RESEARCH ARTICLE

Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing

Sarah M. Hykin[1,3]*, Ke Bi[2,3]*, Jimmy A. McGuire[1,3]

[1] Department of Integrative Biology, 3101 Valley Life Sciences Building, University of California, Berkeley, California, United States of America, [2] Computational Genomics Resource Laboratory (CGRL), California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, California, United States of America, [3] Museum of Vertebrate Zoology, 3101 Valley Life Sciences Building, University of California, Berkeley, California, United States of America

* smhykin@berkeley.edu (SMH); kebi@berkeley.edu (KB)

OPEN ACCESS

Citation: Hykin SM, Bi K, McGuire JA (2015) Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing. PLoS ONE 10(10): e0141579. doi:10.1371/journal.pone.0141579

Editor: Ludovic Orlando, Natural History Museum of Denmark, University of Copenhagen, DENMARK

Received: March 26, 2015

Accepted: October 9, 2015

Published: October 27, 2015

Data Availability Statement: Raw sequence data are available on the NCBI Short Read Archive. Sample MVZ 214979, Accession number: SAMN04094110. Sample MVZ 43405, Accession number: SAMN04161375.

Funding: This research is funded by the Gordon and Betty Moore Foundation, grant number 2983 awarded to Dr. Charles Marshall and the Berkeley Initiative for Global Change Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

For 150 years or more, specimens were routinely collected and deposited in natural history collections without preserving fresh tissue samples for genetic analysis. In the case of most herpetological specimens (i.e. amphibians and reptiles), attempts to extract and sequence DNA from formalin-fixed, ethanol-preserved specimens—particularly for use in phylogenetic analyses—has been laborious and largely ineffective due to the highly fragmented nature of the DNA. As a result, tens of thousands of specimens in herpetological collections have not been available for sequence-based phylogenetic studies. Massively parallel High-Throughput Sequencing methods and the associated bioinformatics, however, are particularly suited to recovering meaningful genetic markers from severely degraded/fragmented DNA sequences such as DNA damaged by formalin-fixation. In this study, we compared previously published DNA extraction methods on three tissue types subsampled from formalin-fixed specimens of Anolis carolinensis, followed by sequencing. Sufficient quality DNA was recovered from liver tissue, making this technique minimally destructive to museum specimens. Sequencing was only successful for the more recently collected specimen (collected ~30 ybp). We suspect this could be due either to the conditions of preservation and/or the amount of tissue used for extraction purposes. For the successfully sequenced sample, we found a high rate of base misincorporation. After rigorous trimming, we successfully mapped 27.93% of the cleaned reads to the reference genome, were able to reconstruct the complete mitochondrial genome, and recovered an accurate phylogenetic placement for our specimen. We conclude that the amount of DNA available, which can vary depending on specimen age and preservation conditions, will determine if sequencing will be successful. The technique described here will greatly improve the value of museum collections by making many formalin-fixed specimens available for genetic analysis.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The primary goal of natural history museums is to preserve a biological record of the natural world for scientific study [1]. Museum collections have long provided geographic, morphological, and life history data for biologists. During approximately the past four decades, museums have also become the repositories of choice for tissue samples for molecular genetic analyses, especially for non-model organisms. These tissue collections are the necessary source materials for a tremendous diversity of biological studies. Given that the vast majority of museum specimens were collected before the advent of molecular genetics and routine collection of tissue samples, researchers have long been interested in developing protocols that would allow for the successful collection of historical DNA (hDNA) sequence data directly from museum specimens, even when properly prepared tissue samples were not available [2]. The traditional taxon-specific methods by which museum specimens have been prepared were a critical factor in this effort. Organisms prepared as study skins or dry preps, including birds, mammals and herbarium specimens are not typically exposed to formalin during preparation, and have been found to be highly amenable to hDNA data collection using traditional Sanger sequencing [3–5]. Indeed, hDNA extraction and Sanger sequencing from museum skins of birds, mammals and ancient human remains has not only been used in numerous routine studies, but also made possible molecular studies of extinct species [6–8], as well as studies of historical populations spanning both time and space [9–11].

Recent advances in High-Throughput Sequencing (HTS) data collection have revolutionized molecular genetic studies by making it possible to rapidly and efficiently obtain data sets composed of hundreds or even thousands of loci. Unsurprisingly, HTS sequencing efforts using museum study skins as source material have been very successful, and it is now possible to obtain not just DNA sequence data sets but genomic-scale DNA sequence data sets from these samples [12,13]. Of course many museum specimens are not routinely prepared as study skins or dry preps, but rather are formalin-fixed and stored in ethanol as fluid specimens. The organisms most often prepared in this manner include fish, amphibians, reptiles, and various invertebrate taxa. The extraction of usable DNA sequence data from these materials has proven much more challenging and, indeed, largely intractable. Studies that have successfully recovered DNA sequences from formalin-fixed samples generally obtained only short fragments (often for mitochondrial genes) by stitching together very small sequence fragments (typically just 50–100 base pairs in length) painstakingly obtained using custom-designed primers for each short read. Thus, developing effective and consistent protocols for the successful extraction and sequencing of historical formalin-fixed samples has been elusive, leaving millions of formalin-fixed museum specimens collected over the course of decades largely unavailable for molecular genetic analysis. Furthering progress in unlocking this potential treasure trove is the primary objective of this study.

Formalin-fixation of specimens damages DNA in three ways: (1) fragmentation, (2) base modification, and (3) cross-linkage within the DNA itself or between DNA and proteins [14–17]. Though DNA is still present, stretches that can be sequenced are heavily and randomly fragmented, posing a challenge for Sanger-sequencing techniques. Sanger sequencing relies on targeting specific regions of the genome to accurately copy long (typically 300–1500 bp) stretches of DNA, rendering the random and fragmented formalin-fixed DNA particularly unsuitable. Illumina high-throughput sequencing, however, typically sequences as few as 50–150 contiguous nucleotides per read, and can produce several hundred million such reads spanning an entire genome. This, in conjunction with the bioinformatics techniques for assembling reads and aligning them to a reference genome, makes the Illumina platform a promising one for sequencing DNA from formalin-fixed tissues.

Attempts to extract DNA from formalin-fixed museum specimens (FFMS) [16,18] have been successful, but the methods were relatively destructive to specimens—often requiring removal and destruction of skeletal elements—and labor-intensive when using a Sanger-sequencing platform. More recently, attempts to sequence formalin-fixed, paraffin-embedded (FFPE) human and cancer-cell lines have been successful using HTS [19–21]. However, there are notable differences in the protocols employed for formalin-fixing and paraffin-embedding cell cultures versus formalin-fixation of museum specimens. In particular, cell lines are usually exposed to a 2%—10% formalin solution for mere minutes (typically 20 min or less) before paraffin embedding [15,19], whereas FFMS are injected with and then soaked in 10% (or more) formalin solution for anywhere from 12 hours to several weeks. In addition, the vast majority of FFMS are prepared under conditions known only to the researchers who prepare them, as these data are not routinely recorded. In addition to the age of the specimen, this introduces a wide range of variables (e.g. light exposure, temperature, formalin concentration, whether or not the formalin was buffered) that could affect DNA quality and sequencing success for any particular specimen.

