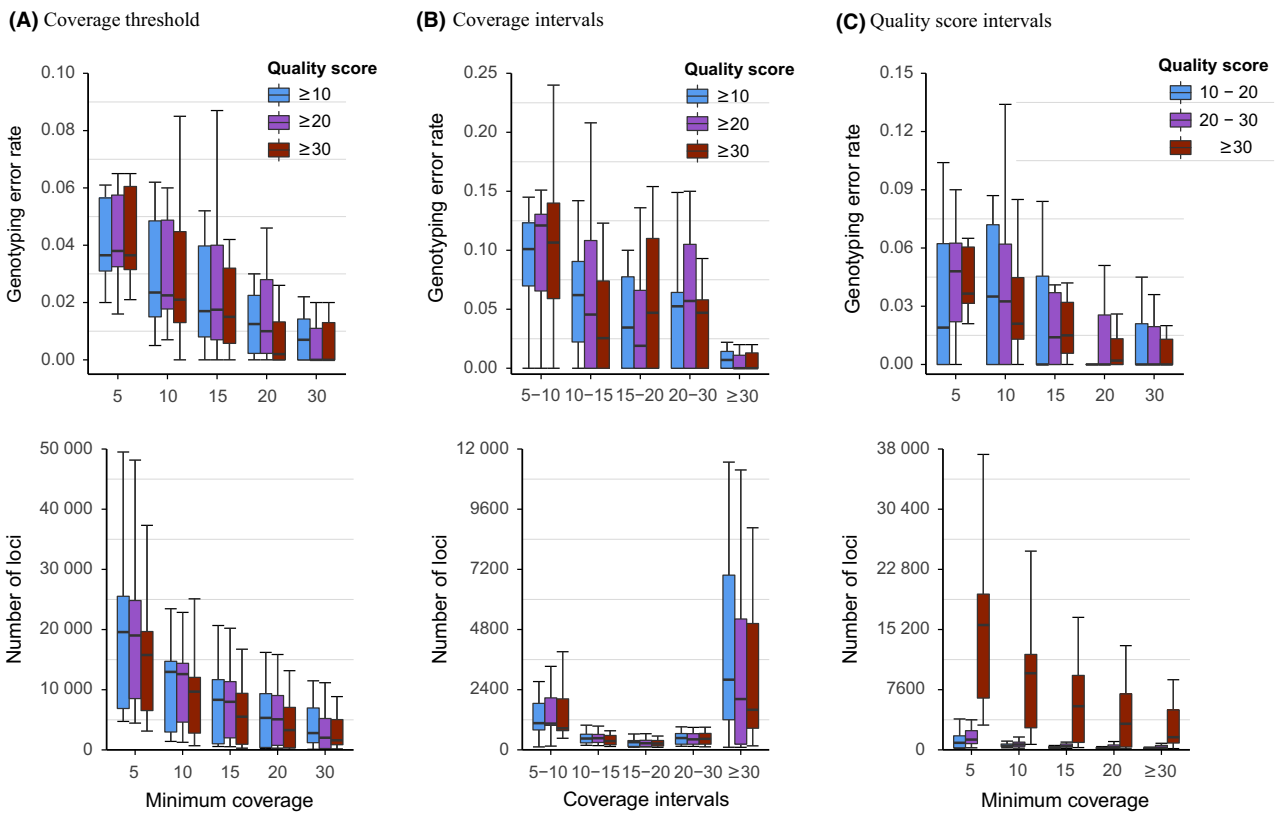

Fig. 3 Genotyping error rates (top panel) and number of loci shared by dyad (bottom panel) as a function of quality score and coverage in the reference-aligned *full* data set. Box plots represent the median, 25th and 75th percentiles, with whiskers extending to a maximum of 1.5 × IQR. (A) Genotyping error rates by coverage and quality score thresholds, (B) genotyping error rates by coverage interval and quality score threshold, and (C) genotyping error rates by coverage threshold and quality interval.

