

Article

Not peer-reviewed version

The Genome Assembly and Annotation of the Chinese Cobra, *Naja atra*

[Jiangang Wang](#), Yuxin Wu, Shiqing Wang, Weiwu Mu, Wenmei Zeng, Xi Chen, Kangfeng Jiang, Liangyu Yang, Guohai Hu*, Fengping He*

Posted Date: 9 November 2023

doi: 10.20944/preprints202311.0597.v1

Keywords: Genetics and Genomics, Zoology, Animal Genetics



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

The Genome Assembly and Annotation of the Chinese Cobra, *Naja atra*

Jiangang Wang ^{*,a,b} [ORCID:0009-0004-9731-2576], Yuxin Wu ^a, Shiqing Wang ^{*,b,c}, Weiwu Mu ^a, Wenmei Zeng ^a, Xi Chen ^a, Kangfeng Jiang ^a, Liangyu Yang ^a, Guohai Hu ^{*,d} and Fengping He ^{*,a}

^a College of Veterinary Medicine, Yunnan Agricultural University, Kunming, 650231, China

^b State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China

^c College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

^d China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

* Correspondence: Jiangang Wang: 18576792453@163.com and Shiqing Wang: wangshiqing@genomics.cn

Abstract: In China, 65 types of venomous snakes exist, with the Chinese Cobra *Naja atra* being prominent and a major cause of snakebites in humans. Furthermore, *N. atra* is a protected animal in some areas, as it has been listed as vulnerable by the International Union for Conservation of Nature. Recently, due to the medical value of snake venoms, venomomics has experienced growing research interest. In particular, genomic resources are crucial for understanding the molecular mechanisms of venom production. Here, we report a highly continuous genome assembly of *N. atra*, based on a snake sample from Huangshan, Anhui, China. The size of this genome is 1.67 Gb, while its repeat content constitutes 37.8% of the genome. A total of 26,432 functional genes were annotated. This data provides an essential resource for studying venom production in *N. atra*. It may also provide guidance for the protection of this species.

Keywords: genetics and genomics; zoology; animal genetics

Introduction

Elapidae is a family of snakes divided into three subfamilies (Bungarinae, Elapinae and Notochilinae), with 44 genera and around 186 described species distributed widely [1]. The front of the mouth of an elapid has permanently erect tusks, which are his distinguishing features. Elapids include terrestrial and sea snakes. Terrestrial elapids, a family of venomous snakes, are distributed across the globe in tropical and subtropical regions, with most species inhabiting the Southern Hemisphere. Elapid sea snakes are mainly distributed in the Indian Ocean and the Southwest Pacific Ocean [2].

The Chinese cobra, or *Naja atra* (NCBI: txid8656) (Figure 1), is a species of cobra from the family Elapidae. Chinese cobras are usually between 1.2 and 1.5 m long [3] and they are among the most prevalent cobra species in China. The Chinese cobra likes to inhabit plains, hills and low mountains [4]. Humans often encounter Chinese cobras, although these snakes usually escape to avoid confrontation with humans. Chinese cobras can be observed hunting during daylight hours from March to October and up to 2-3 hours after sunset at temperatures of 20-32°C [5]. They have a widely varied diet and prey on rodents, frogs, toads and other snakes.

The Chinese cobra is highly poisonous, its venom consisting mainly of postsynaptic neurotoxins and cardiotoxins [6]. Their venom offers them protection from predation to a certain extent; however, populations of Chinese cobra have declined by 30% to 50% due to habitat loss and hunting by humans. The venom of Chinese cobras can be used to extract anti-cobra snake venom, which is used to treat cobra snake bites. Although the Chinese cobra is currently listed as a Vulnerable species on the International Union for Conservation of Nature Red List [7], its numbers in the wild have declined from Vulnerable to Endangered due to continued hunting.



Figure 1. *N. atra*. ([https://commons.wikimedia.org/wiki/File:View_of_deadly_cobra_snake_\(Unsplash\).jpg](https://commons.wikimedia.org/wiki/File:View_of_deadly_cobra_snake_(Unsplash).jpg)).