In this study, we attempted to develop an extraction and HTS protocol for DNA from formalin-fixed, ethanol-preserved herpetological museum specimens. We conducted a parallel set of comparative extraction experiments on two formalin-fixed, ethanol-preserved Anolis carolinensis specimens: one collected and preserved ~100 years ago, the other ~30 years ago. We subsampled liver, leg muscle, and tail-tip from each specimen and performed DNA extractions following two different protocols to determine (i) which tissue yielded a larger quantity of DNA, and (ii) which protocol performed better for each tissue type. This was followed by Illumina HTS of the best extraction for each specimen. After processing the resulting data for quality, the sequences were aligned to the A. carolinensis genome to determine if accurate and phylogenetically informative sequence data could be recovered. In this paper, we report the results of these experiments and outline a minimally-destructive protocol for obtaining phylogenetically informative sequence data from formalin-fixed museum specimens.

Materials and Methods

Tissue Collection and DNA Extraction

We subsampled liver, leg muscle, and tail-tips from two specimens of Anolis carolinensis from the University of California Museum of Vertebrate Zoology (MVZ) at Berkeley. These specimens were MVZ 214979, collected from Louisiana and prepared in 1985, and MVZ 43405, collected from Louisiana and prepared in 1917. These specimens were chosen for subsampling according to two criteria: each was formalin-fixed and preserved in ethanol, and both were large enough, approximately 80–100 mm snout-to-vent length (SVL), to allow subsampling of approximately 0.05 g of leg muscle tissue from the inguinal region without severely damaging the specimen. Subsampling was performed with standard, non-sterile, steel forceps and scissors. Tissues were stored separately in 70% ethanol in sterile 1.5 ml microcentrifuge tubes. Samples ranged in mass from 0.01 g to 0.5 g.

To limit potential contamination, extractions were performed in a room used exclusively for DNA extraction from historical specimens. There had been no previous extractions of Anolis performed in this laboratory space. The two extraction protocols performed were adapted from [19,22,23]. Both protocols begin with a series of ethanol washes followed by treatment in a heated alkali buffer solution. While heat and alkali degrade DNA, limited exposure to a combination of both has been demonstrated to be effective in breaking protein-DNA cross-linkages caused by formalin-exposure [16]. The hot alkali treatment was either followed by a phenol-chloroform extraction [16], or extraction using a standard Qiagen kit [22]. For extraction of tail-tips, we used a protocol for decalcification of formalin-fixed skeletal elements [23], followed by phenol-chloroform extraction. The set of extraction protocols that we tested are provided in the Supporting Information S1 File. After extraction, we quantified DNA yield for each extraction using a Nanodrop 1000 (Thermo Fisher Scientific Products) to measure the concentration of nucleic acids, a Qubit 2.0 fluorometer (Invitrogen, Life Technologies) to measure the concentration of double-stranded DNA, and an Agilent Bioanalyzer 2100 low-sensitivity chip (Aligent Technologies, Inc.) with DNA standards at 15 and 1500 bp to quantify DNA concentration and fragment size.

Library Preparation and Sequencing

The extraction that yielded the largest quantity of high-quality DNA for MVZ 214979 was from liver tissue using the phenol-chloroform protocol (provided as S2 File). This was prepared for sequencing using the standard TruSeq protocol for DNA (Illumina, Paired-End Sample Preparation Guide, document # 1005063 Rev. D) and NEB (New England Biolabs # E6006s) reagents. We also prepared a library for MVZ 43405 from the phenol-chloroform extraction of liver despite its failure to yield measurable amounts of quality (i.e. double-stranded) DNA. We modified the Illumina protocol for both extractions following [24] to account for the fragmented nature of formalin-fixed DNA by omitting the initial DNA fragmentation step. Instead, we proceeded immediately to end-repair, adenylation of the 3' end of DNA fragments, and adapter ligation according to the Illumina protocol. We used Agencourt Ampure XP (Beckman Coulter) magnetic-bead purification for nucleotide recovery and purification between steps in the Illumina protocol. This was followed by 16 PCR amplification cycles using Phusion PCR High Fidelity master mix (NEB, F-531S). After bead purification of the PCR'd libraries, they were analyzed for quantity and quality of DNA present on an Agilent Bioanalyzer 2100 low-sensitivity chip with two replicates for each library. The library for the older specimen, MVZ 43405, did not contain a sufficient amount of library product and underwent an additional six cycles of PCR amplification, followed by purification and re-analysis. We then performed 100-bp paired-end Illumina sequencing, pooling both samples on one lane of a HiSeq2000 at the Vincent J. Coates Genomics Sequencing Laboratory (Q3B, University of California, Berkeley)

Data processing

Upon receipt of raw data, pre-processing and alignment largely followed [24] as outlined below, with the following exceptions: Bowtie2 [25] was used instead of Bowtie [26] for contaminant filtration; Bowtie2 and Novoalign (www.novocraft.com) were used for alignment to the Anolis carolinensis genome (Anocar2.0, downloaded from the UCSC Genome Browser, http://watson.compbio.iupui.edu/cgi-bin/hgGateway); and updated versions of in-house scripts (https://github.com/MVZSEQ/) were employed throughout the process.

DNA recovered from ancient and museum historic specimens is often characterized by various types of postmortem nucleotide damage (e.g. [8,27,28]), and formalin-fixation can cause base modifications in the DNA fragment [15]. The specimens used in this study had been fixed in formalin for at least 30 years, thus damage to the DNA could be the result of post-mortem denaturation or subsequent exposure to formalin. To inspect potential base misincorporation in sequence reads, we first aligned the untrimmed raw paired-end reads against the Anolis carolinensis reference genome with Bowtie2. By parsing the SAM output, we generated base mismatch frequency plots by plotting the frequency of all 12 possible mismatches against distance from 5′ and 3′ ends of reads, respectively. We observed a sharp increase in mismatch frequencies of almost all types at both ends, and particularly at the 3' end of reads (Fig 1). We then

