Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens

SARA RUANE* and CHRISTOPHER C. AUSTIN†
*Museum of Natural Science, Louisiana State University, 119 Foster Hall, Baton Rouge LA, 70803, USA, †Department of Biological Sciences, Louisiana State University, 202 Life Science Building, 119 Foster Hall, Baton Rouge LA, 70803, USA

Abstract

Museum specimens provide a wealth of information to biologists, but obtaining genetic data from formalin-fixed and fluid-preserved specimens remains challenging. While DNA sequences have been recovered from such specimens, most approaches are time-consuming and produce low data quality and quantity. Here, we use a modified DNA extraction protocol combined with high-throughput sequencing to recover DNA from formalin-fixed and fluid-preserved snakes that were collected over a century ago and for which little or no modern genetic materials exist in public collections. We successfully extracted DNA and sequenced ultraconserved elements (x = 2318 loci) from 10 fluid-preserved snakes and included them in a phylogeny with modern samples. This phylogeny demonstrates the general use of such specimens in phylogenomic studies and provides evidence for the placement of enigmatic snakes, such as the rare and never-before sequenced Indian Xylophis stenorhynchus. Our study emphasizes the relevance of museum collections in modern research and simultaneously provides a protocol that may prove useful for specimens that have been previously intractable for DNA sequencing.

Keywords: museum collections, next-generation sequencing, phylogeny, sequence capture, snakes, ultraconserved elements

Received 11 October 2016; revision received 10 January 2017; accepted 19 January 2017

Introduction

Natural history specimens provide irreplaceable information on the morphologies, diseases, diets and population changes of organisms (Suarez & Tsutsui 2004; Holmes et al. 2016). Excitingly, for some types of dry preparations, genetic data can be recovered from such specimens, even those over 100 years old (Payne & Sorenson 2002; Lim & Braun 2016; McCormack et al. 2016). But for formalin-fixed and fluid-stored organisms, obtaining these genetic data remains challenging, with much work and little reward (Simmons 2014; Hykin et al. 2015). This is unfortunate because for many taxa, tissue samples ideal for molecular phylogenetic use are unavailable (hereafter referred to as DNA-grade tissue).

One vertebrate clade that would greatly benefit from efficient mechanisms for collecting genetic data from formalin-fixed and fluid-stored intractable specimens is snakes. Despite being a significant part of vertebrate biodiversity, with 3600+ described species, snakes are among the most poorly understood vertebrate groups with respect to basic biology and systematics. This lack of understanding results from multiple factors, including the following: (i) a secretive nature that makes snakes difficult to collect; (ii) no genetic resources for species that were only collected before taking genetic samples was standard; and (iii) regulations that make collection challenging, for example limitations due to endangerment.

While such hindrances make the procurement of DNA-grade tissue samples for phylogenetic work difficult, museums can be excellent resources for species that are hard to obtain. However, such specimens are often fixed and preserved in alcohol or initially fixed using formalin (Simmons 2014). Formalin fixation makes the extraction of DNA for gene amplification problematic, partly due to extensive DNA shearing and damage resulting from the initial formalin treatment (Simmons 2014). However, there is promise with next-generation technologies, which specifically require sheared DNA for sequencing. Therefore, a next-generation sequencing procedure that takes advantage of fluid-preserved museum snakes is timely. Although at least one prior study has successfully sequenced DNA from a ~30-year-old formalin-fixed lizard using a next-generation
approach (Hykin et al. 2015), no study to date has generated large-scale phylogenomic data from multiple, older fluid-preserved specimens.

Here, we successfully sequence as many as 3433 ultra-conserved elements (UCEs; Faircloth et al. 2012) from 10 snakes, including formalin-fixed and 100+ year-old specimens. We integrate these data into a larger UCE data set generated from modern DNA-grade samples to estimate a phylogeny. This phylogeny provides insight into the placement of several enigmatic taxa, as only two of the species had any genetic resources available prior. In addition to demonstrating the utility of our approach for collecting genetic data from previously intractable, formalin-fixed specimens, our results further emphasize the importance of museum collections in modern genome-scale studies.

Materials and methods

Using the web portals VertNet, IUCN Redlist and the Reptile Database, we compiled a list of 21 rare snakes (17 species) with estimated collection dates from the 1870s to 1990 and obtained liver tissue samples through museum loans. Most of these snakes (16 specimens) were formalin fixed, but several of the specimens were fixed with alcohol (see Appendix S1, Supporting information, for details on samples). For all of the fluid-preserved specimens, we used a Qiagen DNeasy Kit® and modified existing protocols for the extraction of degraded tissues (Gilbert et al. 2007; Hykin et al. 2015; McCormack et al. 2016). We initially tried a standard phenol–chloroform extraction in a pilot study of several Nerodia spp. preserved in formalin from the 1930s, 1940s and 1950s (vouchered at the Museum of Natural Science at Louisiana State University) and compared this to results from a Qiagen DNeasy Kit® approach; we found no obvious difference in the amount of DNA recovered from these specimens using each method, and so we focused our extraction efforts using just the Qiagen DNeasy Kit®.

