Simultaneously Collecting Coding and Noncoding Phylogenomic Data Using Homemade Full-length cDNA Probes, Tested by Resolving the High-level Relationships of Colubridae

Jiaxuan Li[1], Dan Liang[1], and Peng Zhang[1]

[1]Sun Yat-Sen University

November 16, 2020

Abstract

Resolving intractable phylogenetic relationships often requires simultaneously analyzing a large number of coding and noncoding orthologous loci. To gather both coding and noncoding data, traditional sequence capture methods require custom-designed commercial probes. Here, we develop a cost-effective sequence capture method based on homemade probes, to capture thousands of coding and noncoding orthologous loci simultaneously, suitable for all organisms. This approach, called "FLc-Capture", synthesizes biotinylated full-length cDNAs from mRNA as capture probes, eliminates the need for costly commercial probe design and synthesis. To demonstrate the utility of FLc-Capture, we prepared full-length cDNA probes from mRNA extracted from a common colubrid snake. We performed capture experiments with these homemade cDNA probes and successfully obtained thousands of coding and noncoding genomic loci from 24 Colubridae species and 12 distantly related snake species of other families. The average capture specificity of FLc-Capture across all test snake species is 35%, similar to the previously published EecSeq method. We constructed two phylogenomic data sets, one including 1,075 coding loci (˜817,000 bp) and another including 1,948 noncoding loci (˜1,114,000 bp), to study the phylogeny of Colubridae. Both data sets yielded highly similar and well-resolved trees, with 85% of nodes having > 95% bootstrap support. Our experimental tests indicated that FLc-Capture is a flexible, fast, and cost-effective sequence capture approach for simultaneously gathering coding and noncoding phylogenomic data sets to study intractable phylogenetic questions. We anticipate that this method can provide a new data collection tool for the evolutionary biologists working in the era of high-throughput sequencing.

Title: Simultaneously Collecting Coding and Noncoding Phylogenomic Data Using Homemade Full-length cDNA Probes, Tested by Resolving the High-level Relationships of Colubridae

Authors: JiaXuan Li[1], Dan Liang[1], Peng Zhang[1,*]

Address:

[1]State Key Laboratory of Biocontrol, College of Ecology and Evolution, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

***Corresponding author:**

Peng Zhang. #434, School of Life Sciences, Sun Yat-Sen University, Higher Education Mega Center, Guangzhou 510006, China. Tel: 86-20-39332782; Email: zhangp35@mail.sysu.edu.cn

Running heads: Sequence capture with full-length cDNAs

Abstract

Resolving intractable phylogenetic relationships often requires simultaneously analyzing a large number of coding and noncoding orthologous loci. To gather both coding and noncoding data, traditional sequence

capture methods require custom-designed commercial probes. Here, we develop a cost-effective sequence capture method based on homemade probes, to capture thousands of coding and noncoding orthologous loci simultaneously, suitable for all organisms. This approach, called "FLc-Capture", synthesizes biotinylated full-length cDNAs from mRNA as capture probes, eliminates the need for costly commercial probe design and synthesis. To demonstrate the utility of FLc-Capture, we prepared full-length cDNA probes from mRNA extracted from a common colubrid snake. We performed capture experiments with these homemade cDNA probes and successfully obtained thousands of coding and noncoding genomic loci from 24 Colubridae species and 12 distantly related snake species of other families. The average capture specificity of FLc-Capture across all test snake species is 35%, similar to the previously published EecSeq method. We constructed two phylogenomic data sets, one including 1,075 coding loci (˜817,000 bp) and another including 1,948 noncoding loci (˜1,114,000 bp), to study the phylogeny of Colubridae. Both data sets yielded highly similar and well-resolved trees, with 85% of nodes having $> 95\%$ bootstrap support. Our experimental tests indicated that FLc-Capture is a flexible, fast, and cost-effective sequence capture approach for simultaneously gathering coding and noncoding phylogenomic data sets to study intractable phylogenetic questions. We anticipate that this method can provide a new data collection tool for the evolutionary biologists working in the era of high-throughput sequencing.

Keywords: high-throughput sequencing, sequence capture, transcriptome, snake, phylogeny

Introduction

Resolving difficult phylogenetic questions usually requires genome-scale data. However, large data sets do not necessarily lead to correct results because accurate phylogenetic inference relies on the correct evolution model. A subtle model violation may be sufficient to mislead phylogenetic inference when data is big (Hahn & Nakhleh 2016). Therefore, when addressing difficult phylogenetic questions, to avoid highly supported but wrong phylogenetic inference, in addition to careful model selection and data refinement, it is often desirable to analyze several independent phylogenomic data sets for consistency. In genomes, coding sequences and noncoding sequences have different evolutionary characteristics and are relatively independent data sources. As a result, it is becoming increasingly popular to simultaneously analyze both coding and noncoding genomic data in many recent phylogenomic studies (Chen et al., 2017; Jarvis et al., 2014; Reddy et al., 2017).

Whole-genome shotgun (WGS) sequencing is the simplest way to obtain coding and noncoding phylogenomic data simultaneously, but it is still cost-prohibitive to sequence dozens or hundreds of full genomes despite the rapid progress of sequencing technology. In fact, because phylogenomic studies do not need fully-assembled genomes but only phylogenetically informative loci, low-coverage WGS sequencing is generally sufficient to meet the basic requirements for phylogenomic studies. Until now, there are three main approaches for extracting phylogenetically informative loci from low-coverage WGS data. The first approach, called "automated Target Restricted Assembly Method (aTRAM)," assembles WGS data into predefined targeted regions by selecting reads with iterative BLAST searches (Allen et al., 2017). This method has been demonstrated to be able to extract over a thousand loci from 5-10× coverage WGS data of sucking lice (genome size 100-150Mbp). However, this method is more suitable for species with small genomes, since iterative BLAST searches will be too computationally intensive with large datasets. The second approach directly extracts phylogenomic data (coding and noncoding) from low-coverage WGS data by assembling entire genomes (Allio et al., 2019; Hughes & Teeling, 2018; Zhang et al., 2019). Zhang et al. (2019) showed that, for species with small genomes (0.1-1 G), 10-20× coverage WGS data is sufficient to extract hundreds to thousands of phylogenetic loci. However, this method is still not suitable for organisms with large genomes (> 1 G) because *de novo* genome assembly is highly difficult under this situation. The third approach does not extract phylogenomic loci by assembling genomes but extract single nucleotide polymorphisms (SNPs) from low-coverage WGS data by mapping reads to reference genomes. Olofsson et al. (2018) used this strategy to study the phylogeny of the olives that have relatively large genomes (˜ 1.5 G). The shortcoming of this method is that it requires annotated reference genomes and tends to perform relatively poorly across highly divergent lineages. Currently, although low-coverage WGS sequencing has shown great promise in constructing phylogenomic data sets, it is still somewhat challenging to apply it in organisms with large

genomes.

Two sequencing methods perform better than genome shotgun sequencing in generating phylogenomic data from large genome species: transcriptome sequencing (Morozova et al., 2009; Wang et al., 2009) and sequence capture (Faircloth et al., 2012; Glenn et al., 2016; Jones et al., 2016; Lemmon et al., 2012; Lemmon & Lemmon 2013). The target of transcriptome sequencing is expressed mRNAs whose size does not vary significantly, no matter how large the genome size is. Because mRNAs contain both open reading frames (ORFs) and untranslated regions (3' UTR and 5' UTR), transcriptome sequencing can enable researchers to obtain a large amount of coding and noncoding sequences simultaneously (Garrison et al., 2016; Misof et al., 2014; Oakley et al., 2012). However, transcriptome sequencing requires fresh or properly stored tissues to provide high-quality RNA, which often limits the number of taxa included in such phylogenomic studies (Lemmon & Lemmon 2013; McCormack et al., 2013). Sequence capture uses biotinylated probes to enrich the target regions of the genome of interest selectively. It allows researchers to attain higher sequencing depth over a predefined subset of the genome for a given cost, particularly helpful to species with large genomes (Mccartney-melstad et al., 2016). An advantage of sequence capture is that it does not require high-quality DNA samples and can handle highly degraded DNA extracted from old museum specimens (e.g., Blaimer et al., 2016; Guschanski et al., 2013). This property can greatly increase the sampling number of taxa in a phylogenomic study. Moreover, sequence capture is also very flexible. Many capture methods have been developed for various purposes, such as ultra-conserved element (UCE) sequencing (Faircloth et al., 2012) for collecting noncoding sequences, anchored hybrid enrichment (AHE; Lemmon et al., 2012) and exon capture (Albert et al., 2007; Bi et al., 2012; Ng et al., 2009) for collecting coding sequences, and a combination of AHE and UCE for collecting both coding and noncoding sequences simultaneously (Singhal et al., 2017). However, most current sequence capture methods require the researcher to have prior genomic information for probe design and then to synthesize the probes through commercial companies. For nonmodel species, the probe design is often difficult due to a lack of genome information. Also, the cost of using commercial probes will be high when a research project has hundreds of samples or more, probably reaching several thousands of dollars.

