MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) **15**, 329–336 doi: 10.1111/1755-0998.12314

Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy

M. M. Y. TIN,* F. E. RHEINDT,† E. CROS† and A. S. MIKHEYEV*
*Ecology and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan, †Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore

Abstract

RAD-tag is a powerful tool for high-throughput genotyping. It relies on PCR amplification of the starting material, following enzymatic digestion and sequencing adaptor ligation. Amplification introduces duplicate reads into the data, which arise from the same template molecule and are statistically nonindependent, potentially introducing errors into genotype calling. In shotgun sequencing, data duplicates are removed by filtering reads starting at the same position in the alignment. However, restriction enzymes target specific locations within the genome, causing reads to start in the same place, and making it difficult to estimate the extent of PCR duplication. Here, we introduce a slight change to the Illumina sequencing adaptor chemistry, appending a unique four-base tag to the first index read, which allows duplicate discrimination in aligned data. This approach was validated on the Illumina MiSeq platform, using double-digest libraries of ants (*Wasmannia auropunctata*) and yeast (*Saccharomyces cerevisiae*) with known genotypes, producing modest though statistically significant gains in the odds of calling a genotype accurately. More importantly, removing duplicates also corrected for strong sample-to-sample variability of genotype calling accuracy seen in the ant samples. For libraries prepared from low-input degraded museum bird samples (*Mixornis gularis*), which had low complexity, having been generated from relatively few starting molecules, adaptor tags show that virtually all of the genotypes were called with inflated confidence as a result of PCR duplicates. Quantification of library complexity by adaptor tagging does not significantly increase the difficulty of the overall workflow or its cost, but corrects for differences in quality between samples and permits analysis of low-input material.

Keywords: genotyping, methodology, next-generation sequencing, RAD-seq, RAD-tag

Received 29 November 2013; revision received 6 August 2014; accepted 6 August 2014

Introduction

Reduced representation sequencing, particularly restriction site-associated DNA sequencing, known as RAD-tag or RAD-seq, has allowed large-scale cost-effective genotyping of a wide range of model and nonmodel organisms (Miller *et al.* 2007; Baird *et al.* 2008). New methodological advances that improve the performance of this technique appear regularly in the literature (Peterson *et al.* 2012; Wang *et al.* 2012; Stolle & Moritz 2013). RAD-tag analysis benefits from existing bioinformatic pipelines for SNP processing, and from new software developed specifically for this application (Catchen *et al.* 2011; Chong *et al.* 2012). The chemistry of RAD-tags, which relies on PCR amplification of the template, can introduce a number of significant sources of error into next-generation sequencing data (Kozarewa *et al.* 2009). In particular, because PCR duplicates arise from the same template molecule and are statistically nonindependent, duplicate reads will artificially inflate the confidence of genotype calls at a site. For example, ten reads all resulting from the same template molecule by PCR duplication will provide a genotype caller with sufficient evidence to call a homozygote allele, although in reality there is only one data point. In shotgun sequencing, these problems are dealt with either by eliminating PCR entirely, or by bioinformatically filtering out reads that start at the same location, and are possible duplicates. Both of these strategies are currently not applicable to RAD-tags.

There are two classes of errors that can be introduced by PCR. First, the polymerase can introduce copying

Correspondence: Alexander S. Mikheyev, Fax: 81-098-966-8889; E-mail: sasha@homologo.us

errors, which are propagated through future amplification cycles, and may lead to incorrect base calls. Second, and perhaps most importantly for RAD-tag data, sequencing depth varies greatly across loci, resulting in some with relatively lower coverage than others. These low-coverage loci are particularly vulnerable to PCR duplicate inflation, which may result in some of them being called confidently, when in reality they do not contain enough information for a genotype call.

Consequently, PCR duplicate removal is important for accurate SNP calls from short read-based sequence data and is a recommended preprocessing step in SNP calling pipelines (Auwera *et al.* 2013). However, because reads from homologous RAD-tag loci share the same start position, it is currently impossible to distinguish PCR duplicates using single-end sequencing in RAD-tag data. This problem can be mitigated using paired-end sequencing of mechanically sheared DNA libraries (Baxter *et al.* 2011; Davey *et al.* 2013). Unfortunately, on the commonly used Illumina platforms, paired-end sequencing is significantly more expensive relative to single-end sequencing. Furthermore, even paired-end sequencing fails to remove duplicates for popular RAD-tag chemistries, such as double-digest, or IIB-type tags, where restriction sites flank both ends of a read (Peterson *et al.* 2012; Wang *et al.* 2012).