A 100–200-mg piece of liver tissue was removed from the preserved snakes via a small incision on the ventral surface. Many snakes have incisions already along the ventral surface to allow for better penetration of preservation fluids, and many times an additional incision was unnecessary. All tools and work surfaces were sterilized between specimens using a 10% bleach solution; tools, such as pipettes, were also sterilized using a UV box. After the initial removal of the liver sample, the tissue was cut into multiple, smaller pieces and placed into a 50-mL Falcon® tube and soaked in distilled water for 6 h to help remove residual ethanol. Each tissue sample was further cut up into ~15–25-mg pieces with a sterile razor blade, and a total of ~50–100 mg of tissue was placed into a 2-mL microcentrifuge tube with 300 µL of buffer ATL (preheated to 98 °C), which was then incubated at 98 °C for 15 min. The tube was then cooled on ice for 2 min. We added 40 µL of Protinease K and incubated the tube at 65 °C for 48 h, and vortexed it two or more times per 24-h period. Additional Protinease K (25 µL at a time and up to three additional times) was added during the incubation period if lysis was not progressing via visual assessment (i.e. obvious pieces of tissue remained in the tube).

After 48 h, DNA was extracted following the Qiagen DNeasy Kit® instructions, but with a 30-min room temperature incubation with 140 µL of elution buffer AE. Extractions were quantified using a Qubit® high-sensitivity assay kit. A negative control was simultaneously included for DNA extraction and taken all the way through to sequencing to ensure that positive results were not due to contamination. Extracts were stored in a −20 °C freezer.

Extracts with quantifiable DNA and the negative control were sent to MYcroarray® for UCE library preparation using the MYbaits® tetrapod 5k kit, a set of 120-mer baits that target 5060 UCEs from amniotes (Faircloth et al. 2012). Samples were not further sheared, and libraries were captured individually and, following enrichment, were sent to Oklahoma Medical Research Facility for sequencing in a single lane on an Illumina Hiseq 3000, with 75-bp paired ends. Sequences were cleaned using illumiprocessor, contigs assembled using ABYSS with a 60 kmer length, and matched to the UCE probes with a coverage cut-off of 40×, all executed in PHYLUCe (Simpson et al. 2009; Del Fabbro et al. 2013; Faircloth 2015). Additional details on the library preparation, including barcodes, are available in the Appendix S1 (Supporting information).

These data were combined with UCEs from 28 snakes from a previous study (Streicher & Wiens 2016), and we included UCEs mined from the genome of the snake Python molurus (Castoe et al. 2013) and the outgroup Anolis carolinensis (Alföldi et al. 2011). Due to the short lengths of resulting UCEs (see ‘Results’), which are probably inappropriate for species tree analyses (Hosner et al. 2016; Manthey et al. 2016), we focused on concatenated analyses to provide a phylogeny. Concatenated maximum-likelihood trees were estimated with matrices composed of 50% and 35% missing data (see Streicher & Wiens 2016 for justification of missing data matrices) using RAxML v.8 with a single partition under the GTRGAMMA model, with 100 bootstrap replicates (Stamatakis 2014).

Results

Quantifiable DNA was recovered from 13 samples (with the negative control showing no DNA) (Table 1). During
library construction, size-appropriate inserts indicating success were detected in only 10 samples, but all 13 samples were enriched and sequenced, along with the negative control. Ultimately, only the samples with the detectable inserts resulted in viable UCEs postprocessing (Table 1). Successfully sequenced individuals were from nine formalin-fixed samples and one that was likely an alcohol-fixed sample. Postassembly, a mean of 2318 UCEs were collected from these successfully sequenced snakes (Table 1). Resulting topologies were generally well supported and similar across trees; we show the 50\% matrix (3544 UCEs; 968 500 bp) RAXML tree here (Fig. 1; additional results in Appendix S1, Supporting information, and Dryad http://dx.doi.org/10.5061/dryad.cm28r.2). Raw sequence reads available on the NCBI Sequence Read Archive SAMN06142703–SAMN06142712. The UCEs from the modern samples vs. the fluid-preserved, intractable samples were comparable in most ways. For example, the average number of UCEs collected from intractable samples was 2318 vs. 2669 UCEs for modern samples and the average number of parsimony informative sites across all UCEs for the intractable samples 24 915 and slightly higher at 31 658 for the modern samples. This was also reflected by the number of gaps/ambiguities in the intractable samples with 65.2\% gaps/ambiguities vs. 38.8\% gaps/ambiguities for the modern samples (additional information for these resulting UCE details in Appendix S1, Supporting information, and Dryad http://dx.doi.org/10.5061/dryad.cm28r.2).