Main Content

Context

Snakebite is a serious threat to human life as it kills around 100,000 people annually. Genome-enabled research of toxin genes may facilitate the development of effective antivenoms. Here, we present a highly continuous reference genome assembly of *N. atra*. While there is a reference genome for the Indian cobra (*Naja naja*) [20], this is the first for the Chinese cobra. This resource may also provide valuable information for the conservation of this vulnerable species, which can be used for targeted protection and breeding.

Methods

The detailed methods used in this study are available via a protocol collection hosted in protocols.io [9] (Figure 2).



Protocols for the assembly and annotation of snake genomes V.2

DOI
dx.doi.org/10.17504/protocols.io.5jyl8j6e9g2w/v2

Boyang Liu¹, Liangyu Cui¹, Zhangwen Deng², Yue Ma¹,
Diancheng Yang^{3,4}, Yanan Gong^{3,4}, Yanchun Xu¹, Shuhui Yang¹,
Song Huang^{3,4}

¹College of Wildlife and Protected Area, Northeast Forestry University, Harbin 150040, China;
²Guangxi Forest Inventory and Planning Institute, Nanning 530011, China;
³Anhui Province Key Laboratory of the Conservation and Exploitation of Biological Resource, College of Life Sciences, Anhui Normal University, Wuhu 241000, China;
⁴Huangshan Noah Biodiversity Institute, Huangshan 245000, China

VERSION 2 ▾
FEB 09, 2023

SHARE

WORKS FOR ME 1

COMMENTS 0

BOOKMARK

COPY / FORK

MORE ↓

博洋 Boyang Liu

Figure 2. A protocols.io collection of the standard protocols for sequencing snake genomes [9].

Sample collection and sequencing

The *N. atra* sample used in this study was captured in Huangshan, Anhui, China, in 2021. After collection, the specimen was quickly frozen to -80°C using drikkold dry ice for storage and transport. Methods for DNA extraction, library construction and sequencing were identical those used by Liu et al in a previous study.

Sample collection, experiments and research design were all authorized by the Institutional Review Board of BGI (BGI-IRB E22001). In this research, all the procedures have been operated abiding to the guidelines from BGI-IRB strictly.

Genome survey, assembly, annotation and assessment

The single-tube long fragment read sequencing data were assembled using Supernova (v2.1.1, RRID:SCR_016756) [10]. NextPolish (v1.0.5) [10] was then used to perform a second round of correction and a third round of polishing of this assembly using the Whole Genome Sequencing data. To get a haploid representation of the genome, duplicates were purged from the genome using the purge_dups pipeline (RRID:SCR_021173) [11]. The completeness of the genome was evaluated using sets of BUSCO (v5.2.2, RRID:SCR_015008) with genome mode and lineage data from vertebrata_odb10 [12].

