MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2011) **11**, 1082–1092 doi: 10.1111/j.1755-0998.2011.03052.x

Museum genomics: low-cost and high-accuracy genetic data from historical specimens

KEVIN C. ROWE,*† SONAL SINGHAL,†‡ MATTHEW D. MACMANES,†‡ JULIEN F. AYROLES,§ TONI LYN MORELLI,† EMILY M. RUBIDGE,† KE BI† and CRAIG C. MORITZ†‡

**Sciences Department, Museum Victoria, GPO Box 666, Melbourne, Vic. 3001, Australia, †Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720, USA, ‡Department of Integrative Biology, 3060 Valley Life Sciences Building, University of California, Berkeley, CA 94720, USA, §Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA*

Abstract

Natural history collections are unparalleled repositories of geographical and temporal variation in faunal conditions. Molecular studies offer an opportunity to uncover much of this variation; however, genetic studies of historical museum specimens typically rely on extracting highly degraded and chemically modified DNA samples from skins, skulls or other dried samples. Despite this limitation, obtaining short fragments of DNA sequences using traditional PCR amplification of DNA has been the primary method for genetic study of historical specimens. Few laboratories have succeeded in obtaining genome-scale sequences from historical specimens and then only with considerable effort and cost. Here, we describe a low-cost approach using high-throughput next-generation sequencing to obtain reliable genome-scale sequence data from a traditionally preserved mammal skin and skull using a simple extraction protocol. We show that single-nucleotide polymorphisms (SNPs) from the genome sequences obtained independently from the skin and from the skull are highly repeatable compared to a reference genome.

Keywords: historical DNA, natural history collections, next-generation sequencing, *Rattus*

Received 1 April 2011; revision received 16 June 2011; accepted 24 June 2011

Introduction

By preserving large numbers of specimens from across the geographical range of a great breadth of species, natural history collections (NHC) are repositories of an immense record of the change in biodiversity conditions over historical time (Shaffer *et al.* 1998; Graham *et al.* 2004; Suarez & Tsutsui 2004). Traditional specimens (e.g. whole animals, skulls and skins) dating back decades, centuries in some cases, provide a wealth of data on the geographical distribution, phenotypic variation and identity of species (Goldstein & Desalle 2003; Austin & Melville 2006; Murphy *et al.* 2010). These long-term time series have proven invaluable in tracking species' and population-level changes in response to environmental change (Moritz *et al.* 2008). However, until the last few decades, NHCs did not preserve materials (e.g. blood or tissue) that are appropriate for molecular studies. Thus, studies of the genetic changes in populations have required the laborious and technically challenging tasks of extracting and PCR-amplifying degraded genetic material from traditional specimens (Taberlet *et al.* 1996; Cooper & Poinar 2000; Paabo *et al.* 2004). The resulting products have been short (<500 bp) sequences of mitochondrial DNA or small microsatellite fragments (<300 bp) (Thomas *et al.* 1990; Ellegren 1991; Cooper *et al.* 1992) These snippets of genetic data, obtained at great effort, have proven extremely informative, especially when contemporary populations do not exist or have experienced dramatic changes in population size or connectivity (Cooper *et al.* 1992; Taylor *et al.* 1994; Bouzat *et al.* 1998; Peery *et al.* 2010; reviewed in Wandeler *et al.* 2007).

Skins, bones, teeth, nails and other dried material are the most common types of historical specimens preserved for mammals and birds. Exposure to air, light and chemicals, sometimes for decades, has often resulted in extensive degradation of DNA from historical specimens (i.e. postmortem damage; see (Willerslev & Cooper 2005) for thorough review). This degradation is particularly challenging for amplification and accurate sequencing by traditional PCR and Sanger sequencing. One of the biggest challenges is that DNA from historical specimens has been sheared into small fragments, usually on the order of a few hundred base pairs in length or less,

Correspondence: Kevin C. Rowe, Fax: 61 3 8341 7421; E-mail: krowe@museum.vic.gov.au

that require the designing of specific PCR primers and the stitching together of sequences from multiple small PCR products. Another major challenge for obtaining reliable genetic data from historical DNA is that chemical damage can occur during specimen storage (i.e. such as deamination of cytosine caused by UV exposure). These resulting mutations can deteriorate PCR priming sites, leading to PCR failure for specimens of interest. Perhaps even more troublesome is that even if PCR of historical DNA is successful, it often succeeds for only one or a small number of DNA template copies and thus can propagate postmortem mutations into the resulting sequence trace. Thus, traditional PCR amplification and sequencing of historical specimens require considerable resources to obtain reliable results.

Next-generation sequencing (NGS) methods, originally developed for the shotgun sequencing of whole genomes at low cost, have several advantages that are directly applicable to obtaining DNA sequences from highly fragmented and degraded DNA templates. First, NGS methods rely on sequencing large quantities of short fragments (100–400 bp), and so the fragmented nature of historical specimen DNA does not pose a major problem. Second, NGS library preparation involves ligation of adapter sequences to both ends of the template DNA fragment, thus avoiding the problems of template–primer mismatch. This latter result could lead to increased amplification success and reduced propagation of postmortem mutations occurring on individual strands. While the application of NGS technologies to ancient DNA from palaeo-specimens has had some success (Gilbert *et al.* 2007; Lindqvist *et al.* 2010; Rohland *et al.* 2010) and the potential has been reviewed (Miller *et al.* 2008; Knapp & Hofreiter 2010), its application to historical NHC specimens is still rare and limited in scope (Miller *et al.* 2009).

The development of techniques for sequencing historical genomes comes at a critical time in museum bioinformatics. NHC that include specimens of extirpated populations or species can contribute significantly to studies of biodiversity loss, conservation and population genetics (Roy *et al.* 1994; Nielsen *et al.* 1997; Pichler *et al.* 2001; Hansen 2002; Martinez-Cruz *et al.* 2007; Peery *et al.* 2010). Screening of a large number of loci spread throughout the genome (Vera *et al.* 2008; Meyer *et al.* 2009) will greatly improve the accuracy and power to infer population demographic history (Gilad *et al.* 2009). In addition, genomic data may allow researchers to identify the genetic basis of evolutionary adaptation and facilitate inferences about recent selection (Luikart *et al.* 2003). This is especially meaningful in the face of rapid human-induced environmental change and for predicting consequences of future change (Suarez & Tsutsui 2004; Robbirt *et al.* 2011).

