MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1204–1223 doi: 10.1111/1755-0998.12568

SPECIAL ISSUE: SEQUENCE CAPTURE

High-throughput SNP genotyping of historical and modern samples of five bird species via sequence capture of ultraconserved elements

HAW CHUAN LIM and MICHAEL J. BRAUN
Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA

Abstract

Sample availability limits population genetics research on many species, especially taxa from regions with high diversity. However, many such species are well represented in museum collections assembled before the molecular era. Development of techniques to recover genetic data from these invaluable specimens will benefit biodiversity science. Using a mixture of freshly preserved and historical tissue samples, and a sequence capture probe set targeting >5000 loci, we produced high-confidence genotype calls on thousands of single nucleotide polymorphisms (SNPs) in each of five South-East Asian bird species and their close relatives (N = 27–43). On average, 66.2% of the reads mapped to the pseudo-reference genome of each species. Of these mapped reads, an average of 52.7% was identified as PCR or optical duplicates. We achieved deeper effective sequencing for historical samples (122.7×) compared to modern samples (23.5×). The number of nucleotide sites with at least 8× sequencing depth was high, with averages ranging from 0.89 × 10^6 bp (*Arachnothera*, modern samples) to 1.98 × 10^6 bp (*Stachyris*, modern samples). Linear regression revealed that the amount of sequence data obtained from each historical sample (represented by per cent of the pseudo-reference genome recovered with ≥8× sequencing depth) was positively and significantly ($P \leq 0.013$) related to how recently the sample was collected. We observed characteristic post-mortem damage in the DNA of historical samples. However, we were able to reduce the error rate significantly by truncating ends of reads during read mapping (local alignment) and conducting stringent SNP and genotype filtering.

Keywords: ancient DNA, historical DNA, Indo-Burma, Sundaland, target enrichment, ultraconserved elements

Received 25 September 2015; revision received 12 July 2016; accepted 15 July 2016

Introduction

The advent of high-throughput 'next-generation' DNA sequencing methods enables biologists to generate enormous amounts of genetic data from both model and nonmodel organisms (Church 2006). For studies of intraspecific genetic variation, the ideal data set would be a well-assembled genome and whole-genome resequencing data from numerous individuals. Indeed, this approach is starting to be applied to some long-studied nonmodel organisms (e.g. Ellegren 2014). However, whole-genome resequencing is prohibitively expensive for many systems and inefficient for many questions where a sampling of loci from across the genome would suffice. In these situations, methods that rapidly and reliably generate data from a reproducible subset of the genome are desirable, especially when they are scalable to large numbers of individuals.

The class of techniques that have received the most attention in this regard are collectively called genotyping by sequencing or restriction-site-associated DNA sequencing (RAD-Seq). These techniques are relatively inexpensive, simple to implement, and can deliver data on tens of thousands of markers from hundreds of individuals (Davey *et al.* 2013). However, genotype calls from these techniques can be sensitive to changes in the bioinformatic pipelines used to produce them, partly because they typically lack a reference genome or any prior sequence information to aid in locus assembly and orthology determination (Harvey *et al.* 2015; Leache *et al.* 2015). Another problem is allelic dropout. As sequence divergence accumulates, some restriction sites are lost, and it becomes impossible to sequence the associated alleles. Further, because most RAD-Seq approaches use fragment size selection to achieve complexity reduction (Puritz *et al.* 2014), they may not work effectively on

Correspondence: Haw Chuan Lim, Fax: +202-633-1133; E-mail: limhc@si.edu

© 2016 John Wiley & Sons Ltd

degraded DNA samples. Finally, because RAD-Seq determines targets using only restriction sites, it is less suitable for samples likely to contain heterogeneous mixtures of DNA, such as historical samples contaminated with exogenous materials, faecal samples or environmental samples (Jones & Good 2016).

An alternative approach to reducing genomic complexity is DNA sequence capture or target enrichment (Gnirke *et al.* 2009; Jones & Good 2016). Essentially, sequence capture methods use the affinity of RNA or DNA probes to isolate complementary sequences out of a larger pool of DNA fragments, thus facilitating rapid, high coverage sequencing of numerous targeted loci of specific interest. These techniques have been applied in many ways. Examples include estimation of genetic diversity of extinct species (Briggs *et al.* 2009); simultaneous identification of hosts, parasites and pathogens (Campana *et al.*, 2016); and estimation of deep phylogenetic relationships (McCormack *et al.* 2012; Faircloth *et al.* 2015). For nonmodel organisms, much of the developmental work has focused on exon-based capture probe sets (Bi *et al.* 2013; Good *et al.* 2013; Li *et al.* 2013), or probe sets based on highly conserved genomic elements that may have wider phylogenetic applicability (Faircloth *et al.* 2012; Lemmon *et al.* 2012; Penalba *et al.* 2014).

One recent development in sequence capture is to use conserved loci as probes to generate data for investigating intraspecific processes, such as population structure and divergence time (Harvey *et al.* 2013; Smith *et al.* 2014; McCormack *et al.* 2015). Using such 'universal' probe sets to generate population genomic data will greatly benefit and simplify studies of species with few genomic resources, because it alleviates the need to create custom species- or genus-specific probe sets (e.g. Gardiner *et al.* 2014). While this strategy of generating data is promising and the above studies showed ample genetic variation in flanking regions of conserved loci, investigators should keep in mind that the data may be affected by molecular signatures of natural selection when making population genetic inferences (Jones & Good 2016).

The ability to generate population genetic data from historical samples using sequence capture is significant because it can facilitate study of historical populations that predate the Anthropocene (Bi *et al.* 2013), as well as species that occupy regions that are currently closed to collection of modern museum specimens and genetic resources (e.g. Afrotropics; Stoeckle & Winker 2009). For the study of contemporary populations, lack of well-preserved genetic samples from some regions hampers our abilities to decipher region-wide phylogeographic patterns, detect hidden diversity and delineate zones of genetic transition. This is unfortunate because many biologically diverse regions also suffer from high rates of habitat loss (Hansen *et al.* 2010; Miettinen *et al.* 2011). Without knowledge of the distribution and distinctiveness of a region's biological diversity, creation of sound conservation strategies is fraught with difficulties. In such cases, museum specimen collections can be a valuable source of DNA if various technical challenges are overcome (Houde & Braun 1988). By facilitating phylogeographic and population genomic studies in many heretofore poorly known regions, the emerging field of archaeogenomics will also help shed light on mechanisms that underlie differences in diversification on a global scale (Shapiro & Hofreiter 2014; Hofman *et al.* 2015).

Here, we test the utility of sequence capture from both modern and historical samples with a probe set based on genomic elements conserved across the amniote tree [ultraconserved elements (UCEs)] (Faircloth *et al.* 2012). Our goal was to conduct single nucleotide polymorphism (SNP) discovery and genotyping in five species of South-East (SE) Asian birds. While using historical samples helps to fill in sampling gaps, it also poses challenges, because DNA derived from historical samples is more fragmented and suffers from various other post-mortem damages. Principal types of damage include preferential occurrence of strand breaks immediately 3′ of purine residues (A or G) and a higher rate of C to T misincorporation at the 5′ end of reads (and G to A misincorporation at the 3′ end) due to increased deamination of C residues along single-stranded overhangs (Briggs *et al.* 2007; Brotherton *et al.* 2007). Read mapping and variant filtering strategies that help to control these errors are described. We further describe the amounts of data recovered at each stage of the bioinformatic pipeline and test how they are influenced by different experimental factors. Our study shows a diminishing return in per cent of target recovered as sequencing depth increases, especially in historical samples. Nonetheless, a large number of high-confidence SNPs were recovered from both modern and historical samples (>3000 SNPs per sample). This study illustrates the promising strategy of rapidly investigating heretofore poorly studied species and biogeographic regions with thousands of markers using a combination of powerful sequence capture techniques and challenging museum samples.

Materials and methods

Sampling and DNA extraction

For this study, we used 185 samples from the following five bird species: *Arachnothera longirostra* (Nectariniidae), *Irena puella* (Irenidae), *Niltava grandis* (Muscicapidae), *Pycnonotus atriceps* (Pycnonotidae) and *Stachyris nigriceps*

© 2016 John Wiley & Sons Ltd

(Timaliidae). We further included nine outgroup samples that belong to the same or closely related genera (Appendix). Each focal species was represented by 30–47 samples (average = 37). Overall, 70.3% (N = 137) of the samples were historical, consisting of ~2 × 2 mm slices from toe pads of museum specimens. Five historical specimens were of unknown age; the rest ranged from 27 to 142 years (average = 77 years). The remaining samples were modern muscle (N = 55) or blood (N = 2) specimens, freshly preserved in the field and stored as genetic resource materials in freezers. In the field, these fresh materials were first preserved either in 90–95% EtOH (muscle) or lysis buffer (blood) (Seutin *et al.* 1991). All modern muscle samples were associated with museum voucher specimens; the two blood samples were unvouchered.

DNA was extracted from historical samples in a dedicated, PCR-free extraction facility at the Smithsonian Conservation Biology Institute (Washington, DC). This facility has been used successfully to conduct genetic studies using various degraded materials such as subfossils, animal faeces and bird toe pads (Cooper *et al.* 1996; Paxinos *et al.* 1997; Dumbacher *et al.* 2003). During extraction, strict protocols were followed to minimize cross-contamination among samples (e.g. UV irradiation of equipment and buffer before each session, bleaching work surfaces and changing gloves between samples, use of filtered tips in all pipetting steps). No more than 22 samples were processed during each laboratory session, and they were always accompanied by two negative controls. Each toe pad was cut into finer pieces, placed in 600 μL extraction buffer (0.01 M Tris-EDTA, 0.02 M EDTA, 0.01 M NaCl, 1% w/v sodium dodecyl sulphate, 0.1 mg/mL Proteinase K and 10 mg/mL dithiothreitol) and incubated on a rotator at 55 °C for 24 h. DNA was then isolated from the digested products using a combination of phenol–chloroform extraction and ultrafiltration (Amicon Ultra-4 filter units, 30 kDa). Negative controls from each extraction batch were tested for contamination via PCR using bird-specific primers that target a 307-bp portion of the cytochrome *b* gene (Paxinos *et al.* 2002). No false positives were detected. DNA of modern samples was extracted with the Qiagen DNeasy Blood and Tissue kit or with an Autogen Gene Prep machine using the 'Animal Tissue' protocol and later stored at −20 °C.

Library preparation and sequencing

Prior to library preparation, DNA from modern samples was sheared in a Qsonica Q800R1 sonicator (25% amplitude, 10 s on to 10 s off, and a total shearing time of 1 min). Historical DNA samples were not subjected to shearing because they were already fragmented. Both sets of DNA were then quantified using Qubit 2.0 Fluorometer (Thermo Fisher Scientific). We prepared Illumina TruSeq sequencing libraries using the KAPA Biosystem Library Preparation Kit. For each library, we used 100 μL of DNA as starting material. We diluted modern DNA down to 10 ng/μL if the original concentration exceeded this amount, to achieve a maximum starting quantity of 1000 ng. Dilution was not done for historical DNA, or modern DNA with concentration lower than 10 ng/μL. Thus, the amount of starting DNA from modern and historical samples averaged 845.5 and 448.4 ng, respectively, and was more variable for historical samples (Appendix). The protocols we followed for library preparation can be found in http://ultraconserved.org/#protocols. Briefly, we conducted on-bead ligation of universal Y-adaptors followed by 10 cycles of PCR to attach Illumina binding sequences and two 8-bp index sequences that uniquely identified each sample (Glenn *et al.* 2016). We used two 25 μL reactions per sample instead of the usual one to reduce experimental error from biased PCR amplification and other sources. Library preparation for historical samples was carried out in the above-mentioned historical DNA laboratory up to the limited cycle PCR step, while libraries for modern samples were prepared in a conventional molecular laboratory.

