

DR. SARA RUANE (Orcid ID : 0000-0002-9543-1297)

Received Date : 11-Oct-2016

Revised Date : 10-Jan-2017

Accepted Date : 19-Jan-2017

Article type : Resource Article

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

Sara Ruane^{1,2} and Christopher C. Austin^{2,3}

¹Current Address: Department of Biological Sciences, Rutgers University-Newark, 409 Boyden Hall, Newark, NJ 07102

²Museum of Natural Science, Louisiana State University, 119 Foster Hall, Baton Rouge, LA 70803

³Department of Biological Sciences, 202 Life Science Building, Louisiana State University, 119 Foster Hall, Baton Rouge, LA 70803

Corresponding author:

Sara Ruane, Department of Biological Sciences, Rutgers University-Newark, 409 Boyden Hall, Newark, NJ 07102; Email: sara.ruane@rutgers.edu

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:

10.1111/1755-0998.12655

This article is protected by copyright. All rights reserved.

Key words: Museum collections; Next-generation sequencing; Sequence capture; Phylogeny; Snakes; Ultraconserved elements

Running title: Phylogenomics using fluid-preserved specimens

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 ($\bar{x} = 2,318$ 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.

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; McCormack *et al.* 2016; Lim & Braun 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: 1) a secretive nature which makes snakes difficult to collect; 2) no genetic resources for species that were only collected before taking genetic samples was standard; 3) 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 3,433 ultraconserved 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 dataset generated from modern DNA-grade samples to estimate a phylogeny. This phylogeny provides insight on 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–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 online Appendix 1 for details on samples). For all of the fluid preserved specimens, we used a Quiagen 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 µg 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 hours to help remove residual ethanol. Each tissue sample was further cut up into ~15–25 µg pieces with a sterile razor blade and a total of ~50–100 µg 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 minutes. The tube was then cooled on ice for 2 minutes. We added 40 µl of Protinease K and incubated the tube at 65° C for 48 hours, and vortexed it two or more times per 24 hour 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 hours, DNA was extracted following the Qiagen DNeasy Kit® instructions, but with a 30 minute 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 120mer baits that target 5,060 UCEs from amniotes (Faircloth *et al.* 2012). Samples were not further sheared, 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 60kmer length, and matched to the UCE probes with a coverage cutoff of 40x, 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 online supporting material Appendix 1.

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 likely 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 insert 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 post processing (Table 1). Successfully sequenced individuals were from nine formalin-fixed samples and one that was likely an alcohol-fixed sample. Post-assembly, a mean of 2,318 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 (3,544 UCEs; 968,500 bp) RAxML tree here (Fig. 1; additional results in online Appendix 1 and Dryad <http://dx.doi.org/10.5061/dryad.cm28r> ; raw sequence reads available on the NCBI Sequence Read Archive SAMN06142703–SAMN06142712). The UCEs from the modern samples versus the fluid-preserved, intractable samples were comparable in most ways. For example the average number of UCEs collected from intractable samples was 2,318 vs. 2,669 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 parsimony informative sites for modern samples. Furthermore, a posthoc examination in the program Geneious® for potential DNA damage also showed little difference between the samples based on base composition content (an excess of C>T artefacts may indicate DNA damage in formalin-fixed or ancient samples, e.g., Do & Dobrovic 2009; Hofreiter *et al.* 2001), with the intractable specimen UCEs having average base pairs contents (with intractable specimens listed first) of 31% vs. 31.5% T, 32.3% vs. 31.2% A, 18.3% vs. 19.1% G, and 17.9% vs. 18.8% C. The most apparent difference and indication of higher levels of degradation for the fluid-preserved specimens was with respect to UCE length, with a 164 bp average UCE length for intractable samples and a 468 bp average UCE length 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 online Appendix 1 and Dryad <http://dx.doi.org/10.5061/dryad.cm28r>).

Posthoc 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 genomes for these taxa, in some cases identifiable genes (e.g., cytochrome b) and which could be “blasted” back via the NCBI webserver “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.

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 high coverage of thousands of phylogenetically informative loci, even for low-quality samples (McCormack *et al.* 2016; Harvey *et al.* 2016). Using UCEs also allows us to integrate the dataset of snakes sequenced here with previously generated tetrapod UCE datasets (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 (2,318 UCEs vs. 2,669 UCEs), with some fluid-preserved samples exceeding the number of UCEs collected using DNA-grade tissues (Streicher & Wiens 2016; see online Appendix 1). Longer UCE lengths for the intractable

specimens could likely be obtained by using a more densely tiled UCE probe set (B. Faircloth, Pers. Comm.).