With these advantages in mind, we set out to determine the feasibility and reliability of obtaining genome-scale sequences from historical museum specimens at a cost that is comparable to standard PCR-based research projects. We compared high-coverage reads from the same extraction to estimate sequencing error rates, and we compared high-coverage reads from two different source materials (skin and skull) from the same individual to determine whether this approach recovers consistent sequences from the degraded DNA in historical specimens. Finally, we discuss the prospects for population genomic analyses of museum specimens.

Methods

Museum specimens and DNA extraction

We extracted total genomic DNA from two traditionally preserved specimens of *Rattus norvegicus* containing a skin and a skull stored separately in the teaching collection in the Department of Integrative Biology (formerly Department of Zoology) at the University of California, Berkeley. Specimen DZ397 was a male *R. norvegicus* collected on 12 June 1939 from Alameda, CA. Specimen DZ762 was a female *R. norvegicus* collected on 2 January 1963 from the UC Berkeley campus. From the skin of each specimen, we sampled (i) a 5 × 5 mm section of skin at the base of the lip, (ii) a toe from the right hind foot, including the first and second phalange, and (iii) the ankle from the right forefoot, including the carpals. For each sample, we removed the hair and the outer layer of skin using a sterile, disposable surgical blade and stored the sample in a sterile 1.5-mL microcentrifuge tube. From the skull of each specimen, we collected a sample from the second and third right maxillary molars by drilling into the roots of the molars from the upper surface of the maxillary using a hand-held dremmel tool and a sterilized drill bit, 2 mm in diameter. We discarded the powder resulting from drilling into the surface bone (~2 mm). We collected the remaining powder onto a sterilized weigh boat and stored the sample in a sterile 1.5-mL microcentrifuge tube. We collected all samples following sterile procedures that included wearing gloves, eye shield and mask. We weighed samples to the nearest 0.01 g. Samples ranged in mass from 0.02 to 0.09 g.

We conducted all DNA extractions in a room used exclusively for extraction of museum specimen DNA. No previous studies in this laboratory space involved any *Rattus* specimens, and no PCR or extractions of *Rattus norvegicus* had been carried out in adjacent laboratories. All stages of the extraction process included a negative control run in parallel. We extracted total genomic DNA using a slight modification of the protocol of Mullen & Hoekstra (2008) that was itself a modification of the Qiagen DNeasy Blood & Tissue Kit extraction protocol (Qiagen). Our two main modifications to Mullen and Hoekstra's published protocol were (i) grinding of our samples in liquid nitrogen and (ii) substitution of Qiaquick® columns (Qiagen) for DNeasy columns, which allow the collection of smaller DNA fragments. Extractions resulted in a final volume of 100 μL. Our full extraction protocol is available from the authors upon request. We ran aliquots (10 μL) of the extractions alongside a 100-bp ladder on a 2% agarose gel by electrophoresis. We stained gels with GelStar Nucleic Acid Gel Stain (Cambrex Bio Science Rockland, Inc.) and visualized total nucleotide concentrations and fragment lengths by UV illumination. We also quantified concentrations of DNA extractions on a Bioanalyzer 2100 with DNA standards at 15 and 1500 bp.

Library and sequencing

We prepared the DNA extractions for sequencing by Illumina technology following the standard DNA protocol (Illumina, Paired-End Sample Preparation Guide, document # 1005063 Rev. D) with reagents provided by NEB (New England Biolabs # E6006s); protocol modification is described below. Typically, the first step in DNA library construction involves the fragmentation of DNA molecules into pieces <800 bp using either a mechanical or enzymatic process. Because the DNA from museum skins is already fragmented (Fig. 1), we omitted this step. From here, we completed steps including end-repair, adenylation of 3′ ends and adapter ligation. For nucleotide recovery and purification between steps of the Illumina protocol, we used the Agencourt Ampure XP (Beckman Coulter) magnetic-bead purification protocol, which has a higher sample recovery rate than standard