To enrich for UCEs, we used MYbaits tetrapods 5K capture kits from MYcroarray, Inc. This probe set uses 5 472 120-mer RNA baits to target 5060 UCE loci. Enrichment experiments for historical and modern samples were conducted separately. For each enrichment, we pooled libraries from four historical or 6–8 (average = 7.1) modern samples in equal mass ratios to yield a total of 500 ng starting material. We followed the manufacturer's protocol (version 2.3.1) for the enrichment experiments, but replaced the kit's blocking agent (human Cot-1 DNA) with chicken Cot-1 DNA (Applied Genetics Laboratory). After enrichment, we conducted 18-cycle PCR to increase the amount of products and validated enrichment using qPCR with primer sets that target eight randomly selected UCE loci. Next, we conducted size selection on the enriched products (350–650 bp for libraries derived from modern DNA and 200–500 bp for libraries from historical DNA) using BluePippin 2% agarose gel cassettes (Sage Science). Quality and quantity of the libraries were checked on agarose gels, an Agilent Bioanalyzer (Agilent Technologies; Fig. S1, Supporting information) and a Qubit Fluorometer (Thermo Fisher Scientific). Finally, libraries were pooled in equal mass ratios and sent to BGI Americas for 100-bp paired-end sequencing on an Illumina HiSeq 2000 [two lanes for historical samples and one partial lane (90%) for modern samples].

© 2016 John Wiley & Sons Ltd

Assembly and variant calling

Reads were first demultiplexed using Casava version 1.8 (Illumina, Inc.). We then employed illumiprocessor version 2 (a trimmomatic version 0.3 based wrapper) to conduct batch quality control (e.g. read trimming, adaptor sequence removal) on the raw reads using default settings (Faircloth 2013; Bolger *et al.* 2014). Cleaned reads (reads that passed trimmomatic filtering) were inspected with FastQC version 0.10.1 (www.babraham.ac.uk) and Geneious version 7.0.6 (www.geneious.com) to ensure that reads were of high quality and adaptor read-through was absent. We then assembled reads from a subset of individuals of each target species to create a pseudo-reference genome for that species. This strategy alleviated the computational demand of having to conduct *de novo* assembly of all individuals and is a commonly adopted strategy when creating pseudo-reference genomes (e.g. Hird *et al.* 2011).

We used the program ABySS version 1.3.7 wrapped inside the python script assemblo_abyss.py (a component of the Phyluce version 2 pipeline) to individually assemble reads from 10 to 15 modern and historical samples from each species (Simpson *et al.* 2009; Faircloth *et al.* 2012). As shown in Results, a subset of individuals per species was sufficient to create a pseudo-reference genome that contained most (>97.5%) of the targeted UCE loci because of the large amount of sequencing done per individual. Prior to assembly, we tested a range of kmer sizes (25–55 bp) and selected the optimal size (ranging from 35 to 39 bp) based on assembly continuity statistics output by the perl script ABySS-fac. The assembled reads were matched against probe sequences using a Phyluce version 2 python script, match_contigs_to_probes.py, to remove probes and assembled contigs that did not have one-to-one relationships. Sequences of contigs that passed this test were exported in fasta format and aligned locus by locus for each species using the Phyluce script seqcap_align.py (using MAFFT version 7.13 as the aligner, –no-trim option selected, minimum of two individuals per alignment). Alignments were exported into Geneious and those with low (<95%) pairwise sequence identity were manually inspected and corrected or discarded. Consensus sequences of these alignments were then generated using 0% threshold and combined to produce pseudo-reference genomes for use in subsequent mapping steps (one pseudo-reference genome per species). When constructing a consensus sequence from an alignment using a 0% threshold, a variable site was either called as the most common nucleotide in the alignment, or given an IUPAC ambiguity code if two or more nucleotides were equally common.

Mapping of cleaned paired-end reads to the pseudo-reference genome was conducted using bowtie2 version 2.2 and the following options: local alignment (ends of reads can be soft-clipped), very sensitive alignment (a preset option) and maximum of one mismatch allowed in a seed alignment (Langmead & Salzberg 2012). We then sorted the generated BAM files and used Picard tools version 1.122 to identify optical and PCR duplicates based on the coordinates of reads on the pseudo-reference genome (http://broadinstitute.github.io/picard/). The Picard tools program MarkDuplicates uses 5′ coordinates and mapping orientations of read pairs to identify duplicates, taking into account existing clippings and gaps in the alignment. Next, the GATK version 3.2 IndelRealigner tool was used to improve local alignments in BAM files and better account for indels by minimizing mismatches across all reads in a multisample BAM file (McKenna *et al.* 2010). Finally, we used GATK HaplotypeCaller and GenotypeGVCFs tools to conduct single-sample and joint genotype calling, respectively.

Data evaluation and statistical analysis

A number of metrics were calculated to evaluate how experimental variables affected the amount of quality data generated. First, we calculated the number of cleaned read pairs per sample as a measure of the sequencing effort expended per sample. Using analysis of variance (ANOVA) and analysis of covariance (ANCOVA) scripts (R Development Core Team, 2014), we then tested whether the amount of starting DNA or variation among enrichment pools affected the proportion of reads mapping to pseudo-reference genomes (on-target reads) or the rate of read duplication. To calculate depth of sequencing at every pseudo-reference genome site, we applied the GATK DepthofCoverage tool to the deduplicated and indel-adjusted mapped reads of each sample. When calculating sequencing depth, we ignored bases with Phred quality score below 10 and reads with root-mean-square (RMS) mapping quality below 30. For each historical sample, we further evaluated the proportion of the pseudo-reference genome that had ≥8× sequencing depth (hereafter as target coverage) and determined, using linear regression, whether it correlated with the age of the sample. For each statistical analysis, diagnostic plots (e.g. residuals vs. fitted values, normal Q–Q) were inspected to ensure that there was no significant violation of assumptions.

Post-mortem damage in historical DNA, such as high rates of C → Tor G → A misincorporation at read termini, may have negative effects on the accuracy of variant calling, especially at low levels of sequencing depth (Parks & Lambert 2015). We therefore assessed DNA damage patterns in mapped reads of historical samples

© 2016 John Wiley & Sons Ltd

using mapDamage version 2.0 by looking at (i) frequency of different nucleic acid residues at positions upstream and downstream of the start/end of reads; and (ii) empirical and posterior probability of different base substitution types at the beginning and end of each read (Jonsson *et al.* 2013). When running mapDamage, we used default settings and 60 000 MCMC iterations (first 10 000 iterations discarded as burnin). Trace plots of various model parameters were inspected to ensure that convergence was reached.

Variant filtering and DNA damage assessment

Using Variant Call Format (VCF) files produced by the GenotypeGVCFs tool, we carried out site and genotype filtering using GATK version 3.2 VariantFiltration tool and VCFtools version 0.1.13 (Danecek *et al*. 2011), respectively. We removed sites where reference and alternate alleles have statistically different locations along reads (GATK ReadPosRankSum test; e.g. testing if alternate alleles tend to be found at the ends of reads) or statistically different base quality (GATK BaseQRankSum test), as well as sites whose reference and alternate alleles occurred in reads with statistically different mapping qualities (GATK MQRankSum test). As these rank sum tests produced *z*-scores, we removed sites with scores falling outside of the 2.5th ($z < -1.96$) and 97.5th ($z > 1.96$) percentile. Additionally, we removed sites whose reference and alternate alleles show significant read strand bias (GATK FisherStrand test, removing sites with Phred-scale *P*-value >20). Finally, we removed sites supported by reads whose average RMS mapping quality fell below a Phred score of 30. At the level of individual genotypes, we removed those (rendering a previously called genotype uncalled) that had a genotype quality (GQ) falling below 13 and those with (deduplicated) sequencing depth (DP) falling below eight.

We wished to ascertain whether post-mortem damage in historical samples had a major impact in elevating C → T and G → A substitutions compared to other substitution types (e.g. C → A, A → T). After variant filtering for each species group, we compared the number of C → T and G → A substitutions between modern and historical samples, with C and G being the reference alleles, while T and A represent the corresponding alternate alleles. These ratios were then compared against those for all other substitution types combined using *z*-test of proportions. An individual heterozygous for the alternate allele (e.g. genotype C/T) is considered to represent one C → T substitution if the reference allele is C. On the other hand, a homozygous T/T individual is considered to represent two C → T substitutions if the reference allele is C. The comparison of C → T (or G → A) ratio between modern and historical samples against the corresponding ratio for all other substitution types controls for differences in sample make-up (i.e. the number of modern samples vs. historical samples) among species groups and assumes that there is no interaction between sample age and substitution type on genotyping success.

Results

We obtained 169 million pairs of raw reads for the 57 modern samples from one partial Illumina lane. The two lanes containing 68 and 69 historical samples produced 222 million and 202 million pairs of reads, respectively. Overall, 28.0% of modern and 11.7% of historical libraries failed, not producing enough reads (less than a

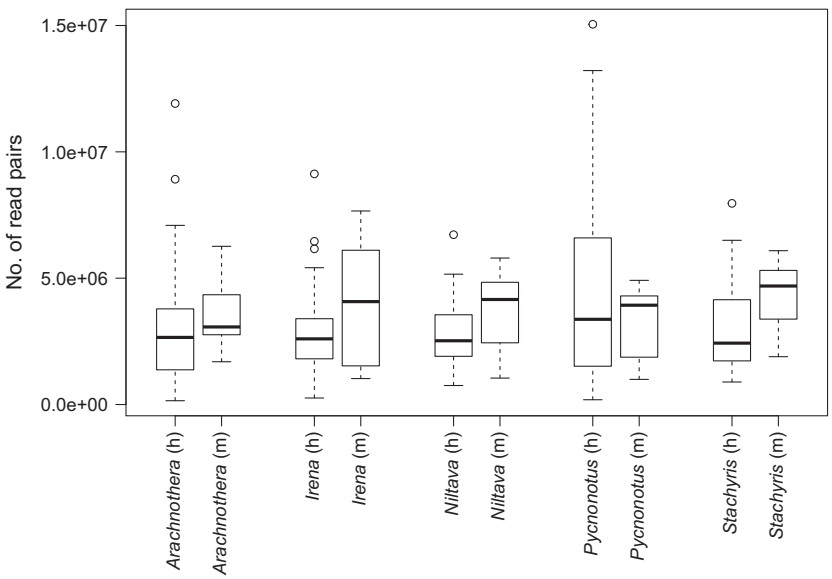

Fig. 1 Box and whisker plots showing medians and quartiles for the number of cleaned read pairs for different species groups and sample types (historical = h, modern = m). Circles indicate all outlying values 1.5× interquartile range more or less than the third or first quartile, respectively.

© 2016 John Wiley & Sons Ltd

few hundred thousand per sample) for genotyping many SNPs present in other samples Appendix). The quantity of starting DNA was not an important factor behind library failure for either modern (Mann–Whitney *U*-test, $U = 384.0$, $P = 0.324$) or historical samples (Mann–Whitney *U*-test, $U = 786.5$, $P = 0.182$). Therefore, failure could be related to other factors such as degradation of previously extracted DNA (some modern samples), lack of endogenous DNA, contamination of historical samples with exogenous DNA or biased enrichment of libraries during pooled probe DNA hybridization reactions. Both the average number of raw read pairs and read pairs that passed trimmomatic filtering were slightly higher in modern samples compared to historical samples, which is expected given the slightly higher sequencing effort for modern samples (Fig. 1, Table 1 and Appendix). There was not a strong pattern of sequencing bias among libraries of the same enrichment pool (Figs. S2 and S3). The average coefficient of variation (CV) in the number of cleaned read pairs produced per library was lower for enrichment pools containing historical samples (average = 73.31%) compared to those for modern samples (average = 86.50%), but the difference was not significant statistically (*t*-test, $P = 0.149$).