Post hoc, we also attempted to extract mitochondrial loci from each of the intractable samples by mapping the raw reads to related snake mitogenomes downloaded from GenBank in Geneious®. While we were unable to reconstruct full mitogenomes for these taxa, in some cases identifiable genes (e.g. cytochrome \( b \)) could be recovered and ‘blasted’ back via the NCBI web server ‘BLAST’ search to the most closely related taxon available. We did not include these mtDNA loci in any analyses here, but mention it to highlight the additional data that may be garnered from such samples.

Table 1 Summary of results for fluid-preserved specimens sequenced for this study

<table>
<thead>
<tr>
<th>Sample</th>
<th>Family</th>
<th>Collected</th>
<th>ng/( \mu )L DNA</th>
<th># Reads</th>
<th># Contigs</th>
<th># UCEs</th>
<th>( x ) UCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAS44118 Cacophis harriettae</td>
<td>Elapidae</td>
<td>1913</td>
<td>1.53</td>
<td>38 470 164</td>
<td>73 530</td>
<td>2855</td>
<td>152</td>
</tr>
<tr>
<td>CAS84059 Cacophis kreftii</td>
<td>Elapidae</td>
<td>1948</td>
<td>1.72</td>
<td>34 568 255</td>
<td>77 382</td>
<td>2520</td>
<td>145</td>
</tr>
<tr>
<td>CAS84060 Cacophis kreftii</td>
<td>Elapidae</td>
<td>1948</td>
<td>0.89</td>
<td>25 987 774</td>
<td>33 899</td>
<td>2664</td>
<td>138</td>
</tr>
<tr>
<td>CAS84057 Cacophis kreftii</td>
<td>Elapidae</td>
<td>1948</td>
<td>4.97</td>
<td>26 933 634</td>
<td>71 235</td>
<td>2060</td>
<td>224</td>
</tr>
<tr>
<td>MCZ R-65918 Paraparadactylus hedigeri</td>
<td>Elapidae</td>
<td>1961</td>
<td>2.04</td>
<td>29 472 343</td>
<td>126 674</td>
<td>3443</td>
<td>294</td>
</tr>
<tr>
<td>CAS121098 Antaioserpens warro</td>
<td>Elapidae</td>
<td>1968</td>
<td>11.5</td>
<td>30 952 427</td>
<td>60 594</td>
<td>2818</td>
<td>150</td>
</tr>
<tr>
<td>MCZ R-23814 Thalassophis anomalus</td>
<td>Elapidae</td>
<td>&lt;1958</td>
<td>1.33</td>
<td>28 727 457</td>
<td>12 053</td>
<td>995</td>
<td>123</td>
</tr>
<tr>
<td>MCZ R-145946 Toxicocalamus Ernestmayri</td>
<td>Elapidae</td>
<td>1969</td>
<td>0.71</td>
<td>30 710 619</td>
<td>5793</td>
<td>359</td>
<td>125</td>
</tr>
<tr>
<td>MCZ R-177248 Brachyophis albus</td>
<td>Homolopsidae</td>
<td>1906</td>
<td>1.82</td>
<td>32 315 527</td>
<td>60 863</td>
<td>2918</td>
<td>155</td>
</tr>
<tr>
<td>CAS17199 Xylophis stenorhynchus</td>
<td>Uncertain</td>
<td>1878–1911</td>
<td>2.00</td>
<td>32 236 948</td>
<td>23 499</td>
<td>2546</td>
<td>135</td>
</tr>
<tr>
<td>CAS15967 Lycocton truncoracicus*</td>
<td>Colubridae</td>
<td>1941</td>
<td>0.69</td>
<td>10 778 312</td>
<td>362</td>
<td>Failed</td>
<td>N/A</td>
</tr>
<tr>
<td>LSUMNS-41802 Stegonotus muelleri*</td>
<td>Colubridae</td>
<td>1986</td>
<td>1.76</td>
<td>1 707 518</td>
<td>12</td>
<td>Failed</td>
<td>N/A</td>
</tr>
<tr>
<td>CAS244490 Uropeltis nitidus*</td>
<td>Uropeltidae</td>
<td>1990</td>
<td>1.24</td>
<td>2 653 679</td>
<td>87</td>
<td>Failed</td>
<td>N/A</td>
</tr>
<tr>
<td>Negative control*</td>
<td>N/A</td>
<td>N/A</td>
<td>‘too low’</td>
<td>23 614 868</td>
<td>17</td>
<td>Failed</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\( x \) UCE = mean UCE length.

* Inserts not detected during library preparation.