To mitigate these difficulties, we developed a simple extension of Illumina's combinatorial sequencing chemistry, where a four-base sequence of equally mixed DNA nucleotides is appended to the fixed sequence of the first index read (Fig. 1). This approach is similar to those that have been used for deep high-accuracy sequencing (Jabara *et al.* 2011; Kinde *et al.* 2011; Schmitt *et al.* 2012). From this pool of possible adaptor sequences, a unique barcode is ligated to every template molecule. This barcode is copied by PCR, allowing the identification of all resulting amplicons after alignment. The four-nucleotide sequence produces 256 possible variant sequences, which we term 'adaptor tags', which are sequenced along with the rest of the Illumina barcode. Adding the barcode does not alter the library preparation protocol and costs only four additional sequencing cycles, making it easy to implement for routine use. We validate this approach using a series of MiSeq experiments using samples with known genotypes, showing that duplicate removal can eliminate sample-to-sample variation in genotype accuracy calls. We then show the particular importance of duplicate detection for the analysis of low-complexity libraries (with few starting template molecules) made from degraded samples obtained from museum specimens. Low-complexity libraries can occur in more typical samples due to reaction failure and variability DNA quality.

Fig. 1 Overview of adaptor and sequencing primer configuration. The adaptor sequences (blue and yellow) are ligated onto the sticky ends of the target DNA molecule (orange). Sequences filled in later by PCR are shown in light grey. The EcoRI adaptor index has a run of degenerate nucleotides that constitute the adaptor tag. Black arrows show the placement of the custom sequencing primers, which fully overlap the synthetic adaptor sequences, while indexing primer positions are shown in blue.

Materials and methods

Oligonucleotide design and synthesis

We conducted three experiments, two 'controls' involving double digests of samples with known genotypes and a single-enzyme digestion of a low-input badly degraded museum sample. The control digestions were made using the combination of EcoRI and MseI, while the museum specimen was digested using NotI. Consequently, different adaptors were used for the two experiment types.

For the EcoRI bottom strands that contain the degenerate adaptor tag sequence, the four standard bases were hand mixed at equal frequencies during synthesis by Integrated DNA Technologies, USA. Sequencing primers and bottom strand oligos were PAGE purified. For the museum specimen libraries, the four degenerate bases following the index 1 sequence of the PCR primers were also hand mixed at equal frequencies but with HPLC purification and sodium salt exchange. HPLC purification was used to increase the yield of oligonucleotides after purification, relative to PAGE. The other oligonucleotides were desalted. The oligonucleotide and custom primer sequences can be found in Table 1.

Control experiments

Ant samples. We used five worker larvae from the Hawaii population of the little fire ant *Wasmannia auropunctata*. This species has an unusual reproductive system where males and females are clonal and produce sterile worker offspring sexually; in Hawaii, there is just one pair of queen and male clones (Fournier *et al.* 2005a; Mikheyev *et al.* 2009). We sequenced both parental genomes as a part of other ongoing work, allowing us to predict offspring genotypes. Genome sequencing data are available from the authors upon request. The offspring were not the direct progeny of the sequenced

Table 1 Custom nucleotide sequences used in this study in 5′ -> 3′ orientation. We used combinatorial Hamming (7,4) barcodes (Bystrykh 2012), with a unique combination for each of the ten libraries made. The adaptors are prepared by separately hybridizing the restriction enzyme primer pairs, and using them in the restriction/ligation reactions. Figure 1 shows the arrangement of the primer pairs in the sequencing construct

Name	Sequence
EcoRI top	Phosphate-AATTGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
EcoRI bottom[1,2]	CAAGCAGAAGACGGCATACGAGATNNNNXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
MseI top	CCCTACACGACGCTCTTCCGATC
MseI bottom[1]	Phosphate-TAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTXXXXXXXGTGTAGATCTCGGTGGTCGCCGTATCATT
MseI read 1 sequencing primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAA
EcoRI read 2 sequencing primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAATTC

[1]XXXXXXX: specific barcode sequence.
[2]NNNN: degenerate adaptor tag sequence (an equal mixture of A, C, G and T nucleotides at each site).

clones, but of clones with the same microsatellite profile (Fournier *et al.* 2005b), so some genetic differences between the resequenced genomes and those of the test samples were possible.