was ≥0.04 at coverage ≥30 (Fig. 4A, top panel, and S3, Supporting information). High median genotyping error rates in the *de novo*-assembled data set were mainly due to high error rates (>0.20) in three dyads (dyads 13, 15 and 16; Fig. 1, Tables S1–S3, Supporting information). While removing these dyads reduced the median genotyping error rate (Fig. S4, Supporting information), these dyads were retained in the estimation of genotyping error rates for the reasons described above. The *de novo*-assembled data set contained more loci than reference-aligned *full* data set for all coverage criteria except at quality score ≥30 and coverage ≥30 (Figs 3 and 4). However, like the reference-aligned *full* data set, the number of loci declined substantially with increased coverage; for example, loci declined from 15377 to 1156 at coverage >5 to coverage ≥30 when quality score was ≥30 (Fig. 4A, bottom panel). Also, minimum quality score had little impact on genotyping error rates (Fig. 4A, top panel) and genotyping error rates increased at low coverage intervals (Fig. 4B, top panel) in the *de novo*-assembled data set.

Rxstacks correction algorithm

Genotyping error rates were significantly lower with than without the correction for both reference-aligned *full* and *de novo* data sets for low coverage (≥5 and ≥10, Table 2). However, applying the `rxstacks` correction did not always lower the median genotyping error rate (Figs 5 and S5, Supporting information). In the reference-aligned *full* data set, the median genotyping error rate was greater with the correction than without the correction at higher coverage thresholds (Fig. 5A, top panel). The correction process resulted in a significant decrease in the number of loci (Table 3). The median genotyping error rate was lower and the number of loci was greater in the uncorrected data set when coverage ≥10 than in the corrected data set when coverage ≥5 (Fig. 5). The effectiveness of the correction process in reducing genotyping error rates varied among dyads (Fig. 6A,B). Moreover, some dyads were removed from the analysis because they shared <100 loci after the correction (Fig. 6A,B).

Parentage analysis

Varying the assumed genotyping error rate (0.01 or 0.05) had no effect on the maternity assignments (data not shown). The number of assignments and the proportion of correct assignments were identical between the uncorrected reference-aligned *reduced* and the *de novo*-assembled data sets, but the proportion of correct assignments was slightly lower when the `rxstacks` correction was applied to the reference-aligned data set. A greater number of maternities were assigned at lower coverage, but the proportion of incorrect assignments was higher at low coverage (e.g. between 0.20 and 0.30 for coverage ≥5 vs. 0.14 for coverage ≥30; Fig. 7).

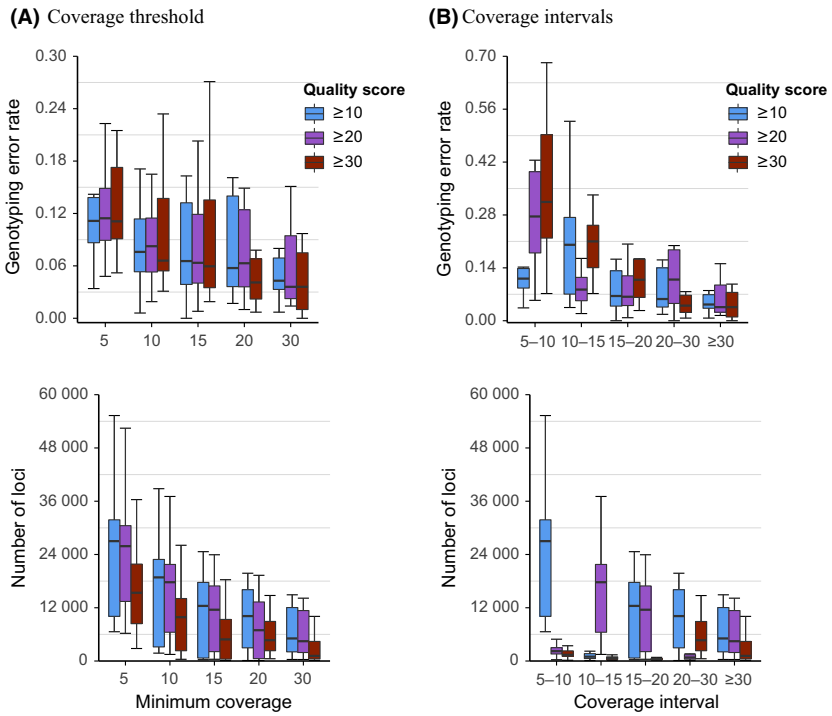

Fig. 4 Genotyping error rates (top panel) and number of loci shared by dyad (bottom panel) as a function of quality score and coverage in the *de novo*-assembled data set. Box plots represent the median, 25th and 75th percentiles, with whiskers extending to a maximum of 1.5 × IQR. (A) Genotyping error rates by coverage and quality score thresholds, (B) genotyping error rates by coverage interval and quality score threshold.