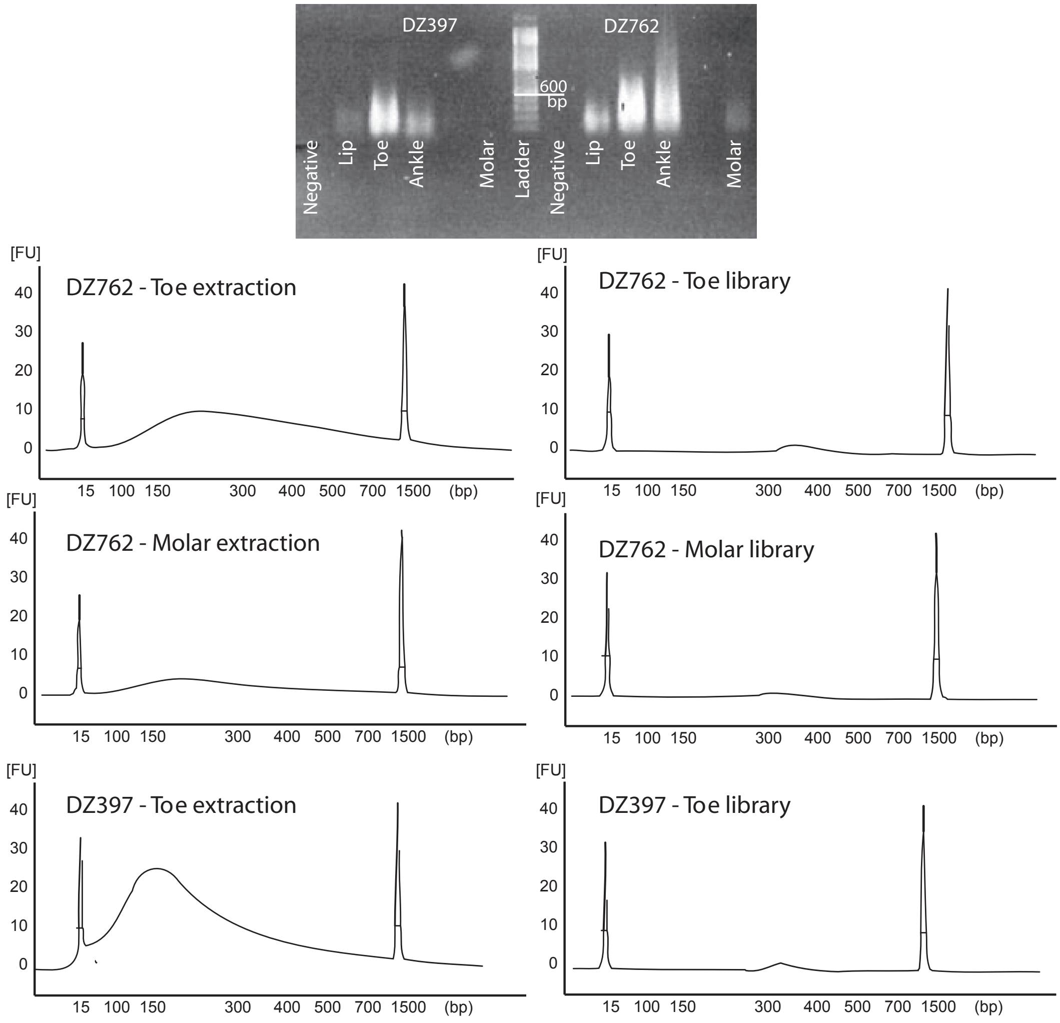

Fig. 1 Electrophoresis gel of all DNA extractions and bioanalyzer traces for extractions and Illumina library preparations.

silica column purification (Quail *et al.* 2008). We finished the library prep with a PCR amplification step and, to reduce error propagated downstream, we used a Phusion PCR High Fidelity master mix (NEB, F-531S). This master mix does not incorporate dUTPs, reducing the propagation of deamination residues that result from conversion of cytosine to uracil and are common with historical specimens. Finally, because the physical process of sequencing (as well as downstream bioinformatics) is optimized when using a known fragment size with narrow size distribution, we size-selected fragments from about 200 to 400 bp using gel electrophoresis and recovery. We purified the final product and analysed it on the Agilent Bioanalyzer 2100 (Fig. 1). Once satisfied with the quality of the final library, we sent the purified libraries to Vincent J. Coates Genomics Sequencing Laboratory (Q3B, University of California, Berkeley) for Illumina sequencing on a GAIIx. We requested one lane of 101-bp paired-end sequencing per library.

Data filtration

To preprocess the raw data, we first removed any sets of paired reads that were identical in both the forward and reverse direction; these pairs are probably the result of overamplification of the library (i.e. PCR duplicates) and can artificially inflate coverage estimates. We trimmed the resulting unique reads for adaptor and low-quality sequence using in-house Perl scripts (available from S. Singhal upon request). We then aligned the trimmed reads to the human (hg19) and *Escherichia coli* (NCBI st. 536) genomes using the short-read aligner, Bowtie (Langmead *et al.* 2009), and removed any reads that were likely sourced from these common contaminants.

Alignment, SNP detection and analyses

Using Bowtie (Langmead *et al.* 2009) and a single-end mapping strategy, we aligned all reads to the *Rattus* RGSC 3.4 genome. We also attempted a paired-end strategy but found that this reduced alignment efficiency (see Results). We used the samtools suite to parse resulting alignments and to call single-nucleotide polymorphisms (SNPs) in the mitochondrial genome (Li *et al.* 2009). We used the samtools output to calculate coverage and to estimate error rates. In particular, to calculate error rate, we assumed that any nucleotide discrepancies between mapped reads and the specimen's consensus mitochondrial sequence were because of either DNA damage or sequencing error (Miller *et al.* 2008). Finally, to determine the source of the reads that did not align to *R. norvegicus*, we queried a subsample of unaligned reads (N = 100 000) against the nonredundant nucleotide database hosted by NCBI using 'blastn' and BioPerl (Stajich *et al.* 2002; Johnson *et al.* 2008). We prepared all statistical analyses and graphs using R (R Development Core Team 2010).

To evaluate the accuracy of SNP detection, we aligned our inferred mitochondrial genome with five other full mitochondrial genomes from *R. norvegicus* (NCBI IDs: AC_000022, AJ428514, DQ673917, NC001665, DQ673916; (Gadaleta *et al.* 1989; Nilsson *et al.* 2003; Schlick *et al.* 2006) and with an outgroup sequence (*R. rattus*; NCBI ID: NC_012374; Robins *et al.*, 2008) using MUSCLE (Edgar 2004). We built a neighbour-joining tree using a HKY model implemented in Geneious (Drummond *et al.* 2011).

Assembly

As many species of evolutionary and ecological interest lack a reference genome, we evaluated the efficacy of NGS data from historical samples for *de novo* assembly. To this end, we assembled both lanes of our sequence data using the short-read de novo assembly program ABySS, using a range of k-mer and coverage sizes to optimize contiguity (Simpson *et al.* 2009). To measure assembly accuracy, we used BLAT to align all assembled contigs >200 bp against the rat RGSC 3.4 genome (Kent 2002). We performed all analyses using in-house BioPerl scripts run on the Texas Advanced Computing Center Ranger cluster (http://www.tacc.utexas.edu/).

Results

Museum specimens and DNA extraction

We obtained sufficient DNA quantities from all specimen types, except the molar extraction from our 1939 specimen (DZ397), to be visually detectable on an agarose gel (Fig. 1). As expected for traditionally preserved specimens, all extractions resulted in highly fragmented DNA with most DNA fragment sizes <1000 bp in length. Bioanalyzer results were consistent with the visualization of the DNA on the agarose gel; DNA concentrations ranged from 10.6 to 86.6 ng/μL, and modes of fragment lengths ranged from 100 to 2500 bp (Fig. 1). We obtained the largest fragment lengths and total nucleotide concentrations from our samples that contained bone (toe and ankle), with distributions of fragment lengths extending above 1000 bp. We recovered the smallest fragment lengths from our two lip samples, with no detectable concentrations of fragments longer than 300 bp. We obtained the lowest concentration of total nucleotides from our one successful molar sample but recovered fragment lengths extending above 1000 bp in length. We did not recover detectable nucleotide levels from our negative controls either on the agarose gel or on bioanalyzer. We selected the toe and molar samples from specimen DZ762 and the toe sample from specimen DZ397 to develop a genomic library for NGS. We developed high-quality libraries in the 300- to 400-bp range from all three samples, whereas we did not produce any detectable library from our negative controls (Fig. 1). The toe and molar samples from specimen DZ762 represented independently stored genomic specimens from the same individual, allowing us to assess the reliability of our sequence results in spite of any possible postmortem mutations that would be specific to each sample. Therefore, we selected the toe and the molar libraries from specimen DZ762 to submit for Illumina GAIIx sequencing, with each library run in a separate lane.