Alignment of assembled contigs of a subset of 10–15 individuals from each species produced pseudo-reference genomes of the following sizes: *A. longirostra* (4934 UCE loci, average locus length = 432.8 bp), *I. puella* (4968 UCE loci, average locus length = 348.0 bp), *N. grandis* (4950 UCE loci, average locus length = 719.5 bp), *P. atriceps* (4933 UCE loci, average locus length = 346.0 bp) and *S. nigriceps* (4946 UCE loci, average locus length = 705.1 bp). The proportion of on-target reads ranged from 15.0% to 94.1%, with historical samples having a higher average rate 70.2% (±17.8%, SD) than modern samples 54.5% (±28.2%) (Fig. S4, Table 1, Appendix). Many off-target reads assembled into contigs that mapped to mitogenomes of closely related species (data not shown). Nested ANOVA showed that both sample type (modern vs. historical) and enrichment pool were highly significant factors behind on-target rate (Table 2). However, after two outlier modern pools were excluded, only the latter remained a significant factor; historical samples still possessed a higher on-target rate, but the difference was not significant statistically (Table 2). For historical samples, the amount of starting DNA used during library preparation did not affect on-target rate, but enrichment pool was an important factor (Table 2).

The proportion of reads detected as duplicates was high. The overall proportion of duplicated reads was 57.2%, and historical samples had a lower average (48.2 ± 17.3%) than modern samples (83.9 ± 7.6%) (Fig. S5, Table 1, Appendix). Based on nested ANOVA,

Table 1 Metrics showing sample sizes and the amount of data recovered at different stages of the bioinformatic pipeline. Size of pseudo-reference genomes are as follows: *A. longirostra* 2 135 583 bp, *I. puella* 1 728 808 bp, *N. grandis* 3 561 348 bp, *P. atriceps* 1 707 079 bp and *S. nigriceps* 3 487 311 bp

Species group	Sample type	Starting sample size	Final sample size (successful libraries)	Avg. no. of cleaned read pairs	SD of no. of cleaned read pairs	Avg. bowtie2 mapping rate (%)	Avg. read duplication rate (%)	Avg. sequencing depth	Avg. target coverage (%)	Avg. target coverage (sites)
Arachnothera	Historical	22	19	3 358 941	3 052 345	69.8	44.2	92.3	57.6	1 229 421
	Modern	16	11	3 549 985	1 350 492	39.8	87.1	12.8	41.9	893 839
Irena	Historical	43	38	2 877 027	1 765 766	65.2	48.8	90.7	59.5	1 028 277
	Modern	6	5	4 079 165	2 860 605	62.7	90.9	22.4	58.0	1 003 054
Niltava	Historical	23	19	2 847 856	1 564 885	77.8	50.7	55.7	35.8	1 276 650
	Modern	9	8	3 714 949	1 629 690	65.7	82.2	16.9	51.5	1 832 314
Pycnonotus	Historical	25	24	4 607 250	4 029 231	74.5	46.3	189.9	60.4	1 030 649
	Modern	11	7	3 169 559	1 590 730	59.4	84.5	22.6	58.4	997 666
Stachyris	Historical	24	21	3 098 871	2 059 999	67.6	50.4	56.1	33.7	1 174 227
	Modern	14	10	4 394 515	1 355 616	54.4	77.7	27.4	56.9	1 983 234

© 2016 John Wiley & Sons Ltd

sample type and enrichment pool were again significant factors behind read duplication rate (Table 2). Analysis of covariance showed that, in historical samples, higher starting DNA quantities resulted in lower duplication rates, and enrichment pool remained an important factor (Table 2). We did not test modern samples with similar ANCOVA because the amount of starting DNA was less variable; most modern sample libraries (80.5%) had close to the maximal amount (>950 ng) of starting DNA.

Target region size (i.e. total length of the pseudo-reference genome) for the five study species ranged from 1.70×10^6 bp (*P. atriceps*) to 3.56×10^6 bp (*N. grandis*) and was closely correlated with the average length of loci assembled for each species reported above (Table 1). Because these pseudo-reference genomes encompassed both the UCEs and their flanking regions, they were 2.6–5.5× larger than the region covered by the UCE probes alone (~650 000 bp). As expected, individual sites within a target region varied greatly in their sequencing depth (Figs 2 and S6). Averaged across the entire target region, historical samples possessed higher per-site sequencing depth (across-sample average = 122.7×) than modern samples (across-sample average = 23.5×; Table 1, Appendix). However, although historical samples, on average, had at least twice the per-site sequencing depth compared to modern samples of the same species group, target coverage in historical samples was generally similar to or smaller than that of their modern counterparts (Table 1). This is likely a result of sequencing depth in historical samples being less even across each locus. In other words, mapped reads generated from historical samples tended to concentrate around the baited regions (covered by actual probes) because of the smaller library insert sizes, resulting in lower sequencing depth in the flanking regions (Fig. 2 and Fig. S7). While increasing the sequencing depth of a sample increased target coverage, the relationship was not linear (Fig. 3). Instead, we obtained a diminishing return in the extent of the target region that was covered as sequencing depth went up. Nonetheless, sequence capture was able to achieve a very large number of sites sequenced 8 or more times in every species group, with averages ranging from 0.89×10^6 bp (*Arachnothera*, modern samples) to 1.98×10^6 bp (*Stachyris*, modern samples) (Table 1). Historical samples of all species showed significant positive relationships between how recently samples were collected and the amount of quality sequence data obtained from each sample (represented by target coverage) (Fig. 4 and Table 3). Goodness of fit of regression, as measured by R^2, was highest in *A. longirostra* (0.93), followed by the other species (R^2 = 0.30–0.52).

Based on mapDamage analysis of the fully processed and mapped reads of historical samples, the rates for C to T substitution at the 5′ end of reads and G to A substitution at the 3′ end of reads were low (<2%) and do not show an increasing trend towards the ends of each read

Table 2 Results of nested ANOVA and ANCOVA tests showing factors that influenced various data recovery metrics

Metric	Statistical test	Factors	Degree of freedom	Mean square	*F* value	*P*
On-target rate	Nested ANOVA	Sample type (modern vs. historical)	1	7480.0	211.5	<0.0001
		Enrichment pools nested within sample type	40	1638.0	46.3	<0.0001
		Residuals	120	35.0		
	Nested ANOVA*	Sample type (modern vs. historical)	1	6.90	0.179	0.673
		Enrichment pools nested within sample type	38	1130.4	29.468	<0.0001
		Residuals	110	38.4		
	ANCOVA (historical samples only)	Input DNA amount	1	1.50	0.083	0.774
		Enrichment pools	33	1101.7	62.6	<0.0001
		Residuals	86	17.6		
Read duplicate rate	Nested ANOVA	Sample type (modern vs. historical)	1	3.91	538.4	<0.0001
		Enrichment pools nested within sample type	40	0.074	10.24	<0.0001
		Residuals	120	0.0073		
	ANCOVA (historical samples only)	Input DNA amount	1	0.19	19.63	<0.0001
		Enrichment pools	33	0.078	7.81	<0.0001
		Residuals	86	0.0099		

*Two outlier enrichment pools (8.22.14.pool1 and 8.22.14.pool2) composed of modern samples were excluded.

© 2016 John Wiley & Sons Ltd

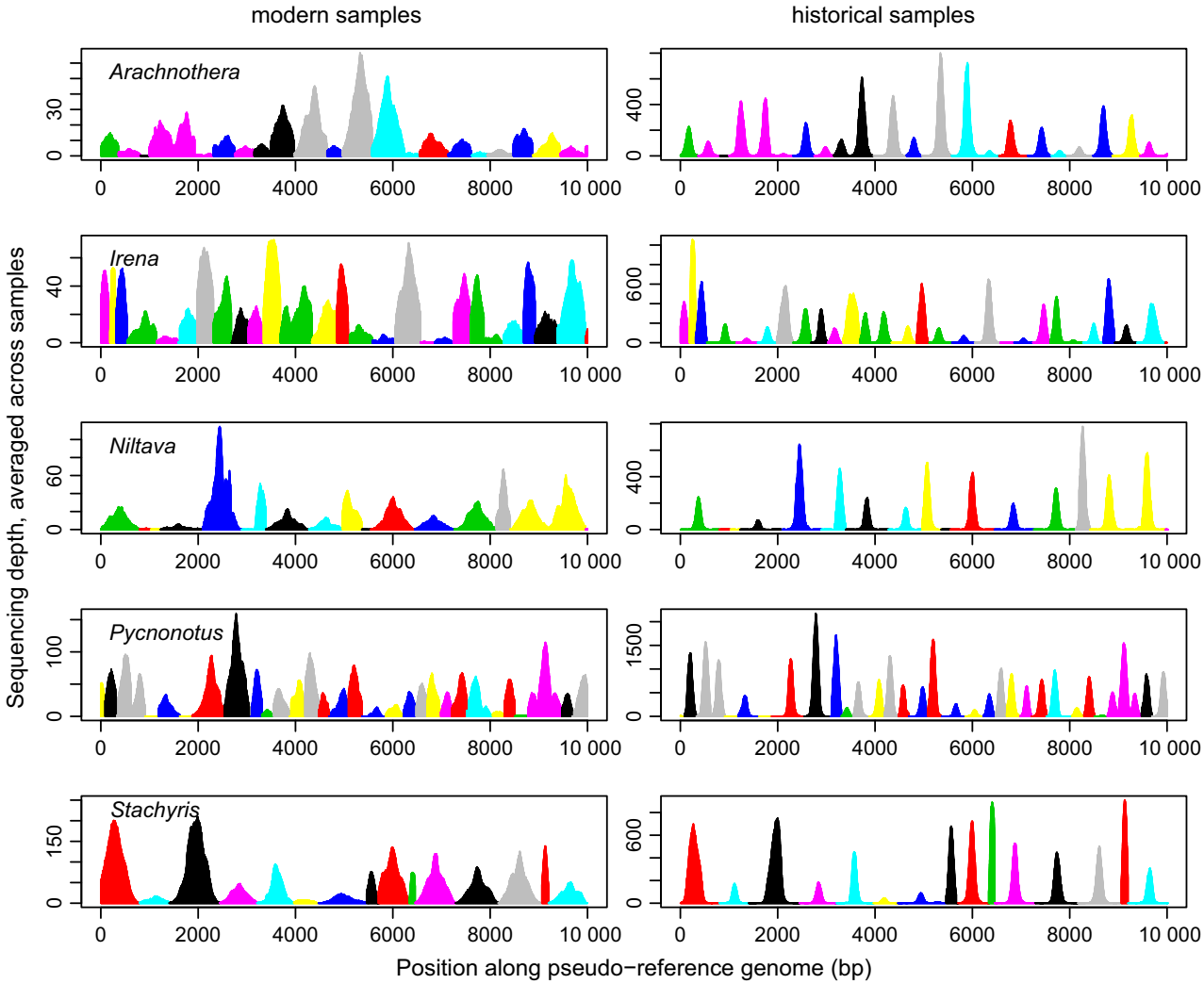

Fig. 2 Sequencing depth of sites (averaged over samples) along a randomly chosen 10 000-bp stretch of each pseudo-reference genome. UCE loci are arranged end to end (represented by different colours) along each stretch. The same order of loci is maintained within each row (species group), but not among rows. Charts are sorted according to species group (rows) and age of samples (columns).

(see Fig. S8 for plots based on five exemplar samples, and Fig. S9 that includes approximate Bayesian estimates of C → T, G → A and other substitutions). On the other hand, if bases at the ends of reads soft-clipped by bowtie2's local alignment algorithm are taken into account, the expected trend of increasing C → T or G → A substitutions at the ends of reads is evident (Fig. S10). Additionally, damage assessment reveals a characteristic increase in purine frequency just before the start of reads (due to strand breakage 3′ of purines) and a corresponding decline in pyrimidine frequency at the same position (Fig. S10).