Yeast samples. We used derivatives of commonly used SK1 and SC228c strains of *Saccharomyces cerevisiae*, which differ from each other in about 0.7% of their genome (Heck *et al.* 2006). Libraries were prepared from pure cultures of either haploid parents and of three diploid crosses between the two strains. As in the case of the ants, diploids should be heterozygous at all loci.

Bird samples. We used three individuals of *Mixornis gularis* (previously *Macronous gularis* or *Macronus gularis*), an extremely common avian insectivore from forest and edge habitat throughout South-East Asia and immediately adjacent regions. These individuals were collected 70, 57 and 39 years ago, respectively (Table 3), and their dry skins are deposited as museum vouchers in the Raffles Museum of Biodiversity Research in Singapore. DNA was extracted from their toe pads. Although highly degraded in DNA content, toe pads are generally the best source of DNA from old museum bird skins.

Library preparation. DNA was extracted from ants using Qiagen Micro kits, producing a total yield of 38 ± 9.0 ng per sample. Dr. GenTLE (from Yeast) High Recovery kit (Takara) was used for DNA extraction from yeast cultures. The SC228c culture had a small cell pellet, so DNA yield (31 ng) was lower compared to the SK1 culture (802 ng) and diploid offspring (average 869 ng). DNA concentration was measured with Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). Every sequencing library was prepared separately with a unique combination of barcodes.

The digestion and ligation reactions were performed sequentially. The 10 μL digestion reaction contained 1 μL of 10× T4 ligase buffer (NEB), 1 μL of 0.5 M sodium chloride (NaCl), 0.5 μL of 1 mg/μL bovine serum albumin (NEB), 0.25 μL of 20 units (U)/μL EcoRI (NEB), 0.1 μL of 10 U/μL MseI (NEB) and 5 ng genomic DNA. The reaction was performed at 37 °C for 1 h and then heat inactivated at 65 °C for 20 min. Then, a 4 μL ligation mix consisting of 0.4 μL of 10× T4 ligase buffer (NEB), 0.2 μL of 400 U/μL of T4 DNA ligase (NEB), 1 μL of 5 μM EcoRI adaptor (EcoRI top and bottom) and 1 μL of 50 μM MseI adaptor (MseI top and bottom) were added to the digestion mixture. The ligation reaction was incubated at room temperature for 1 h.

Size selection was performed on ligation products, and purified products were used as templates for PCR enrichment of markers. Dynabeads MyOne Carboxylic Acid (Invitrogen) were washed twice in EB buffer (Qiagen) and then resuspended in the same volume of EB buffer. Each sample was adjusted to 50 μL with MilliQ water before size selection. In the first selection step, 100 μL of 13% PEG-6000 (with 0.9 M NaCl and 10 mM Tris, pH 6) and 10 μL washed Dynabeads were added to the library and resuspended. The mixture was incubated for 5 min. The tube was then placed on a magnetic stand for 5 min. 150 μL supernatant was transferred to a new tube while the beads were discarded. In the second selection, 100 μL of 13.5% PEG-6000 (with 0.9 M NaCl and 10 mM Tris, pH 6) and 10 μL washed Dynabeads were added to the supernatant and mixed. The mixture was incubated for 5 min followed by bead separation on the magnetic stand. This time, the supernatant was discarded and the beads were saved. The beads were washed twice with 70% ethanol (with 10 mM Tris, pH

6) and dried for 5 min. The tubes were then taken off the magnetic stand, and DNA was eluted from the beads by resuspending them in 40 μL EB. After 5-min incubation, beads were separated from the DNA solution on the magnetic stand. The size-selected ligation products resulted in libraries ranging between 300 and 700 bp after PCR amplification.