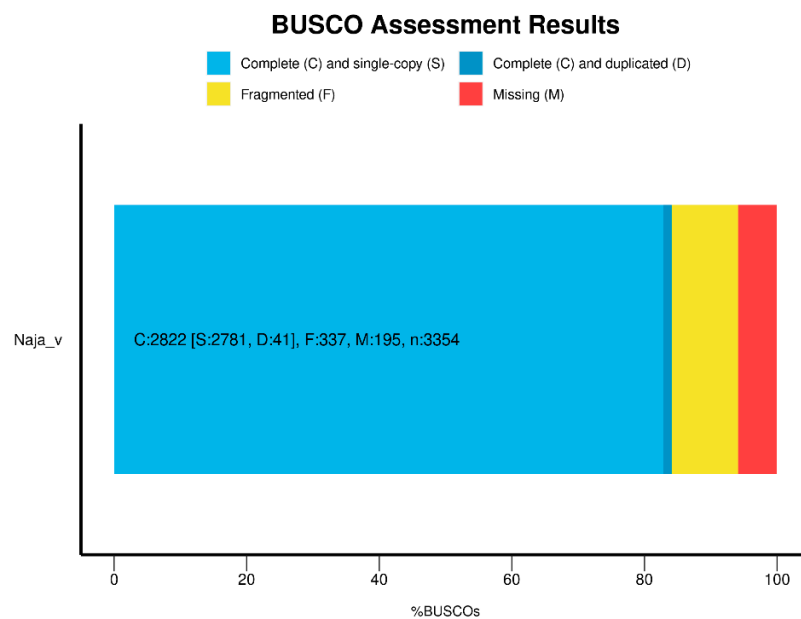
In order to detect the presence of known repeat elements in the genome of the many-banded *P. mucosa*, the following approach was employed. To identify the known repetitive elements in the genome of the many-banded krait, we used Tandem repeats Finder [13], LTR_Finder (RRID:SCR_015247) [14] and RepeatModeler (v2.0.1, RRID:SCR_015027) [15]. RepeatMasker (v3.3.0, RRID:SCR_012954) [16] and RepeatProteinMask v3.3.0 [17] were used to search the genome sequences for known repeat elements. The BRAKER2 pipeline (RRID:SCR_018964) [18] was used for gene prediction. Then, the gene sets were aligned against several known databases, including SwissProt [19], TrEMBL [19], Kyoto encyclopedia of genes and genomes (KEGG) [8], gene ontology (GO), and the Non-Redundant Protein Sequence Database [8] database.

Results

We present a draft genome sequence of *N. atra*. The size of this genome is 1.67 Gb (Table 1), similar to the previously published 1.79 Gb genome of *N. naja* [8]. The scaffold N50 length is 234.17 Kb, and the CG content reached 37.8%. The maximal scaffold length is 2,929,773 bp, demonstrating that the reference is highly continuous according to the characteristics of the genome sequence. In addition, the integrity of the genome was assessed at 84.1% using BUSCO (Figure 3).

Table 1. Summary of the features of our *N. atra* genome.

	contig	Scaffold
Maximal length (bp)	271,789	2,929,773
N90 (bp)	4,371	7,368
N50 (bp)	33,081	234,173
Number \geq 100bp	194,909	106,418
Number \geq 2kb	113,570	54,157
GC content (%)	40.3	37.8
Genome size (bp)	1,671,178,062	

**Figure 3.** BUSCO assessment result of our *N. atra* genome.

In the *N. atra* genome, the content of repetitive elements is up to 40.26%, and the total length of the genome is 672 Mb (Table 2, Table 3). After we counted all repeat elements, we found that long interspersed nuclear elements (LINEs) accounted for 30.63%, long terminal repeats (LTRs) accounted for 14.03% and DNA accounted for 4.27% (Figure 4).

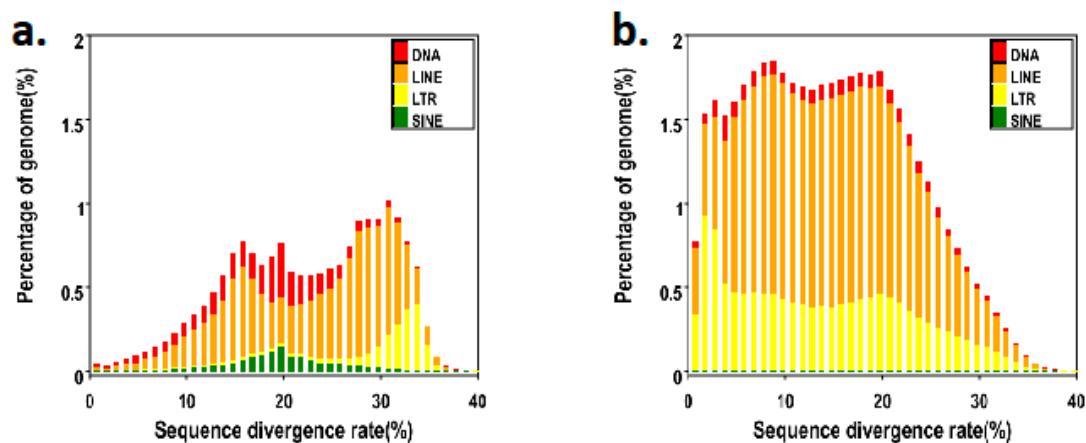
**Figure 4.** Distribution of transposable elements (TEs) in our *N. atra* genome. The TEs include DNA transposons (DNA) and RNA transposons (i.e., DNAs, LINEs, LTRs, and short interspersed nuclear elements (SINEs)). (a) Known sequences divergence rate (b). *De novo* sequences divergence rate.