For each species group, after SNP filters were applied, 61.5–79.8% of the SNPs were removed compared to the number of raw SNPs. After filtering, the number of variable UCE loci in each species group ranged from 753 (*Irena*) to 4051 (*Niltava*), and the total number of SNPs per species group ranged from 3919 (*Irena*) to 18 472 (*Niltava*) (Table 4). Within each species group, modern and historical samples generally had similar numbers of high-confidence, filtered genotype calls (Fig. S11). The highest number of called genotypes per sample was found in *Niltava* (average = 14 707.9) and the lowest in *Irena* (average = 3310.8). The level of data completeness was generally high, and the number of SNPs represented in 80% or more of the samples within a species group ranged from 1748 (*Arachnothera*) to 10 060 (*Niltava*) (Table 4). We found a slight to moderate increase in C → T (0.8–13.0%) or G → A (1.6–10.3%) substitutions in historical samples compared to the level expected

© 2016 John Wiley & Sons Ltd

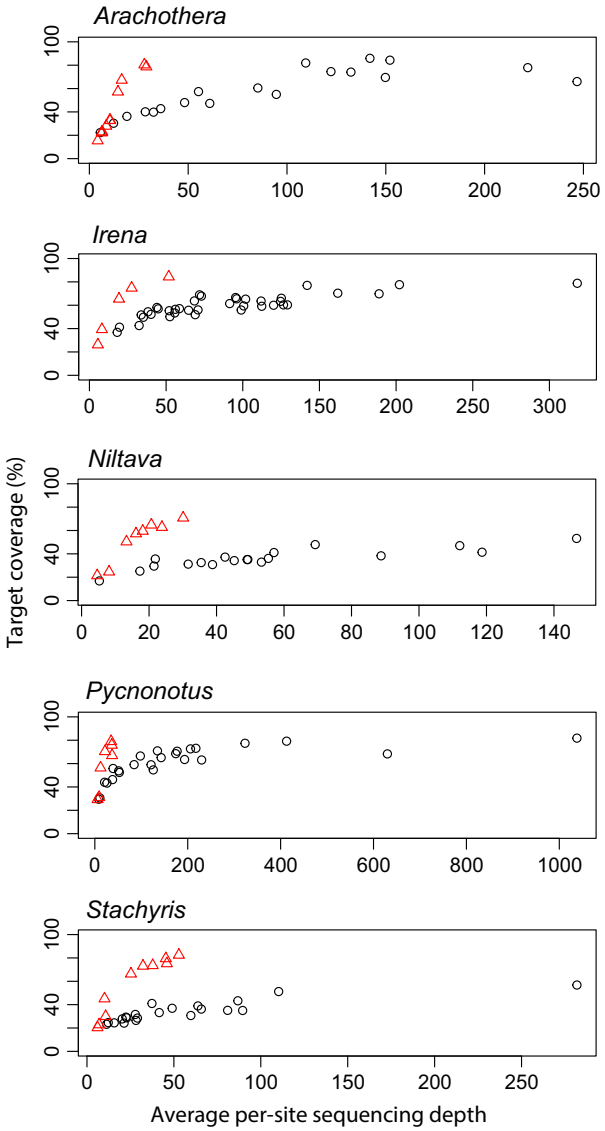

Fig. 3 Per cent target coverage (*y*-axis) vs. average sequencing depth for each sample (red triangles – modern samples; black circles – historical samples).

based on other substitution types (Table 5). Based on *z*-tests of proportions, all but two of the increases were significant at $P = 0.01$.

Discussion

This study is similar to a recently published paper that applied the UCE probe set to toe pad samples of the western scrub jay (*Aphelocoma californica*) (McCormack *et al.* 2015), but our paper contains some important differences. To assess the amount of data recovered, we used metrics based on uniquely mapped reads – depth of sequencing and proportion of target region covered – whereas McCormack *et al.* (2015) used length and number of *de novo* assembled UCE loci. Additionally, we included modern samples in our study, thus allowing direct comparison of data recovery between modern and historical samples. Finally, we also conducted an in-depth analysis of post-mortem damage of historical DNA.

Our study echoes McCormack *et al.*'s (2015) findings, showing that it is possible to obtain accurate genotype data on thousands of SNPs from hundreds to thousands of independent UCE loci at the intraspecific level. Compared to studies that used the same or a smaller UCE probe set (Smith *et al.* 2014; McCormack *et al.* 2015), our study reveals a slightly higher number of SNPs per UCE locus (their studies: 1.9–3.2 per locus; our study: 2.6–5.2 per locus), which is related to the inclusion of outgroup taxa. With the closely related outgroup taxa removed, the number of SNPs per locus dropped to 2.4–3.6. Smith *et al.* (2014) also found 52.7–77.1% of loci to be polymorphic in each of five bird species, similar to the 35.0–81.8% observed in our study. Success in generating genotype data is evident in both modern and historical samples. Not only do modern and historical samples of the same species have similar number of called genotypes, many of the SNPs are well represented across modern and historical individuals of the same group. This high-quality genotype information can subsequently be outputted in the form of independent SNPs, phased haplotypes or unphased diploid genotypes. These data have been shown to be useful in a variety of downstream population-level analyses, such as identification of population structure and admixture (Harvey *et al.* 2013; McCormack *et al.* 2015) and estimation of population divergence history (Smith *et al.* 2014).

While using a widely applicable, conserved probe set to genotype many species reduces the need to design custom probe sets, more also needs to be learned about how UCEs and their flanking regions behave evolutionarily. It is possible that the flanking regions are subjected to some background selection, resulting in reduced variation and biased allele frequencies (Hahn 2008). Just as the lack of strict neutrality is common to many molecular markers (e.g. Bazin *et al.* 2006; Hodgkinson *et al.* 2013), we recommend that population-level studies using conserved loci as probes include tests of signatures of molecular selection or use other marker types as well, so that concordance of inferences can be assessed (Reed *et al.* 2005; Harvey *et al.* 2013; Vitti *et al.* 2013). To circumvent the inability of RAD-Seq techniques to handle degraded DNA or heterogeneous DNA mixtures, and the cost associated with designing and manufacturing custom sequence capture probes, some authors have recently developed creative 'hybrid' approaches that are both easy to implement and highly specific. One is to create

© 2016 John Wiley & Sons Ltd

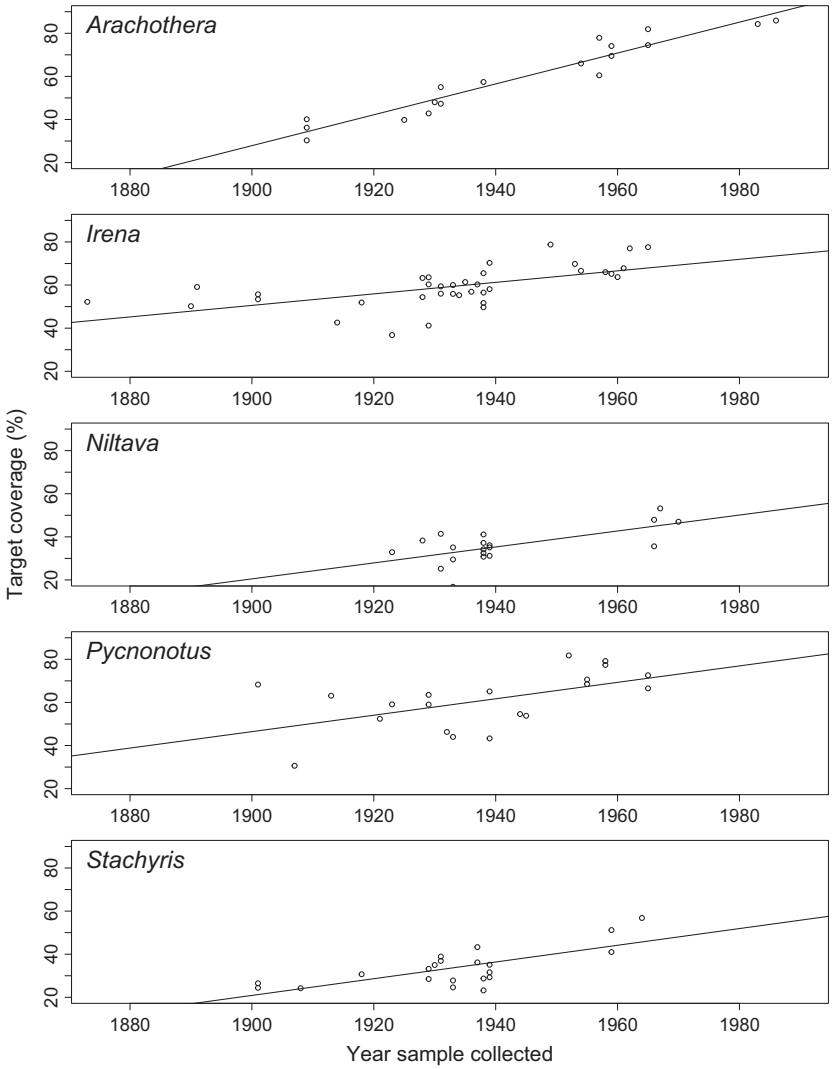

Fig. 4 Per cent target coverage (*y*-axis) vs. year of collection for each historical sample. Linear regression-based lines of best fit are also shown.

Table 3 Results of linear regression between the year a sample was collected and the percentage of target region with 8× sequencing depth. Only historical samples with known collection year were used in this analysis. One outlier sample was removed from *Arachnothera*, two samples with unknown collection year were removed from *Irena*, and two samples with unknown collection year were removed from *Pycnonotus*. Results in which the dependent variable (target coverage) was subjected to arcsine square root transformation were very similar and as such are not shown

Species group	Sample size	Estimated coefficient	SE	*t*	*P*	R^2	adjusted R^2
Arachnothera	18	0.716	0.051	14.1	<0.001	0.926	0.921
Irena	36	0.267	0.061	4.42	<0.001	0.365	0.346
Niltava	19	0.370	0.107	3.47	0.003	0.415	0.380
Pycnonotus	20	0.381	0.139	2.75	0.013	0.296	0.257
Stachyris	21	0.388	0.085	4.55	<0.001	0.521	0.496

probes in-house by simply using PCR products of targeted loci as probes (Maricic *et al.* 2010; Penalba *et al.* 2014). Another is to add biotin to enzyme-digested DNA sequences (the RAD sequences) and apply them as sequence capture probes on more degraded samples (Suchan *et al.* 2015).

Experimentally, the ability to generate unique mapped reads is dependent on many factors, such as sequencing effort, amount of starting DNA, DNA fragment size, level of library multiplexing during enrichment (probe:DNA ratio) and enrichment conditions (Mamanova *et al.* 2010; Paijmans *et al.* 2016). While

© 2016 John Wiley & Sons Ltd

evaluating the impact of many of these factors is beyond the scope of our study, analysis of performance metrics provides useful insight into how to improve data yield. First, given a fixed overall sequencing effort, it is important to ensure that the number of reads generated from different samples is as uniform as possible. Other than having similar number of samples per sequencing lane and normalizing library concentration before pooled enrichment, another step that can be adopted to ensure even sequencing is to have fewer libraries per enrichment pool (Hawkins *et al*., 2015). When capturing mammalian mitogenomic sequences, Hawkins *et al.* (2015) demonstrated that pooling more libraries per enrichment increased skew in the number of reads generated per sample, possibly a result of the increased significance of pipetting error when pooling smaller volumes of liquid and/or increased intersample competition for capture baits. Our results support this observation, showing that enrichment pools containing a higher number of libraries (modern samples) have higher CV of read output.

Discrepancies in the proportion of on-target reads between modern and historical samples were not statistically significant after two outlier modern enrichment pools were removed. Nonetheless, historical samples still have a marginally higher on-target rate, and we believe that the shorter DNA molecules of historical samples actually improved enrichment specificity and on-target rate. This is because a short fragment contains a lower proportion of off-target sequence, and this reduces opportunities to cross-hybridize with other DNA molecules during enrichment (Hodges *et al.* 2009; Lee *et al*. 2009; Mamanova *et al*. 2010). Additionally, the smaller number of libraries per enrichment probably contributed to the lower read duplication rate of historical samples. By having a better per-sample probe:DNA ratio, more DNA could be captured, thus increasing the complexity of the enriched products of historical samples. Because each enrichment was followed by a PCR, having a more complex library improved the chance that sequenced DNA fragments were not each other's duplicate. The importance of having a more complex input template is also supported by the significant effect starting DNA quantity has on reducing read duplication rate. The overall result of having higher on-target and lower duplicate rates is that historical samples have greater effective sequencing depths. The rate of read duplication for modern samples is high (~84%) compared to the levels seen in studies using well-optimized commercial kits (~20–30%; e.g. Bodi *et al*. 2013). To correct this problem, we recommend the following measures: increase the amount of input DNA from ~1000 ng to higher, reduce the number of libraries pooled during hybridization reactions and reduce the per-sample sequencing effort (i.e. multiplex more samples per sequencing lane).