The 30 μL PCR contained 1× Phusion HF buffer (Thermo Scientific), 200 μM dNTP (Promega), 0.5 μM PCR 1 primer: 5′-AATGATACGGCGACCACCGA, 0.5 μM PCR 2 primer: 5′-CAAGCAGAAGACGGCAT ACGA, 0.3 μL of 2 U/μL Phusion DNA polymerase (Thermo Scientific) and 5 μL size-selected ligation products. The PCR was carried out with the following conditions: initial denaturation at 98 °C for 30 s, with either 18 or 25 cycles of denaturation at 98 °C for 10 s, 72 °C for 30 s, followed by final extension at 72 °C for 5 min. The PCR products were purified using Dynabeads MyOne Carboxylic Acid. The PCRs were adjusted to 50 μL with MilliQ water before purification. 100 μL of 15% PEG-6000 (with 0.9 M NaCl and 10 mM Tris, pH 6) and 10 μL of the washed Dynabeads were added to each sample and resuspended. The mixtures were incubated for 5 min. The tubes were then placed on a magnetic stand for 5 min. The supernatant was discarded, and the beads were washed twice with 70% ethanol (with 10 mM Tris, pH 6) and dried for 5 minutes. The tubes were then taken off the magnetic stand, and DNA was eluted from the beads by resuspending them in 15 μL EB buffer. After 5-min incubation, beads were separated from the DNA on the magnetic stand. The eluant contained the purified library.

DNA concentration was measured with the Quant-iT PicoGreen dsDNA Assay Kit. Equal amounts of DNA, normalized to the sample with the lowest total amount, were pooled and concentrated with an Amicon Ultra-0.5 column (Millipore). The concentrated pool was subjected to 1% agarose gel electrophoresis, and DNA in the range 400–500 bp was excised from the gel and purified with a MinElute Gel Extraction Kit (Qiagen). Libraries were sequenced on the Illumina MiSeq platform using paired-end mode on a 50-cycle flow cell using custom sequencing adaptors (Table 1).

Yeast libraries were prepared according to the same procedures with 14 and 18 cycles PCR amplification without gel extraction on the pooled library as the yeast genome is small with much fewer fragments after restriction digestion. Lower PCR cycles were used for the same reason. At 25 cycles, discrete bands were observed after agarose gel electrophoresis of the PCR products. This indicated overamplification of a small number of fragments had occurred, and cycle number should be reduced. The libraries were sequenced on the Illumina MiSeq for 50 cycles in single-end mode.

Bioinformatic analysis of control experiments

Ant RAD-tag libraries. Variants found in parental genotypes (a diploid queen and a haploid male) were filtered to include only high-quality sites that were homozygous and different between the male and female clones. These should have been heterozygous in the offspring. We used a 99.9% quality-filtering threshold for female SNPs and chose only those that were homozygous. The male genome was sequenced using Roche 454 technology and used as a reference for resequencing the female genome, so the quality of its bases could not be accurately ascertained.

Read sequence IDs were modified to include the four-base adaptor tag. The paired-end reads were then mapped to the *W. auropunctata* reference genome using bowtie2, and duplicate reads were removed using a custom script, keeping the read with the highest mean quality. From this point, we ran the Genome Analysis Toolkit (GATK) pipeline, including base quality recalibration (McKenna *et al.* 2010). This analysis was conducted separately on libraries with and without duplicate removal. Output files were scored based on whether a genotype predicted from parental genomes was called correctly (as a heterozygote). Effects of duplicate removal, PCR cycle number, genotype quality and DNA extract on the probability of correctly calling a genotype were assessed using logistic regression.

Yeast RAD-tag libraries. Most of the scripts used for ants were slightly modified for yeast, which had a different sequencing configuration. Single-end reads were mapped to the S288c assembly (Goffeau *et al.* 1996). Instead of GATK, we used SAMtools (Li *et al.* 2009) for SNP calling to diversify our analysis. All samples were amplified to 14 and 18 cycles; we used only loci where the 14- and 18-cycle genotypes were concordant for the parental clones. The rest of the analysis remained the same as in the ants.