Recently, Puritz and Lotterhos (2017) demonstrated that cDNA fragments could be used as capture probes to capture coding sequences from genomes. Using cDNAs by reverse transcription from mRNAs as probes to sequence capture genomes can avoid using commercial probes, thus greatly reduce the cost of experiments. The method of Puritz and Lotterhos (EecSeq) only focuses on capturing coding regions, and the experiment design and bioinformatic pipeline all revolve around how to obtain exonic SNPs. In fact, full-length cDNA sequences consist of coding ORFs and noncoding UTRs. If both ORFs and UTRs are considered in the cDNA probe preparation, genomic DNA of both coding and noncoding regions can be captured and sequenced simultaneously. The direct use of full-length cDNAs as probes for sequence capture can produce transcriptome-level data and skips the step of probe design, which is particularly suitable for nonmodel organisms lacking of genomic information. Moreover, it can allow investigators for simultaneously obtaining coding and noncoding phylogenomic data, and thus will be helpful for studying difficult phylogenetic questions.

In this study, we present a novel sequence capture method based on homemade probes, called "full-length cDNA capture sequencing" (FLc-Capture). It is a universal, flexible, and cost-effective sequence capture method that works for all organism groups. The most distinctive feature of this method is to use the SMART technology (Clontech Inc.) to synthesize full-length cDNAs and then created biotinylated probes from cDNAs. The specially designed bioinformatics analysis scheme enables users to extract a large number of genomic loci (both coding and noncoding) from the capture data without any genome knowledge of the taxa been investigated. To demonstrate the utility of the FLc-Capture method, we used it to study the phylogeny of the family Colubridae (Serpentes: Caenophidia), a rapid radiation lineage with large genomes (~2 G). We successfully obtained hundreds to thousands of coding and noncoding genomic loci from dozens of colubrid and distantly related outgroup snake species from the FLc-Capture data. These coding and noncoding phylogenomic data were able to reconstruct a robust phylogeny of Colubridae and addressed the long-debated relationships among subfamilies. We anticipate the method presented in this study can provide a new high-throughput sequencing approach for studies seeking to resolve difficult phylogenetic questions.

Materials and methods

Experimental overview

FLc-Capture sequencing is designed with two specific goals: (a) to eliminate the need for expensive capture probe synthesis and (b) to obtain genome-scale data of both coding and noncoding regions simultaneously. To this end, researchers should choose one common species readily available from their taxonomic group of interest to extract high-quality RNA and synthesize full-length cDNAs. These full-length cDNAs are then amplified by biotinylated primers to generate the homemade capture probes for subsequent sequence capture experiments. The steps for probe preparation and sequence capture are visualized in Figure 1.

Taxon sampling, DNA extraction, and library preparation

The family Colubridae is a widespread snake group with relatively large genomes (˜2 G). It is a rapid radiation lineage and the subfamily relationships within this family has historically proven difficult to resolve. Therefore, the family Colubridae is a good case study for demonstrating the utility of the FLc-Capture method for generating genome-scale coding and noncoding data sets to resolve difficult phylogenetic questions. Based on the latest phylogenies of Colubridae (Li et al., 2020), we sampled 24 colubrid species representing 23 genera, seven subfamilies (Dipsadinae, Pseudoxenodontinae, Natricinae, Sibynophiinae, Calamariinae, Ahaetuliinae, Colubrinae) and 12 distantly related outgroup snake species from five families (Xenodermatidae, Pareatidae, Viperidae, Elapidae, Homalopsidae). The detailed information of these samples, such as taxonomy, collection locality, and voucher, is given in Table 1.

Total genomic DNA was extracted from ethanol-preserved liver or muscle tissue of each sample using a TIANamp Genomic DNA Kit (Tiangen, Beijing). All DNA extracts were measured using an ND-2000 spectrophotometer and diluted to a concentration of 50 ng/μl with 1x TE. For each sample, 250 ng of its genomic DNA was randomly fragmented to 200-400 bp using NEBNext dsDNA Fragmentase (NEB). The fragmented DNA was purification with AMpure XP beads (Beckman Coulter). The purified DNA was used for Illumina library preparation with NEBNext Ultra DNA Library Prep Kit (New England Biolabs) (Fig. 1a). Each sample was labeled with a unique 8-bp index sequence. Three or four libraries were mixed into a pooled library in equal concentrations for subsequent hybridization capture.

Low-coverage whole-genome sequencing (WGS) and data analysis

In order to test whether it is possible to extract phylogenomic data from low-coverage WGS data for organisms with large genomes (> 1 G), we selected two colubrid species (*Amphiesma stolatum* and *Heterodon platirhinos*) as the test samples. We sequenced their DNA libraries on an Illumina HiSeq X-ten lane using paired-end 150-bp mode. We obtained ˜40 G sequence data per sample corresponding to a sequencing depth of about 20×. The genome sizes of the two colubrid species were estimated from the WGS data by using Jellyfish version 2.3.0 (Guillaume & Carl 2011). As a comparison, we also downloaded the WGS data of four insects with relatively small genomes from NCBI: *Pediculus humanus* (108M), *Phoebis sennae* (287M), *Zootermopsis nevadensis* (485M), and*Halyomorpha halys* (996M). The WGS data resources of these four insect species are given in Appendix S2.

We adopted the method of Zhang et al. (2019) to directly extract phylogenetic loci from the WGS data through *de novo* genome assembling. The raw reads were first filtered to remove adapter sequences and low-quality nucleotides. The filtered reads of each species were assembled into scaffolds using the SPAdes version 3.8.1 genome assembler, using an auto K-mer mode (–cov-cutoff auto). We downloaded a vertebrate core database comprising 2,586 genes (a total length of ˜3,280 K) and an insect core database composed of 1,367 genes (a total length of ˜1,285 K) from the OrthoDB database as targeted gene clusters and used BUSCO v3.0.2 (Waterhouse et al., 2018) to extract orthologous sequences from the genome scaffolds. The genome assembly and gene extraction process were repeated at a sequencing depth of 1×, 5×, 10×, 20×, respectively. To compare the effect of extract phylogenomic loci from low-coverage WGS data of small and large genome species, we calculated the proportion of complete/fragmented genes (gene recovery rates) and the proportion of sites across the total length of reference genes.

Full-length cDNA probe preparation

The workflow of producing full-length cDNA probes is illustrated in Figure 1b. We extracted high-quality RNA from fresh liver tissue of *Ptyas korros* , a common colubrid species, using the RNA prep Pure Tissue Kit (Tiangen, Beijing). The quality of total RNA was assayed using an Agilent Bioanalyzer 2100 and the RNA integrity number (RIN) was greater than seven. The synthesis of full-length cDNAs was performed using SMARTer PCR cDNA Synthesis Kit (Clontech Inc.) based on the manufacturer's protocols. 4.5 μl of cDNA synthesis mixture containing 1 μg total RNA and 2.67 μM universal tail primer I (included in the kit) was incubated for 3 min at 72 °C and 2 min at 42 °C. The volume was then adjusted to 10 μl with the following reagents: 1× First-Strand Buffer, 2.5 mM DTT, 1mM dNTP Mix, 1.2 μM SMARTer II A Oligonucleotide (included in the kit), 0.25 μl RNase Inhibitor and 10 U SMARTScribe Reverse Transcriptase. Tailing, template switching, and extension were carried out for 90 min at 42°C. The reaction was terminated for 10 minutes at 72°C. The first-strand synthesis product was diluted with 40 μl 1× TE, and used as templates for cDNA amplification. The SMART technology can ensure the synthesized cDNAs are in full-length, and both ends of the synthesized cDNAs contain universal tail sequences.