Table 4 The numbers of variable UCE loci and SNPs before and after SNP filtering. Also shown are the numbers of postfiltering SNPs with various levels of representation across samples within a species group

Species group	Prefiltering	Postfiltering	Proportion left after filtering	No missing data	Max. 20% missing data	Max. 50% missing data
Arachnothera						
No. of loci with ≥1 SNP	2378	1726	0.73			
Total no. of SNP	15 849	5890	0.37	332	1748	5886
No. of SNP per locus	6.66	3.41				
Irena						
No. of loci with ≥1 SNP	854	753	0.88			
Total no. of SNP	10 186	3919	0.38	594	2766	3864
No. of SNP per locus	11.93	5.20				
Niltava						
No. of loci with ≥1 SNP	4845	4051	0.84			
Total no. of SNP	84 983	18 472	0.22	2304	10 060	18 467
No. of SNP per locus	17.54	4.56				
Pycnonotus						
No. of loci with ≥1 SNP	2655	2432	0.92			
Total no. of SNP	30 622	11 626	0.38	1554	6957	11 624
No. of SNP per locus	11.53	4.78				
Stachyris						
No. of loci with ≥1 SNP	4838	3506	0.72			
Total no. of SNP	43 758	9154	0.21	1225	5658	9154
No. of SNP per locus	9.04	2.61				

© 2016 John Wiley & Sons Ltd

While we were able to sequence historical samples more deeply, this study identifies a downside in that sequences generated from historical samples tended to not go as far outside of the baited regions (i.e. recovering less flanking DNA) as modern samples. This is because the extent of coverage of flanking regions is a function of library insert size (Jones & Good 2016), which is smaller in historical samples. To achieve the same target coverage as modern samples, historical samples of some species needed to be sequenced more deeply (~150×) compared to modern samples (~50×). For other species (*N. grandis* and *S. nigriceps*) in our study, increasing sequencing depth in historical samples did not cause target coverage to converge to those seen in their modern counterparts. These results are consistent with those from McCormack *et al.* (2015). They found that sequences of UCE loci generated from older historical samples were shorter, and deeper sequencing of historical samples resulted in more recovered UCE loci, but not necessarily longer sequences. To improve target coverage, one solution would be to cover each targeted locus with more probes (overlapping or nonoverlapping). However, if having more independent loci is more desirable than having longer haplotypes, then the extra investment should be targeted towards probes for more loci, rather than more probes for each locus. Another possible way to match the width of coverage of modern and historical samples is to fine tune hybridization temperature when enriching each type of library (Paijmans *et al.* 2016). Nevertheless, irrespective of the discrepancies between modern and historical samples, the number of sites with ≥8× sequencing depth in any sample often exceeded one million.

Beyond increased DNA fragmentation, the other prime concern with the use of historical DNA is the elevation of C → T misincorporation at the 5′ end of reads and G → A misincorporation at the 3′ end of reads. In our study, the use of a mapping strategy that incorporates high sensitivity and local alignment (allowing ends of reads to be trimmed to maximize alignment scores) has reduced the impact of these errors significantly. Compared to other studies, which show up to 60% C → T or G → A substitution rate towards the termini of reads (Sawyer *et al.* 2012; Bi *et al.* 2013), our study shows rates that are more than an order of magnitude lower (observed rate = ~1.5–2%). After variant filtering (e.g. removing SNPs whose alternate alleles tend to occur towards the end of reads), we observe a slight to moderate (<13%) increase in C → T or G → A substitutions in historical samples compared to baseline (other types of substitutions). This suggests that not all C → T or G → A misincorporations have been removed, but the problem has been significantly attenuated. Better control of this type of error may justify the inclusion of C → T

Table 5 Count of reference and alternate alleles for modern and historical samples to determine whether C → T or G → A substitutions in historical samples occurred more frequently than expected based on frequencies of other types of substitutions. *P*-values are from *z*-tests of proportions

	C → T					G → A					Other substitutions		
Species group	Number of T allele (modern samples)	Number of T allele (historical samples)	% T from historical samples	% increase compared to expectation based on other substitutions	*P*-value	Number of A allele (modern samples)	Number of A allele (historical samples)	% A from historical samples	% increase compared to expectation based on other substitutions	*P*-value	Number of alternate allele (modern samples)	Number of alternate allele (historical samples)	% alternate allele from historical samples
Arachnothera	1598	2464	60.7	4.6	<0.01	1680	2410	58.9	1.6	0.24	5331	7363	58.0
Irena	560	1666	74.8	0.8	0.56	572	1913	77.0	3.6	<0.01	2191	6326	74.3
Niltava	4226	5625	57.1	13.0	<0.01	3910	4926	55.7	10.4	<0.01	14 265	14 565	50.5
Pycnonotus	1925	6197	76.3	8.3	<0.01	1929	5858	75.2	6.8	<0.01	6785	16 161	70.4
Stachyris	1725	4456	72.1	7.7	<0.01	1645	4072	71.2	6.4	<0.01	5063	10 251	66.9

© 2016 John Wiley & Sons Ltd

or G → A SNPs in downstream analyses, rather than have them removed completely (e.g. Bi *et al.* 2013). Another common issue with the use of historical DNA is that strand breaks tend to occur after purine bases due to depurination of the residues followed by hydrolysis of the phosphate-sugar backbone (e.g. Briggs *et al.* 2007). Because this does not involve nucleotide misincorporation, it has little negative effect on the accuracy of genotype calls. Instead, it causes slight bias in coverage and increases the likelihood of spuriously identifying independent paired-end reads as PCR duplicates (because reads have a greater chance of having the same starting points). The negative impact of this type of damage has not been thoroughly surveyed in the literature, but we believe that it can be at least partially alleviated by deeper sequencing.

Our study uses various filters at the levels of SNPs and individual genotypes to improve the quality of our variant call set – an approach that has been shown to be effective by other studies (Carson *et al.* 2014; Li 2014). Another effective approach is to recalibrate confidence of each variant by comparing the annotation profile of each variant against a well-validated set of variants (e.g. human HapMap) and selecting 'true' variants based on the desired level of sensitivity and specificity (Van der Auwera *et al.* 2013). This method, however, is not applicable to species lacking substantial genomic resources and high-quality sets of known variants. A third approach is to incorporate uncertainties in genotype likelihood directly into the estimates of population genetic parameters of interest (e.g. θ), which bypasses the need to generate high-confidence genotype calls (Nielsen *et al.* 2012; Fumagalli 2013). This approach is designed to handle low coverage data, and benefits from having large sample sizes, but requires populations to be known a priori, making it unsuitable for species whose population structure is poorly known.

Conclusions

We used sequence capture and a general probe set to generate large amounts of SNP data from multiple species of tropical birds. The feasibility to utilize degraded museum samples with this approach opens the door for similar studies focusing on geographic regions or clades that have poor modern sampling. Our effective sequencing coverage is both deep and wide. Combining the use of local alignment (vs. end-to-end alignment) during read mapping with SNP/genotype filtering, we generated high-confidence genotype calls that can be used in a variety of downstream analyses. Our approach of using a widely applicable probe set to generate population-level polymorphism data for multiple species is a viable and cost-effective alternative over one that creates a custom probe set for each study species.

Acknowledgements

This research was partially funded by National Museum of Natural History Small Grants Program to MJB and HCL, and American Ornithologists' Union Wetmore Research Award to HCL. HCL also received fellowship and research funding from Smithsonian Institution Molecular Evolution Fellowship. Portions of the laboratory and computer work were conducted in and with the support of the L.A.B. facilities of the National Museum of Natural History; we thank Jeffrey Hunt, Margaret Halloran, Brian Coyle, Noor White, Matthew Kweskin and Paul Frandsen for logistics and bioinformatics support. Additional computer analyses were conducted on Smithsonian Institution High-Performance Computing Cluster, Hydra-2. Extraction of historical DNA was conducted at the Smithsonian Conservation Biology Institute; we thank Nancy Rotzel McInerney, Rob Fleischer, Courtney Hofman and Taylor Callicrate for providing training and logistics support. We are indebted to the following natural history museums and personnel for facilitating tissue and toe pad loans: American Museum of Natural History (Paul Sweet), Academy of Natural Sciences of Drexel University (Nate Rice), Burke Museum of Natural History and Culture (Sharon Birks), Field Museum of Natural History (Ben Marks), Kansas University Museum of Natural History (Mark Robbins), Louisiana State University Museum of Natural Science (Donna Dittmann), Museum of Comparative Zoology at Harvard (Jeremiah Trimble), Museum of Vertebrate Zoology at Berkeley (Carla Cicero), Smithsonian Institution National Museum of Natural History (Brian Schmidt) and YPM Peabody Museum of Natural History (Kristof Zyskowski). We are grateful to Melissa Hawkins, Caroline Judy and two anonymous reviewers for providing helpful comments on the manuscript.

References

Bazin E, Glemin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**, 570–572.

Bi K, Linderoth T, Vanderpool D *et al.* (2013) Unlocking the vault: next-generation museum population genomics. *Molecular Ecology*, **22**, 6018–6032.

Bodi K, Perera AG, Adams PS *et al.* (2013) Comparison of commercially available target enrichment methods for next-generation sequencing. *Journal of Biomolecular Techniques*, **24**, 73–86.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Briggs AW, Stenzel U, Johnson PL *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences USA*, **104**, 14616–14621.

Briggs AW, Good JM, Green RE *et al.* (2009) Targeted retrieval and-analysis of five Neandertal mtDNA genomes. *Science*, **325**, 318–321.

Brotherton P, Endicott P, Sanchez JJ *et al.* (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base