Discussion

Our study provides the first empirical estimates of genotyping error rates in SNPs identified with ddRAD-based methods and highlights the importance of using sufficiently stringent coverage criteria to achieve 'reasonable' genotyping error rates. What constitutes a reasonable genotyping error rate will ultimately depend on the objective of the study using SNP genotypes, but using coverage criteria ≥5 and ≥10 led to genotyping error rates that reduced the proportion of correct maternity assignments in this study. Adopting higher coverage thresholds increased the proportion of correct assignments but resulted in the loss of loci and individuals, as well as fewer total correct assignments. Consequently, a trade-off exists between sample sizes and genotyping error rates, and reducing the number of loci to achieve a lower error rate could affect the strength of inference. Conducting additional sequencing runs or including fewer individuals in a library may be required to obtain the desired number of loci and individuals at sufficiently stringent coverage criteria. By comparison, genotyping error rates were relatively insensitive to sequence quality when coverage was ≥10, and applying a relaxed quality criteria (e.g. ≥10) may in some cases be a reasonable approach for maintaining the sample size of loci and individuals.

Approximately 20% of the raw reads did not align to the reference genome in either reference-aligned data sets. Raw reads can be discarded during alignment as reference genomes are often incomplete as is the case with the *Choloepus hoffmanni* genome. We suspect that the reference-aligned *full* data set contained considerably more loci than the reference-aligned *reduced* data set (Fig. 2) because many (35–48%) reads were secondary alignments. These secondary reads could have aligned to multiple locations because they contained repetitive

Table 2 Comparison of genotyping error rates for single nucleotide polymorphisms genotyped with and without the correction algorithm, `rxstacks` for a quality score ≥20 at four coverage thresholds (5, 10, 15 and 20) for reference-aligned *full* and *de novo*-assembled data sets. The number of dyads in the uncorrected data was restricted to match the dyads present in the corrected data. A higher number for the sum of positive ranks, that is the sum of ranks assigned to the differences as greater than zero, indicate that the corrected data have lower genotyping error rates

	Sum of positive ranks	Sum of negative ranks	*Z*	*P*
Reference aligned				
Coverage ≥5	120	16	−2.69	<0.01
Coverage ≥10	105	0	−3.30	<0.01
Coverage ≥15	61	17	−1.72	0.08
Coverage ≥20	48	18	−1.33	0.18
De novo-assembled				
Coverage ≥5	124	12	−2.90	<0.01
Coverage ≥10	96	40	−1.45	0.15
Coverage ≥15	74	46	−0.80	0.42
Coverage ≥20	51	40	−0.38	0.70