Sequencing and preprocessing

Sequencing resulted in 46 million reads (4.6 billion base pairs, Gbp) for the toe-derived library and 48 million reads (4.8 Gbp) for the molar-derived library (Fig. S1). These read counts are typical of those generated by the Illumina GAIIx at the QB3 sequencing facility for other genome projects (L. Tonkin, personal communication). Duplicate reads represented a small fraction of the data set (1.1% and 2.6% for the molar and toe libraries, respectively), and contamination by *E. coli* or *H. sapiens* was negligible (<0.1% for both libraries). After extensive trimming for low quality and adapter sequences, each library contained 4.1 Gbp of sequence data (data available from DRYAD entry doi: 10.5061/dryad.1fm3f).

Alignment and SNP detection

We aligned each of the libraries using both a paired-end and single-end strategy but ultimately found the paired-end strategy to be inefficient. In most cases (82.3%), only one end of the pair mapped to the genome, while the other end of the pair generally aligned to the genome with low mapping quality, despite being of good sequencing quality. Predominantly, the low-quality read was the 'reverse' read. For a substantial number of paired reads, each half either mapped to different chromosomes or mapped to the same chromosome at distances much greater than our insert size (14.3%, Fig. S2). These mismapped pairs may reflect chimeras of different genomic segments from our specimen. While such chimeras are rare when sequencing fresh tissue (McKernan *et al.* 2009), the highly fragmented condition of historical DNA may result in ligation of random DNA fragments during library preparation (Willerslev & Cooper 2005). While paired-end information was unreliable, the sequence data from these reads were still reliable, and therefore, we used our results from the single-end mapping strategy in subsequent analyses (data available from DRYAD entry doi:10.5061/dryad.1fm3f).

Using single-end mapping strategy, 38.0% (1.75 Gbp) of our reads from the toe library and 21.8% (1.05 Gbp) of our reads from the molar library aligned uniquely. Average coverage of the nuclear genome was 0.642× and 0.380× for the toe and molar libraries, respectively. Coverage was unbiased among chromosomes (Fig. S3), but was unevenly distributed along individual chromosomes (Fig. 2). As expected, repetitive elements were under-represented; aligning uniquely to these regions without paired-end information is challenging (Li *et al.* 2010). Overall, 45.8% and 30.9% of the nuclear genome of *R. norvegicus* were represented by at least one read in the toe and molar libraries, respectively. Not surprisingly, we recovered much higher coverage for the mitochondrial genome (14.2× for the toe library, 81.9× for the molar library; Fig. 3) than the nuclear genome, and we sequenced the complete mitochondrial genome of *R. norvegicus*. As in the nuclear genome, sequence coverage was not evenly distributed across the mitochondrion. This bias was seen in both libraries; read density was highly correlated between the two independent libraries (Fig. S4).

At best, 42.1% of our filtered reads (Fig. 4) mapped uniquely to the *R. norvegicus* genome. For comparison, studies that sequence DNA from fresh tissue typically align >80% of reads to the reference genome (Atanur *et al.* 2010). Our reduced alignment efficiency can be

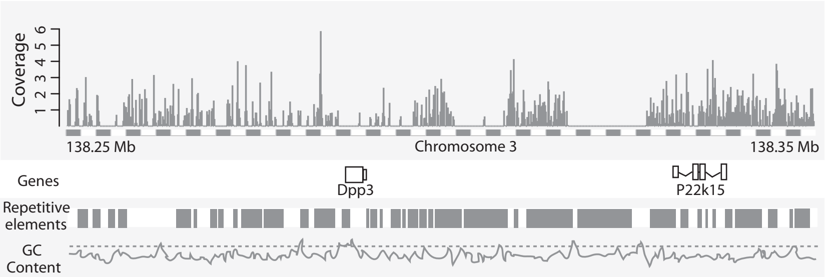

Fig. 2 Coverage (in bins of 100 bp) along a randomly selected 0.1-Mb segment of chromosome 3, using sequence data from the library prepared from the toe sample.

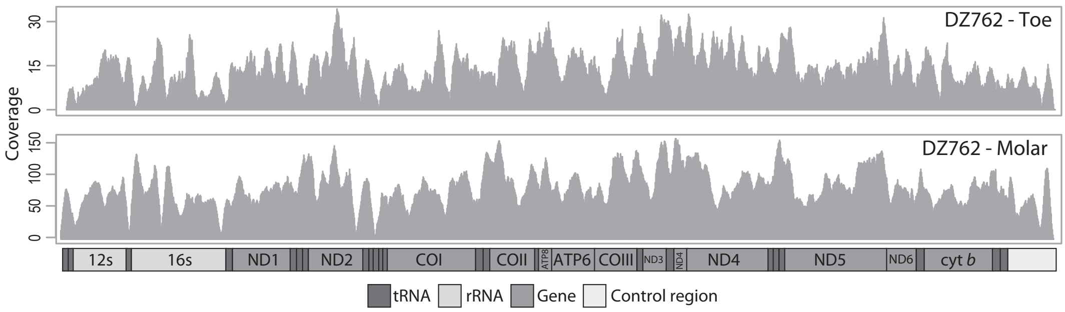

Fig. 3 Average sequence coverage (10-bp bins) of the mitochondrial genome of both the toe and molar libraries. Note the different scales for the toe and molar libraries. The mammalian mitochondrial map is provided for reference on the bottom with gene regions, ribosomal genes, transfer RNA and control region indicated by shades of grey.