© 2016 John Wiley & Sons Ltd

modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research*, **35**, 5717–5728.
Campana M, Hawkins M, Stewardson K *et al.* (2016) Simultaneous identification of host, vector and pathogen DNA via in-solution capture. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12524.
Carson AR, Smith EN, Matsui H *et al.* (2014) Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*, **15**, 125.
Church GM (2006) Genomes for all. *Scientific American*, **294**, 46–54.
Cooper A, Rhymer J, James HF *et al.* (1996) Ancient DNA and island endemics. *Nature*, **381**, 484.
Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.
Dumbacher JP, Pratt TK, Fleischer RC (2003) Phylogeny of the owlet-nightjars (Aves: Aegothelidae) based on mitochondrial DNA sequence. *Molecular Phylogenetics and Evolution*, **29**, 540–549.
Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.
Faircloth BC (2013) Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. doi: 10.6079/J9ILL.
Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.
Faircloth BC, Branstetter MG, White ND, Brady SG (2015) Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, **15**, 489–501.
Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS ONE*, **8**, e79667.
Gardiner LJ, Gawronski P, Olohan L *et al.* (2014) Using genic sequence capture in combination with a syntenic pseudo genome to map a deletion mutant in a wheat species. *Plant Journal*, **80**, 895–904.
Glenn TC, Nilsen R, Kieran TJ *et al.* (2016) Adapterama I: universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). *bioRxiv*, doi: http://dx.doi.org/10.1101/049114.
Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.
Good JM, Wiebe V, Albert FW *et al.* (2013) Comparative population genomics of the ejaculate in humans and the great apes. *Molecular Biology and Evolution*, **30**, 964–976.
Hahn MW (2008) Toward a selection theory of molecular evolution. *Evolution*, **62**, 255–265.
Hansen MC, Stehman SV, Potapov PV (2010) Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences USA*, **107**, 8650–8655.
Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2013) Sequence capture versus restriction site associated DNA sequencing for phylogeography. arXiv:1312.6439v1.
Harvey MG, Judy CD, Seeholzer GF *et al.* (2015) Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ*, **3**, e895.
Hawkins M, Hofman C, Callicrate T *et al.* (2015) In-solution hybridization for mammalian mitogenome enrichment: pros, cons, and challenges associated with multiplexing degraded DNA. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12448.
Hird SM, Brumfield RT, Carstens BC (2011) PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. *Molecular Ecology Resources*, **11**, 743–748.
Hodges E, Rooks M, Xuan Z *et al.* (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, **4**, 960–974.
Hodgkinson A, Casals F, Idaghdour Y *et al.* (2013) Selective constraint, background selection, and mutation accumulation variability within and between human populations. *BMC Genomics*, **14**, 495.
Hofman CA, Rick TC, Fleischer RC, Maldonado JE (2015) Conservation archaeogenomics: ancient DNA and biodiversity in the Anthropocene. *Trends in Ecology & Evolution*, **30**, 540–549.
Houde P, Braun MJ (1988) Museum collections as a source of DNA for studies of avian phylogeny. *The Auk*, **105**, 773–776.
Jones MR, Good JM (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, **25**, 185–202.
Jonsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, **29**, 1682–1684.
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.
Leache AD, Chavez AS, Jones LN *et al.* (2015) Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biol Evol*, **7**, 706–719.
Lee H, O'Connor BD, Merriman B *et al.* (2009) Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. *BMC Genomics*, **10**, 646.
Lemmon AR, Emme S, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.
Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.
Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP (2013) Capturing protein-coding genes across highly divergent species. *BioTechniques*, **54**, 321–326.
Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.
Maricic T, Whitten M, Paabo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*, **5**, e14004.
McCormack JE, Faircloth BC, Crawford NG *et al.* (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, **22**, 746–754.
McCormack JE, Tsai WL, Faircloth BC (2015) Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12466.
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.
Miettinen J, Shi C, Liew SC (2011) Deforestation rates in insular Southeast Asia between 2000 and 2010. *Global Change Biology*, **17**, 2261–2270.
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.
Paijmans JL, Fickel J, Courtiol A, Hofreiter M, Forster DW (2016) Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources*, **16**, 42–55.
Parks M, Lambert D (2015) Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics*, **16**, 19.
Paxinos E, McIntosh C, Ralls K, Fleischer R (1997) A noninvasive method for distinguishing among canid species: amplification and enzyme restriction of DNA from dung. *Molecular Ecology*, **6**, 483–486.
Paxinos EE, James HF, Olson SL *et al.* (2002) mtDNA from fossils reveals a radiation of Hawaiian geese recently derived from the Canada goose (*Branta canadensis*). *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1399–1404.
Penalba JV, Smith LL, Tonione MA *et al.* (2014) Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources*, **14**, 1000–1010.

© 2016 John Wiley & Sons Ltd

Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reed FA, Akey JM, Aquadro CF (2005) Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. *Genome Research*, **15**, 1211–1221.

Sawyer S, Krause J, Guschanski K, Savolainen V, Paabo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE*, **7**, e34131.

Seutin G, White BN, Boag PT (1991) Preservation of avian blood and tissue samples for DNA analyses. *Canadian Journal of Zoology*, **69**, 82–90.

Shapiro B, Hofreiter M (2014) A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, **343**, doi: 10.1126/science.1236573.

Simpson JT, Wong K, Jackman SD *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT (2014) Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, **63**, 83–95.

Stoeckle MY, Winker K (2009) A global snapshot of avian tissue collections: state of the enterprise. *Auk*, **126**, 684–687.

Suchan T, Pitteloud C, Gerasimova N *et al.* (2015) Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on museum collection specimens. *bioRxiv*, doi: http://dx.doi.org/10.1101/025551.

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **11**, 11.10.11–11.10.33.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

H.C.L. conceived and designed the project. H.C.L. performed molecular work and analysed data. H.C.L. and M.J.B. wrote the study.

Data accessibility

Pseudo-reference genome fasta files and BAM files: Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.2c220.

Appendix

Samples with successfully sequenced libraries and various associated data metrics. Species group indicate samples of target species and related outgroup taxa (if successfully sequenced). Institutional source: AMNH = American Museum of Natural History, ANSP = Academy of Natural Sciences of Drexel University, Burke = Burke Museum of Natural History and Culture, FMNH = Field Museum of Natural History, KUMNH = Kansas University Museum of Natural History, LSUMNS = Louisiana State University Museum of Natural Science, MCZ = Museum of Comparative Zoology at Harvard, MVZ = Museum of Vertebrate Zoology at Berkeley, NMNH = Smithsonian Institution National Museum of Natural History and YPM = YPM Peabody Museum of Natural History. Under 'Year collected', the year of collection of museum specimens are given or labelled as '?' if the year is unknown. Samples with frozen genetic materials are labelled as 'modern'

© 2016 John Wiley & Sons Ltd

© 2016 John Wiley & Sons Ltd

Species group	Genus	Species	Institutional source	Year collected	Catalogue number	Latitude	Longitude	DNA used in lib prep (ng)	No. of raw read pairs	Number of cleaned read pairs	Mapping rate (%)	% Duplicated reads	Avg. sequencing depth	% of target region with ≥8× sequencing
a	*Arachnothera*	*crassirostris*	KUMNH	Modern	24 436			991	2 276 797	2 123 960	78.3	88.3	16.4	67.3
a	*Arachnothera*	*longirostra*	AMNH	Modern	10 813	15.18	108.03	1000	6 638 421	6 261 529	15.0	86.5	10.3	32.8
a	*Arachnothera*	*longirostra*	AMNH	Modern	12 308	15.18	108.03	992	2 880 565	2 697 033	18.0	84.2	6.3	22.2
a	*Arachnothera*	*longirostra*	KUMNH	Modern	23 171	11.38	107.06	120	3 142 971	2 858 412	76.9	92.3	14.5	57.2
a	*Arachnothera*	*longirostra*	KUMNH	Modern	23 173	11.38	107.06	1001	5 309 669	4 965 004	77.0	91.1	28.9	78.7
a	*Arachnothera*	*longirostra*	LSUMNS	Modern	36 446	5.58	116.49	1001	3 045 056	2 828 168	18.2	84.8	6.5	22.7
a	*Arachnothera*	*longirostra*	LSUMNS	Modern	38 541	4.40	117.89	1000	1 809 903	1 694 095	19.6	84.6	4.2	15.4
a	*Arachnothera*	*longirostra*	LSUMNS	Modern	52 076	1.72	103.50	998	4 108 567	3 860 197	17.9	84.9	8.5	27.9
a	*Arachnothera*	*longirostra*	LSUMNS	Modern	52 126	1.18	110.20	999	3 284 729	3 071 104	20.2	87.0	6.5	23
a	*Arachnothera*	*longirostra*	LSUMNS	Modern	58 174	4.16	114.03	1001	4 835 755	4 555 713	18.4	85.0	10.6	32.8
a	*Arachnothera*	*robusta*	LSUMNS	Modern	36 483			416	4 430 257	4 134 624	78.2	89.7	27.8	80.4
a	*Arachnothera*	*longirostra*	YPM	1938	18 051	19.20	102.72	189	2 251 910	2 136 162	66.2	47.7	55.2	57.4
a	*Arachnothera*	*longirostra*	YPM	1957	47 450	9.95	98.61	476	9 379 646	8 915 575	88.9	68.8	221.8	77.9
a	*Arachnothera*	*longirostra*	YPM	1957	47 454	7.58	99.58	375	159 233	151 325	90.6	55.2	5.5	22.3
a	*Arachnothera*	*longirostra*	YPM	1957	65 321	18.80	99.00	870	2 635 260	2 470 663	88.8	56.6	85.3	60.5
a	*Arachnothera*	*longirostra*	YPM	1959	68 966	16.75	98.94	533	5 152 543	4 930 547	88.3	61.8	149.8	69.5
a	*Arachnothera*	*longirostra*	YPM	1959	68 967	16.75	98.94	133	3 272 613	3 115 125	75.2	37.3	132.4	74.1
a	*Arachnothera*	*longirostra*	FMNH	1929	80 201	22.54	103.29	266	1 080 661	1 029 991	66.8	39.3	36.1	42.8
a	*Arachnothera*	*longirostra*	FMNH	1931	91 634	15.43	106.38	368	3 301 823	3 134 430	50.8	51.8	60.9	47.3
a	*Arachnothera*	*longirostra*	FMNH	1931	91 636	15.43	106.38	253	7 444 390	7 089 888	48.0	66.3	94.6	55
a	*Arachnothera*	*longirostra*	NMNH	1909	219 071	−6.60	106.80	137	296 763	276 441	64.2	17.1	12.3	30.3
a	*Arachnothera*	*longirostra*	NMNH	1909	219 072	−6.60	106.80	130	2 141 551	2 034 006	88.7	81.1	28.3	40.1
a	*Arachnothera*	*longirostra*	NMNH	1909	220 091	−6.21	106.61	148	471 316	443 979	60.7	16.7	19.0	36.2
a	*Arachnothera*	*longirostra*	NMNH	1930	307 417	12.47	102.39	170	1 858 644	1 718 931	43.2	23.5	48.2	48
a	*Arachnothera*	*longirostra*	NMNH	1925	307 418	13.13	101.04	145	1 038 300	970 657	75.6	44.9	32.5	39.8
a	*Arachnothera*	*longirostra*	NMNH	1954	459 990	17.21	101.21	602	12 424 979	11 912 408	86.4	72.9	246.8	66
a	*Arachnothera*	*longirostra*	NMNH	1965	534 528	14.78	101.12	308	2 805 713	2 656 310	54.9	15.9	109.5	81.9
a	*Arachnothera*	*longirostra*	NMNH	1965	534 529	14.78	101.12	613	4 266 002	4 061 567	44.0	22.9	122.3	74.5
a	*Arachnothera*	*longirostra*	NMNH	1983	585 214	17.83	82.33	652	3 489 794	3 261 131	89.8	42.6	152.1	84.3
a	*Arachnothera*	*longirostra*	NMNH	1986	585 629	9.58	77.32	372	3 712 696	3 510 748	54.9	17.5	142.0	85.9
i	*Chloropsis*	*venusta*	LSUMNS	Modern	70 063			123	6 479 148	6 105 166	64.3	92.3	27.5	74.8
i	*Irena*	*cyanogastra*	LSUMNS	Modern	14 294			1002	8 211 442	7 659 526	64.8	88.8	51.7	84.4
i	*Irena*	*puella*	LSUMNS	Modern	51 044	4.40	117.89	1004	4 400 479	4 073 162	65.5	92.3	19.3	65.4
i	*Irena*	*puella*	LSUMNS	Modern	57 075	2.94	113.03	998	1 104 670	1 027 440	60.8	90.7	5.5	26.2
i	*Irena*	*puella*	YPM	Modern	142 997	4.34	115.26	1000	1 641 642	1 530 533	58.3	90.3	8.1	39.3
i	*Irena*	*puella*	YPM	1949	9680	25.90	91.90	234	9 599 919	9 127 089	85.4	62.2	318.2	78.8
i	*Irena*	*puella*	YPM	1938	17 353	19.20	102.72	186	1 421 024	1 347 945	50.8	53.7	33.7	51.7
i	*Irena*	*puella*	ANSP	1901	38 966	−0.46	100.61	288	2 379 641	2 261 434	56.2	59.3	55.6	53.4
i	*Irena*	*puella*	ANSP	1901	39 172	−4.99	105.21	159	2 506 949	2 387 239	89.8	71.2	64.7	55.7
i	*Irena*	*puella*	MVZ	1923	43 865	5.43	100.27	725	268 366	259 041	88.1	23.7	18.1	36.8
i	*Irena*	*puella*	YPM	?	47 008	7.58	99.58	854	3 261 262	3 070 809	47.5	53.7	71.8	68.9
i	*Irena*	*puella*	ANSP	?	56 492	−7.00	106.50	0	2 981 899	2 838 307	76.4	73.8	58.7	57.1
i	*Irena*	*puella*	YPM	1960	68 529	16.75	98.94	721	1 776 833	1 669 031	65.9	41.8	68.5	63.7
i	*Irena*	*puella*	YPM	1959	68 535	16.75	98.94	321	4 156 085	3 918 620	42.6	42.5	101.9	65.2