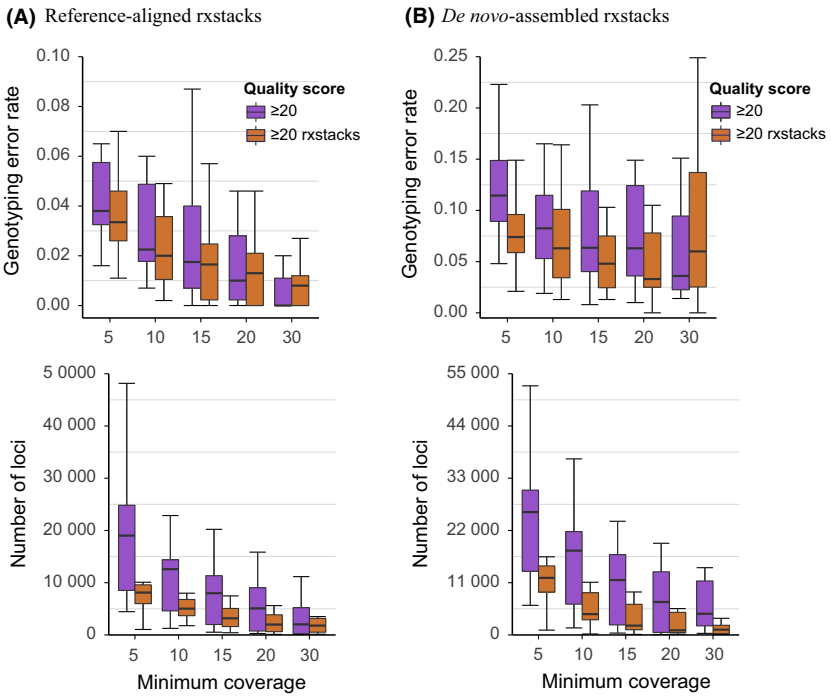

Fig. 5 Genotyping error rates (top panel) and number of loci shared by dyad (bottom panel) as a function of coverage for the `rxstacks` correction algorithm at a quality score ≥20. Box plots represent the median, 25th and 75th percentiles, with whiskers extending to a maximum of 1.5 × IQR. (A) reference-aligned *full* analysis, (B) *de novo*-assembled analysis.

DNA, which can comprise 50% of a genome (Schmid & Deininger 1975). Including secondary alignments may increase the likelihood of calling false SNPs and impact subsequent data analysis (Treangen & Salzberg 2011), but in our case, genotyping error rates were not appreciably different between the reference-aligned *full* and *reduced* data sets (Fig. 2A). As secondary reads had negligible effects on genotyping error rates, and the reference-aligned *full* and *de novo*-assembled data sets included a similar proportion of reads, we limited comparisons of the effect of coverage criteria on genotyping error rates and numbers of loci to these two data sets (i.e. not the reference-aligned *reduced* data set).

The reference-aligned *full* data set yielded lower genotyping error rates than the *de novo*-assembled data set for a given combination of coverage and sequence quality criteria. For example, applying coverage ≥5 yielded median genotyping error rates ≥0.03 and ≥0.11 in reference-aligned *full* and *de novo*-assembled data sets, respectively. Coverage interval analyses for both data sets indicated that loci with the lowest coverage (coverage ≥5 and ≤10) had the highest median genotyping error rates. The coverage interval results suggest that low coverage loci were responsible for the high genotyping error rates when lower coverage thresholds were applied. Increasing coverage reduced error rates to what could be considered low levels in the reference-aligned *full* data set (≤0.01 at coverage ≥30), but error rates remained comparatively high in the *de novo*-assembled data set (≥0.04 at coverage ≥30). Thus, high coverage criteria will likely need to be applied to *de novo*-assembled data sets if low

Table 3 Comparison of mean loci for single nucleotide polymorphisms genotyped with and without the correction algorithm, `rxstacks` for a quality score ≥20 at four coverage thresholds (5, 10, 15 and 20) for reference-aligned *full* and *de novo*-assembled data sets. The number of dyads in the uncorrected data was restricted to match the dyads present in the corrected data. A higher number for the sum of positive ranks, that is the sum of ranks assigned to the differences as greater than zero, indicate that the corrected data have fewer loci

	Sum of positive ranks	Sum of negative ranks	*Z*	*P*
Reference-aligned *full*				
Coverage ≥5	136	0	−3.52	<0.01
Coverage ≥10	105	0	−3.30	<0.01
Coverage ≥15	105	0	−3.30	<0.01
Coverage ≥20	91	0	−3.18	<0.01
De novo-assembled				
Coverage ≥5	130	6	−3.21	<0.01
Coverage ≥10	136	0	−3.52	<0.01
Coverage ≥15	120	0	−3.41	<0.01
Coverage ≥20	91	0	−3.18	<0.01

