Data Descriptor

The Genome Assembly and Annotation of the *Bungarus multicinctus*

Boyang Liu *, Liangyu Cui, Zhangwen Deng, Yue Ma, Diancheng Yang, Yanan Gong, Yanchun Xu, Tianming Lan, Shuhui Yang, Song Huang

Posted Date: 9 June 2023

doi: 10.20944/preprints202306.0658.v1

Keywords: Genetics and Genomics; Evolutionary Biology; Zoology

Data Descriptor

The Genome Assembly and Annotation of the *Bungarus multicinctus*

Boyang Liu [1,†], Liangyu Cui [1,†], Zhangwen Deng [2,†], Yue Ma [1], Diancheng Yang [3,4], Yanan Gong [3,4], Yanchun Xu [1], Tianming Lan [5,6], Shuhui Yang [1,*] and Song Huang [3,4,*]

[1] College of Wildlife and Protected Area, Northeast Forestry University, Harbin 150040, China
[2] Guangxi Forest Inventory and Planning Institute, Nanning 530011, China
[3] Anhui Province Key Laboratory of the Conservation and Exploitation of Biological Resource, College of Life Sciences, Anhui Normal University, Wuhu 241000, China
[4] Huangshan Noah Biodiversity Institute, Huangshan 245000, China
[5] BGI Life Science Joint Research Center, Northeast Forestry University, Harbin 150040, China
[6] State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
* Correspondence: yangshuhui@nefu.edu.cn (S.Y.); snakeman@ahnu.edu.cn (S.H.); ORCID: 0000-0001-6786-8523 (S.H.)
† These authors contributed equally to this work.

Abstract: Snakes are a vital component of wildlife resources and are widely distributed across the globe. *Bungarus multicinctus*, a highly venomous snake, is found in central and southern China. *B. multicinctus* is a highly venomous snake and is distributed in central and southern China. Snakes are an ancient group of reptiles, and their genome resources can provide important clues for understanding the evolutionary history of reptiles. Meanwhile, genomic resources play a crucial role in comprehending the evolution of species. So far, the genomic resources of snakes are a rarity. In 2021, a snake sample was collected from Beiliu Longgukeng, Guangxi, which was identified as *B. multicinctus* through morphological identification. In this study, we present a highly contiguous genome of B. multicinctus with a size of 1.51 Gb. The genome contains a repeat content of 40.15%, with a total length exceeding 620 Mb. Additionally, we annotated a total of 24,869 functional genes. This research is of great significance for comprehending the evolution of *B. multicinctus* and provides a genomic basis for the genes involved in venom gland function.

Keywords: Genetics and Genomics; Evolutionary Biology; Zoology

Introduction

Snakes are a fascinating group of reptiles that exhibit unique and diverse characteristics. With approximately 3070 extant species found in all continents except Antarctica [1], they are known for their lack of limbs, elongated body shape, and exclusively carnivorous diet. Snakes have evolved many specialized adaptations, such as infrared sensing pits and venom apparatus, which provide them with exceptional predatory capabilities [2]. These adaptations have made snakes important model organisms for evolutionary studies, yielding insights into limb development, sex chromosome evolution, and venom evolution. In recent years, genetic approaches have become increasingly important in understanding the evolution and diversity of snakes [3].

By exploring the evolution of venomous snakes, we can gain a deeper understanding of the ecological and evolutionary roles of these fascinating reptiles.

The *Bungarus multicinctus* (NCBI:txid8616), also known as the manybanded krait or umbrella snake, is widely distributed throughout southern Asia, with its range spanning across multiple countries including India, Pakistan, Indonesia, Sri Lanka, Malaysia, Bangladesh, Vietnam, and China [4]. In China, the *Bungarus multicinctus* is recognized as one of the top ten most venomous snakes, with a lethality rate ranging from 26.9% to 33.3% [5].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 June 2023 doi:10.20944/preprints202306.0658.v1

In this study, We collect a muscle sample from the *B.multicinctus* and present a highly contiguous *B. multicinctus* genome with a genome size of 1.51Gb. Its repeat element content reached 41.68%, providing new evidence for understanding the relationship between repeat elements and genome size in Elapidae species.

Main Content

Context

In this study, we have presented a highly continuous genome assembly of *Bungarus multicinctus*. The genome size of *B. multicinctus* was found to be 1.51 Gb, with a GC content of 37.8% (Table 1). The maximal length of scaffold was 39.68 Mb and the N50 length was 6.55 Mb, indicating a highly continuous genome sequence. This draft genome sequence of *B. multicinctus* will serve as an invaluable resource for further research on venomous snakes, enabling a better understanding of their genetic makeup.

Table 1. Summary of the features of the *B. multicinctus* genome.