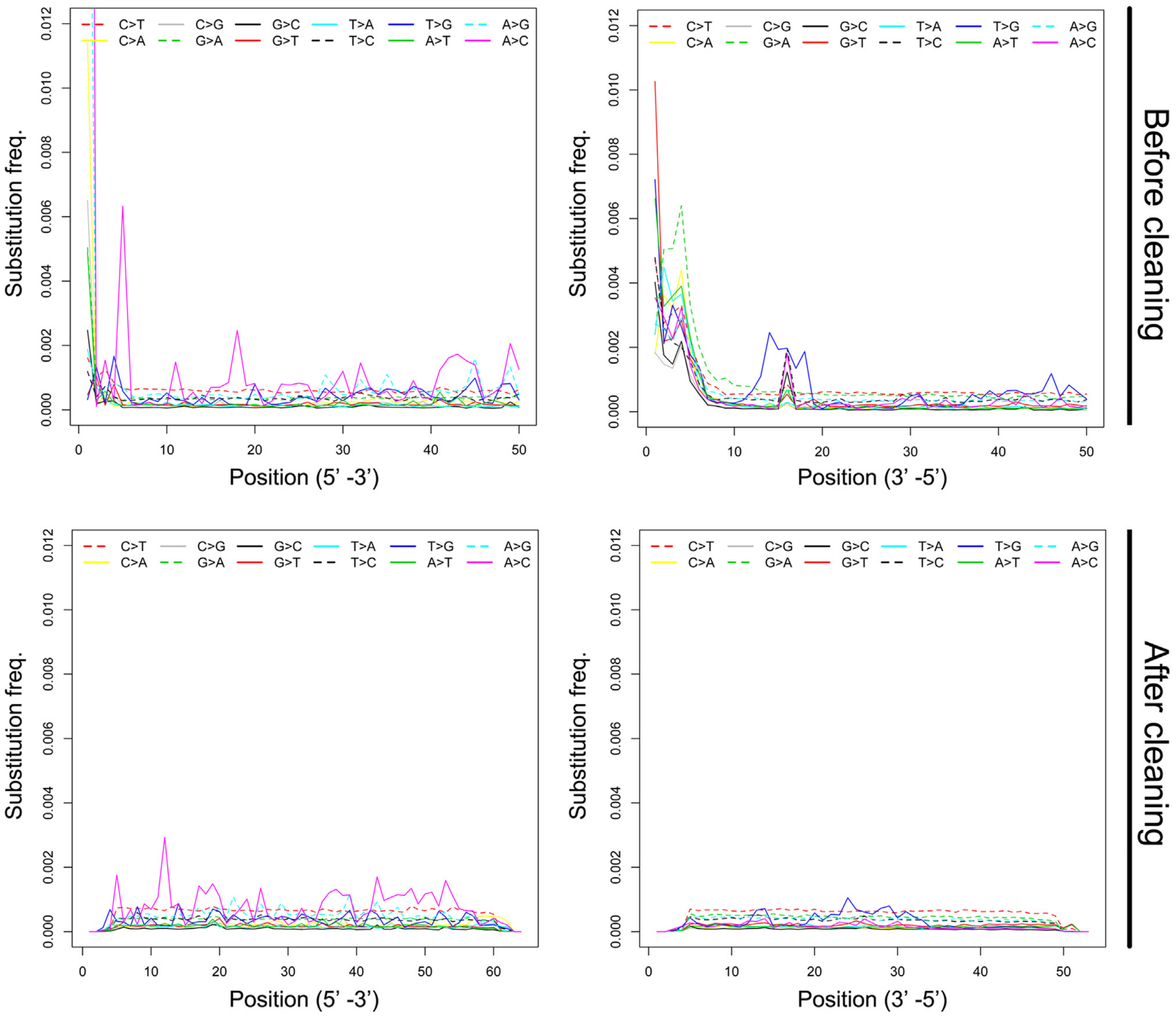

Fig 1. Patterns of mismatches in MVZ 214979 sequences. The frequencies of the 12 types of mismatches (y-axis) are plotted as a function of distance from the 5′ and 3′ ends of the sequence reads (x-axis). The frequencies of each mismatch type are coded in different colors and line patterns. 'After cleaning' shows mismatch frequencies after deleting the first 50 bp form the end of each read.

doi:10.1371/journal.pone.0141579.g001

performed multiple rounds of trimming from both 5′ and 3′ ends of reads, until the frequencies of all 12 types of mismatches were relatively constant and similar along post-trimmed reads (Fig 1). As a final trimming step, we removed 36 bp from the forward reads (6 bp from 5′ and 30 bp from 3′ end) and 47 bp from the reverse reads (17 bp from 5′ end and 30 bp from 3′ end). To evaluate the quality of sequence reads before and after cleaning and trimming, SAMtools [29] and an in-house script were used to estimate empirical error rates, measured as the percentage of mismatched bases out of the total number of aligned bases in the mitochondrial genome [12].

The hard-trimmed raw sequence data were then re-processed to remove exact duplicate reads, adaptors, and low-quality sequences, and to merge overlapping paired-end reads following [24] and [30] using in-house scripts. To remove reads that might result from contamination by organisms other than Anolis, we aligned all adaptor-trimmed reads to the human (hg19) and Escherichia coli (NCBI st. 536) genomes using Bowtie2 [25]. We assumed that reads aligning to these genomes represented contamination and removed them from our data. After cleanup, we mapped the resulting paired-end reads to the *Anolis* reference genome using Novoalign, then applied SAMtools to check mapping efficiency and depth. All cleaned data, including paired-end and unpaired reads, were de novo assembled using ABySS [31] and individual assemblies were generated under a wide range of k-mers as in [32]. We used cd-hit-est [33], Blat [34], and CAP3 [35] to merge raw assemblies and reduce redundancy in our libraries. Contiguous sequences (contigs) less than 200 bp were removed. The resulting contigs were mapped to the *Anolis* reference genome using the BLASTn program [36]. To evaluate coverage of the mitochondrial genome, we mapped cleaned reads to the mitochondrial reference genome of Anolis carolinensis and used SAMtools to reconstruct the coding sequence of the mitochondrial genome from the MVZ 214979 library.

To evaluate if we had accurately recovered phylogenetically useful sequence data, we extracted the consensus sequence from reads mapping to the mitochondrial genome. We aligned our inferred complete mitochondrial sequence to the Anocar2.0 reference genome to assess sequence similarity. We then aligned the NADH dehydrogenase subunit 2 (ND2) sequence recovered from our formalin fixed sample to that available for the *Anolis* genome, as well as to ND2 data from NCBI for an outgroup, Oplurus cyclurus, and eight additional *Anolis* species, including the putative sister taxon of *A. carolinensis*, *A. porcatus*, and three other close relatives, *A. brunneus*, *A. allisoni*, and *A. smaragdinus* (NCBI IDs: OCU39585, AB218960, AY263042, KJ954109, AF337807, AY902412, AY296151, AY902417, and AY296195). We based this analysis on ND2 alone because this gene is widely available for *Anolis* species. We then estimated the phylogeny for this alignment using maximum likelihood under the GTR+I +G model in PAUP (version 4.0a142) [37] and calculated bootstrap values using maximum likelihood with 100 replicates, also under the GTR+I+G model in Garli [38].

Results

DNA extraction and library preparation

We were able to extract DNA from both specimens of Anolis carolinensis, however only phenol-chloroform extraction of liver tissue from MVZ 214979 yielded enough high-quality DNA for Illumina sequencing. According to Nanodrop analysis, concentrations of nucleic acids were generally higher in liver extractions (S1 Table), and according to Qubit quantification, extractions by either phenol-chloroform or Qiagen kit from muscle and tail-tips yielded insufficient quantities of double-stranded DNA for either specimen to proceed with library preparation. According to Qubit quantification, the phenol-chloroform extraction of liver for MVZ 214979 had a DNA concentration of 26.4 ng/μl, but the Qiagen and tail-tip extractions of this specimen and all extractions of MVZ 43405 failed to yield measurable amounts of double-stranded DNA. The Bioanalyzer results were consistent with Qubit quantifications: the phenol-chloroform extraction of liver from MVZ 214979 had a DNA concentration of 27.81 ng/μl, and the Qiagen extraction of MVZ 214979 had a DNA concentration of 1.51 ng/μl. The phenol-chloroform extraction of MVZ 43405 showed a concentration of 0.27 ng/μl, and all other extractions failed to show detectable amounts of DNA. We elected to proceed with library preparation of the phenol-chloroform liver extraction of both specimens to see if we could obtain usable data from MVZ 43405 despite the poor quantification values. Bioanalyzer results for the MVZ 214979 library showed peaks at ~120 bp and ~240 bp (S1 Fig), and a library concentration of 5.69 ng/μl and 7.62 ng/μl (average of two replicates = 6.66 ng/μl). Bioanalyzer results for MVZ 43405's library showed unusual oscillations between ~150 bp and ~410 bp, and a final library concentration of 0.68 ng/μl and 3.04 ng/μl (average of two replicates = 1.86 ng/μl).