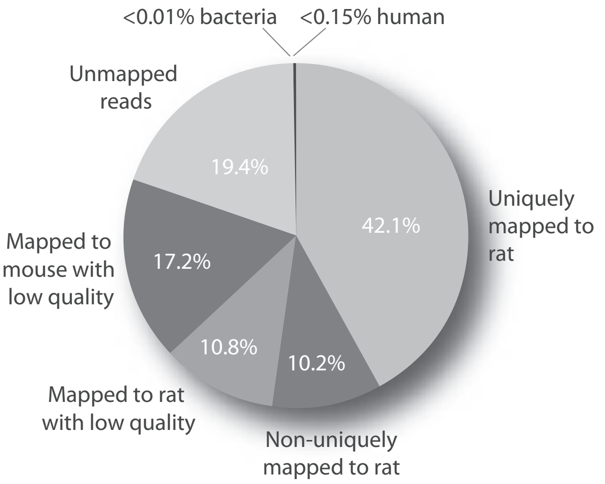

Fig. 4 Percentage of trimmed, unique reads that mapped to the *Rattus norvegicus* genome and other genome references. Results are based on the library derived from the toe sample. The proportions for the molar sample were consistent.

explained by several factors. Unlike many ancient DNA studies, we did not lose many of our reads to contamination as very few mapped to microorganisms or humans (Poinar *et al.* 2006a,b; Miller *et al.* 2008). An additional 10.2% of our final data set mapped accurately to the *R. norvegicus* genome but at multiple, equally likely, often highly repetitive parts of the genome. An additional 28% of our reads mapped with low quality to either the *R. norvegicus* (10.8%) or *Mus musculus* (17.2%) genomes, and the remaining 19.4% could not be matched to any known reference sequences. The low-quality matches to *R. norvegicus* and *M. musculus* largely fell into one of three categories: (i) the read aligned but with overall low identity, (ii) only part of the read aligned with high quality and the rest of the read failed to align or (iii) the read aligned, but with substantial gaps in the alignment. Low-quality matches were much more common in the 'reverse' read and are probably the result of postmortem damage to DNA coupled with common biases related to Illumina sequencing.

We used the high-coverage reads of the mitochondrial genomes from our toe and molar libraries separately to estimate overall sequencing error rates, guanine to adenine damage rates and cytosine to thymine damage rates. Postmortem damage to DNA does not affect all bases equally; guanine to adenine and cytosine to thymine lesions are much more common than others (Hofreiter *et al.* 2001). Error rates for the toe and molar samples were 0.269% and 0.263%, respectively; these rates are comparable with other studies of ancient DNA (Table 1) and with studies of fresh tissue (Bentley *et al.* 2008). Neither library showed a strong bias in guanine to adenine/cytosine to thymine lesions; all damage rates were <0.2% (Table S2). Although these error rates are higher than those seen in Sanger sequencing of historical DNA (0.01%; (Sefc *et al.* 2007), the increased coverage afforded by NGS more than compensates for the difference.

The most important consideration for historical and ancient DNA studies is whether or not sequencing errors resulting from either low-quality DNA template or postmortem substitutions propagate into the resulting consensus sequence. To address this concern, we compared the consensus sequences from the mitochondrial genomes of our toe and molar libraries. While we lacked sufficient coverage to call SNPs and estimate error rates from our nuclear genomes, previous ancient DNA studies have suggested that rates of mitochondrial and nuclear genome damage are not significantly different (Binladen *et al.* 2006). Therefore, we expect that our mitochondrial results should be representative of nuclear results given sufficient sequencing depth. We identified 87 SNPs between our specimen (DZ762) and the reference *R. norvegicus* mitochondrion (NC_001665). Importantly, the toe and molar libraries identified exactly the same set of SNPs. For the consensus sequence of each library, SNPs were covered by an average of 40 reads and could be called confidently with a probability of

Table 1 Summary of genome sequencing projects of ancient and historical samples

Sequenced species	Sequencing method	Study	DNA source	Reference genome	Error rate (%)	% reads mapped	% human contamination
Archaic hominin	Illumina	Reich *et al*. 2010	Phalanx	Human	0.04	–30	<1
Cave bears	Sanger	Noonan *et al*. 2005	Tooth/bone	Dog	–	1.1–5.8	0.03–0.06
Denisova hominin	Illumina	Krause *et al*. 2010	Bone	Human	–	12.40	<0.5
Horse	454/Illumina	Blow *et al*. 2008	Bone	Horse	–	0.70	0.01
Mammoth	454	Miller *et al*. 2008	Hair shaft	Elephant	0.14	58–80	<1.5
Mammoth	454	Poinar *et al*. 2006a, 2006b	Mandible	Elephant	1.40	45.40	1.40
Neandertal	454/Illumina	Green *et al*. 2010	Bone	Human	0.3–5.9	14.70	0.50
Neandertal	454	Briggs *et al*. 2009	Bone	Neandertal genome	–	–	0.2–1.4
Paleo-eskimo	454	Gilbert *et al*. 2008	Hair shaft	Human	0.25	80.72	–
Paleo-eskimo	Illumina	Rasmussen *et al*. 2010	Hair shaft	Human	–0.5	49.20	0.80
Polar bear	454	Lindqvist *et al*. 2010	Jawbone	Extant polar bear	–	40	4.50
Rat	**Illumina**	**This study**	**Toe (bone)**	**Rat**	**0.27**	**38**	**<0.01**
Rat	**Illumina**	**This study**	**Molar**	**Rat**	**0.26**	**21.80**	**<0.01**
Ruminant	454/Illumina	Blow *et al*. 2008	Bone	Cow	–	1.10	0
Thylacine	454	Miller *et al*. 2009	Dry skin	Numbat	0.51	25–40	8.90
Thylacine	454	Miller *et al*. 2009	Ethanol preserved	Numbat	0.60	25–40	4.30
Wolf	454/Illumina	Blow *et al*. 2008	Bone	Dog	–	1.80	0.01

Results of this study in bold.

miscalling of <0.0001. The 0.5% divergence of our sample from the reference genome is comparable to divergence of other *R. norvegicus* strains that have been sequenced, and our specimen is tightly nested within a phylogeny of these strains (Fig. 5; Table S1). The SNPs observed in our specimen did not produce unexpected rates of amino acid substitutions as would be expected if they were random errors. We used the 'yn00' function in PAML v 4.3 (Yang 2007) to calculate the ratio of nonsynonymous to synonymous mutations (ω) for the mitochondrion of each of the *R. norvegicus* strains compared to the mitochondrion of the outgroup, *R. rattus*. For our specimen, we calculated an ω value of 0.0304, which is nearly identical to the rates from other strains (0.0306–0.0320). Together, these results show that, despite the postmortem damage to DNA and the higher error rates in massively parallel sequencing technologies, increased throughput of these same technologies allowed us to identify variation accurately.