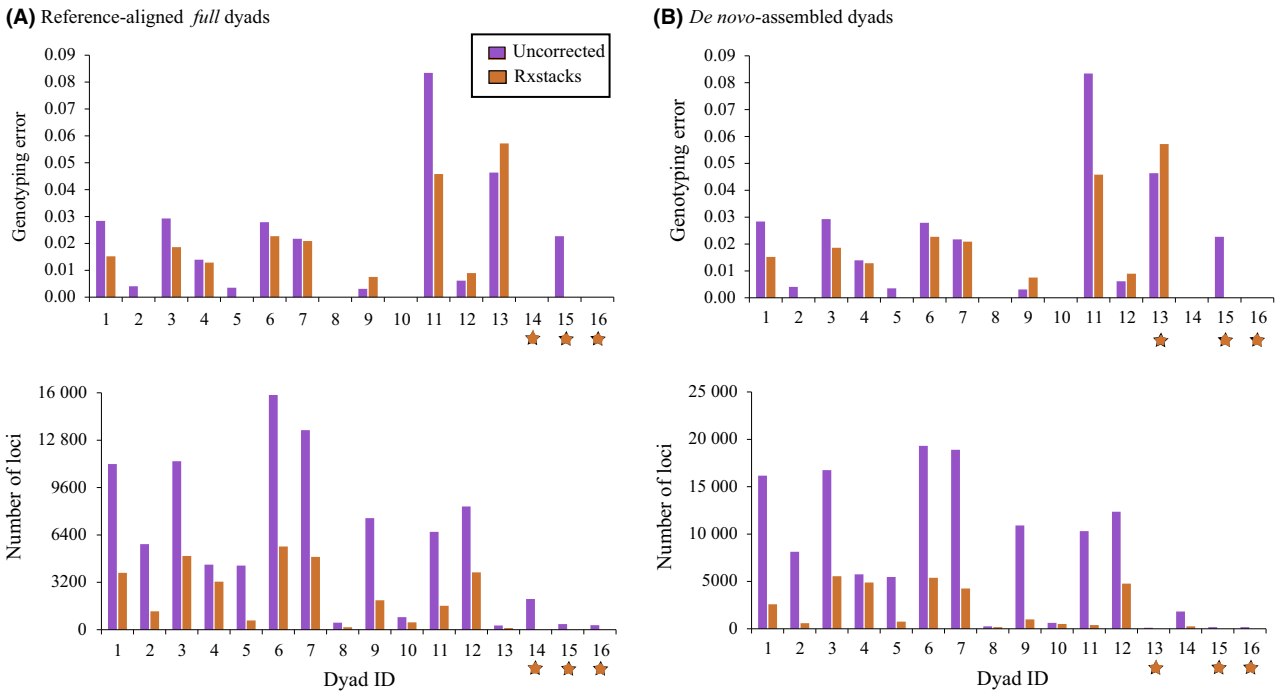

Fig. 6 The differences in genotyping error rate (top panel) and number of shared loci by dyad (bottom panel) without (uncorrected) and with (rxstacks) the correction algorithm, `rxstacks` at quality score ≥20 and coverage ≥20. Stars represent dyads that dropped from the analysis after the correction algorithm was applied. (A) Reference-aligned *full* analysis, (B) *de novo*-assembled analysis.

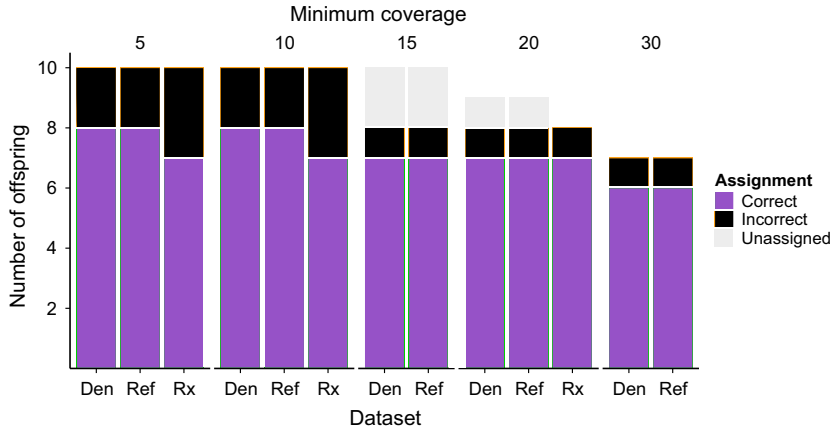

Fig. 7 Number of offspring assigned to either the correct parent, incorrect parent or no parent with 95% confidence for each data set. Data sets are grouped by shared minimum coverage threshold. Minimum coverage ≥30 and ≥20 do not equal ten as some were lost due to low number of loci. One dyad was incorrectly assigned at all coverage thresholds. One dyad was incorrectly assigned at all coverage thresholds. Den = *de novo*-assembled data set; Ref = uncorrected reference-aligned *reduced* data set; Rx = reference-aligned data set with `rxstacks` correction.

genotyping error rates are required. Adopting strict coverage criteria will come at the expense of the number of loci and individuals sampled in both data sets, as we observed approximately 10- and 13-fold decline in the number of loci sampled in the reference-aligned *full* and *de novo*-assembled data sets when coverage was increased from ≥5 to ≥30, respectively.