The synthesized first-strand cDNAs were amplified with a 5'-biotinylated primer (universal tail primer II; see Appendix S1) to generate full-length cDNA probes (Fig. 1b). The PCR reaction mixture contained 1.25 U HiFi Taq DNA Polymerase (TransGen, Beijing), 1× HiFi PCR buffer, 0.2 mM dNTPs, 0.24 μM of universal tail primer II and one μl of diluted first-strand cDNA synthesis product in a total volume of 100 μl. The thermal cycling program is as follows: an initial denaturation for 1 min at 95 °C followed by 19 cycles of 15 s at 95°C, 30 s at 65°C, 6 min at 68 °C. The amplification product was purified by AMpure XP beads and checked on a 1.2% TAE agarose gel. After that, the purified amplification product (full-length cDNA probes) was measured using an ND-2000 spectrophotometer and diluted to a concentration of 50 ng/μl with 1× TE. We did not normalize our full-length cDNA probes to decrease the abundance of highly expressed cDNA, but directly used them for subsequent capture experiments.

Hybridization capture and sequencing

Hybridization capture of the FLc-Capture method is similar to that of previously published capture protocols (Li et al., 2013) with some modifications. For each capture reaction, 500 ng of DNA libraries and 200 ng of cDNA probes are used. In order to increase the capture efficiency, we used a touch-down hybridization program: after denaturation at 94°C for five minutes, the hybridization starts from 65°C decreased by 5°C every 6 hr and ended at 45°C, for a total duration of 30 hours. The hybridized DNA fragments are captured with streptavidin magnetic beads (Dynabeads MyOne bead, Life Technologies). The beads are washed to remove unhybridized DNAs and eluted in 30 μl of 1× TE to release the captured DNA fragments. The captured libraries are amplified with Illumina P5 and P7 universal primers. Finally, the captured libraries of different capture experiments are pooled in equal concentrations and sequenced on three lanes of Illumina HiSeq X-ten with paired-end 150-bp mode (˜400 G of total data). The workflow of the hybridization capture experiment is shown in Figure 1c.

Bioinformatic workflow

Building up reference ORF and UTR sets

The cDNA probes need to be sequenced to provide reference sequence sets for subsequently captured data analysis. To this end, 100 ng of the cDNA probes were used to construct a sequencing library following the same procedure as genomic library preparation. The probe library was sequenced on an Illumina HiSeq X-ten sequencer using paired-end 150-bp mode. The raw reads were first filtered to remove adapter sequences and low-quality nucleotides by using Trimmomatic version 0.36 (Bolger et al., 2014) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Clean reads were assembled into transcripts using TRINITY r20140717 with default parameters (Grabherr et al., 2011). The obtained transcripts were filtered with CD-HIT-EST version 4.6.5 (Fu et al., 2012) to reduce redundancy (95% similarity cutoff). The sequencing depths for filtered transcripts were calculated by SAMtools version 1.4.1 (Li et al., 2009). Only transcripts of average sequencing depth [?] 5x, length [?] 200 bp were retained. TransDecoder, a program in

the TRINITY package, was used to determine the open reading frame (ORF) for each transcript. Based on the position of the ORF, each transcript can be annotated to 5' UTR (untranslated region), coding region, and 3' UTR. The translated protein sequences of the predicted ORFs were searched by BLASTP (NCBI BLAST+ version 2.6.0, Boratyn et al., 2013) against the human proteomes with an e-value threshold of 1E-10. Only transcripts that have BLASTP hits were retained to focus on known vertebrate transcripts. Finally, all ORFs of length > 300 bp and UTRs of length > 100 bp are extracted using a custom Python script to build two reference sets (ORF and UTR) for the subsequent captured data analysis (Fig. 2a).

Sequence capture data analysis

Sequence capture reads were sorted into each species by the 8-bp species index. The raw reads of each species were filtered to remove adapter sequences and low-quality nucleotides. To accelerate assembly and save computing resources, we down-sampled reads over high-depth areas at an average depth of 20x by normalization using BBNORM.SH (BBTools, Bushnell 2014). The normalized read data were then *de novo* assembled using the SPAdes version 3.8.1 genome assembler (Bankevich et al., 2012), using an auto K-mer mode (–cov-cutoff auto). Only contigs longer than 200 bp were retained. The retained contigs were further filtered with CD-HIT-EST to reduce redundancy (95% similarity cutoff). Finally, the contigs were BLASTed against the reference ORF and UTR sets to remove non-target sequences, thus reduce the computational intensity of the subsequent analyses.

Extract ORF (coding) sequences —Each of the reference ORF sequences is typically composed of multiple exons. In the genome, these exons are interrupted by introns. To extract the coding sequence corresponding to the entire reference ORF, we need to identify correct exons from assembled contigs and stitch these identified exons into a complete ORF sequence. We adopted a bioinformatics pipeline called "exon mapping" to fulfill this purpose. For each reference ORF, we used EXONERATE version 2.2.0 (Slater & Birney 2005) to locate its relevant exons from the filtered contigs based on its translated protein sequence. The identified exons are then mapped onto the reference ORF to make an orthologous ORF sequence (missing regions were filled with N), based on the coordinate information from the EXONERATE searches. This exon mapping strategy makes the obtained coding sequences from each sample have the same length to the reference ORF, which reduces the difficulty of sequence aligning. The workflow of extracting the orthologous ORF sequence is illustrated in Figure 2b. The whole bioinformatics procedure is fulfilled by an in-house Python script "Extracting coding sequences.py."

Extract UTR (noncoding) sequences — Unlike the ORF sequences are fractured in the genome, UTR sequences are commonly continuous in the genome and most likely located within single assembled contigs. Based on the reference UTRs, we used a mutual best-hit (MBH) strategy (program = BLASTN, e-value < 1e-10, identity > 70%) to extract UTR orthologous groups (OGs) from all samples. The 1:1 orthology is confirmed if one contig of a sampled species and one reference UTR sequence find each other as the best hit in the bidirectional BLAST. Within each UTR OGs, the extracted sequences are normally different in length, which may make them difficult to be aligned. We used a previously published Python script (Li et al., 2019) to determine the optimal aligning region for all sequences within a UTR OG. In brief, the script uses the mutual-BLAST results to determine the relative position of each sequence to the reference sequence and searches for the optimal upstream and downstream boundaries to trim the sequences. Here, we demand that at least 40% of species have data at both the upstream and downstream boundaries. The workflow of extracting the orthologous UTR sequence is illustrated in Figure 2c.

Phylogenetic analysis

All extracted ORF or UTR sequences with mean sequencing depth < 5 were discarded. The ORF OGs (orthologous groups) were already alignments and do not need to be aligned (see Figure 2b). We used Gblocks version 0.91 (Castresana 2000) under codon mode (-t = c) and half gaps allowed (-b5 = h) to refine these ORF alignments. The UTR OGs were aligned using the program SATe-II version 2.2.2 (Liu et al., 2012) with the "-auto" option. The resulting alignments are also refined using Gblocks with half gaps allowed (-b5 = h). To further reduce possible errors in orthology assignment and alignment, we reconstructed a maximum-

likelihood (ML) tree for every ORF or UTR alignment using RAxML v8.2.2 (Stamatakis 2014) under GTR + GAMMA model. If a tree contained extremely long branches that accounted for > 50% of the total tree length, the corresponding sequences were removed from the alignment. We kept only the alignments that are longer than 100 bp and contain more than seventeen species (at least ten colubrid species and one outgroup species). Gene trees for each refined ORF or UTR alignment were constructed using RAxML with the GTR + GAMMA model and 100 bootstrapping replicates (-f a option). To investigate the degree of incongruence among gene trees, we calculated the pairwise Robinson-Foulds distances (Robinson & Foulds 1981) between gene trees using the python script of Gori et al. (2016). The tree-to-tree distances were visualized using multidimensional scaling (MDS) in R (Hillis et al., 2005).

The filtered ORF and UTR alignments were combined into two concatenated supermatrix, respectively, and subjected to ML analyses using RAxML. Because our data sets included thousands of genes, partitioning our data sets by genes was not possible. Therefore, the ORF data set was partitioned by three codon positions, and the UTR data set was unpartitioned. ML analyses for both data sets were performed with the GTR + GAMMA model. Branch support for the resulting phylogeny was evaluated with 500 rapid bootstrapping replicates implemented in RAxML.

Results

The performance of low-coverage WGS sequencing in species with large genomes

Our WGS sequencing produced 261 and 269 million clean 150-bp paired-end reads for the two colubrid species (*Amphiesma stolatum* and*Heterodon platirhinos*), respectively (˜40 G sequence data per sample). The genome sizes of *Amphiesma stolatum*and *Heterodon platirhinos* estimated by Jellyfish are 1,348 Mbp and 1,686 Mbp, respectively. We compared the performance of the method of Zhang et al. (2019) for extracting phylogenetic loci from low-coverage WGS data in these two colubrid species with large genomes and other four insects with relatively small genomes (Fig. 3). Detailed statistics of genome assembly and locus extraction are summarized in Appendix S3.