Low-complexity libraries from museum specimens

Samples and DNA extraction. We obtained toe pads from *M. gularis* specimens (Museum tissue numbers YPM 19536; YPM 47167; YPM 91434 collected in 1944, 1957 and 1975, respectively) from The Peabody Museum of Natural History (Yale University, Connecticut, USA). The extraction process was carried out in a separate room under a designated Biological Safety Cabinet Class-II with equipment exclusively dedicated to the toe pad extractions to prevent any contaminations with fresh tissues or PCR products. First, the samples were washed twice in 100% ETOH to remove any PCR inhibitor. They were then rehydrated overnight in 200 μL of 10 nM

Tris-HCL (pH 8.0) (Boessenkool *et al.* 2009). We then extracted the DNA using the Exgene Clinic SV mini kit (GeneAll Biotechnology CO., LTD, Seoul, Korea). We followed the manufacturer's protocol G for animal tissue. However, to improve the DNA yield, we modified it and added additional steps. First, 40 μL of proteinase K was added instead of 20 μL (Fulton *et al.* 2012) and the samples were incubated at 56 °C until completely digested. We added 20 μL of proteinase K on any subsequent day needed for the digestion to be completed (Fulton *et al.* 2012). Additionally, before starting the DNA elution step, we incubated the columns with the lids opened at 56 °C for 30 min (Lijtmaer *et al.* 2012). Finally, we eluted the DNA with two 100 μL volumes of double-distilled water.

Library preparation and sequencing. The library preparation procedure was modified from the museum RAD-tag pipeline developed by Tin *et al.* (2014), who found that libraries made from museum samples typically showed low complexity. Specific reaction conditions used in this study can be found as a Appendix S1 for Supporting information.

Bioinformatic analysis of low-complexity libraries. Read names were prepended with adaptor tag sequences and assembled using Stacks (Catchen *et al.* 2011). Reads within a stack that shared the same adaptor tag sequence were then flagged as duplicates. We then removed duplicate reads form the raw data, keeping the read with the highest quality. To examine the effect of duplicates on the outcome of the analysis, Stacks was rerun using original samples with duplicates, and also data with the duplicates removed as separate samples.

Results

Control experiments

Sequencing of adaptor tags allowed for identification of duplicate reads despite their identical read position by filtering reads that resulted from the same template molecule. For ants, the RAD-tag data set contained 2525 and 2331 high-quality SNPs that were polymorphic between the male and female clones, in raw and duplicate-filtered libraries, respectively. For yeast, although the genome was much smaller, there was a higher degree of genomic divergence producing more SNPs overall (3830 total and 2774 duplicate filtered). Statistics on the number of sequenced reads, mapping percentages and the coverages can be found in Table S1 for Supporting Information.

We used two controls with different bioinformatic packages for SNP calling. Yet, they both produced remarkably similar distributions of genotype call quality scores with tails of low- and of high-quality calls (Fig. 2a). Low GQ calls (<20, corresponding to a 99% chance of calling the correct genotype) were often