Sequencing and data pre-processing

Sequencing results showed 9.51 billion base pairs (Gbp) for MVZ 214979 and 16.29 Gbp for MVZ 43405. After data cleanup, alignment to the reference genome for MVZ 43405 indicated extremely high PCR duplication (97.5%) and thus low diversity and a low unique mapping rate (0.23%). The mismatch frequency plot for MVZ 43405 indicated that, of the reads mapped to the reference genome, all types of mismatches across the length of reads showed extremely uneven distributions, making hard-trimming impossible (S2 Fig). For this reason, we concluded that sequencing of MVZ 43405 had failed, and these data were excluded from further analyses.

For the more recently collected sample, MVZ 214979, pre-processing resulted in removal of 64% of reads as duplicates. This is a larger fraction of the data set than is typical, even for hDNA [12], but we attribute this to the relatively low amount of starting DNA that we then PCR-amplified. Contamination by E. coli or Homo sapiens represented 0.27% of reads, and after filtering low quality reads, trimming adapter sequences, and merging overlapping paired-end reads, the library contained 1.27 Gbp of sequence data, accounting for 13.37% of the original data.

Alignment to the *Anolis carolinensis* reference genome and phylogenetic informativeness

Based on the patterns of skewed base misincorporation observed from the mismatch frequency plot, we trimmed 36 bp and 47 bp from the forward and reverse reads, respectively. After data filtration, we aligned paired-end reads and unpaired reads to the Anolis carolinensis genome using Novoalign, which resulted in the unique mapping of 70% of all cleaned reads and 72% of cleaned paired-end reads. Over the entire reference genome, 27.93% (502.7 Mb) mapped to at least one read, with and average mapping depth of 0.5X. The total amount of data aligned to the reference genome was 891.1 Mb, accounting for 9.3% of the total obtained from a half-lane's worth of sequencing effort. Of cleaned reads, 2.9% were aligned to protein coding regions of the genome, at an average depth of 1.2X. We did not observe a bias for sequence coverage towards certain chromosomes (Fig 2a), but mapped reads were unevenly distributed within chromosomes (Fig 2b).

Raw sequence reads had an error rate of 0.61%. After hard trimming and quality filtering, the error rate decreased to 0.45%, and these final, cleaned, datasets were used for mapping, assembly, and reconstructing the complete mitochondrial genome.

Due to the highly degraded nature of DNA from MVZ 214979 and shallow sequencing depth, de novo assembly only yielded 21,394 contigs that were longer than 200 bp with an N50 of 398 bp. A total of 9342 contigs (43.67%) were aligned to the Anolis carolinensis genome with an average sequence similarity of 98.34%. Attempted *de novo* assembly of the mitochondrial genome resulted in 30 contigs representing ~75% of the mitochondrial genome (~12KB). These contigs ranged in length from 38 bp to 5157 bp. We estimated GC content in the mapped, assembled contigs to be 39.03%, comparable to the published GC content of the A. carolinensis genome of 40.30% [32].

Our sequencing depth was too low to allow for the generation of a nuclear gene data set usable for reliable phylogenetic analysis. However, the much greater sequence depth for the mitochondrial genome was more than sufficient to recover the complete mitochondrial genome sequence, with an average depth of 57.9X. This is not surprising given the much higher per cell copy number of the mitochondrial genome as compared to the nuclear genome. As in the nuclear genome, mapped reads were unevenly distributed along the entire mitochondrial genome (Fig 3).

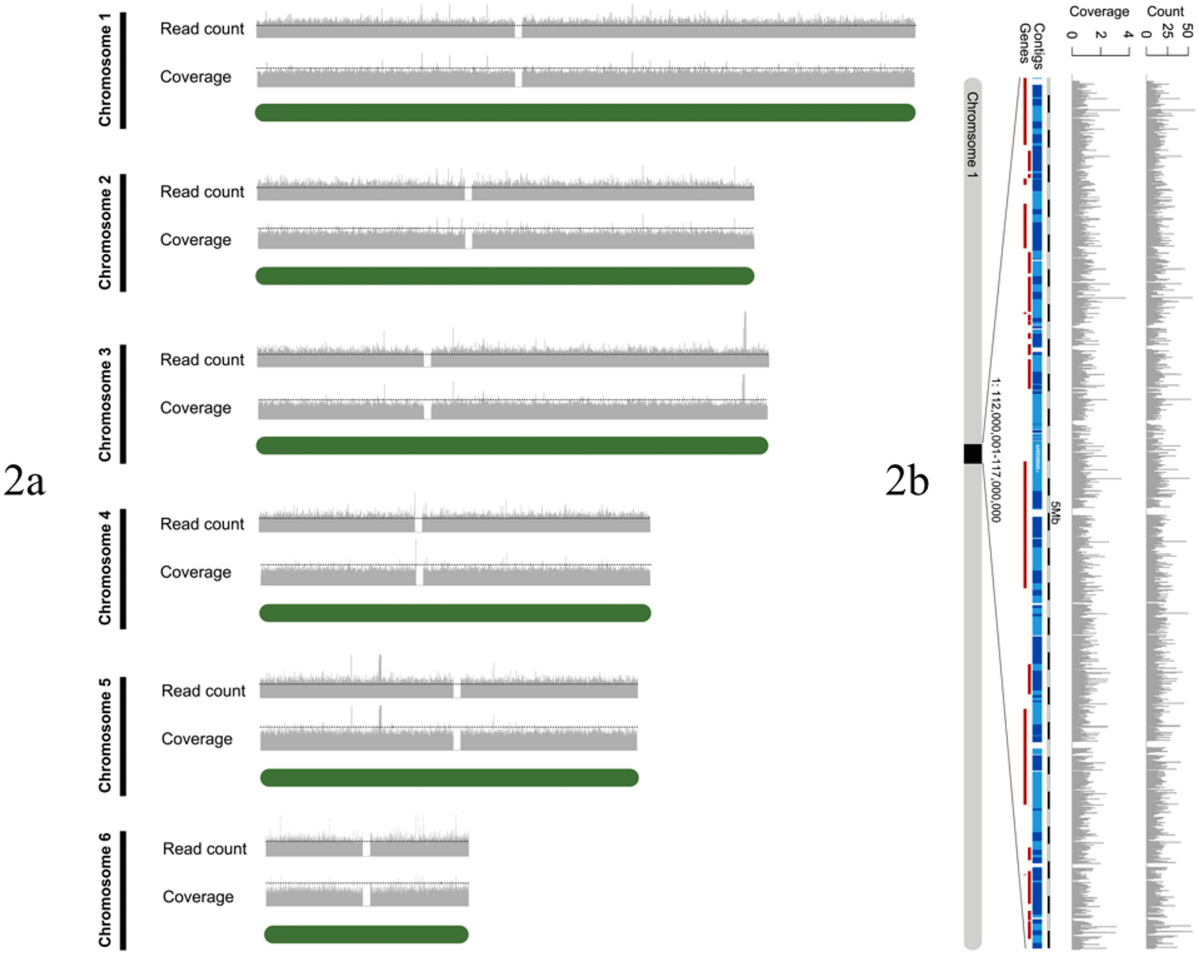

Fig 2. Nuclear coverage. (a) Read count and depth in 10 Kbp bins along the length of the six largest *A. carolinensis* chromosomes (green bars) using the MVZ 214979 library. The green line indicates a read count of 100, and coverage of 1X. (b) Read count and depth is shown in 1 Kb bins along a randomly selected 5 Mbp segment of chromosome 1 using the MVZ 214979 library.