At levels of sequencing depth 5x and less, few genes can be extracted from genome assemblies of both small and large genome species (Fig. 3). At a level of 10x sequencing depth, the recovery rate of genes (complete + fragmented) for small genome species were significantly improved, with a range of 68%˜95%. If the recovered rate was calculated by DNA sites, 41%˜91% of DNA sites were recovered from small genome species. In contrast, the gene recovery rates of the two snake species were still low at 10x sequencing depth, less than 10%, and no more than 5% DNA sites were recovered from the two snake species (Fig. 3). When sequencing depth increased to 20x, the gene recovery rate of small genome species increased to 85%˜98%, and the site recovery rate increased to 63%˜99%, which is consistent with that reported by Zhang et al. (2019). In contrast, at a level of 20x sequencing depth, the gene recovery rate of the two snake species improved to 42%˜59%, but only less than 35% DNA sites were recovered because most of the recovered genes were not complete (Fig. 3). These results suggest that extracting phylogenomic data from low-coverage WGS data in large genome species is much more difficult than in small genome species.

cDNA sequencing and reference ORF and UTR sets

After data quality control, we obtained a total of 45 million clean 150-bp paired-end reads (˜ 6.8 G data) from cDNA sequencing. We assembled these reads using TRINITY and obtained a total of 31,841 cDNA sequences. After filtering by redundancy, sequencing depth ([?] 5x), and length ([?] 200 bp), a total of 26,409 cDNA sequences were retained. Among these sequences, 14,785 contained ORFs with a length greater than 300 bp. Of the protein sequences of these predicted ORFs, 8,429 have orthologous human proteins. Among these 8,429 cDNA sequences, 4000 (47.6%) have both 5' and 3' untranslated regions (UTRs), 3909 (46.3%) contain 3' UTR, 273 (3.2%) contain 5' UTR and 247 (2.9%) only contain coding sequences. From these 8,429 cDNA sequences, we extracted a total number of 8,429 ORFs and 10,665 UTRs (3391 of 5' UTR and 7274 of 3' UTR) as reference sets. The length of reference ORFs ranged from 300 bp to 14,325 bp, with an average length of 1,225 bp; the length of the reference UTRs ranged from 100 bp to 7,694 bp, with an average length of 716 bp. The total length of reference ORFs and UTRs are ˜10,300 K and ˜7,600 K, respectively.

FLc-Capture sequencing results

For the 24 colubrid snake and 12 outgroup snake samples, we obtained a total of 2,577 million quality-filtered 150-bp paired-end reads (˜387 G data) using the FLc-Capture sequencing, ˜10.7 G data per sample (range: 5.5 G - 20.3 G). The numbers of assembled contigs for different samples ranged from 158,424 to 872,008. These contigs were searched against the reference ORF and UTR sets to generate orthologous ORF and UTR sequences for each sample (see Methods). On average, for the 8,429 ORF and 10,665 UTR targets, we could obtain about 6,000 ORF and 5,900 UTR sequences from the ingroup species, and about 5,000 ORF and 4,600 UTR sequences from the outgroup species that are more distantly related to the probe species (Table 2). When the recovery rate is calculated by genes, the recovery rates of ORF (67%) are generally higher than those of UTR (51%) (Fig. 4a). However, when the recovery rate is calculated by nucleotides, the recovery rates of UTR (44%) are, on the contrary, higher than those of ORF (32%) (Fig. 4b), indicating that the recovered sequences of UTRs are more integrated than those of ORFs across reference sequences.

We explored capture specificity, the percentage of reads that can be aligned to target sequences (on-target) across all samples in our experiment (Table 2). The average on-target value of UTRs of all samples (23%) is higher than that of ORFs (12%) (Table 2), indicating that UTR sequences are more easily captured, possibly because they were physically continuous in genomes. However, the fluctuation of the on-target value of UTRs among samples (square deviation = 6.19, n = 36) is higher than that of ORFs (square deviation = 3.44, n = 36), suggesting that the capture efficiency of UTR sequences may be more sensitive to genetic distance. We found that the capture specificity of both ORF and UTR from a sample is negatively related to its genetic distance to the probe species, and the regression slope of UTR (coefficient = -204.39) is much smaller than that of ORF (coefficient = -64.04) (Fig. 5a). These results showed that the capture specificity of UTR decreases more rapidly than that of ORF with the increase of genetic distance from the probe species, in line with that noncoding sequences (UTR) evolve more rapidly than coding sequences (ORF).

In our experiment, we did not perform normalization treatment for our full-length cDNA probes to reduce the abundance of highly expressed transcripts. We thus want to know whether our capture experiment will dominate by those highly abundant cDNA probes. We found that, to some extent, the mean capture depth of each target does appear to relate to the abundance of its probe (Fig. 5b). However, the relationship between the capture depth of one target and the abundance of its probe is not absolute because there are also many ORF and UTR sequences with high capture depth (> 1000) while their corresponding probes are not abundant (< 100) (Fig. 5b). These results indicated that unnormalized cDNA probes do not significantly affect the ability of our method to sequence capture those target sequences corresponding to low-expressed transcripts.

Detailed statistics of the sequencing, contig assembly, gene recovery, nucleotide recovery, and capture specificity for each sample are summarized in Table 2. On the whole, the FLc-Capture method can simultaneously obtain thousands of coding and noncoding sequences from both the ingroup and outgroup samples, which indicates that our experimental design using homemade full-length cDNA probes to capture both of genomic coding and noncoding regions is successful.

The ORF and UTR data sets and the Colubridae phylogeny

From the FLc-Capture sequencing data, we extracted a total of 1,075 ORFs and 1,948 UTRs that have passed our filtering criteria (mean sequencing depth > 5 and containing at least seventeen taxa) and can be used for phylogenetic analysis. A summary of data characteristics for ORFs and UTRs, including length, taxa occupancy, GC content (the average GC content at the third codon position for each ORF and average full GC content for each UTR), percentage of missing data, is given in Appendix S4. The lengths of ORF alignments range from 168 to 9,063 bp (average = 760 bp) and the lengths of UTR alignments range from 107 to 3,011 bp (average = 572 bp) (Fig. 6a). In general, the UTRs have lower mean GC content and lower GC content variation (among genes and among species) than the ORFs (Fig. 6b). The UTR alignments have a higher pairwise distance than the ORF alignments, consistent with the expectation that noncoding sequences evolve more rapidly than coding sequences (Fig. 6c). Multidimensional scaling plots of the RF-

distance among genes (Fig. 6d) indicated that the ORF gene trees were more similar to each other compared with the UTR gene trees, but phylogenetic signals among ORFs or UTRs are overall rather congruent.

The concatenated supermatrix of ORFs is 817,164 bp in length and 72.8% complete by characters, while the concatenated supermatrix of UTRs is 1,114,278 bp in length and 78.2% complete by characters. The ML trees inferred from the ORF and UTR data sets are identical and well resolved, with at least 85% of nodes having > 95% bootstrap (BS) support (Fig. 7). The backbone phylogeny among the snake families sampled in this study is congruent with that reported by many previous studies (e.g., Pyron et al., 2014; Zheng & Wiens 2015). Within the family Colubridae, we recognized three major clades: (A) Dipsadinae + Pseudoxenodontinae, (B) Natricinae, and (C) Sibynophiinae + (Calamariinae + (Ahaetuliinae+ Colubrinae)). These three clades were repeatedly found in previous studies, but the relationship among them was not well-supported and different in those studies (Burbrink et al., 2020; Li et al., 2020; Pyron et al., 2014; Wiens et al., 2012). For example, Wiens et al. (2012) used 44 nuclear genes but did not resolve the relationships among these three clades. Both Pyron et al. (2014) and Burbrink et al. (2020) used hundreds of AHE loci and resolved the relationship as (C,(A,B)) (former: weakly supported; BS = 65%, later: posterior probabilities = 0.88). Li et al. (2020) used 96 mitochondrial and nuclear genes but found the relationship is (A,(B,C)) (weakly supported; BS = 46%). Different from the previous results, both our ORF and UTR data sets favored a relationship of (B,(A,C)), and this result is strongly supported by the ORF data set (BS=100%; Fig. 7).