Sampling more individuals and loci by adopting lower coverage may counteract some of the effects of SNP genotyping errors on ecological inference, particularly with *de novo* assembly (Davey *et al.* 2013). Lowering coverage thresholds has been shown to improve estimates of some population genetic parameters (Buerkle & Gompert 2013; Mastretta-Yanes *et al.* 2015). Although relaxing coverage criteria substantially increased the number of loci in our study, several dyads included in the *de novo*-assembled data set had notably high genotyping error rates at low coverage (e.g. four of 16 dyads had error rates >0.15 at coverage ≥10; Fig. 1). We do not expect using a small sample size inflated genotyping error rates for the *de novo* assembly. Both the *de novo*-assembled and reference-aligned *full* data sets had a similar number of dyads with high genotyping error rates; however, the dyad Mendelian incompatibilities were consistently higher in the *de novo*-assembled data set. Adding more dyads would not influence the overall pattern of higher genotyping errors in the *de novo*-assembled data. Based on the results of our maternity assignments despite higher genotyping error rates, the *de novo*-assembled data set performed on par with the reference-aligned *reduced* data set in a population genetic application. The similar precision may be due to the substantially larger number of loci in the *de novo*-assembled data set; however, if a low genotyping error is required, setting a higher quality and coverage scores for *de novo* assembly may be necessary.

Although the number of sequenced raw reads was similar among individuals, the number of retained reads was highly variable. This variation may have resulted from differences in DNA concentration after the barcode ligation and DNA clean-up (DaCosta & Sorenson 2014). The high level of variation in genotyping error rates among dyads was likely indicative of differences in genotyping error rates among sampled individuals. Two offspring with incorrect maternities (at coverage ≥5 and ≥10) were members of dyads with high genotyping error rates. Samples with high genotyping error rate can be identified and excluded from analyses via resequencing or by checking for Mendelian incompatibilities when parent–offspring data are available. Moreover, removing individuals occurring in dyads with high genotyping error rates may allow for the relaxation of coverage and quality score criteria, in turn, increasing the number of loci and individuals that can be sampled. However, resequencing is often not feasible and parent–offspring data are often not available, such that excluding individuals with consistently low coverage or few genotyped loci may represent the best approach for reducing genotyping error rates (Bradic *et al.* 2013; Franchini *et al.* 2014). Nevertheless, high genotyping error rates in some of the dyads in this study were not directly attributable to sample quality suggesting that 'low-quality' samples may not necessarily need to be excluded from analyses. Alternatively, overall genotyping error rates can be reduced by removing problematic loci rather than excluding individuals. However, while applying a coverage ≥20 lowered the median genotyping error substantially in reference-aligned *full* data set, doing so still led to the

inclusion of some dyads with high genotyping error rates. Adopting a coverage ≥30 removed all dyads with high genotyping errors, in addition to reducing median genotyping error rates to nearly zero, in reference-aligned *full* data sets. Thus, applying relatively strict coverage criteria may be the most effective way to reduce both overall error rates and the prevalence of high-error individuals.

The correction algorithm of `rxstacks` significantly improved genotyping error rates for the reference-aligned *full* data set when low-quality scores and coverage criteria were applied. At higher minimum coverage (≥20), the correction resulted in little improvement because low genotyping error rates had already been achieved. Although the correction improved genotyping error rates when coverage was low, the corrected data set contained 43% fewer loci than the uncorrected data set. Low minimum coverage is often selected to increase the number of loci (Buerkle & Gompert 2013), but this advantage is lost when applying the correction to low coverage data. For example, adopting a coverage ≥10 for the reference-aligned *full* data set without the correction resulted in a lower median genotyping error rate and a greater number of loci than applying the correction at coverage ≥5. Thus, we recommend considering more stringent coverage criteria rather than applying the `rxstacks` correction to low coverage loci, as doing so may result in more loci with acceptable genotyping error rates.