doi:10.1371/journal.pone.0141579.g002

The total number of high quality SNPs detected between the Anolis carolinensis reference genome and this formalin-fixed specimen was 73 (0.53% sequence dissimilarity), with an average depth of 51.4X. Alignment of the 1038 bp ND2 gene from our A. carolinensis mitochondrial genome with orthologous gene regions from the reference *Anolis* genome, and to other taxa, including eight other Anolis species and an outgroup (*Oplurus cyclurus*), resulted in six SNPs between our sample and the A. carolinensis reference genome and strong support for our specimen being more closely related to A. carolinensis than to any other Anolis (Fig 4). This included A. porcatus, the sister species of A. carolinensis. In addition, the recovered ND2 sequence for MVZ 214979 was no more divergent from the reference sequence than any other *A. carolinensis* sequence downloaded from GenBank. This is consistent with our having recovered ND2 sequence data for MVZ 214979 with sufficient accuracy to be phylogenetically informative at the species level.

Discussion

This study highlights both the opportunities and the challenges of obtaining genomic data from formalin-fixed museum specimens. For example, we show that phylogenetically

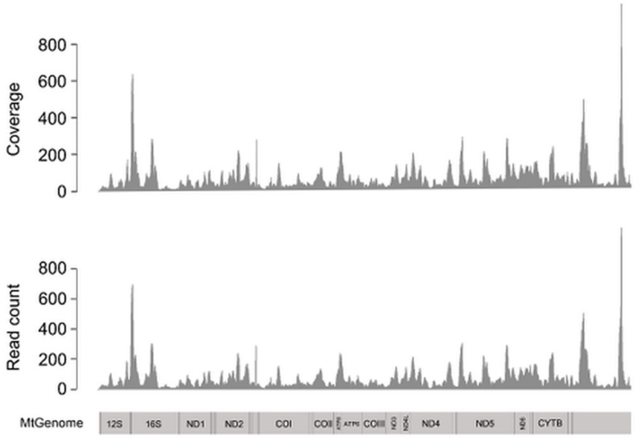

Fig 3. Coverage of the mitochondrial genome. Distribution of read counts (in 10 bp bins) and depth of the mitochondrial genome from the MVZ 214979. A vertebrate mitochondrial map is used for reference on the bottom to label regions of protein coding and rRNA genes. The control region is at the end of the map and is not labelled.

doi:10.1371/journal.pone.0141579.g003

informative DNA sequences can be generated from such specimens. Despite PCR duplications and DNA damage, our conservative approach resulted in genomic sequence data with high alignment quality for a 30-year old *Anolis carolinensis* museum specimen (MVZ 214979). In contrast, our sequencing effort failed with a 100-year old *Anolis carolinensis* specimen (MVZ 43405), which we elaborate on in greater detail below. For MVZ 214979, we obtained ~0.5X sequencing depth across the nuclear genome using half of one lane on an Illumina Hi-Seq sequencing platform. The low-coverage nuclear data were insufficient for SNP or genotype calling or to reconstruct sequence markers, especially given high error rates, but this would be remedied by greater sequencing effort (i.e., using additional sequencing lanes). Furthermore, in conjunction with our extraction protocols (S1 and S2 Files), Illumina sequencing was able to accurately recover the entire mitochondrial genome of MVZ 214979 with 57X average coverage even with the relatively modest sequencing effort employed here.

Formalin-fixation prior to DNA extraction results in extensive DNA damage as well as lower DNA yields, and sequences derived from formalin-fixed samples are likely to require special processing to account for both of these issues. For example, our raw reads required hard trimming to remove extensive base misincorporation at the ends of reads. In addition, a large percentage of our reads were discarded due to high levels of PCR duplication. The high level of PCR duplication was likely due to the low quantity of DNA available for library preparation, which then resulted in a low diversity of DNA fragments for sequencing. To obtain enough library material for sequencing, we followed the Illumina protocol's recommendation to increase the number of PCR cycles, which likely resulted in deep sequencing of the relatively small number of unique fragments that were present in the library compared to what would be expected with a fresh tissue sample. To avoid this problem using standard library-preparation protocols, we suggest using a larger amount of starting material, and performing multiple extractions of multiple tissues with the aim of reducing the number of PCR cycles necessary during library preparation, thereby limiting the level of redundancy in final libraries. Despite requiring extensive processing as described above, the sequencing results for MVZ 214979 were comparable to those of other ancient and museum historic DNA studies for proportion of reads mapped, error rate, and percent contamination (S2 Table). The highly variable depth of coverage for both the nuclear and mitochondrial genomes (Figs 2 and 3, respectively) seen in

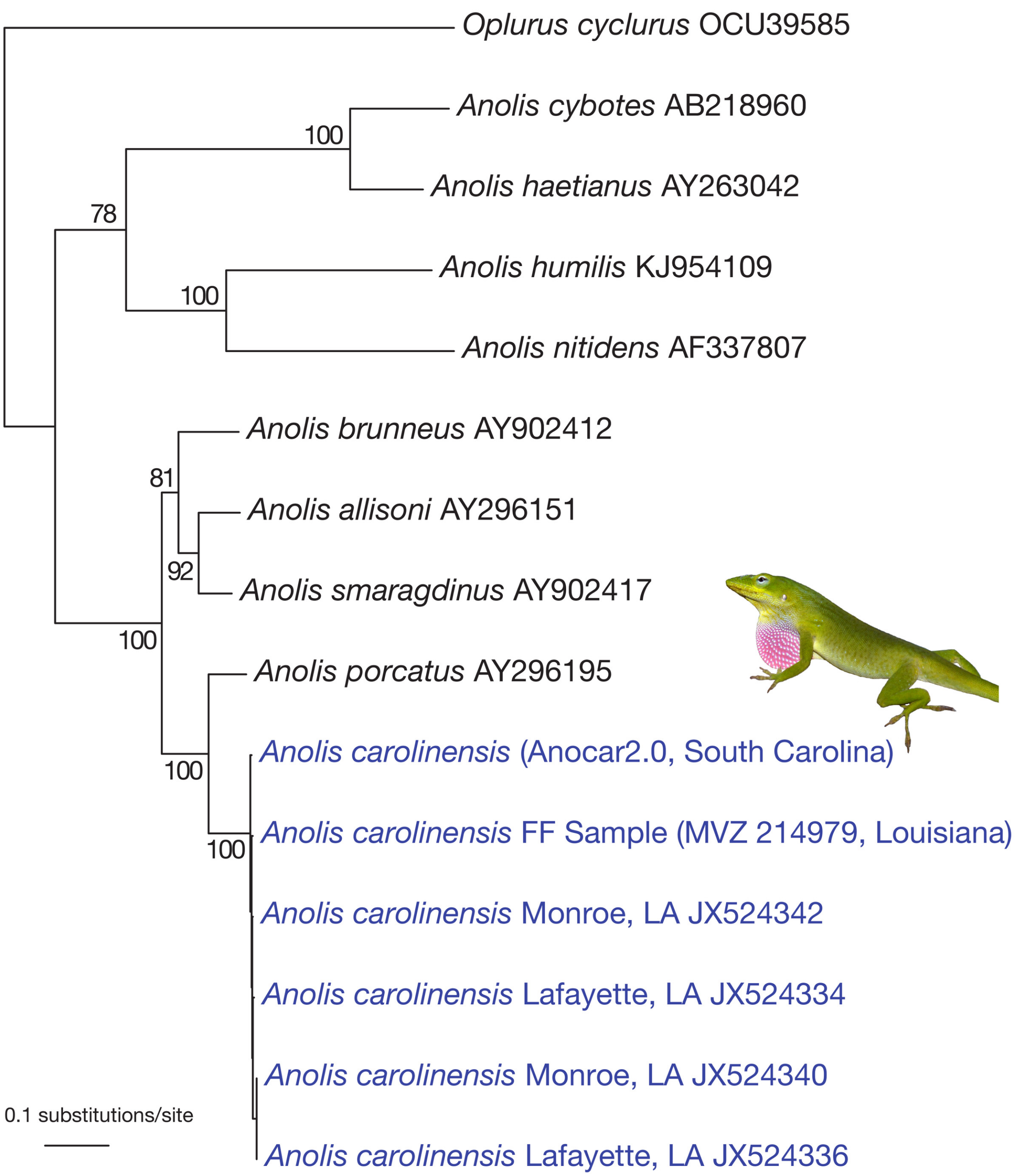
Oplurus cyclurus OCU39585
Anolis cybotes AB218960
Anolis haetianus AY263042
Anolis humilis KJ954109
Anolis nitidens AF337807
Anolis brunneus AY902412
Anolis allisoni AY296151
Anolis smaragdinus AY902417
Anolis porcatus AY296195
Anolis carolinensis (Anocar2.0, South Carolina)
Anolis carolinensis FF Sample (MVZ 214979, Louisiana)
Anolis carolinensis Monroe, LA JX524342
Anolis carolinensis Lafayette, LA JX524334
Anolis carolinensis Monroe, LA JX524340
Anolis carolinensis Lafayette, LA JX524336
100
78
100
81
92
100
100
100
0.1 substitutions/site