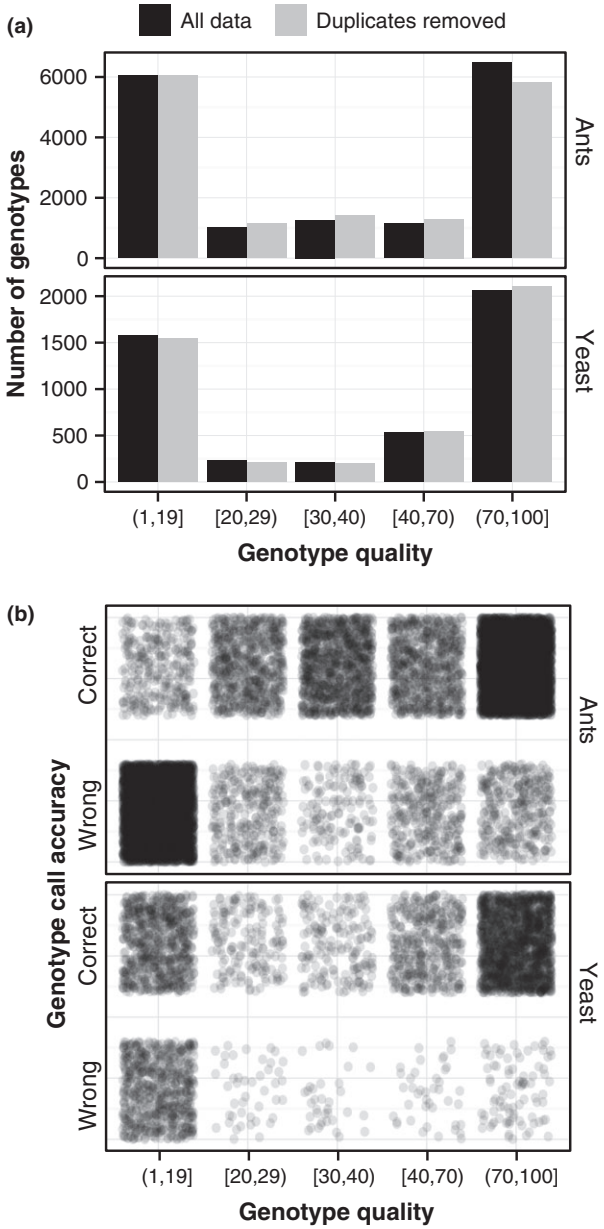

Fig. 2 Relationship between called genotype quality (GQ) and call accuracy. (a) Distribution of genotype calls in the two control experiments, with and without duplicate removal. Removing duplicates slightly affected the overall distribution of quality scores, presumably making them more accurate. (b) Genotype call accuracy as a function of GQ, with points jittered to minimize overlap. Call accuracy was low below 20 GQ, but remained at a similar level above that. Therefore, duplicate removal affects the distribution of GQ, a quality that determines the accuracy of read calls.

inaccurate, although the chances of calling the correct genotype remained approximately the same above that threshold (Fig. 2b). Taking GQ greater than or equal to 20, duplicate removal always increased call accuracy rates (Table 2). This effect was strongly significant when all factors were considered together using logistic regression, providing a modest but significant increase in per genotype call accuracy (Table 2). These effects were robust over the entire range of GQ cut-offs up to the maximum of 99, reliably providing a few percentage points increase in the calling rate accuracy.

Although after extraction libraries were prepared separately for the low and high PCR cycle experiments, there was a strong correlation between the percentages of duplicate reads in the two treatments (Spearman $r = 0.71$, $P = 0.028$, both control experiments pooled), suggesting that the quality of the extract had a strong effect on library complexity. To investigate whether removing duplicates may correct for sample-to-sample variability, we reran the logistic regression on the control data separately for subsets with and without duplicates, using the Wald test to determine whether all individual coefficients were zero. For ants, there was a strong effect of sample on the likelihood of correctly calling genotype when duplicates were included in the data ($X^2 = 133.6$, d.f. = 4, $P = 3.2 \times 10^{-28}$). When duplicates were removed, this effect disappeared ($X^2 = 2.6$, d.f. = 4, $P = 0.64$). For yeast, once the data were split, the sample effect was not significant in either case.

Low-complexity libraries

Removing duplicates from the museum libraries reduced coverage by three- to fivefold per library, as most of the data resulted from read duplicates (vs. 45 and 63% in the ant and yeast libraries, respectively, Table S1). Consequently, duplicate removal had major effects on the ability of Stacks to call genotypes, invalidating all but a few genotype calls (Table 3).

Discussion

Our method successfully detected PCR duplicates in double-digest RAD-tag experiments, where conventional methods for counting duplicates by identification of shear points would not have been an option. This correction led to a significant increase in the likelihood of correctly calling a heterozygous genotype, which is the most difficult genotype to call, as it requires a balanced number of reads from both alleles. Although the improvements were modest in the control experiments, libraries that were prepared from poor-quality DNA were shown to contain mostly PCR duplicates. Without duplicate removal, these libraries produced thousands of confident, but possibly incorrect genotype calls, almost all of which did not hold up once duplicates were removed. False SNPs are particularly problematic for old DNA, where DNA degradation introduces numerous mutations (Pääbo *et al*. 2004). Finally, duplicate removal allowed us to correct for significant sample-to-sample variation in genotype call accuracy, which was likely caused by variability in extract quality.