Species group	Genus	Species	Institutional source	Year collected	Catalogue number	Latitude	Longitude	DNA used in lib prep (ng)	No. of raw read pairs	Number of cleaned read pairs	Mapping rate (%)	% Duplicated reads	Avg. sequencing depth	% of target region with ≥8× sequencing
i	*Irena*	*puella*	YPM	1962	73 857	8.78	117.83	949	6 857 172	6 459 342	60.7	66.0	142.0	77
i	*Irena*	*puella*	YPM	1958	76 836	24.70	91.68	366	3 005 902	2 857 189	80.9	49.7	125.2	66
i	*Irena*	*puella*	FMNH	1929	79 454	22.21	102.91	880	2 885 792	2 743 800	60.9	36.8	111.9	63.6
i	*Irena*	*puella*	FMNH	1929	79 464	22.37	102.85	980	4 721 110	4 416 541	63.0	56.2	126.6	60.3
i	*Irena*	*puella*	FMNH	1929	79 475	21.51	101.85	732	442 765	422 395	76.5	40.0	19.6	41.2
i	*Irena*	*puella*	ANSP	1928	82 751	18.82	98.89	137	2 057 412	1 966 068	46.6	60.3	38.2	54.4
i	*Irena*	*puella*	FMNH	1931	90 853	15.18	106.24	446	3 564 938	3 396 640	51.4	59.8	70.8	56
i	*Irena*	*puella*	FMNH	1931	90 856	15.11	105.80	735	4 037 794	3 828 467	48.4	46.0	100.6	59.4
i	*Irena*	*puella*	ANSP	1933	114 178	18.49	99.30	126	2 335 936	2 236 893	69.8	39.3	99.0	55.9
i	*Irena*	*puella*	ANSP	1933	114 180	20.87	99.94	113	5 684 109	5 415 320	64.1	66.3	120.1	60
i	*Irena*	*puella*	ANSP	1934	123 311	17.00	101.00	200	2 692 073	2 569 841	84.6	77.3	52.0	55.3
i	*Irena*	*puella*	ANSP	1935	124 243	12.68	101.24	113	2 512 102	2 359 566	54.7	32.2	91.5	61.4
i	*Irena*	*puella*	ANSP	1937	130 338	11.64	99.59	175	3 333 296	3 189 463	89.7	57.2	129.1	60.3
i	*Irena*	*puella*	ANSP	1938	131 513	7.65	99.45	214	1 146 090	1 093 034	90.4	66.6	35.2	49.7
i	*Irena*	*puella*	ANSP	1939	139 744	3.51	97.82	219	6 475 832	6 160 412	89.1	72.4	162.0	70.3
i	*Irena*	*puella*	MCZ	1936	177 788	−3.80	102.27	738	2 129 273	2 014 932	36.0	41.8	44.8	56.9
i	*Irena*	*puella*	FMNH	1939	213 980	−6.21	106.85	822	1 860 099	1 746 446	46.9	48.0	44.0	58.1
i	*Irena*	*puella*	FMNH	1938	237 499	15.27	74.51	1000	3 259 533	3 097 265	47.5	35.8	96.0	65.5
i	*Irena*	*puella*	NMNH	1914	249 000	11.66	102.56	226	539 444	519 280	71.6	15.8	32.4	42.6
i	*Irena*	*puella*	NMNH	1918	278 423	11.56	108.99	259	3 236 005	2 934 232	54.1	56.0	69.0	51.9
i	*Irena*	*puella*	NMNH	1928	311 146	12.47	102.39	288	2 818 383	2 685 453	87.9	50.7	124.7	63.3
i	*Irena*	*puella*	NMNH	1953	450 549	16.26	99.61	697	5 323 204	5 046 052	60.1	40.7	188.9	69.8
i	*Irena*	*puella*	NMNH	1954	452 227	16.99	101.08	837	2 763 560	2 629 476	42.2	18.3	95.3	66.6
i	*Irena*	*puella*	AMNH	1938	462 684	15.47	74.52	715	1 939 429	1 832 316	47.7	39.1	56.0	56.5
i	*Irena*	*puella*	NMNH	1961	475 601	11.88	108.20	645	1 682 022	1 600 690	51.6	15.7	73.0	67.8
i	*Irena*	*puella*	NMNH	1965	534 777	8.49	99.73	588	3 962 221	3 775 446	87.3	43.2	202.2	77.6
i	*Irena*	*puella*	AMNH	1890	565 002	17.82	97.68	131	2 071 656	1 970 755	85.3	69.3	52.5	50.2
i	*Irena*	*puella*	AMNH	1891	565 003	17.82	97.68	245	1 731 893	1 669 871	82.1	20.9	112.4	59.1
i	*Irena*	*puella*	AMNH	1873	565 014	11.35	76.80	599	1 966 352	1 810 328	45.9	51.6	40.2	52.2
n	*Niltava*	*davidi*	KUMNH	Modern	11 093			1001	4 463 364	4 207 056	88.1	91.9	16.1	57.3
n	*Niltava*	*grandis*	AMNH	Modern	12 300	15.18	108.03	999	4 339 148	4 106 604	15.7	74.6	8.2	24.6
n	*Niltava*	*grandis*	KUMNH	Modern	15 257	22.00	96.00	1000	5 419 144	4 997 188	90.8	90.3	23.8	62.8
n	*Niltava*	*grandis*	KUMNH	Modern	27 994	21.95	104.26	1004	1 900 816	1 760 254	88.6	78.6	18.1	59.5
n	*Niltava*	*grandis*	KUMNH	Modern	27 995	21.95	104.26	248	1 145 952	1 046 935	91.5	91.0	4.6	21.4
n	*Niltava*	*grandis*	NMNH	Modern	620 545	22.90	93.70	998	6 398 799	5 797 026	35.5	72.8	30.1	70.8
n	*Niltava*	*grandis*	NMNH	Modern	631 832	27.83	97.76	383	3 540 456	3 125 355	25.3	68.1	13.3	50.4
n	*Niltava*	*macgregoriae*	KUMNH	Modern	10 341			1010	5 068 700	4 679 176	89.7	90.8	20.7	64.8
n	*Niltava*	*grandis*	YPM	1938	41 013	4.07	97.23	445	2 617 646	2 467 456	91.5	62.0	45.2	34.2
n	*Niltava*	*grandis*	MVZ	1923	44 126	4.85	100.73	378	2 741 897	2 595 983	70.3	42.8	53.3	32.9
n	*Niltava*	*grandis*	ANSP	1928	87 134	18.82	98.89	139	5 396 119	5 158 196	92.1	65.0	88.8	38.3
n	*Niltava*	*grandis*	FMNH	1931	91 439	18.00	105.00	496	793 799	753 236	78.2	44.2	17.3	25.2
n	*Niltava*	*grandis*	FMNH	1931	91 441	15.43	106.38	1100	4 162 282	3 984 344	74.0	22.8	118.7	41.4
n	*Niltava*	*grandis*	ANSP	1933	112 990	18.83	98.89	155	1 213 776	1 145 585	15.9	39.9	5.3	16.8

© 2016 John Wiley & Sons Ltd

© 2016 John Wiley & Sons Ltd

Species group	Genus	Species	Institutional source	Year collected	Catalogue number	Latitude	Longitude	DNA used in lib prep (ng)	No. of raw read pairs	Number of cleaned read pairs	Mapping rate (%)	% Duplicated reads	Avg. sequencing depth	% of target region with ≥8× sequencing
n	*Niltava*	*grandis*	ANSP	1933	112 994	21.32	98.90	267	1 815 810	1 671 681	46.6	47.1	21.5	29.5
n	*Niltava*	*grandis*	ANSP	1933	114 777	18.79	98.96	161	2 645 925	2 523 878	89.2	59.3	49.1	35.1
n	*Niltava*	*grandis*	ANSP	1938	137 711	22.00	93.50	207	2 659 214	2 474 372	56.9	23.9	57.0	41.1
n	*Niltava*	*grandis*	ANSP	1938	137 713	22.00	93.50	175	3 161 730	3 001 162	88.1	69.0	42.5	37.2
n	*Niltava*	*grandis*	ANSP	1939	139 563	3.92	97.35	161	1 815 090	1 729 493	86.0	60.0	31.6	31.2
n	*Niltava*	*grandis*	ANSP	1939	139 566	3.92	97.35	129	4 331 423	4 137 797	92.2	72.7	55.4	36.1
n	*Niltava*	*grandis*	MVZ	1970	160 375	4.52	101.38	2280	5 355 248	5 068 787	87.9	53.2	112.1	47
n	*Niltava*	*grandis*	AMNH	1938	306 183	21.23	93.92	223	3 290 439	3 121 200	85.2	71.2	38.8	30.8
n	*Niltava*	*grandis*	AMNH	1938	306 186	21.23	93.92	321	2 645 868	2 528 237	86.0	69.3	35.5	32.5
n	*Niltava*	*grandis*	NMNH	1939	359 147	12.00	108.40	388	2 258 447	2 138 937	89.8	51.6	49.3	35.1
n	*Niltava*	*grandis*	NMNH	1967	519 870	27.23	90.65	411	7 098 644	6 721 617	94.1	57.3	146.7	53.2
n	*Niltava*	*grandis*	NMNH	1966	535 501	13.10	102.19	1690	2 194 944	2 087 623	78.7	19.9	69.2	47.9
n	*Niltava*	*grandis*	NMNH	1966	535 506	13.10	102.19	1560	846 714	799 684	74.6	32.3	21.9	35.6
p	*Pycnonotus*	*atriceps*	KUMNH	Modern	12 641	9.84	118.64	648	1 063 091	996 155	66.8	80.7	12.3	56.4
p	*Pycnonotus*	*atriceps*	KUMNH	Modern	12 642	9.84	118.64	1001	1 301 136	1 195 047	62.5	91.7	5.8	29.3
p	*Pycnonotus*	*atriceps*	LSUMNS	Modern	58 540	1.80	109.71	1001	4 130 186	3 930 482	19.1	73.7	9.4	31.1
p	*Pycnonotus*	*atriceps*	BURKE	Modern	116 982	−2.27	101.03	608	5 222 077	4 914 687	64.3	88.2	35.1	79
p	*Pycnonotus*	*atriceps*	ANSP	Modern	tiss_16118	5.23	116.00	1001	4 574 225	4 235 623	67.7	86.5	37.2	67
p	*Pycnonotus*	*atriceps*	ANSP	Modern	tiss_16185	5.23	116.00	998	2 711 503	2 553 741	69.2	78.8	36.4	75.9
p	*Pycnonotus*	*eutilotus*	LSUMNS	Modern	57 023			1001	4 634 881	4 361 179	66.1	92.0	21.8	70.4
p	*Pycnonotus*	*atriceps*	YPM	1945	17 185	16.57	104.75	198	1 989 692	1 885 769	89.2	70.2	51.5	53.8
p	*Pycnonotus*	*atriceps*	YPM	1944	17 186	16.57	104.75	109	2 411 927	2 257 054	66.3	15.0	125.8	54.6
p	*Pycnonotus*	*atriceps*	ANSP	1901	38 984	−0.46	100.61	124	9 489 253	9 131 179	72.5	31.2	629.9	68.3
p	*Pycnonotus*	*atriceps*	MVZ	1921	42 709	5.43	100.27	280	1 198 793	1 146 951	73.5	37.5	52.7	52.4
p	*Pycnonotus*	*atriceps*	MVZ	1923	43 852	5.43	100.27	625	2 342 456	2 216 718	67.7	42.9	84.6	59.1
p	*Pycnonotus*	*atriceps*	YPM	?	47 011	7.58	99.58	273	199 780	190 207	86.4	35.9	8.5	29.5
p	*Pycnonotus*	*atriceps*	YPM	1957	47 021	10.00	98.60	586	6 609 384	6 305 022	86.5	63.1	217.6	73.1
p	*Pycnonotus*	*atriceps*	YPM	?	64 313	18.80	100.80	896	3 322 241	3 143 789	86.1	54.1	134.8	70.9
p	*Pycnonotus*	*atriceps*	YPM	?	64 314	18.80	100.80	442	3 806 495	3 601 933	24.6	44.0	39.2	55.8
p	*Pycnonotus*	*atriceps*	YPM	1958	77 504	23.00	92.25	575	10 381 959	9 879 062	81.3	62.3	323.3	77.4
p	*Pycnonotus*	*atriceps*	YPM	1958	77 511	24.70	91.68	299	13 877 059	13 215 854	77.5	61.9	412.9	79.2
p	*Pycnonotus*	*atriceps*	ANSP	1932	112 576	14.20	101.23	110	1 034 854	977 078	75.9	45.2	38.1	46.3
p	*Pycnonotus*	*atriceps*	ANSP	1933	112 577	8.43	99.96	124	1 160 672	1 098 346	44.9	47.8	21.0	44
p	*Pycnonotus*	*atriceps*	ANSP	1939	139 756	3.69	97.60	237	627 764	570 062	61.4	19.8	26.1	43.3
p	*Pycnonotus*	*atriceps*	ANSP	1939	139 759	3.69	97.60	189	5 142 367	4 886 973	65.5	56.0	142.8	65.1
p	*Pycnonotus*	*atriceps*	MVZ	1965	156 056	10.75	106.67	940	7 221 550	6 883 236	69.2	57.8	205.9	72.6
p	*Pycnonotus*	*atriceps*	MVZ	1965	156 057	10.75	106.67	972	2 518 289	2 399 890	69.9	43.3	98.0	66.5
p	*Pycnonotus*	*atriceps*	NMNH	1907	181 552	−5.80	112.65	130	209 106	199 225	78.7	28.9	10.8	30.6
p	*Pycnonotus*	*atriceps*	NMNH	1913	182 876	1.10	117.90	287	4 156 380	4 020 054	83.6	33.5	229.9	63.1
p	*Pycnonotus*	*atriceps*	NMNH	1929	313 386	17.96	102.61	121	6 563 455	6 248 015	82.3	62.6	193.2	63.5
p	*Pycnonotus*	*atriceps*	NMNH	1929	313 387	17.96	102.61	170	5 963 292	5 689 005	80.7	73.1	121.0	59
p	*Pycnonotus*	*atriceps*	NMNH	1952	450 852	11.63	99.60	450	15 540 432	15 048 523	89.8	27.9	1037.9	81.8
p	*Pycnonotus*	*atriceps*	NMNH	1955	459 653	17.48	101.50	429	7 457 750	7 050 978	86.8	72.9	177.1	70.6