Our phylogenomic analysis provided a highly resolved phylogeny for colubrid snakes, for the first time, based on extensive sampling of both genes and species. Thanks to the characteristics of the FLc-Capture method, we were able to simultaneously collect genome-scale coding and noncoding data to study the phylogeny of Colubridae. Although some nodes of our resulting phylogeny were not conclusively supported in all analyses, they received high support (ML bootstrap > 85%) from at least one type of data set, which shows the benefit of simultaneously using both coding and noncoding data sets for studying difficult phylogenetic questions.

Discussion

The originality of the FLc-Capture method

The first distinctive originality of the FLc-Capture method is to use the SMART technology, which is widely adopted in cDNA cloning researches, to synthesize cDNA probes. The SMART technology guarantees that each of the synthesized cDNA is in full-length, consisting of open reading frames (ORF) and untranslated regions (5' UTR and 3' UTR). This unique feature allows researchers to enrich coding and noncoding sequences from genomes simultaneously. In our demonstrating snake case, the final lengths of ORF and UTR datasets are 817 K and 1,114 K, respectively, relatively close, indicating that FLc-Capture can enrich both coding and noncoding sequences with similar efficiency.

Because cDNA sequences are discontinuous in genomes (interrupted by introns), the direct use of full-length cDNA probes to capture ORF and UTR regions from DNA libraries makes data post-processing more challenging. Therefore, another originality of FLc-Capture is its unique data processing strategy. Considering ORF sequences are fractured in genomes, and UTR sequences are usually continuous in genomes, FLc-Capture adopts two different ways to extract ORF and UTR sequences from capture data, respectively. For ORF, FLc-Capture first assembled reads to contigs, identified exons from contigs, and then mapped identified exons onto the reference coding sequences. Our study demonstrated that this "exon mapping" strategy could extract coding sequences from genetically distant samples (~15% divergence in our study) without the need for highly similar reference sequences. For UTR, because they are most likely within a single assembled contig, FLc-Capture directly adopts a mutual best-hit (MBH) strategy to identify orthologous UTR sequences to the reference UTRs. Our case study showed that these two specially designed bioinformatics pipelines are effective, able to extract thousands of ORF and UTR sequences from both ingroup species and more distantly related outgroup species.

The merits of the FLc-Capture method

The first advantage of the FLc-Capture method is that it saves both cost and time compared to commercially synthesized probe sets. Previous transcriptome-based capture studies normally used transcriptome data to design capture probes and ordered those probe sets from commercial companies (e.g., Bi et al., 2012; Bragg et al., 2016; Portik et al., 2016; Quek et al., 2020). The whole process to synthesize a custom probe kit typically takes several weeks and cost $2,400-$5,000, depending on the supplier (Penalba et al., 2014). Although these commercial probes can be diluted for applying to more samples, the cost of using commercial probes would still be high when a research project has hundreds of samples or more, probably reaching several tens of thousands of dollars. In contrast, the primary initial investment for our method was the SMARTer PCR cDNA synthesis reagent (Clontech Inc.), which costs ˜$80 per reaction. Including the extraction of RNA, the probe preparation can be done within three days. In our lab, one SMARTer PCR cDNA synthesis reaction can produce up to 100 μg of full-length cDNA probes when input RNA is 1 μg. Such amount of cDNA probes is enough to handle at least 2,000 samples.

FLc-Capture has the merit of transcriptome sequencing while largely avoiding its shortcomings. Compared to transcriptome sequencing, FLc-Capture can produce transcriptome-level data (thousands of ORF and UTR sequences) with only DNA samples. Researchers just need to collect one common species of their taxonomic group of interest for RNA extraction, prepare biotinylated full-length cDNA probes from RNA, and then use these probes to capture target regions from their DNA libraries. Except for the probe species used for RNA extraction, FLc-Capture has no strict requirements on the DNA quality of other samples, so highly degraded DNA extracted from old museum specimens can also be analyzed, which can greatly increase the sampling number of taxa in a phylogenomic study.

In addition, FLc-Capture uses cDNA sequencing to provide reference sequences. This feature enables researchers to efficiently capture thousands of coding and noncoding sequences without knowing any genome knowledge of the taxa been investigated, especially suitable for nonmodel organisms. Unlike low-coverage WGS sequencing, which is more suitable for extracting phylogenomic data from small genome species (Fig. 3), FLc-Capture can efficiently collect coding and noncoding phylogenomic data not only from small genome species but also large genome species. These two features make the FLc-Capture method highly versatile and applicable for any organism groups.

Comparison with the EecSeq method

In general, our method has some similarity to the expressed exome capture sequencing (EecSeq) recently presented by Puritz and Lotterhos (2017), because our FLc-Capture method and EecSeq both prepare homemade cDNA probes from expressed mRNAs and use them to capture target sequences from genomic libraries. However, FLc-Capture also has two apparent differences to EecSeq. First, these two methods have a different focus on capturing genomic targets, so that the requirement of the cDNA probes are different. EecSeq only focuses on capturing coding sequences, so using fragmented cDNA probes is sufficient because fragmented cDNAs can essentially cover most coding regions. While FLc-Capture aims to capture both coding and noncoding sequences, so using the SMART technology to synthesize full-length cDNAs is critical because the full-length cDNA contains not only ORF (coding), but also UTR (noncoding) regions. Second, the application scenarios of these two methods are different. EecSeq generates genome-wide exome-derived SNP data, which is more suitable for identifying loci under selection at the population level to understand the genetic basis of adaptation. While FLc-Capture generates genome-scale sequence data, which is better suited to infer the evolutionary relationships for organisms across multiple phylogenetic scales. The unique feature of simultaneously collecting coding and noncoding sequences makes FLc-Capture advantageous in studying difficult phylogenetic questions (e.g., rapid radiation).

Application suggestions of the FLt-Capture method

In our FLc-Capture experiment, we did not perform cDNA normalization to decrease the abundance of highly expressed transcripts, so the cDNA probe pools skew towards highly expressed genes. However, our FLc-Capture result showed that the capture depth of the obtained ORF and UTR sequences is not all related to the abundance of their cDNA probe (Fig. 5b); using unnormalized cDNA probes is still able

to capture thousands of coding and noncoding loci. To obtain more uniform capture coverage across high and low-expressed transcripts, Puritz and Lotterhos (2017) used duplex-specific nuclease (DSN) treatment to prepare normalized cDNA probes for exome capture. They found that RNA sequencing coverage and exome sequencing coverage was still highly correlated (capture coverage will be higher for highly expressed genes) even using normalized cDNA probes. Therefore, it seems that the cDNA normalization is not an indispensable step. For projects focused on obtaining more phylogenetically informative loci for phylogenomic analysis, increasing the diversity of cDNA probes may be more effective than cDNA normalization. In such scenarios, we suggest to pool mRNAs extracted from multiple tissue types to create a high-diversity probe pool rather than using only liver mRNA as in our demonstration case.

Despite the demonstrated effectiveness of FLc-Capture from our snake case, one note should be considered before employing the method. Because the efficiency of sequence capture decreases with increased genetic distance between the probes and the targets, FLc-Capture might be less effective in large and highly divergent organism groups such as insects and other arthropods. In our demonstrating snake case, the maximal sequence difference between our probe species and outgroup species is about 15%, and we finally recovered 65% of the target ORF loci and 35% of the target UTR loci from these outgroup species, respectively. This threshold value (15% genetic difference) can be used as a starting point for other researchers to determine the phylogenetic depth of their FLc-Capture experiments. It has been shown that using DNA mixtures pooled from different representative species to prepare homemade probes is an effective strategy for sequence capture across large phylogenetic scales (Zhang et al., 2019). Therefore, if an investigator wants to apply the FLc-Capture method to a highly divergent organism group, it is possible to use several probe species that cover the entire phylogenetic span, mix their mRNAs to prepare a full-length cDNA probes, which can reduce the sequence divergence between probe and target. Of course, in such circumstances, the reference ORF and UTR sets should also be separated by different probe species, and the bioinformatics pipeline should be adjusted accordingly. This mixing strategy may allow for applying FLc-Capture across large phylogenetic scales but needs to be tested in the future.

Conclusion

In this study, we demonstrated that the FLc-Capture method could efficiently capture and enrich a large number of coding and noncoding loci for nonmodel organisms without any prior genome information. The direct use of homemade full-length cDNA probes in FLc-Capture can skip the expensive commercial probe design and synthesis, significantly reducing experimental cost. FLc-Capture can generate transcriptome-level data just based on DNA samples, which facilitates including more number of taxa in a phylogenomic study. In summary, FLc-Capture holds substantial promise in phylogenomic researches as a universally applicable and cost-effective sequence capture method of simultaneously collecting genome-level coding and noncoding orthologous loci for any organism.