Table 2. Statistics for repetitive sequences identified in our *N. atra* genome.

Type	Length (bp)	% in genome
DNA	37,917,702	2.269170
LINE	449,338,074	26.890460
SINE	2,779,035	0.166310
LTR	224,765,038	13.450975
Other	0	0
Satellite	632,498	0.037852
Simple_repeat	5,080,994	0.304070
Unknown	7,924,824	0.474258
Total	672,795,525	40.263183

Table 3. Summary of the TEs in our *N. atra* genome.

Repbase TEs			TE proteins		<i>De novo</i>		Combined TEs	
Type	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	44,907,141	2.57	3,638,477	0.20	41,761,899	2.39	81,259,555	4.66
LINE	170,663,721	9.79	140,023,530	8.03	581,624,764	33.36	619,156,475	35.51
SINE	25,759,131	1.47	0	0	8,061,060	0.46	32,226,226	1.84
LTR	22,468,876	1.28	30,088,483	1.72	149,994,747	8.60	159,624,403	9.15
Other	23,680	0.001	0	0	0	0	23,680	0.001
Unknown	0	0	0	0	5,653,213	0.32	5,653,213	0.32
Total	251,569,212	14.43	173,669,200	9.96	722,435,038	41.44	752,340,302	43.15

Finally, 29,063 functional genes were annotated. Through KEGG annotation, we found that the genes related to signal transduction are essential in *N. atra* (Figure 5). Furthermore, through a pathway enrichment analysis, we found that the number of Human Diseases pathways is the highest. Environmental Information Processing and Organismal systems also account for a relatively large proportion. According to the annotation and enrichment in the GO database, 6,292 genes are enriched in cellular process and 6,734 in binding.



Figure 5. Gene annotation information of *N. atra*. (a) KEGG enrichment of *N. atra*. (b) GO enrichment of *N. atra*.

Data Availability: The data supporting the findings of this study have been deposited into the CNGB Sequence Archive (or CNSA) of China National GeneBank DataBase (or CNGBdb) with the accession number CNP0004141. Raw reads are available in the SRA via bioproject PRJNA955401. Additional data is in the GigaDB repository [21].

Editor's note: This paper is part of a series of Data Release papers presenting the genomes of different snake species [22].

Ethics approval and consent to participate: The authors declare that ethical approval was not required for this type of research.

Author Contributions: TL designed and initiated the project. JW performed DNA extraction, library construction and data analysis. JW wrote the manuscript. All authors read and approved the final manuscript.

Funding: Our project was financially supported by the Guangdong Provincial Key Laboratory of Genome Read and Write (grant no. 2017B030301011). This work was also supported by China National GeneBank (CNGB).

Acknowledgments: Yunnan Agricultural University collected the samples.

Conflicts of Interest: The authors declare no conflict of financial interests.

Abbreviations

GO, gene ontology; KEGG, Kyoto encyclopedia of genes and genomes; LINE, long interspersed nuclear element; LTR, long terminal repeat; SINE, short interspersed nuclear element; TE, transposable element.