Fig 4. Phylogenetic inference using ND2 sequence data. Maximum-likelihood tree inferred from ND2 sequence alignment of the formalin-fixed sample (MVZ 214979), the Anocar2.0 reference genome (Anocar2.0), four *Anolis carolinensis* collected from Louisiana, USA, eight other *Anolis* species, and *Oplurus cyclurus*. *A*. *carolinensis* image printed under a CC BY license with permission of the original photographer and copyright owner J. Losos.

doi:10.1371/journal.pone.0141579.g004

our data is similar to that obtained in other aDNA/hDNA studies that employed non-targeted HTS [24,32]. For this reason, we suspect that this observed unevenness in coverage is more an artefact of PCR and/or sequencing and not other factors such as the composition of the *Anolis* genome or the exposure of DNA to formalin.

Since this experiment was performed, several new library preparation procedures have been developed that may be better suited to damaged or single-stranded DNA than those implemented here, and these protocols may increase the viability of otherwise marginal samples such as MVZ 43405. For example, the protocols of [39] and [40] omit the blunt-end repair step employed in this experiment, thereby maintaining the integrity of the DNA sequence at fragment-ends. Employing one of these alternate methods of library preparation and using a lower-fidelity polymerase (in order to avoid PCR bias [40,41]) might similarly improve the quality of libraries made from formalin-damaged DNA. Also, the pattern of DNA damage we observed in our formalin-fixed specimens differs in important ways from the patterns documented in prior ancient and historic DNA studies involving non-formalin-fixed specimens, suggesting that improved methods for modelling formalin damage to DNA could significantly improve sequencing success. With non-formalin-fixed samples, elevated rates of C to T misincorporated substitutions occur at the 5' ends of DNS strands, whereas G to T transitions are elevated at the 3' ends [28,32]. In contrast, we observed a sharp increase in mismatch frequencies of almost all types at both ends, and particularly at the 3' ends of reads. New approximate Bayesian methods for modelling ancient and historic DNA damage, such as implemented in the program mapDamage (available at http://ginolhac.github.io/mapDamage/ [42]), will likely result in less data-loss, especially if targeted-sequencing can be used to attain sufficient coverage of regions of interest. These potential refinements of our protocol offer the potential to reduce the amount of data loss resulting from hard trimming and should be considered in future studies.

As part of this study, we compared two basic extraction protocols on three tissue types to determine what, if any, combination of protocols and tissue types would yield sufficient quantities of double-stranded DNA for successful HTS. In choosing to test liver, muscle, and tail-tip, our expectation was that tail-tip would yield the most double-stranded DNA because previous studies of fluid preserved museum specimens [18,23] found that bone tissue is likely to protect DNA from degenerative forces, and DNA can be found in larger quantities in skeletal elements. We also tested muscle tissue because excising muscle is less destructive to the specimen than taking a tail-tip, digits, or teeth. Of all tissue types considered here, liver was the only one that yielded a sufficient quantity of double-stranded DNA. This result is encouraging because, in addition to being an easy, abundant source of tissue for subsampling, taking liver is minimally destructive to fluid-preserved specimens.

MVZ 43405, collected in 1917, did not yield a sufficient quantity of high-quality DNA to warrant continuing with library preparation and sequencing under typical circumstances. Nevertheless, we attempted (unsuccessfully) to obtain genomic sequence data from this sample. Why sequencing of MVZ 43405 failed is not entirely clear. One possibility is that our attempt to systematically test alternative extraction protocols for the two specimens was the decisive factor. As noted above, only liver samples returned measureable concentrations of double-stranded DNA. For each specimen, we partitioned the liver into two subsamples for extraction, one using the Qiagen protocol and the other using the more effective modified phenol-chloroform (PC) extraction protocol. For MVZ 214979 (which was sequenced successfully), the larger piece of liver (0.46g) was extracted using PC, whereas a much smaller liver sample (0.04g) was extracted using the Qiagen protocol. This relationship was reversed for MVZ 43405, with the larger liver sample (0.33g) extracted using the less effective Qiagen protocol and a much smaller liver sample (0.017g) extracted using phenol-chloroform. Notably, the PC extraction of the much smaller liver sample taken from MVZ 43405 yielded a larger quantity of double-stranded DNA than did the Qiagen extraction of a sample ~20 times larger (S1 Table). Had we known at the outset that PC extraction would be more effective and applied that extraction protocol to the larger liver sample from the older specimen, our sequencing effort may have succeeded. Another potential explanation for the sequencing failure is the condition of the formalin used to initially fix the specimens. It is known that MVZ herpetological specimens from the early 1900s were fixed in unbuffered formalin (David B. Wake, pers. comm.). This changed around 1970, when buffering formalin became standard practice to better preserve tissues for histological studies. Because we do not know whether the two MVZ specimens were fixed with buffered or unbuffered formalin, we cannot rigorously evaluate the importance of this variable here. We can, however, note that the data obtained in our generally unsuccessful attempt to sequence MVZ 43405 are consistent with expectations for DNA damage resulting from exposure to unbuffered formalin. The sequence data that we did obtained showed an extremely uneven distribution of mismatch frequencies that could be the result of high rates of base misincorporation across the length of all reads. Paireder et al. (2013) [43] systematically compared DNA yield from tissues fixed in either buffered or unbuffered formalin for known periods of time and found that DNA yield from tissues fixed in unbuffered formalin were significantly lower than those fixed in buffered formalin after two years, which they attributed to accelerated DNA degeneration due to the higher amount of formalin per unit of fixative and to the lower pH of unbuffered formalin [43]. While the small sample size of our study prevents us from testing the effect of buffered vs. unbuffered formalin, as well as many other potential variables (e.g. temperature, sunlight exposure, conditions of long-term storage), on hDNA from museum specimens, these are important considerations and worthy of further investigation. Future studies should seek to establish heuristics on endogenous DNA content of formalin-fixed liver according to the age and preservation conditions of the specimen.