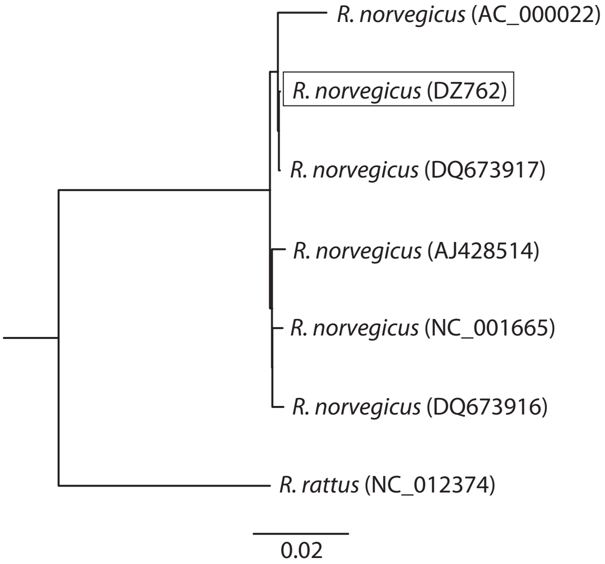
R. norvegicus (AC_000022)
R. norvegicus (DZ762)
R. norvegicus (DQ673917)
R. norvegicus (AJ428514)
R. norvegicus (NC_001665)
R. norvegicus (DQ673916)
R. rattus (NC_012374)
0.02

Fig. 5 Neighbour-joining tree of complete mitochondrial genomes from *Rattus rattus*, five strains of *R. norvegicus* and the specimen sequenced in this study. Individual sequenced in this study is boxed.

De novo assembly

De novo assembly of the reads resulted in 4662 contigs >200 bp in length (n50: 19 394 bp; mean: 1202 bp); 83% of these contigs mapped to the *R. norvegicus* genome with average 97.0% identity, representing about 0.1% of the *R. norvegicus* genome. Importantly, we recovered the full mitochondrial genome, and its sequence was identical to that inferred via alignment (data available from DRYAD entry doi:10.5061/dryad.1fm3f).

Discussion

We have demonstrated that genome-scale data sets can be generated efficiently and accurately from historical specimens given sufficient sequencing and a reference genome. Our study produced reliable sequence data from throughout the nuclear and mitochondrial genomes from the skin and the skull of a specimen that was collected nearly 50 years before present. Our results compared favourably to ancient DNA studies, in terms of human contamination, per cent of reads from target genome and error rate (Table 1). A majority of these studies sequenced subfossil or ancient frozen remains, for which time postmortem, and thus opportunity for degradation, is significantly greater than our specimens. Still, we recovered considerably more sequence data with less contamination than the most comparable study, in which researchers sequenced dried museum skins from the extinct Thylacine (Miller *et al.* 2009). Our error rates, which include both postmortem DNA mutations and sequencing error, were low and comparable to other sequencing projects of both modern and historical samples (Table 1). However, we observed reduced alignment efficiency compared to other studies sequencing modern samples; only ~20–40% of our reads aligned, whereas 80% of reads align typically. Reduced alignment efficiency may be the result of postmortem damage to DNA altering sequences such that they were still of high sequence quality but low alignment quality.

Although prospects are good for NGS to open research opportunities for using NHC, the application of NGS to the study of historical samples has several disadvantages. One limitation of our study is our inability to call SNPs in the nuclear genome because of low coverage. If our threshold for SNP discovery is the widely accepted 5× coverage (Yi *et al.* 2010) and if we assume sequencing additional lanes of our same libraries will result in similar sequencing and mapping efficiency, we estimate that we would need about eight lanes of sequencing to obtain sufficient coverage of the nuclear genome to call SNPs confidently. Eight lanes of sequencing per individual represent a substantial cost that is out of reach for most population genomic projects, particularly those focused on population-level changes in nonmodel organisms. However, two emerging developments offer promise. First, the cost of sequencing per unit continues to drop quickly. Second, new methods for preparing reduced representation libraries (RRL) could reduce genome complexity by 100-fold or more and greatly increase sequencing coverage (Altshuler *et al.* 2000; Young *et al.* 2010). Most RRL approaches rely on shearing or tagging of long fragments of DNA and are not appropriate for the highly damaged condition of historical specimen DNA. However, an emerging RRL approach, targeted enrichment capture technology (TECT), uses arrays of small genome tags to capture a select fraction of the genome such as the exome (Hodges *et al.* 2007, 2009). TECT is an ideal solution for highly fragmented DNA and has been applied successfully to ancient DNA studies (Burbano *et al.* 2010). Arrays can be scaled to population studies by barcoding individual samples uniquely and then pooling multiple samples prior to targeted capture and sequencing (Kim *et al.* 2010).