Discussion

There are multiple high-throughput sequencing techniques now available, but targeted capture, in this case of UCEs, seems ideal for fluid-preserved intractable specimens. It takes advantage of already fragmented DNA, and unlike restriction-site approaches, target capture is likely to result in replicable results across taxa, with a high coverage of thousands of phylogenetically informative loci, even for low-quality samples (Harvey et al. 2016; McCormack et al. 2016). Using UCEs also allows us to integrate the data set of snakes sequenced...
here with previously generated tetrapod UCE data sets (Streicher & Wiens 2016), as well as with UCEs from genomes, at no additional cost. Although the average UCE length from fluid-preserved specimens was lower (164 vs. 468 bp), the average number of UCEs collected from the fluid-preserved specimens is comparable to those from contemporary snake studies (2318 UCEs vs. 2669 UCEs), with some fluid-preserved samples exceeding the number of UCEs collected using DNA-grade tissues (Streicher & Wiens 2016; see Appendix S1, Supporting information). Longer UCE lengths for the intractable specimens could probably be obtained using a more densely tiled UCE probe set (B. Faircloth, personal communication).

The successfully sequenced fluid-preserved samples included eight species representing several snake families. These samples have the expected taxonomic relationships to the snakes previously sequenced (Streicher & Wiens 2016) and even corroborate phylogenetic hypotheses based on morphology from prior work. For example, we inferred the never-before sequenced Para-pistocalamus hedigeri as the sister taxon to all other hydrophiine elapids, as has been previously suggested (Fig. 1; Strickland et al. 2016). Our data set included three intraspecific samples of the elapid Cacophis kreffti, the intrageneric C. harriettae, plus five other intrafamilial species, illustrating the success using fluid-preserved specimens across multiple phylogenetic levels (Fig. 1).

Fig. 1 Maximum-likelihood tree of 3544 UCEs of 39 snakes + outgroup. Fluid-preserved taxa in pink bold font; bootstrap support shown <100; *familial affiliation for Xylophis stenorhynchus uncertain, see text for details. Illustration used under creative commons license (https://doi.org/10.5962/bhl.title.4679).

© 2017 John Wiley & Sons Ltd
Whether this approach will work consistently for such specimens across different phylogenetic scales or taxonomic groups is unknown and remains an area for future research.

While this is not the first study to successfully generate DNA sequences from formalin-fixed and fluid-preserved specimens, prior work has relied mainly on time-intensive Sanger sequencing, producing only a few loci (Friedman & DeSalle 2008; Dubey et al. 2011; Simmons 2014 and citations within). In contrast, our protocol is simple, cost and time-effective, and generates thousands of loci. We expect this will be especially useful for museum specimens that are the only samples available for certain species or specific localities. For example, we successfully obtained 2546 UCEs from the enigmatic Indian snake *Xylophis stenorhynchus*, collected between 1878 and 1911 from the Travancore of India. Collecting in India is challenging, and this is one of the few specimens of *X. stenorhynchus* available in public collections.

There is uncertainty regarding both the number of species and the phylogenetic placement of *Xylophis* among snakes; it has been previously suggested to have close affinity to the Asian Xenodermatidae (Gower & Winkler 2007) or to natricines (Simões et al. 2016). However, our results provide strong support for a sister relationship with the Asian Pareatidae (Fig. 1). A robust phylogenetic assessment of *Xylophis* could be conducted by combining target-capture data sets from DNA-grade *Xylophis* with those from historic specimens and including additional paratids and xenodermatids.

As recently as a decade ago, the possibility of collecting large quantities of genetic data from formalin-fixed, fluid-preserved museum specimens seemed exceedingly unlikely. This study highlights the potential uses in the contemporary phylogenomic age for museum specimens that were previously intractable. We suggest that not only do these historical specimens have importance with respect to both traditional and modern applications, but that additional uses for these specimens will become apparent as technology advances and underscores the pertinence of the continuance of biological collections.

**Acknowledgements**

Thanks to J. Vindum (California Academy of Sciences) and J. Rosado (Harvard Museum of Comparative Zoology) for specimen loan; J. Streicher for UCE contigs; A. Devault and J. Enk (MYcroarray) for sequencing suggestions; F. Alda for computational aid and moral support; B. Faircloth and R. Brumfield for review and suggestions on earlier versions of the manuscript; and the three anonymous reviewers whose comments considerably improved our manuscript. Funding was provided by the Louisiana State University Museum of Natural Science to S.R. and NSF DEB-1146033 to C.C.A.

**References**


Data accessibility

Data supporting this article are available as Supporting information, at Dryad http://dx.doi.org/10.5061/dryad.cm28r.2, and raw sequence reads are available on the NCBI Sequence Read Archive, samples SAMN06142703–SAMN06142712.

Supporting Information

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

Appendix S1 Specimen details and library prep details.

© 2017 John Wiley & Sons Ltd