References

1. Slowinski JB, Keogh JS (April 2000). "Phylogenetic relationships of elapid snakes based on cytochrome b mtDNA sequences". *Molecular Phylogenetics and Evolution*. 15 (1): 157–64. doi:10.1006/mpev.1999.0725. i
2. Chan, S. (2006). *A field guide to the venomous land snakes of Hong Kong*. Cosmos Books Ltd., Hong Kong. ISBN 988-211-326-5.
3. Snake of medical importance. Singapore: Venom and toxins research group. 1990. ISBN 9971-62-217-3.
4. "Naja atra – General Details, Taxonomy and Biology, Venom, Clinical Effects, Treatment, First Aid, Antivenoms". WCH Clinical Toxinology Resource. University of Adelaide. Retrieved 18 December 2011.
5. Zhao, EM; Adler, K (1993). *Herpetology of China*. United States: Society for the Study of Amphibians and Reptiles. ISBN 0-916984-28-1.
6. Wang, AH; Yang, CC (10 September 1981). "Crystallographic studies of snake venom proteins from Taiwan cobra (*Naja nana atra*). Cardiotoxin-analogue III and phospholipase A2". *Journal of Biological Chemistry*. 256 (17): 9279–9282. doi:10.1016/S0021-258(19)52542-X
7. "Naja atra". IUCN Red List of Threatened Species. 2014: e.T192109A2040894. doi:10.2305/IUCN.UK.2014-3.RLTS.T192109A2040894.en
8. Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schröder MS, Wu M, Kuriakose B, Mathew OK, Perumal RC, Koludarov I, Goldstein LD, Senger K, Dixon MD, Velayutham D, Vargas D, Chaudhuri S, Muraleedharan M, Goel R, Chen YJ, Ratan A, Liu P, Faherty B, de la Rosa G, Shibata H, Baca M, Sagolla M, Ziai J, Wright GA, Vucic D, Mohan S, Antony A, Stinson J, Kirkpatrick DS, Hannoush RN, Durinck S, Modrusan Z, Stawiski EW, Wiley K, Raudsepp T, Kini RM, Zachariah A, Seshagiri S. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. *Nat Genet*. 2020 Jan;52(1):106-117. doi: 10.1038/s41588-019-0559-8.
9. Liu B, Cui L, Deng Z, Ma Y, Yang D, Gong Y, et al. Protocols for the assembly and annotation of snake genomes. 2023. *Protocols.io* <https://dx.doi.org/10.17504/protocols.io.5jyl8j6e9g2w/v2>
10. Guan D, McCarthy SA, Wood J, Howe K, Wang Y and Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36 9:2896-8. doi:10.1093/bioinformatics/btaa025
11. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, 2019; 8: 2138. doi:10.12688/f1000research.21782.4.
12. Benson and G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*. 27 2:573-80
13. Zhao X and Hao W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*. 2007 35(suppl_2): W265–W268..
14. Smit A, Hubley R and Green P. RepeatModeler Open-1.0. 2008–2015. Seattle, USA: Institute for Systems Biology Available from: <http://www.repeatmasker.org>, Last Accessed May. 2015;1:2018.

15. Tarailo-Graovac M and Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current protocols in bioinformatics* / editorial board, Andreas D Baxeavanis [et al]. 2009;Chapter 4 Unit 4:Unit 4.10.
16. Tempel S. Using and understanding RepeatMasker. *Mobile Genetic Elements*. Springer; 2012. p. 29-51 doi:10.1007/978-1-61779-603-6_2 .
17. Bruna T, Hoff KJ, Lomsadze A, Stanke M and Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*. 2021;3 1:lqaa108.
18. Amos B and Rolf A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*. 2000; 1:45
19. Pitk E. KEGG database. *Novartis Foundation Symposium*. 2006;247:91-103.
20. Kushal Suryamohan, Sajesh P. Krishnankutty, Joseph Guillory, Matthew Jevit, Markus S. Schröder, Meng Wu, Boney Kuriakose The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins doi:doi.org/10.1038/s41588-019-0559-8 .
21. Cite GigaDB DOI when ready
22. Snake Genomes. GigaByte. 2023; https://doi.org/10.46471/GIGABYTE_SERIES_0004

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.