Appropriate coverage thresholds and the effect of genotyping errors will vary by application. For example, relaxed coverage criteria are more likely to miss rare variants than common ones and will thus be more likely to impact studies of demographic history than population genetic structure (Johnson & Slatkin 2008). Different types of genotyping errors can affect population genetic analyses differently. In genomewide association studies that involve case–control populations, classifying a more common homozygote as a rare homozygote is the most deleterious genotyping error (Kang *et al.* 2004). An incorrectly classified rare allele may indicate an association with a trait or disease within the case population. Additionally, high genotyping error rates may result in calling excess homozygous individuals, which may lower estimates of population-level heterozygosity (e. g. F_{st} estimates, Pompanon *et al.* 2005). Not all incorrect genotypes will likely impact population genetic inferences and importantly, we were unable to determine whether estimates of genotyping error rates were a result of heterozygote to homozygote errors or vice versa given the use Mendelian incompatibilities. In general, median genotyping error rates and variation in genotyping error among dyads were lowest with high coverage (≥20 and ≥30), and adopting a conservative strategy may benefit many applications. If relaxed coverage criteria are applied to increase the sample size of loci and individuals, the degree to which the application in question is sensitive to the rate and type of genotyping errors could be assessed (for example) through simulation analyses.

Restriction-enzyme-based NGS methods allow for the genotyping of a large number of SNPs, making them an attractive choice for many applications in molecular ecology (Davey *et al.* 2011). However, our results suggest that care should be given to decisions regarding coverage criteria and quality score when applying these methods. Although our estimates of genotyping error rates apply primarily to RAD-seq analysis (particularly with STACKS), increases in genotyping error rates as coverage decreases are probably the case for RAD-based methods in general. Moreover, we suggest that reporting genotyping error rates in RAD-based studies become standard (Pompanon *et al.* 2005), as has been called for in microsatellite-based studies (Lampa *et al.* 2013). Resequencing a subset of individuals or characterizing Mendelian incompatibilities is necessary to estimate genotyping error rates; if both approaches are infeasible, we recommend applying stringent quality scores and coverage criteria to reduce error rates. Equally important, we recommend that the trade-off between number of loci and genotyping error rates be considered in the design phase of the study rather than in downstream analyses.

Acknowledgements

We thank Gustavo Gutierrez for assistance with permits and Geovanny Herrera for fieldwork. This work was supported by NSF grant #1257535 to M.Z.P., J.N.P. and P.J.P. Sloth field work was conducted as stipulated and authorized by IACUC protocol A01424 issued by the University of Wisconsin-Madison. We also thank the University of Wisconsin-Madison for support to publish this manuscript.

References

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

Bansal V (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Bradic M, Teotónio H, Borowsky RL (2013) The population genomics of repeated evolution in the blind cavefish *Astyanax mexicanus*. *Molecular Biology and Evolution*, **30**, 2383–2400.

Buerkle AC, 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.

Chain PSG, Grafham DV, Fulton RS *et al.* (2009) Genome project standards in a new era of sequencing. *Science (New York, N.Y.)*, **326**, 236–237.

Chen N, Van Hout CV, Gottipati S, Clark AG (2014) Using Mendelian Inheritance to improve high throughput SNP discovery. *Genetics*, **198**, 847–857.

Corrales C, Höglund J (2012) Maintenance of gene flow by female-biased dispersal of Black Grouse *Tetrao tetrix* in northern Sweden. *Journal of Ornithology*, **153**, 1127–1139.

DaCosta JM, Sorenson MD (2014) Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PLoS ONE*, **9**, e106713.

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.

Douglas JA, Skol AD, Boehnke M (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics*, **70**, 487–495.

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

Franchini P, Fruciano C, Spreitzer ML *et al.* (2014) Genomic architecture of ecologically divergent body shape in a pair of sympatric Crater Lake cichlid fishes. *Molecular Ecology*, **23**, 1828–1845.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, **27**, 489–496.

Gnerre S, MacCallum I, Przybylski D *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences, USA*, **108**, 1513–1518.

Haaland ØA, Skaug HJ (2013) Estimating genotyping error rates from parent–offspring dyads. *Statistics & Probability Letters*, **83**, 812–819.

Hao K, Li C, Rosenow C, Hung Wong W (2004) Estimation of genotype error rate using samples with pedigree information—an application on the GeneChip Mapping 10K array. *Genomics*, **84**, 623–630.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, R32.

He Z, Li X, Ling S *et al.* (2013) Estimating DNA polymorphism from next generation sequencing data with high error rate by dual sequencing applications. *BMC Genomics*, **14**, 535.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Jezkova T, Riddle BR, Card DC *et al.* (2015) Genetic consequences of post-glacial range expansion in two co-distributed rodents (genus *Dipodomys*) depend on ecology and genetic locus. *Molecular Ecology*, **24**, 83–97.