Assuming that degenerate nucleotides are well mixed, the probability of detecting a false positive at a given

Table 3 Summary statistics for museum bird libraries. In each category, the top and bottom rows give values with and without duplicates, respectively. The apparent yield was much greater in the original data set, leading to potentially incorrect genotype inference

	YPM 19536	YPM 47167	YPM 91434
Called	1554	4916	8041
genotypes	67	49	21
Reads mapped	107 725 (24%)	215 771 (39%)	232 454 (45%)
(%)	32 376	49 363	46 971
Coverage	6.38	7.67	8.30
	1.92	1.75	1.68

Table 2 Logistic regression of genotype call accuracy on duplicate removal, PCR cycle and genotype quality score. Significant effects are highlighted in ***bold italic***. The number of correctly called genotypes increased from 85.8% to 89.3% for the ant data, and from 92.4% to 94.3% for the yeast data. Adding 8 and 4 PCR cycles to ant and yeast libraries, respectively, had no effect on the rate of number of additional duplicates. The logistic regression also included effects of individual samples on genotype call accuracy, but they are not presented here. Instead, see Results text for a separate and more comprehensive analysis of sample effect

	Ants		Yeast	
Coefficient	95% C.I.	*P*-value	95% C.I.	*P*-value
Intercept	−0.03, 0.26	0.13	***−0.03, 0.26***	***4.4×10^{-5}***
Duplicates removed	***0.34, 0.52***	***$< 2 \times 10^{-16}$***	***0.34, 0.52***	***0.0046***
PCR cycles	−0.05, 0.13	0.35	−0.05, 0.36	0.14
Genotype quality	***0.02, 0.03***	***$< 2 \times 10^{-16}$***	***0.02, 0.03***	***$< 2 \times 10^{-16}$***

locus is equivalent to the classical birthday problem in probability theory and is distributed as $n!\binom{256}{n}/256^n$, where n is the read depth (Gorroochurn 2012). False-positive rates will not be significant at moderate read depths. At high read depths, where there will be many false positives, adequate information for accurate base calls will still remain. If a lower false-positive rate is desired, the adaptor tag can easily be extended by one or more additional degenerate nucleotides.

The number of duplicates and call accuracy were both related to DNA quality and may not be knowable a priori. For instance, although our DNA extracts were prepared in the same way and had essentially the same quality control metrics, they nonetheless varied in the number of PCR duplicates by about 17% within the same number of PCR cycles (Table S1). Although all steps of library preparation were performed separately for the low- and high-cycle experiments, the number of PCR duplicates between them was strongly correlated with the identity of the starting DNA extract. The sample effect was not due to genetic heterogeneity in the samples used, as the ants were full siblings sharing 75% of their DNA (ants are haplodiploid), and the yeast were genetically identical clones. This suggests that PCR duplicate number may vary between samples and that it should be accounted for, lest there be sample-to-sample variation in PCR-driven bias.

Although paired-end sequencing of double-digest fragments is typically much more expensive than single-end sequencing, it can be employed if more sequence is desired. In this case, several bases of the read are consumed by sequencing the restriction sites. Using a custom sequencing primer that includes the restriction site, we were able to avoid this wasted sequence, more than offsetting the loss of four bases for sequencing the adaptor tag.

In our experiment, we used a small amount of input DNA (5 ng), which is orders of magnitude less than typically used (*e.g.* Barchi *et al*. 2011; Scaglione *et al*. 2012; Palaiokostas *et al*. 2013). For some of our samples, which were insect larvae less than 2 mm long, obtaining substantially more DNA is not possible. Although such small DNA inputs require relatively large numbers of PCR amplification cycles, by quantifying PCR duplicates, it is possible to control for the extent of this bias, even at 25 cycles, when the number of duplicates becomes significant. The ability to identify PCR duplicates even with a large number of PCR cycles can allow RAD-tags to be used for increasingly small amounts of input material, potentially expanding this technique's utility to novel taxa or samples. The ability to control for variation in complexity between samples should be useful for any experiment.

Acknowledgements

We thank Steven D. Aird for editing the manuscript, and Misato Okamoto and Yutaka Watanabe for providing the ant and yeast samples, respectively. We are grateful to Miguel Grau-Lopez for carrying out the Stacks analysis. This project emerged after a fruitful discussion with Jeff Hussmann and Carly D. Kenkel at the Okinawa Integrative Biology Course 2013. Three anonymous reviewers made excellent comments on earlier versions of this manuscript.