Acknowledgments

We thank Song Huang, Peng Guo, and YingYong Wang for sharing valuable snake samples with us. This work was supported by National Natural Science Foundation of China (grants No. 31872205 and No. 32071611) to P. Zhang and National Natural Science Foundation of China (grants No. 31601847) to D. Liang.

References

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., ... Gibbs, R. A. (2007). Direct selection of human genomic loci by microarray hybridization. *Nature Methods* , **4** (11), 903–905. https://doi:10.1038/nmeth1111

Allen, J. M., Boyd, B., Nguyen, N. P., Vachaspati, P., Warnow, T., Huang, D. I., ... Johnson, K. P. (2017). Phylogenomics from whole genome sequences using aTRAM. *Systematic Biology* , **66** (5), 786–798. https://doi.org/10.1093/sysbio/syw105

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A. L., Sperling, F. A., Condamine, F. L. (2019). Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution.*Systematic*

Biology , **69** (1), 38–60. https://doi:10.1093/sysbio/syz030

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* , **19** (5), 455–477. https://doi:10.1089/cmb.2012.0021

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* , **13** (1), 403. https://doi:10.1186/1471-2164-13-403

Blaimer, B. B., Lloyd, M. W., Guillory, W. X., & Brady, S. G. (2016). Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS One* ,**11** (8), e0161531. https://doi:10.1371/journal.pone.0161531

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* ,**30** (15), 2114–2120. https://doi:10.1093/bioinformatics/btu170

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., ... Zaretskaya I. (2013). BLAST: A more efficient report with usability improvements. *Nucleic Acids Research* , **41** , W29–W31.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: efficacy across scales of divergence. *Molecular Ecology Resources* , **16** (5), 1059–1068. https://doi:10.1111/1755-0998.12449

Burbrink, F. T., Grazziotin, F. G., Pyron, R. A., Cundall, D., Donnellan, S., Irish, F., ... Zaher, H. (2020). Interrogating Genomic-Scale Data for Squamata (Lizards, Snakes, and Amphisbaenians) Shows no Support for Key Traditional Morphological Relationships.*Systematic Biology* , **69** (3), 502–520. https://doi.org/10.1093/sysbio/syz062

Bushnell, B. (2014). BBtools. Retrieved from https://sourceforge.net/projects/bbmap/

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *MolecularBiology and Evolution* , **17** (4), 540–552. https://doi:10.1093/oxfordjournals.molbev.a026334

Chen, M. Y., Liang, D., & Zhang, P. (2017). Phylogenomic resolution of the phylogeny of Laurasiatherian mammals: exploring phylogenetic signals within coding and noncoding sequences. *Genome Biology and Evolution* , **9** (8), 1998–2012. https://doi:10.1093/gbe/evx147

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* , **61** (5), 717–726. https://doi:10.1093/sysbio/sys004

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data.*Bioinformatics* , **28** (23), 3150–3152. https://doi:10.1093/bioinformatics/bts565

Garrison, N. L., Rodriguez, J., Agnarsson, I., Coddington, J. A., Griswold, C. E., Hamilton, C. A., ... Bond, J. E. (2016). Spider phylogenomics: untangling the spider tree of life. *PeerJ* ,**4** , e1719. https://doi:10.7717/peerj.1719

Glenn, T. C., & Faircloth, B. C. (2016). Capturing Darwin's dream.*Molecular Ecology Resources* , **16** (5), 1051–1058. https://doi:10.1111/1755-0998.12574

Gori, K., Suchan, T., Alvarez, N., Goldman, N., & Dessimoz, C. (2016). Clustering genes of common evolutionary history. *MolecularBiology and Evolution* , **33** (6), 1590–1605. https://doi:10.1093/molbev/msw038

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis X., Fan, L., ... Regev A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome.*Nature Biotechnology* , **29** (7), 644–652. https://doi:10.1038/nbt.1883

Guillaume, M., & Carl, K. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.*Bioinformatics* , **27** (6), 764–770. https://doi:10.1093/bioinformatics/btr011

Guschanski, K., Krause, J., Sawyer, S., Valente, L. M., Bailey, S., Finstermeier, K., ... Savolainen, V. (2013). Next-generation museomics disentang one of the largest primate radiations. *Systematic Biology* , **62** (4), 539–554. https://doi:10.1093/sysbio/syt018

Hahn, M. W., & Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution* , **70** (1), 7–17. https://doi:10.1111/evo.12832

Hillis, D. M., Heath, T. A., & St John, K. (2005). Analysis and visualization of tree space. *Systems Biology* , **54** (3), 471–482. https://doi:10.1080/10635150590946961

Hughes, G. M., & Teeling, E. C. (2018). AGILE: an assembled genome mining pipeline. *Bioinformatics* , **35** (7), 1252–1254. https://doi:10.1093/bioinformatics/bty781

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., ... Zhang G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* , **346** (6215), 1320–1331. https://doi:10.1126/science.1253451

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology* , **25** (1), 185–202. https://doi:10.1111/mec.13304

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* , **61** (5), 727–744. https://doi:10.1093/sysbio/sys049

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology Evolution and Systematics* , **44** , 99–121. https://doi:10.1146/annurev-ecolsys-110512-135822

Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. (2013). Capturing protein-coding genes across highly divergent species.*BioTechniques* , **54** (6), 321–326. https://doi:10.2144/000114039

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools.*Bioinformatics* , **25** (16), 2078–2079. https://doi:10.1093/bioinformatics/btp352

Li, J., Liang, D., Wang, Y., Guo, P., Huang, S., & Zhang, P. (2020). A large-scale systematic framework of Chinese snakes based on a unified multilocus marker system. *Molecular Phylogenetics and Evolution* ,**148** , 106807. https://doi:10.1016/j.ympev.2020.106807

Li, J., Zeng, Z., Wang, Y., Liang, D., & Zhang, P. (2019). Sequence capture using AFLP-generated baits: A cost-effective method for high-throughput phylogenetic and phylogeographic analysis. *Ecology and Evolution* , **9** (10), 5925–5937. https://doi:10.1002/ece3.5176

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., & Linder, C. R. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* , **61** (1), 90–106.

McCartney-Melstad, E., Mount, G. G., & Shaffer, H. B. (2016). Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resource* , **16** (5), 1084–1094. https://doi:10.1111/1755-0998.12538

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* , **66** (2), 526–538. https://doi:10.1016/j.ympev.2011.12.007

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., ... Zhou X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science* , **346** (6210), 763–767. https://doi:10.1126/science.1257570

Morozova, O., Hirst, M., & Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Geneticsis* , **10** , 135–151. https://doi:10.1146/annurev-genom-082908-145957

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., ... Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* ,**461** (7261), 272–276. https://doi:10.1038/nature08250

Oakley, T. H., Wolfe, J. M., Lindgren, A. R., & Zaharoff, A. K. (2012). Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement and pancrustacean phylogeny.*Molecular Biology and Evolution* , **30** , 215–233. https://doi:10.1093/molbev/mss216

Olofsson, J. K., Cantera, I., Paer, C. V. D., Hong-Wa, C., Zedane, L., Dunning, L. T., ... Besnard, G. (2019). Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe.*Molecular Ecology Resource* , **19** (4), 877–892.

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., ... Moritz, C. (2014). Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources* , **14** (5), 1000–1010. https://doi:10.1111/1755-0998.12249

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources* , **16** (5), 1069–1083. https://doi:10.1111/1755-0998.12541

Puritz, J. B., & Lotterhos, K. E. Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. 2018.*Molecular Ecology Resources* , **18** (6), 1209–1222. https://doi:10.1111/1755-0998.12905

Pyron, R. A., Hendry, C. R., Chou, V. M., Lemmon, E. M., Lemmo, A. R., & Burbrink, F. T. (2014). Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). *Molecular Phylogenetics and Evolution* , **81** , 221–231. https://doi:10.1016/j.ympev.2014.08.023

Quek, R. Z. B., Jain, S. S., Neo, M. L., Rouse, G. W., & Huang, D. W. (2020). Transcriptome-based target-enrichment baits for stony corals (Cnidaria: Anthozoa: Scleractinia). *Molecular Ecology Resources* ,**20** (3), 807–818. https://doi:10.1111/1755-0998.13150

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., ... Braun, E. L. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the Avian tree of life more than taxon sampling. *Systems Biology* , **66** (5), 857–879. https://doi:10.1093/sysbio/syx041

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences* , **53** (1–2), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Singhal, S., Grundler, M., Colli, G., & Rabosky, D. L. (2017). Squamate Conserved Loci (SqCL): A unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Molecular Ecology Resources* , **17** (6), e12–e14. https://doi:10.1111/1755-0998.12681