The main limitation of TECT is that it relies on knowledge of the sequence of genome target of interest. In this study, we sequenced historical DNA from a species for which a reference genome was already available. However, many species of ecological and evolutionary interest currently lack a reference genome. Despite our success in obtaining accurate genome-scale data from a museum specimen, we remain cautious about the prospects for *de novo* assembly of nuclear genomes from historical specimens. We were able to generate a complete and accurate mitochondrial genome without a reference genome, but our ability to assemble the rest of the genome was limited. This is not surprising because we (i) did not have sufficient raw sequence data and (ii) lacked accurate pair-end read data. By sequencing the same individual to greater coverage, researchers should be able to improve the quality of their final assembly. How much sequencing is necessary is unclear as contiguity of genome assembly and amount of raw data do not scale linearly (Zerbino & Birney 2008). Even with more data, however, without paired-end data, accurate assembly of eukaryotic genomes will be extremely challenging, because of the large amount of repetitive elements and the complexity of these genomes (Li *et al.* 2010). Prospects for improved paired-end sequencing are not favourable as *de novo* genome projects typically sequence paired ends from short DNA fragments (~300 bp) and from longer inserts (~10K bp) (Li *et al.* 2010). This is not feasible for historical samples as long DNA fragments are rarely recovered (Fig. 1) and chimeras of small fragments may occur (Fig. S2). For museum scientists, there are two promising, cost-effective solutions. First, researchers could use the genome of a related species as the reference for their target species. This was performed effectively for studies of mammoth DNA using the elephant genome (Poinar *et al.* 2006b). Second, if a relevant reference genome does not already exist, researchers could first sequence a modern specimen to assemble a genome and then use it as a reference for historical samples. With these reference genomes, museum scientist can then also build TECT arrays to target select genome regions for population-level sampling.

Our goal in this study was to assess the reliability and feasibility of NGS technologies for obtaining genome-scale data from common historical museum specimens. We extracted DNA from various sources and used those samples that produced the best nucleotide concentrations and fragment length distributions. We did not attempt to minimize damage to specimens; indeed, we removed toes and feet and drilled into skulls. NHC preserve irreplaceable records of historical faunal conditions, and much care and consideration should be applied to any destructive sampling. The policies pertaining to the destructive sampling of specimens are at the discretion of individual NHC, and many institutions will consider each request on a case-by-case basis. We suggest that any researcher planning to sample specimens destructively is obligated minimally (i) to define a problem of substantial scientific merit and (ii) to document that the methods and the samples requested are both sufficient and necessary to address the problem defined. We encourage future research to determine whether less destructive uses of the specimen (e.g. preparing libraries from the lip or venter) can also produce reliable libraries and sequence data. After all, the goal is to access the genomic data each specimen represents without compromising its integrity. This study shows that the prospects for doing so are promising and that, with improved sampling methodologies and new sequencing technologies, we are at the cusp of accessing the vast array of historical genomic data in NHC while preserving these collections for future research.

Acknowledgements

The authors would like to acknowledge the the MVZ community for helpful suggestions and advice. We also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper. MDM would like to thank his wife and kids for continued support.

References

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Atanur SS, Birol I, Guryev V *et al.* (2010) The genome sequence of the spontaneously hypertensive rat: analysis and functional significance. *Genome Research*, **20**, 791–803.

Austin JJ, Melville J (2006) Incorporating historical museum specimens into molecular systematic and conservation genetics research. *Molecular Ecology Notes*, **6**, 1089–1092.

Bentley D, Balasubramanian S, Swerdlow H *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Binladen J, Wiuf C, Gilbert MT *et al.* (2006) Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics*, **172**, 733–741.

Blow MJ, Zhang T, Woyke T *et al.* (2008) Identification of ancient remains through genomic sequencing. *Genome Research*, **18**, 1347–1353.

Bouzat JL, Lewin HA, Paige KN (1998) The ghost of genetic diversity past: historical DNA analysis of the greater prairie chicken. *The American Naturalist*, **152**, 1–6.

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

Burbano HA, Hodges E, Green RE *et al.* (2010) Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, **328**, 723–725.

Cooper A, Poinar G (2000) Ancient DNA: do it right or not at All. *Science*, **289**, 1139.

Cooper A, Mourer-Chauvire C, Chambers GK *et al.* (1992) Independent origins of New Zealand moas and kiwis. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 8741–8744.

Drummond A, Ashton B, Cheung M *et al.* (2011) Geneious v5.4. Available from http://www.geneious.com/.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Ellegren H (1991) DNA typing of museum birds. *Nature*, **354**, 113.

Gadaleta G, Pepe G, Decandia G *et al.* (1989) The complete nucleotide-sequence of the *Rattus norvegicus* mitochondrial genome—cryptic signals revealed by comparative-analysis between vertebrates. *Journal of Molecular Evolution*, **28**, 497–516.

Gilad Y, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics*, **25**, 463–471.

Gilbert MTP, Tomsho LP, Rendulic S *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science*, **317**, 1927–1930.

Gilbert MTP, Kivisild T, Gronnow B *et al.* (2008) Paleo-eskimo mtDNA genome reveals matrilineal discontinuity in Greenlan. *Science*, **320**, 1787–1789.

Goldstein PZ, Desalle R (2003) Calibrating phylogenetic species formation in a threatened insect using DNA from historical specimens. *Molecular Ecology*, **12**, 1993–1998.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Hansen MM (2002) Estimating the long-term effects of stocking domesticated trout into wild brown trout (*Salmo trutta*) populations: an approach using microsatellite DNA analysis of historical and contemporary samples. *Molecular Ecology*, **11**, 1003–1015.

Hodges E, Xuan Z, Balija V *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*, **39**, 1522–1527.

Hodges E, Smith AD, Kendall J *et al.* (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Research*, **19**, 1593–1605.

Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Paabo S (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, **29**, 4793–4799.

Johnson M, Zaretskaya I, Raytselis Y *et al.* (2008) NCBI BLAST: a better web interface. *Nucleic Acids Research*, **36**, W5–W9.

Kent WJ (2002) BLAT—The BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Kim DW, Nam SH, Kim RN, Choi SH, Park HS (2010) Whole human exome capture for high-throughput sequencing. *Genome*, **53**, 568–574.

Knapp M, Hofreiter M (2010) Next generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes*, **1**, 227–243.

Krause J, Fu Q, Good JM *et al.* (2010) The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, **464**, 894–897.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

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

Li R, Fan W, Tian G *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Lindqvist C, Schuster SC, Sun Y *et al.* (2010) Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 5053–5057.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics*, **4**, 981–994.

Martinez-Cruz B, Godoy JA, Negro JJ (2007) Population fragmentation leads to spatial and temporal genetic structure in the endangered Spanish imperial eagle. *Molecular Ecology*, **16**, 477–486.