For population genetic applications, which require accurate SNP/genotype calling, two approaches may be feasible using short sequence reads derived from formalin-fixed samples. The first, a whole genome shotgun approach as employed in this study, could be sufficient to attain good alignment with the presence of a pre-existing reference genome (S2 Table), although successful application for nuclear genomic data would require greater sequencing effort per specimen than was achieved here. Lacking genomic resources, de novo genome assemblies from modern specimens of closely related species could be used for alignment given sufficient sequencing depth. Unfortunately, in cases where formalin-fixed specimens are the only available genetic material for a project, generating a sound de novo assembly seems unlikely without a prohibitive amount of sequencing effort. Not only does the sequencing depth necessary for accurate base-calling present a challenge, formalin-fixation ultimately results in severely fragmented DNA. Increasing data yield might not sufficiently improve assembly quality due to the lack of long-insert genomic libraries available for genome scaffolding. For population genetic studies in which reduced representation approaches are used, the obstacles to accurate base-calling presented by formalin-induced base misincorporation may be circumvented by targeted capture of specific genomic regions. If genomic resources are available for marker development, we suspect that exon-capture [12] or similar methods will be able to achieve the depth of coverage necessary for these applications.

Although we were able to reconstruct the complete mitochondrial genome by mapping back to the reference genome, the results of this pilot study demonstrate the difficulty in generating high quality assemblies of the nuclear genome. This was due to the inherently fragmented nature of the DNA, and the low diversity of raw sequence data. Acknowledging that Whole Genome Sequencing is not necessary for phylogenetic studies, our results suggest that target enrichment approaches such as exon-capture—successful in other museum hDNA studies of non-fluid preserved specimens [12]—could be effective for targeting nuclear genomic regions for formalin-fixed museum specimens. Attempting target enrichment was deemed outside the scope of this exploratory study, but clearly would be an excellent avenue for future work. In this vein, exon-capture requires a genomic reference with which to design probes for targeted regions, and studies lacking modern genomic resources will have to contend with the difficulty posed by de novo assembly of formalin-fixed hDNA, as addressed above. In these situations, de novo sequencing of transcriptomes and whole genomes of modern samples of the same or closely related species may be an effective alternative. The nature of formalin-induced DNA damage we observed, however, recommends against the use of restriction enzyme-based reduced representation library approaches such as RADSeq (e.g [44,45]). The likelihood of severe degradation and base modification at recognition sites in formalin-fixed DNA may cause extensive data-loss across samples.

While this method will make many formalin-fixed specimens available for genetic study, we emphasize that there are two major differences between this extraction and sequencing protocol and those used to generate data from modern samples. The first is that the amount of starting material necessary to extract sufficient quantities of double-stranded DNA for library preparation is likely to be much greater on average for formalin-fixed samples. The ratio of starting material to DNA yield, however, is not likely to be consistent between samples due to variables such as specimen age, concentration and type of formalin used for preparation, and duration of formalin exposure. Correspondingly, this method will be most appropriate for specimens for which no fresh tissue is available, such as older type specimens or those representing species that are extirpated in the wild or otherwise unavailable for resampling. We encourage researchers not only to obtain fresh tissue samples when preparing specimens (flash-frozen or preserved in an appropriate medium, such as RNALater or 95% ethanol), but also to record the protocol used for preservation and use buffered formalin when fixing specimens. The second major difference is that an informed trimming approach is necessary when working with data generated from formalin-fixed specimens. Standard HTS quality control as used for modern (fresh tissue) samples is not likely to be sufficient for formalin-fixed material.

While hurdles remain regarding the wide-scale application of HTS data collection to formalin-fixed samples, this proof-of-concept study indicates that such samples can retain extractable and usable genomic sequence data, and that these data can be mined using available short-read sequencing platforms such as Illumina. Indeed, even without applying a reduced-representation or targeted sequencing approach, we have shown that direct sequencing of low-quality formalin-fixed specimens can be used to generate substantial nuclear sequence data and a high-coverage complete mitochondrial genome. Given the large number of species that are only represented by formalin-fixed museum specimens without corresponding tissues—including species known from one or a few specimens, or known to be extinct—just the ability to generate complete mitochondrial data alone is transformative. Key questions remain to be answered, of course, including (1) which characteristics determine the quality of the historical specimens for HTS data collection (possibilities include age, whether the specimen was fixed in unbuffered formalin, how long the specimen was exposed to formalin prior to transition to ethanol, etc.), and (2) whether targeted sequencing approaches will return high quality genome-scale data for formalin-fixed specimens as they can for specimens prepared as study skins. Despite the work that still needs to be done to answer these questions and streamline genomic sequencing of formalin-fixed museum specimens, the progress made here constitutes a significant step forward. High-throughput sequencing has the potential to unlock a treasure trove of genetic and genomic information for millions of museum specimens, and bring a large fraction of many museum collections into the age of genomics.

Supporting Information

S1 Fig. Quantification of MVZ 214979. Bioanalyzer trace of MVZ 214979 library prepared from liver extraction by phenol-chloroform.
(TIFF)

S2 Fig. Patterns of mismatches in MVZ 43405 sequences. The frequencies of the 12 types of mismatches (y-axis) are plotted as a function of distance from the 5′ and 3′ ends of the sequence reads (x-axis). The frequency of each mismatch type is coded in different colors and line patterns. Before cleaning the first 50 bp are shown from each end of the read.
(TIF)

S1 File. Protocol for experimental extraction of hDNA from formalin-fixed herpetological museum specimens.
(PDF)

S2 File. Protocol for extraction of hDNA from formalin-fixed museum specimens: liver extraction by phenol-chloroform.
(PDF)

S1 Table. Summary of DNA yield by tissue type and extraction protocol. Extractions used in library preparation and sequencing in bold. The abbreviation "TL" indicates that amounts of DNA were too low to be quantified. DNA quantification for all assays are given in units of ng/ μl.
(DOCX)

S2 Table. Summary of results from ancient and historic samples sequenced on the Illumina platform. Results of this study in bold.
(TIF)

Acknowledgments

We would like to thank the University of California Museum of Vertebrate Zoology for permitting us to sample tissues from the specimens used in this study, Lydia Smith for technical support, Sonal Singhal and Tyler Linderoth for developing scripts used in the bioinformatics pipelines, the members of the McGuire and M. H. Wake laboratories for help in editing this manuscript, and Jonathan Losos for contributing the photograph of *Anolis carolinensis*. This work also used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. We would also like to thank three anonymous reviewers for their helpful comments.

Author Contributions

Conceived and designed the experiments: SMH JAM. Performed the experiments: SMH. Analyzed the data: SMH KB JAM. Wrote the paper: SMH. Read and approved the final manuscript: SMH KB JAM.