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

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.

Kai W, Nomura K, Fujiwara A *et al.* (2014) A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. *BMC Genomics*, **15**, 233.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099–1106.

Kang SJ, Gordon D, Finch SJ (2004) What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology*, **26**, 132–141.

Lamichhaney S, Barrio AM, Rafati N *et al.* (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences, USA*, **109**, 19345–19350.

Lampa S, Henle K, Klenke R, Hoehn M, Gruber B (2013) How to overcome genotyping errors in non-invasive genetic mark–recapture population size estimation—a review of available methods illustrated by a case study. *The Journal of Wildlife Management*, **77**, 1490–1511.

Larson WA, Seeb JE, Pascal CE, Templin WD, Seeb LW (2014) Single-nucleotide polymorphisms (SNPs) identified through genotyping-by-sequencing improve genetic stock identification of Chinook salmon (*Oncorhynchus tshawytscha*) from western Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, **71**, 698–708.

Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, **21**, 952–960.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, **18**, 1851–1858.

Li R, Yu C, Li Y *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Lozier JD (2014) Revisiting comparisons of genetic diversity in stable and declining species: assessing genome-wide polymorphism in North American bumble bees using RAD sequencing. *Molecular Ecology*, **23**, 788–801.

Malhis N, Jones SJM (2010) High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics*, **26**, 1029–1035.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

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

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, **7**, e37558.

Palti Y, Gao G, Liu S *et al.* (2015) The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources*, **15**, 662–672.

Peery MZ, Pauli JN (2012) The mating system of a 'lazy' mammal, Hoffmann's two-toed sloth. *Animal Behaviour*, **84**, 555–562.

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.

Pujolar JM, Jacobsen MW, Frydenberg J *et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources*, **13**, 706–714.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Recknagel H, Elmer KR, Meyer A (2013) A hybrid genetic linkage map of two ecologically and morphologically divergent midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3: Genes|Genomes|Genetics*, **3**, 65–74.

Ruegg K, Anderson EC, Boone J, Pouls J, Smith TB (2014) A role for migration-linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, **23**, 4757–4769.

Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.

SAS Institute Inc. (2013) *What's New in SAS® 9.4*. SAS Institute Inc., Cary, North Carolina.

Saunders IW, Brohede J, Hannan GN (2007) Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*, **90**, 291–296.

Schmid CW, Deininger PL (1975) Sequence organization of the human genome. *Cell*, **6**, 345–358.

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, **13**, 36–46.

Wagner CE, Keller I, Wittwer S *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, **38**, 95–109.

Zhou X, Xia Y, Ren X *et al.* (2014) Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics*, **15**, 351.

E.D.F., J.N.P., P.J.P. and M.Z.P. designed research; E.D.F. and B.N.R. performed laboratory work and data analysis; and E.D.F., J.N.P., P.J.P., B.N.R. and M.Z.P. wrote the manuscript.

Data accessibility

SAS script to calculate genotyping error and SNP input files: DRYAD entry doi: 10.5061/dryad.8g470.

Supporting Information

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

Table S1 Total number of raw and retained reads by individuals, sample quality and the sequencing library for quality score ≥10.

Table S2 Total number of raw and retained reads by individuals, sample quality and the sequencing library for quality score ≥20.

Table S3 Total number of raw and retained reads by individuals, sample quality and the sequencing library for quality score ≥30.

Fig. S1 Figure of median genotyping error for reference-aligned *full* reads using all 16 dyads vs. dataset with the three low-quality dyads removed.

Fig. S2 Reference-aligned *full* dataset changes relative to that observed at minimum coverage (≥5) and quality score (≥10).

Fig. S3 *de novo*-assembled dataset changes relative to that observed at minimum coverage (≥5) and quality score (≥10).

Fig. S4 Figure of median genotyping error for *de novo*-assembled dataset using all 16 dyads vs. dataset with the three low-quality dyads removed.

Fig. S5 `rxstacks` corrected dataset changes relative to that observed at minimum coverage (≥5) and quality score (≥10) for reference-aligned *full* and *de novo*-assembled datasets.

Appendix S1 Detailed methods for laboratory protocols, *de novo* assembly parameter tests and parentage assignments.