Slater, G. S. C., & Birney, E., (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* , **6** , 31. https://doi:10.1186/1471-2105-6-31

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* ,**30** (9), 1312–1313. https://doi:10.1093/bioinformatics/btu033

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., ... Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics.*Molecular Biology and Evolution* , **35** (3), 543–548. https://doi:10.1093/molbev/msx319

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* ,**10** (1), 57–63. https://doi:10.1038/nrg2484

Wiens, J. J., Hutter, C. R., Mulcahy, D. G., Noonan, B. P., Townsend, T. M., Jr, J. W. S., & Reeder, T. W. (2012). Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. *Biology Letters* , **8** (6), 1043–1046. https://doi:10.1098/rsbl.2012.0703

Zhang, F., Ding, Y., Zhu, C., Zhou, X., Orr, M. C., Scheu, S., & Luan, Y. X. (2019). Phylogenomics from low-coverage whole-genome sequencing.*Methods in Ecology and Evolution* , **10** (4), 507–517. https://doi:10.1111/2041-210X.13145

Zhang, Y., Deng, S., Liang, D., & Zhang P. (2019). Sequence capture across large phylogenetic scales by using pooled PCR–generated baits: A case study of Lepidoptera. *Molecular Ecology Resource* ,**19** (4), 1037–1051. https://doi:10.1111/1755-0998.13026

Zheng, Y., & Wiens, J. J. (2015). Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species.*Molecular Phylogenetics and Evolution* , **94** , 537–547. https://doi.org/10.1016/j.ympev.2015.10.009.

Data accessibility

Raw read data were deposited in NCBI SRA (accession PRJNA668136, PRJNA668154). Python scripts, the concatenated data matrix, and the resulting phylogenetic trees were deposited in Dryad XXX.

Author contributions

P.Z., D.L., and J.X.L. designed the research. J.X.L. performed the experiment and analyzed the data. P.Z., D.L. and J.X.L. wrote the paper.

Supplementary Information

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

Appendix S1. Primers used in the cDNA probe preparation of FLc-Capture.

Appendix S2. WGS data resource of the four insect species.

Appendix S3. Statistics of genome assembly and loci extracted from 1×, 5×, 10× and 20× whole-genome sequencing data for four insect and two snake species.

Appendix S4. A summary of data characteristics for ORFs and UTRs, including length, taxa occupancy, GC content, percentage of missing data.

Figure Legends

Figure 1. Schematic overview of the FLc-Capture sequencing method. a) Prepare shotgun genomic DNA library. b) Prepare full-length cDNA probes from high-quality RNA using SMARTer PCR cDNA Synthesis Kit (Clontech Inc.). The full-length cDNA probes contain both UTR and ORF regions. c) Hybridize the probes to the shotgun genomic library. Both coding and noncoding loci bind to the cDNA probes.

Figure 2. Details of the FLc-Capture data processing. a) Creating reference ORF and UTR sets by sequencing the cDNA probes. b) Extracting ORF (coding) sequences. First, FLc-Capture assembled reads into contigs, and identified exons from contigs using EXONERATE based on the reference ORF sets. The identified exons are then mapped onto the reference ORF sequences to make the orthologous ORF sequences, based on the position information resulted from the EXONERATE searches. c) Extracting UTR (noncoding) sequences. Based on the reference UTRs, FLc-Capture uses a mutual best-hit (MBH) strategy to identify orthologous UTR sequences and truncate sequences according to its optimal aligning region.

Figure 3. The results of extracting phylogenetic loci from 1x, 5x, 10x and 20x whole-genome sequencing data for four insect and two snake species. For each species, at each sequencing depth, the left bar depicts

the recovery rate of genes classified as complete (blue) and fragmented (yellow); the right gray bar depicts the recovery rate of DNA sites. The genome size of each species is shown under the species name.

Figure 4. The results of extracting UTR and ORF sequences from the FLc-Capture data for the ingroup Colubridae species and the outgroup species. Bars show a) gene recovery rates b) nucleotide recovery rates.

Figure 5. a) Plots of linear regressions of capture specificity (on-target) across all samples in FLc-Capture experiments (blue = ORF, red = UTR), using the average pairwise distance from probe species as the independent variable. b) Relationship between the mean capture depth of each target and the abundance of its corresponding probe (blue = ORF, red = UTR).

Figure 6. Characteristics of the ORF and UTR data sets (blue = ORF, red = UTR). Boxplots show a) distribution of locus length (bp), b) distribution of GC content (among genes and among species), and c) distribution of evolutionary rates of loci (measured by the mean pairwise distance of each locus). d) Visualization of ML tree space using multidimensional scaling plot of 1,075 ORF gene trees (left) and 1,948 UTR gene trees (right); each dot represents a tree inferred from one gene. Distances between dots represent Robinson–Foulds distances between gene trees.

Figure 7. Phylogenetic relationships among the 25 Colubridae species (including the probe species *Ptyas korros*) and 12 outgroup species inferred from the ORF (1,075 ORFs; ˜817 K) and UTR (1,948 UTRs ˜1,114 K) data sets. The trees are inferred with RAxML, and the two data sets produce identical phylogeny. Branch support values are indicated beside nodes in order of ORF ML bootstrap and UTR ML bootstrap from left to right. The filled circles represent ML bootstrap support [?] 95% (both ORF and UTR). The three hotly debated nodes (A, B and C) within the Colubridae family are indicated by filled circles with letters. The bars right to the species name represents the integrity of the data set for each species (calculated by loci).

Hosted file

`Table 1.pdf` available at https://authorea.com/users/376239/articles/493275-simultaneously-collecting-coding-and-noncoding-phylogenomic-data-using-homemade-full-length-cdna-probes-tested-by-resolving-the-high-level-relationships-of-colubridae

Hosted file

`Table 2.pdf` available at https://authorea.com/users/376239/articles/493275-simultaneously-collecting-coding-and-noncoding-phylogenomic-data-using-homemade-full-length-cdna-probes-tested-by-resolving-the-high-level-relationships-of-colubridae

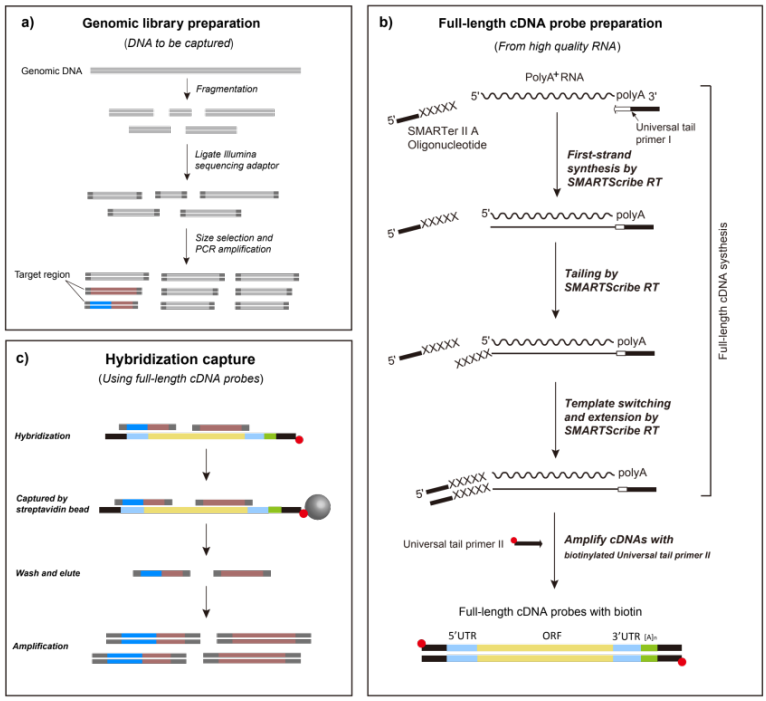

Figure 1.