References

1. Grinnell J (1910) The methods and uses of a research museum. Popular Science Monthly 77: 163–169.
2. Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312: 282–284. PMID: 6504142
3. Ristaino JB, Groves CT, Parra GR (2001) PCR amplification of the Irish potato famine pathogen from historic specimens. Nature 411: 695–697. doi: 10.1038/35079606 PMID: 11395772
4. Jønsson KA, Irestedt M, Fuchs J, Ericson PGP, Christidis L, Bowie RKC, et al. (2008) Explosive avian radiations and multi-directional dispersal across Wallacea: Evidence from the Campephagidae and other Crown Corvida (Aves). Molecular Phylogenetics and Evolution 47: 221–236. doi: 10.1016/j.ympev.2008.01.017 PMID: 18295511

5. Moyle RG, Jones RM, Andersen MJ (2013) A reconsideration of Gallicolumba (Aves: Columbidae) relationships using fresh source material reveals pseudogenes, chimeras, and a novel phylogenetic hypothesis. Molecular Phylogenetics and Evolution 66: 1060–1066. doi: 10.1016/j.ympev.2012.11.024 PMID: 23220516

6. Fulton TL, Wagner SM, Fisher C, Shapiro B (2012) Nuclear DNA from the extinct Passenger Pigeon (Ectopistes migratorius) confirms a single origin of New World pigeons. Annals of Anatomy—Anatomischer Anzeiger 194: 52–57.

7. Jønsson KA, Fabre P-H, Ricklefs RE, Fjeldså J (2011) Major global radiation of corvoid birds originated in the proto-Papuan archipelago. Proceedings of the National Academy of Sciences 108: 2328–2333. doi: 10.1073/pnas.1018956108

8. Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M (2009) Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Research 19: 1843–1848. doi: 10.1101/gr.095760.109 PMID: 19635845

9. O'Keefe K, Ramakrishnan U, Van Tuinen M, Hadly EA (2009) Source-sink dynamics structure a common montane mammal. Molecular Evolution 18: 4775–4789. doi: 10.1111/j.1365-294X.2009.04382.x

10. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463: 757–762. doi: 10.1038/nature08835 PMID: 20148029

11. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, et al. (2011) The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. Science 334: 89–94. doi: 10.1126/science.1209202 PMID: 21868630

12. Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. (2013) Unlocking the vault: next-generation museum population genomics. Molecular Evolution 22: 6018–6032. doi: 10.1111/mec.12516

13. Burrell AS, Disotell TR, Bergey CM (2015) Journal of Human Evolution. Journal of Human Evolution 79: 35–44.

14. Wong SQ, Li J, Tan AY-C, Vedururu R, Pang J-MB, Do H, et al. (2014) Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. 7: 1–10. doi: 10.1186/1755-8794-7-23

15. Quach N, Goodman MF, Shibata D (2004) In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. BMC Clin Pathol 4: 1. doi: 10.1186/1472-6890-4-1 PMID: 15028125

16. Campos P, Gilbert TP (2012) DNA Extraction from Formalin-Fixed Material. In: Shapiro B, Hofreiter M, editors. Methods in Molecular Biology. Humana Press, Vol. 840. pp. 81–85. doi: 10.1007/978-1-61779-516-9_11

17. Do H, Dobrovic A (2012) Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. Oncotarget 3: 546–558. PMID: 22643842

18. Stuart BL, Dugan KA, Allard MW, Kearney M (2006) Extraction of nuclear DNA from bone of skeletonized and fluid-preserved museum specimens. Systematics and Biodiversity 4: 133–136.

19. Loudig O, Brandwein-Gensler M, Kim RS, Lin J, Isayeva T, Liu C, et al. (2011) Illumina whole-genome complementary DNA—mediated annealing, selection, extension and ligation platform: assessing its performance in formalin-fixed, paraffin-embedded samples and identifying invasion pattern—related genes in oral squamous cell carcinoma. Human Pathology 42: 1911–1922. doi: 10.1016/j.humpath.2011.02.011 PMID: 21683979

20. Kerick M, Isau M, Timmermann B, Sültmann H, Herwig R, Krobitsch S, et al. (2011) Targeted high throughput sequencing in clinical cancer Settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. BMC Medical Genomics 4: 68. doi: 10.1186/1755-8794-4-68 PMID: 21958464

21. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, et al. (2014) DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome Research 24: 2022–2032. doi: 10.1101/gr.175141.114 PMID: 25236618

22. Steinau M, Patel SS, Unger ER (2011) Efficient DNA Extraction for HPV Genotyping in Formalin-Fixed, Paraffin-Embedded Tissues. JMDI 13: 377–381. doi: 10.1016/j.jmoldx.2011.03.007

23. Kearney M, Stuart BL (2004) Repeated evolution of limblessness and digging heads in worm lizards revealed by DNA from old bones. Proceedings of the Royal Society B: Biological Sciences 271: 1677–1683. doi: 10.1098/rspb.2004.2771 PMID: 15306287

24. Rowe KC, Singhal S, MacManes MD, Ayroles JF, Morelli TL, Rubidge EM, et al. (2011) Museum genomics: low-cost and high-accuracy genetic data from historical specimens. Molecular Ecology Resources 11: 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x PMID: 21791033

25. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Meth 9: 357–359. doi: 10.1038/nmeth.1923

26. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25. doi: 10.1186/gb-2009-10-3-r25 PMID: 19261174

27. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104: 14616–14621. doi: 10.1073/pnas.0704665104 PMID: 17715061

28. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S (2012) Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLoS ONE 7: e34131.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

30. Singhal S (2013) De novotranscriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Molecular Ecology Resources 13: 403–416.

31. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. (2009) ABySS: A parallel assembler for short read sequence data. Genome Research 19: 1117–1123. doi: 10.1101/gr.089532.108 PMID: 19251739

32. Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13: 1–1.

33. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150–3152. doi: 10.1093/bioinformatics/bts565 PMID: 23060610

34. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Research 12: 656–664.

35. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Research 9: 868–877. PMID: 10508846

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. doi: 10.1016/S0022-2836(05)80360-2 PMID: 2231712

37. Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6: Unit6.4. doi: 10.1002/0471250953.bi0604s00

38. Bazinet AL, Zwickle DJ, Cummings MP (2104) A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Sys Bio 63(5):812–818. doi: 10.1093/sysbio/syu031

39. Gansauge M-T, Meyer M (2013) Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc 8: 737–748. doi: 10.1038/nprot.2013.038 PMID: 23493070

40. Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl E-M, Grange T. (2014) Library construction for ancient genomics: single strand or double strand? BioTechniques 56: 289–90–292–6–298–passim.

41. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, et al. (2012) Optimal enzymes for amplifying sequencing libraries. Nat Meth 9: 10–11. doi: 10.1038/nmeth.1814

42. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. J Gerontol 29: 1682–1684. doi: 10.1093/bioinformatics/btt193

43. Paireder S, Werner B, Bailer J, Werther W, Schmid E, Patzak B, et al. (2013) Comparison of protocols for DNA extraction from long-term preserved formalin fixed tissues. Analytical Biochemistry 439: 152–160. doi: 10.1016/j.ab.2013.04.006 PMID: 23603300

44. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17: 240–248. doi: 10.1101/gr.5681207 PMID: 17189378

45. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6: e19379. doi: 10.1371/journal.pone.0019379.s004 PMID: 21573248