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 *Parapistocalamus hedigeri* as the sister taxon to all other hydrophiine elapids, as has been previously suggested (Fig. 1; Strickland *et al.* 2016). Our dataset included three intraspecific samples of the elapid *Cacophis krefftii*, the intrageneric *C. harriettae*, plus five other intrafamilial species, illustrating the success using fluid-preserved specimens across multiple phylogenetic levels (Fig. 1). 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 2,546 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 familial placement of *Xylophis* among snakes; based on morphology, it has been previously suggested to have close

affinity to the Asian Xenodermatidae (Gower & Winkler 2007). However, our results provide strong support for a sister relationship with the Asian Preatidae (Fig.1). A robust phylogenetic assessment of *Xylophis* could be conducted by combining target-capture datasets from DNA-grade *Xylophis* with those from historic specimens and including additional preatids 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; the three anonymous reviewers whose comments considerably improved our manuscript. Funding provided by the Louisiana State University Museum of Natural Science to S.R. and NSF DEB-1146033 to C.C.A.

Literature Cited

- Alföldi J, Di Palma F, Grabherr M *et al.* (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477**, 587–591.
- Castoe TA, de Koning APJ, Hall KT *et al.* (2013) The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences*, **110**, 20645–20650.
- Do H, Dobrovic A (2009) Limited copy number-high resolution melting (LCN-HRM) enables the detection and identification by sequencing of low level mutations in cancer biopsies. *Molecular Cancer*, **8**, 82.
- Dubey B, Meganathan PR, Haque I (2011) DNA mini-barcoding: An approach for forensic identification of some endangered Indian snake species. *Forensic Science International: Genetics*, **5**, 181–184.
- Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*, **8**, e85024.
- Faircloth BC (2015) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, doi:10.1093/bioinformatics/btv646.
- Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, **61**, 717–726.

Friedman M, DeSalle R (2008) Mitochondrial DNA extraction and sequencing of formalin-fixed archival snake tissue: Short Communication. *DNA Sequence*, **19**, 433–437.

Gilbert MTP, Haselkorn T, Bunce M *et al.* (2007) The isolation of nucleic acids from fixed, paraffin-embedded tissues—which methods are useful when? *PLoS One*, **2**, e537.

Gower DJ, Winkler JD (2007) Taxonomy of the Indian snake *Xylophis Beddome* (Serpentes: Caenophidia), with description of a new species. *Hamadryad*, **31**, 315–329.

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2016) Sequence Capture Versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Systematic Biology*, **65**, 910–924.

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

Holmes MW, Hammond TT, Wogan GOU *et al.* (2016) Natural history collections as windows on evolutionary processes. *Molecular Ecology*, **25**, 864–881.

Hosner PA, Faircloth BC, Glenn TC, Braun EL, Kimball RT (2016) Avoiding missing data biases in phylogenomic inference: An empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, **33**, 1110–1125.

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**, e0141579.

Lim HC, Braun MJ (2016) High-throughput SNP genotyping of historical and modern samples of five bird species via sequence capture of ultraconserved elements. *Molecular Ecology Resources*, **16**, 1204–23.

Manthey JD, Campillo LC, Burns KJ, Moyle RG (2016) Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: *Piranga*). *Systematic Biology*, **65**, 640–650.

McCormack JE, Tsai WLE, Faircloth BC (2016) Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, **16**, 1189–1203.

Payne RB, Sorenson MD (2002) Museum collections as sources of genetic data. *Bonner Zoologische Beiträge*, **51**, 97–104.

Simmons JE (2014) *Fluid preservation: a comprehensive reference*. Rowman & Littlefield, Lanham, MD.

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

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–3.

Streicher JW, Wiens JJ (2016) Phylogenomic analyses reveal novel relationships among snake families. *Molecular Phylogenetics and Evolution*, **100**, 160–169.

Strickland JL, Carter S, Kraus F, Parkinson CL (2016) Snake evolution in Melanesia: origin of the Hydrophiinae (Serpentes, Elapidae), and the evolutionary history of the enigmatic New Guinean elapid *Toxicocalamus*. *Zoological Journal of the Linnean Society*, doi:10.1111/zoj.12423.