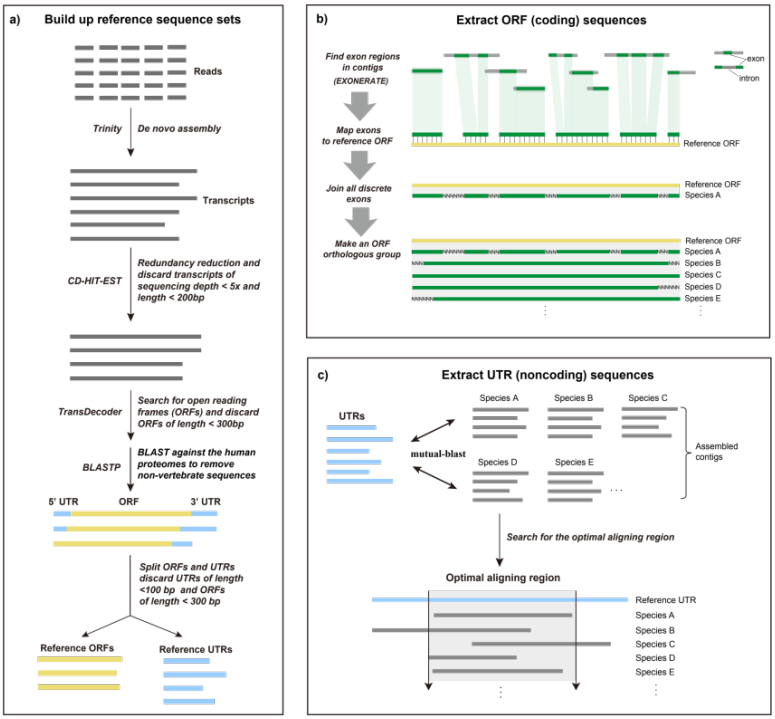

Figure 2.

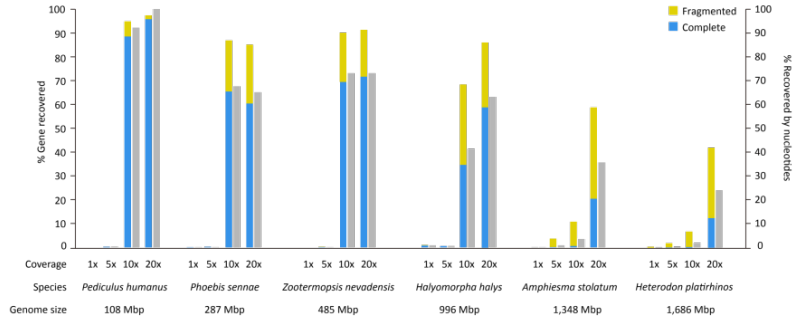

Figure 3.

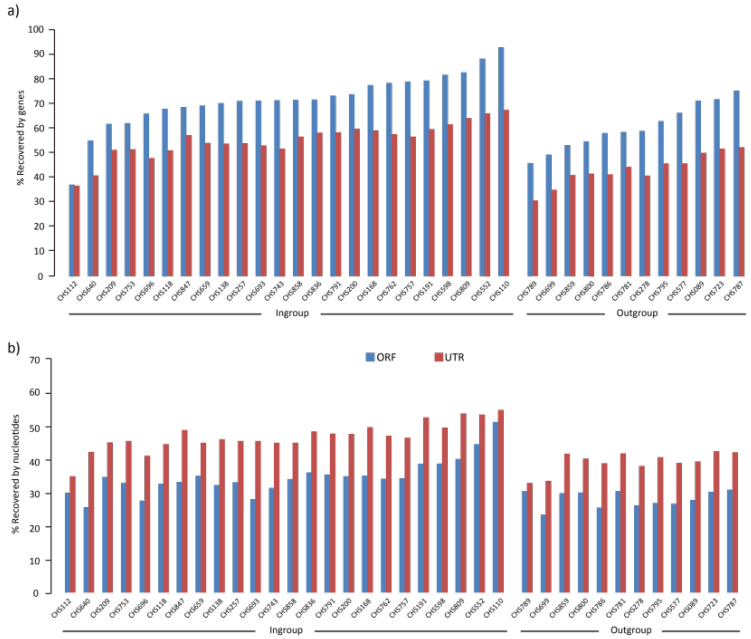

Figure 4.

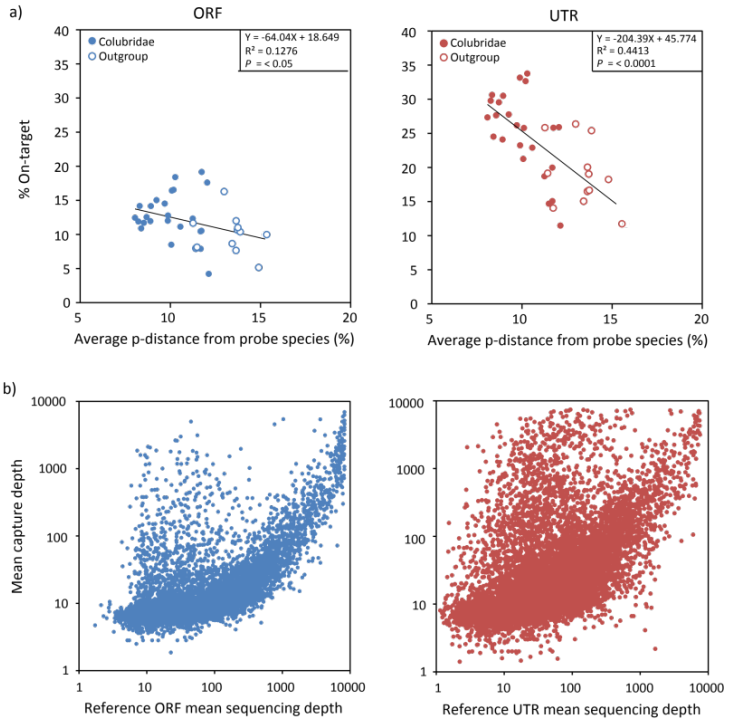

Figure 5.

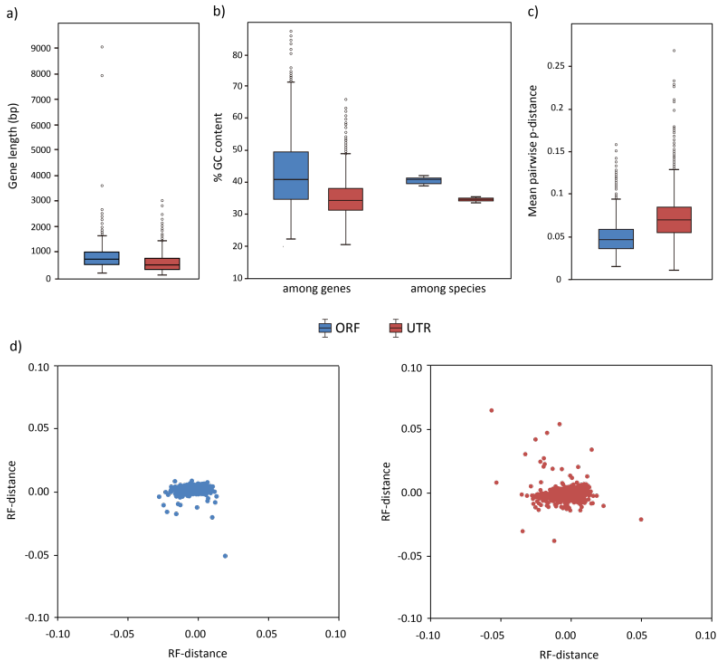

Figure 6.

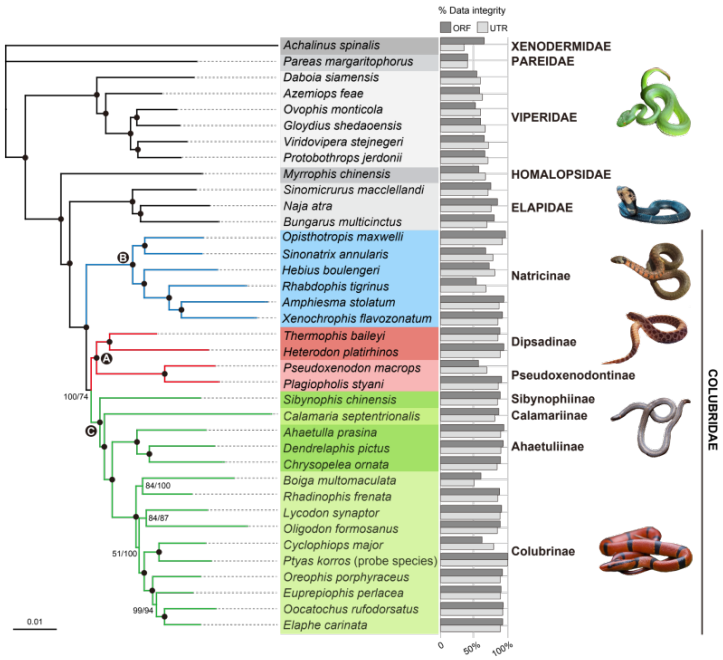

Figure 7.