McKernan KJ, Peckham HE, Costa GL *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, **19**, 1527–1541.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219–236.

Miller W, Drautz DI, Ratan A *et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, **456**, 387–392.

Miller W, Drautz DI, Janecka JE *et al.* (2009) The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Research*, **19**, 213–220.

Moritz C, Patton J, Conroy C *et al.* (2008) Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science*, **322**, 261–264.

Mullen LM, Hoekstra HE (2008) Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution*, **62**, 1555–1569.

Murphy MA, Dezzani R, Pilliod DS, Storfer A (2010) Landscape genetics of high mountain frog metapopulations. *Molecular Ecology*, **19**, 3634–3649.

Nielsen EE, Hansen MM, Loeschcke V (1997) Analysis of microsatellite DNA from old scale samples of Atlantic salmon *Salmo salar*: a comparison of genetic composition over 60 years. *Molecular Ecology*, **6**, 487–492.

Nilsson MA, Gullberg A, Spotorno AE, Arnason U, Janke A (2003) Radiation of extant marsupials after the K/T boundary: evidence from complete mitochondrial genomes. *Journal of Molecular Evolution*, **57**, S3–S12.

Noonan JP, Hofreiter M, Smith D *et al.* (2005) Genomic sequencing of Pleistocene cave bears. *Science*, **309**, 597–600.

Paabo S, Poinar H, Serre D *et al.* (2004) Genetic analyses from ancient DNA. *Annual Review of Genetics*, **38**, 645–679.

Peery MZ, Hall LA, Sellas A *et al.* (2010) Genetic analyses of historic and modern marbled murrelets suggest decoupling of migration and gene flow after habitat fragmentation. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 697–706.

Pichler FB, Dalebout ML, Baker CS (2001) Nondestructive DNA extraction from sperm whale teeth and scrimshaw. *Molecular Ecology Notes*, **1**, 106–109.

Poinar H, Schwarz C, MacPhee R, Miller W, Schuster S (2006a) Mammoth metagenomics: new technologies, new insights, new possibilities. *Journal of Vertebrate Paleontology*, **26**, 110A.

Poinar HN, Schwarz C, Qi J *et al.* (2006b) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.

Quail MA, Kozarewa I, Smith F *et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, **5**, 1005–1010.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/

Rasmussen M, Li Y, Lindgreen S *et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, **463**, 757–762.

Reich D, Green RE, Kircher M *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**, 1053–1060.

Robbirt KM, Davy AJ, Hutchings M, Roberts DL (2011) Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid *Ophrys sphegodes*. *Journal of Ecology*, **99**, 235–241.

Robins JH, McLenachan PA, Phillips MJ, Craig L, Ross HA, Matisoo-Smith E (2008) Dating of divergences within the Rattus genus phylogeny using whole mitochondrial genomes. *Molecular Phylogenetics and Evolution*, **49**, 460–466.

Rohland N, Siedel H, Hofreiter M (2010) A rapid column-based ancient DNA extraction method for increased sample throughput. *Molecular Ecology Resources*, **10**, 677–683.

Roy MS, Girman DJ, Taylor AC, Wayne RK (1994) The use of museum specimens to reconstruct the genetic-variability and relationships of extinct populations. *Experientia*, **50**, 551–557.

Schlick NE, Jensen-Seaman MI, Orlebeke K *et al.* (2006) Sequence analysis of the complete mitochondrial DNA in 10 commonly used inbred rat strains. *American Journal of Physiology-Cell Physiology*, **291**, C1183–C1192.

Sefc KM, Payne RB, Sorenson MD (2007) Single base errors in PCR products from avian museum specimens and their effect on estimates of historical genetic diversity. *Conservation Genetics*, **8**, 879–884.

Shaffer HB, Fisher RN, Davidson C (1998) The role of natural history collections in documenting species declines. *Trends in Ecology & Evolution*, **13**, 27–30.

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

Stajich JE, Block D, Boulez K *et al.* (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Research*, **12**, 1611–1618.

Suarez AV, Tsutsui ND (2004) The value of museum collections for research and society. *BioScience*, **54**, 66–74.

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

Taylor AC, Sherwin WB, Wayne RK (1994) Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat Lasiorhinus krefftii. *Molecular Ecology*, **3**, 277–290.

Thomas WK, Pääbo S, Villablanca FX, Wilson AC (1990) Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens. *Journal of Molecular Evolution*, **31**, 101–112.

Vera JC, Wheat C, Fescemyer H *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wandeler P, Hoeck PEA, Keller LF (2007) Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution*, **22**, 634–642.

Willerslev E, Cooper A (2005) Ancient DNA. *Proceedings of the Royal Society B-Biological Sciences*, **272**, 3–16.

Yang Z (2007) PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yi X, Liang Y, Huerta-Sanchez E *et al.* (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, **329**, 75–78.

Young AL, Abaan HO, Zerbino D *et al.* (2010) A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome Research*, **20**, 249–256.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Data Accessibility

DRYAD entry doi:10.5061/dryad.1fm3f includes:

1 'assembleMito.fa'—The final *de novo* mitochondrial genome assembly using ABySS for each sample.

2 'library*.sorted.bam'—Trimmed reads for the total genomic data for each sample mapped with Bowtie to the *Rattus norvegicus* reference RGSC3.4. Associated index (*.bai) files are also included.

3 'seq/'—Folder containing unassembled reads for each sample after trimming for contamination, sequencing

adapters and low-quality reads. Library5* is derived from the molar. Library6* is derived from the toe. *_1_* reads are forward reads; *_2_* reads are reverse reads.

Supporting Information

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

Table S1 Mitochondrial position and average sequence coverage for 87 mitochondrial SNPs identified in our specimen, DZ762, compared to the *R. norvegicus* reference genome.

Table S2 Calculation of general error rate and biases in error rates.

Fig. S1 Flow diagram of data cleaning and alignment from raw reads to mapped reads.

Fig. S2 (A) Mapping of paired reads. (B) Distance between reads mapped to the same chromosome.

Fig. S3 Broad scale coverage of nuclear genome calculated for 10-kb windows of the chromosome.

Fig. S4 Correlation between coverage of the mitochondrial genomes for the two libraries.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.