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

Data accessibility

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

Author's contributions

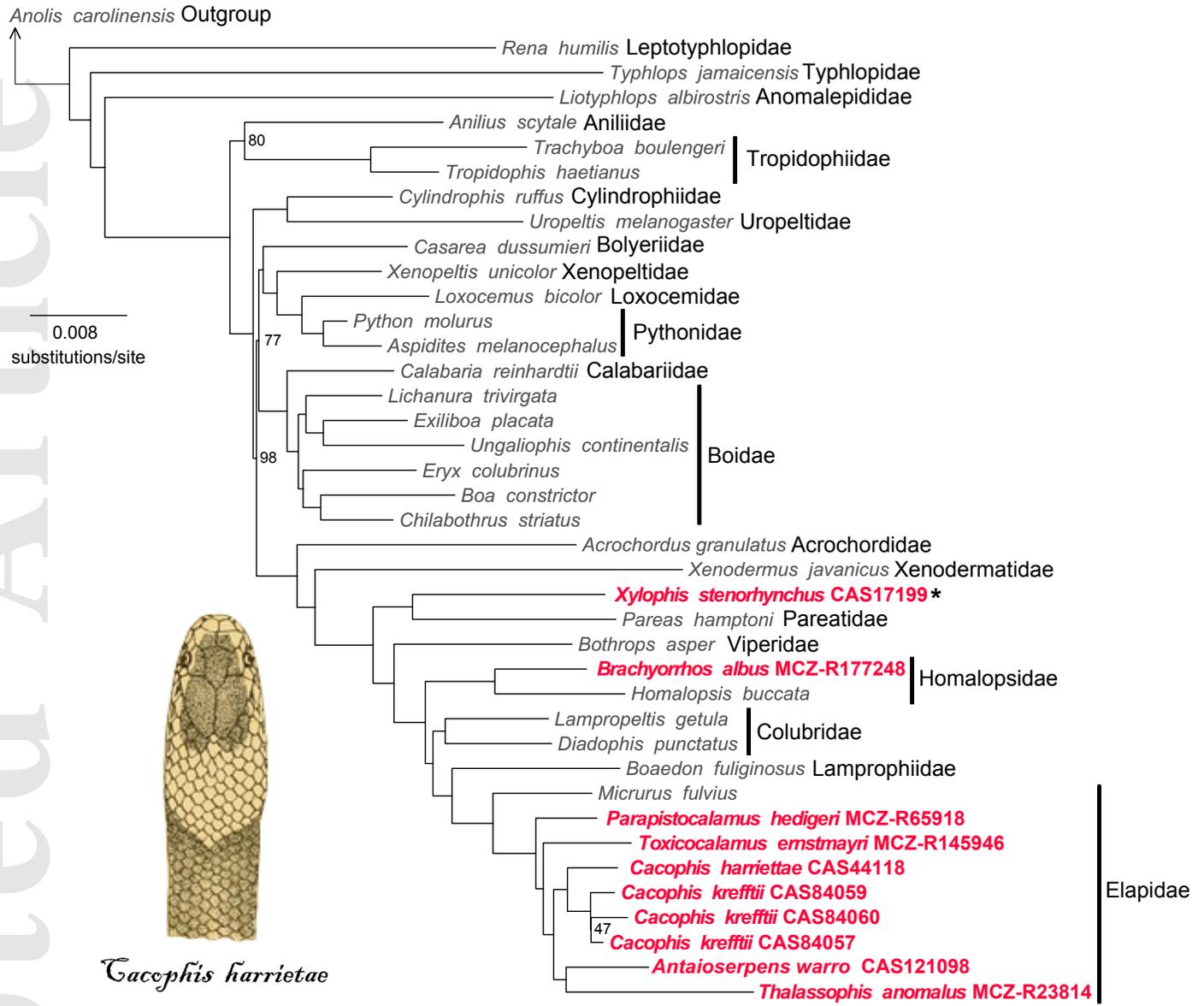
S.R. performed laboratory work and analyses. Both authors contributed intellectually to the study design and writing.

Figure 1. Maximum-likelihood tree of 3,544 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 (<http://dx.doi.org/10.5962/bhl.title.4679>).

Table 1. Summary of results for fluid-preserved specimens sequenced for this study; \bar{x} UCE=mean UCE length; *inserts not detected during library preparation.

Table 1. Summary of results for fluid-preserved specimens sequenced for this study; \bar{x} UCE=mean UCE length; *inserts not detected during library preparation.

Sample	Family	Collected	ng/ μ L DNA	# Reads	# Contigs	# UCEs	\bar{x} UCE
CAS44118 <i>Cacophis harriettae</i>	Elapidae	1913	1.53	38,470,164	73,530	2,855	152
CAS84059 <i>Cacophis krefftii</i>	Elapidae	1948	1.72	34,568,255	77,382	2,520	145
CAS84060 <i>Cacophis krefftii</i>	Elapidae	1948	0.89	25,987,774	33,899	2,664	138
CAS84057 <i>Cacophis krefftii</i>	Elapidae	1948	4.97	26,933,634	71,235	2,060	224
MCZ R-65918 <i>Parapistocalamus hedigeri</i>	Elapidae	1961	2.04	29,472,343	126,674	3,443	294
CAS121098 <i>Antaioserpens warro</i>	Elapidae	1968	11.5	30,952,427	60,594	2,818	150
MCZ R-23814 <i>Thalassophis anomalus</i>	Elapidae	<1958	1.33	28,727,457	12,053	995	123
MCZ R-145946 <i>Toxicocalamus ernstmayri</i>	Elapidae	1969	0.71	30,710,619	5,793	359	125
MCZ R-177248 <i>Brachyorrhos albus</i>	Homolopsidae	1906	1.82	32,315,527	60,863	2,918	155
CAS17199 <i>Xylophis stenorhynchus</i>	Uncertain	1878–1911	2.00	32,236,948	23,499	2,546	135
CAS15967 <i>Lycodon travancoricus</i> *	Colubridae	1941	0.69	10,778,312	362	Failed	N/A
LSUMNS-41802 <i>Stegonotus muelleri</i> *	Colubridae	1986	1.76	1,707,518	12	Failed	N/A
CAS244490 <i>Uropeltis nitidus</i> *	Uropeltidae	1990	1.24	2,653,679	87	Failed	N/A
Negative Control*	N/A	N/A	“too low”	23,614,868	17	Failed	N/A



Cacophis harrietae