Species group	Genus	Species	Institutional source	Year collected	Catalogue number	Latitude	Longitude	DNA used in lib prep (ng)	No. of raw read pairs	Number of cleaned read pairs	Mapping rate (%)	% Duplicated reads	Avg. sequencing depth	% of target region with ≥8× sequencing
p	*Pycnonotus*	*atriceps*	NMNH	1955	459 657	17.06	101.09	1750	2 631 472	2 529 077	88.2	25.4	174.2	68.5
s	*Stachyris*	*nigriceps*	KUMNH	Modern	9957	22.80	108.30	1000	3 010 458	2 811 802	89.0	76.6	32.2	73.1
s	*Stachyris*	*nigriceps*	AMNH	Modern	10 737	15.18	108.03	1006	6 424 583	6 087 682	17.5	79.5	10.8	30
s	*Stachyris*	*nigriceps*	NMNH	Modern	15 183	27.50	97.80	1000	3 752 271	3 380 382	46.5	70.2	25.3	66.4
s	*Stachyris*	*nigriceps*	KUMNH	Modern	15 245	22.00	96.00	1000	6 550 641	5 894 063	44.3	67.6	46.0	75.2
s	*Stachyris*	*nigriceps*	KUMNH	Modern	15 246	22.00	96.00	1003	5 467 961	5 114 238	41.3	67.0	37.8	73.6
s	*Stachyris*	*nigriceps*	LSUMNS	Modern	51 000	4.85	115.70	999	5 050 862	4 750 042	89.0	80.5	45.4	79.5
s	*Stachyris*	*nigriceps*	LSUMNS	Modern	51 017	4.85	115.70	1000	5 687 755	5 308 452	89.5	79.8	52.9	82.5
s	*Stachyris*	*nigriceps*	LSUMNS	Modern	78 736	3.78	115.48	999	4 882 109	4 630 716	18.8	83.3	7.0	22.9
s	*Stachyris*	*nigriceps*	LSUMNS	Modern	78 756	3.78	115.48	1008	4 307 397	4 073 643	18.6	83.2	6.1	20.4
s	*Stachyris*	*nigriceps*	BURKE	Modern	117 024	−0.84	100.53	273	2 029 960	1 894 130	89.3	89.1	10.1	45.1
s	*Stachyris*	*nigriceps*	YPM	1959	38 021	27.37	97.45	407	4 406 930	4 148 111	29.2	41.9	37.4	41
s	*Stachyris*	*nigriceps*	YPM	1959	68 729	16.75	98.94	520	6 705 078	6 328 877	48.8	34.4	110.3	51.2
s	*Stachyris*	*nigriceps*	FMNH	1929	78 922	21.68	102.10	259	1 763 423	1 676 870	71.4	52.5	29.0	28.5
s	*Stachyris*	*nigriceps*	FMNH	1929	78 924	21.39	101.97	345	1 875 988	1 788 618	79.1	45.0	41.6	33.2
s	*Stachyris*	*nigriceps*	FMNH	1931	91 010	15.43	106.38	1060	3 093 949	2 910 319	68.0	38.9	63.8	38.9
s	*Stachyris*	*nigriceps*	FMNH	1931	91 030	15.43	106.38	391	1 841 115	1 756 448	79.1	35.3	49.0	36.9
s	*Stachyris*	*nigriceps*	ANSP	1933	112 341	20.70	100.11	173	942 799	892 031	44.0	40.9	12.2	24.6
s	*Stachyris*	*nigriceps*	ANSP	1933	112 342	20.70	100.11	228	1 966 290	1 865 242	37.2	43.9	20.3	27.8
s	*Stachyris*	*nigriceps*	ANSP	1939	139 856	3.81	97.28	243	3 417 849	3 262 057	71.6	35.0	80.9	35.1
s	*Stachyris*	*nigriceps*	MCZ	1937	196 421	18.58	98.48	399	5 416 997	5 156 856	46.8	33.3	86.8	43.3
s	*Stachyris*	*nigriceps*	MCZ	1937	196 424	18.58	98.48	321	2 744 283	2 603 230	93.4	50.2	65.8	36.2
s	*Stachyris*	*nigriceps*	MCZ	1938	265 882	21.23	93.92	233	2 005 249	1 908 483	85.6	73.3	22.9	28.7
s	*Stachyris*	*nigriceps*	MCZ	1938	265 883	21.23	93.92	174	1 822 931	1 727 774	73.7	82.3	11.3	23.2
s	*Stachyris*	*nigriceps*	MCZ	1939	267 896	19.88	102.14	345	2 557 232	2 430 253	56.3	62.3	27.8	31.6
s	*Stachyris*	*nigriceps*	MCZ	1939	267 907	21.92	102.10	433	1 418 996	1 336 364	57.1	45.6	22.4	29.3
s	*Stachyris*	*nigriceps*	NMNH	1930	330 586	19.00	99.42	238	6 868 647	6 499 553	82.1	67.7	89.6	35
s	*Stachyris*	*nigriceps*	NMNH	1964	534 904	6.94	100.26	538	8 270 920	7 961 597	88.2	25.2	281.9	56.8
s	*Stachyris*	*nigriceps*	AMNH	1908	589 674	24.27	97.23	403	1 136 110	1 070 732	68.1	42.2	21.4	24.2
s	*Stachyris*	*nigriceps*	AMNH	1901	589 678	4.63	102.24	110	6 011 991	5 714 506	65.8	83.8	28.2	26.5
s	*Stachyris*	*nigriceps*	AMNH	1901	589 679	4.63	102.24	130	1 216 692	1 148 991	84.6	68.4	15.7	24.4
s	*Stachyris*	*nigriceps*	AMNH	1918	589 692	3.27	98.55	317	3 054 942	2 889 372	88.7	55.5	59.8	30.7

© 2016 John Wiley & Sons Ltd

Supporting Information

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

Fig. S1. Agilent Bioanalyzer results showing fragment size of pooled enriched products to be sequenced. Products are separated into three pools, one containing enriched products of modern samples (modDNA), and two containing enriched products of historical samples (pool 7-40 and pool 23-39).

Fig. S2. Number of cleaned read pairs per sample. Each sample is represented by one dot and *x*-axis indicates arbitrarily numbered enrichment pools. The last eight pools to the right contain libraries of modern samples while the remainders are pools containing libraries of historical samples. Sequencing output (zero or close to zero) of failed libraries are shown.

Fig. S3. Bar charts of the number of cleaned read pairs per sample. Each bar chart shows one enrichment pool; numbering scheme follows that of Fig. S2. Sequencing output (zero or close to zero) of failed libraries are not shown.

Fig. S4. Bowtie mapping rate of each sample. Each sample is represented by one dot and *x*-axis indicates arbitrarily numbered enrichment pools. Numbering scheme follows that of Fig. S2.

Fig. S5. Read duplication rate of each sample. Each sample is represented by one dot and *x*-axis indicates arbitrarily numbered enrichment pools. Numbering scheme follows that of Fig. S2.

Fig. S6. Number (averaged over samples) of sites (*y*-axis, in $\log_{10}$ scale) with various sequencing depth (*x*-axis). Bar charts are sorted according to species group and age of samples (modern vs. historical).

Fig. S7. Count of UCE loci with various proportion of sites with at least 8× average sequencing depth. Frequency histograms are sorted by species groups (rows) and age of sample (modern – left column, historical – right column). *A. longirostra* group (A and B); *I. puella* group (C and D); *N. grandis* group (E and F); *P. atriceps* group (G and H); and *S. nigriceps* group (I and J).

Fig. S8. Rate of C to T (left column) and G to A (right column) substitutions of five exemplar historical samples (one from each of the five study species). Rates are shown for each of the first 25 bases from the 5′ (left column) or 3′ end (right column).

Fig. S9. Empirical values (line) and Bayesian estimates (filled circle, error bars = 95% posterior prediction intervals) of rates of various substitutions at the first 11 base positions from the start of the 5′ end (positive values on *x*-axis) and 3′ end (negative values on *x*-axis) of each read. Red = C to T substitution, green = G to A substitution, blue = other substitutions. Results from five exemplar historical samples are shown.

Fig. S10. The top four panels show frequency of each of the four nucleotides within reads (demarcated by grey boxes, first 10 and last 10 positions are shown), and just upstream or downstream of reads (based on pseudo-reference genome). Each dot represents the average frequency for each UCE locus at each position, and solid lines show the 'genome-wide' values. The bottom left panel shows: 1) observed C to T substitution rate if soft-clipped bases are included (yellow line), and when soft-clipped bases are excluded (red line); 2) G to A substitution (blue line). The bottom right panel shows: 1) observed G to A substitution rate if soft-clipped bases are included (yellow line), and when soft-clipped bases are excluded (blue line); 2) C to T substitution (red line). Positive *x*-axis labels are base position from the 5′ end of each read (going downstream); negative *x*-axis labels are base position from the 3′ end of each read (going upstream).

Fig. S11. Average number of called genotypes per sample for each of the five species groups, sorted according to age of samples (hist. = historical, mod. = modern). Labels for the five species groups are: A- *Arachnothera*; I - *Irena*; N - *Niltava*; P - *Pycnonotus*; S - *Stachyris*. Error bars show standard deviations.

© 2016 John Wiley & Sons Ltd