References

Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **43**, 11.10.1–11.10.33.

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

Barchi L, Lanteri S, Portis E *et al.* (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, **12**, 304.

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.

Boessenkool S, Austin JJ, Worthy TH *et al.* (2009) Relict or colonizer? Extinction and range expansion of penguins in southern New Zealand. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 815–821.

Bystrykh LV (2012) Generalized DNA barcode design based on Hamming codes. *PLoS ONE*, **7**, e36852.

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

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

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

Fournier D, Estoup A, Orivel JEROM *et al.* (2005a) Clonal reproduction by males and females in the little fire ant. *Nature*, **435**, 1230–1234.

Fournier D, Foucaud J, Loiseau A *et al.* (2005b) Characterization and PCR multiplexing of polymorphic microsatellite loci for the invasive ant Wasmannia auropunctata. *Molecular Ecology Notes*, **5**, 239–242.

Fulton TL, Wagner SM, Shapiro B (2012) Case study: recovery of ancient nuclear DNA from toe pads of the extinct passenger pigeon. In: *Ancient DNA* (eds Shapiro B, Hofreiter M), pp. 29–35. Humana Press.

Goffeau A, Barrell BG, Bussey H *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546–563.

Gorroochurn P (2012) *Classic Problems of Probability*. John Wiley & Sons, Hoboken, NJ.

Heck JA, Argueso JL, Gemici Z *et al.* (2006) Negative epistasis between natural variants of the Saccharomyces cerevisiae MLH1 and PMS1 genes results in a defect in mismatch repair. *Proceedings of the National Academy of Sciences, USA*, **103**, 3256–3261.

Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences, USA*, **108**, 20166–20171.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel

sequencing. *Proceedings of the National Academy of Sciences, USA*, **108**, 9530–9535.
Kozarewa I, Ning Z, Quail MA *et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, **6**, 291–295.
Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Lijtmaer DA, Kerr KCR, Stoeckle MY, Tubaro PL (2012) DNA barcoding birds: from field collection to data analysis. In: *DNA Barcodes: Methods and Protocols* (eds Kress WJ, Erickson DL), pp. 127–152. Springer Science+Business Media.
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.
Mikheyev A, Bresson S, Conant P (2009) Single-queen introductions characterize regional and local invasions by the facultatively clonal little fire ant Wasmannia auropunctata. *Molecular Ecology*, **18**, 2937–2944.
Miller M, Dunham J, Amores A, Cresko W, Johnson E (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240.
Pääbo S, Poinar H, Serre D *et al.* (2004) Genetic analyses from ancient DNA. *Annual Review Of Genetics*, **38**, 645–679.
Palaiokostas C, Bekaert M, Davie A *et al.* (2013) Mapping the sex determination locus in the Atlantic halibut (Hippoglossus hippoglossus) using RAD sequencing. *BMC Genomics*, **14**, 566.
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.
Scaglione D, Acquadro A, Portis E *et al.* (2012) RAD tag sequencing as a source of SNP markers in Cynara cardunculus L. *BMC Genomics*, **13**, 3.
Schmitt MW, Kennedy SR, Salk JJ *et al.* (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences, USA*, **109**, 14508–14513.
Stolle E, Moritz RFA (2013) RESTseq–efficient benchtop population genomics with RESTriction Fragment SEQuencing. *PLoS ONE*, **8**, e63960.
Tin M, Economo EP, Mikheyev AS (2014) Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. *PLoS ONE*, **9**, e96793.
Wang S, Meyer E, McKay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, **9**, 808–810.

M.M.Y.T. performed the performed the experiments. F.E.R. and E.C. selected the museum bird specimens and extracted their DNA. A.S.M. designed the study, analyzed the data and wrote the manuscript.

Data Accessibility

All of the source code used in the analysis can be found at https://github.com/mikheyev/barcodes2. Alignment files of mapped and un-mapped reads can be found on DataDryad (doi:10.5061/dryad.34dt1).

Supporting Information

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

Table S1 Reads, mapping statistics and duplicates for the control experiments.

Appendix S1 Library preparation for museum bird samples.