	Contig	**Scaffold**
Maximal length (bp)	468983	41606426
N90 (bp)	5806	30083
N50 (bp)	33081	6870761
number>=100bp	163090	82383
number>=2kb	81775	22350
Ratio of Ns	0.045	0.045
GC content (%)	39.6	37.8
Genome size (bp)	1548488562	1621955402

The content of repetitive elements in the *B. multicinctus* genome is surprisingly large, reaching 41.68% with a total length of 675Mb (Table 2). We analyzed the content of various repeating elements, of which unknown types of repeating elements accounted for 51%, while LINE and DNA accounted for 10% and 8%, respectively (Figure 1). Research indicates that although snake species have similar genome sizes, they exhibit significant differences in TE content, with low diversity in the types of TEs present [6,7].Species with a longer evolutionary history tend to have higher TE diversity [8]. These results suggest that the significant expansion of repeating elements is an important manifestation of species differences.

Table 2. Summary of transposable elements (TEs) in the *B. multicinctus* genome.

	Repbase TEs		**TE protiens**		**De novo**		**Combined TEs**	
Type	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	32816331	2.02	2921569	0.18	112067211	6.91	129267220	7.97
LINE	174481405	10.76	154961354	9.60	276722230	17.07	301624987	18.61
SINE	13524698	0.83	0	0	39754823	2.45	43837124	2.70
LTR	23313679	1.44	30431704	1.88	52496522	3.24	60898786	3.76
Other	16171	0.01	243	0.01	0	0	16414	0.01
Unknown	0	0	0	0	182574604	11.26	182574604	11.26
Total	234804260	14.49	188249038	11.61	645464460	39.82	675577436	41.68

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 June 2023 doi:10.20944/preprints202306.0658.v1

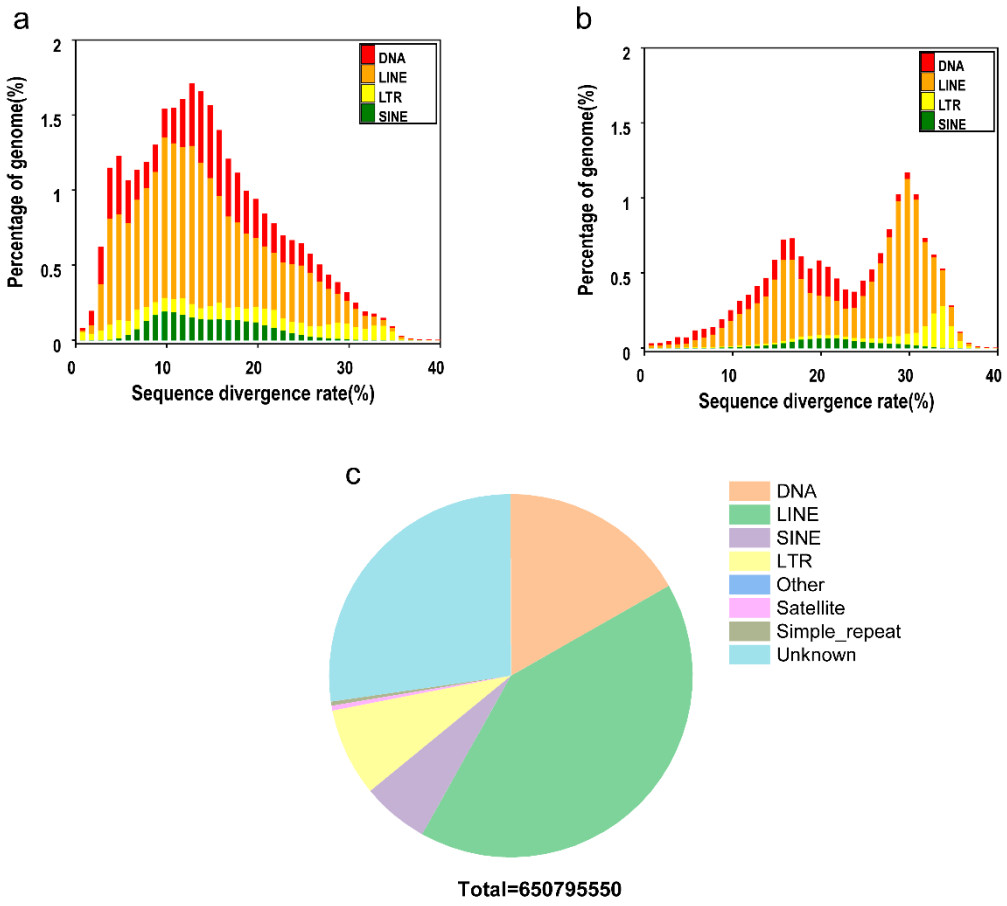

Figure 1. Distribution of transposable elements (TEs) in the *B. multicinctus* genome. The TEs include DNA transposons (DNA) and RNA transposons (i.e., DNAs, LINEs, LTRs, and SINEs). (a) De novo sequence divergence rate distribution. (b) Known sequence divergence rate distribution. (c) Proportion and distribution of repeating elements.

A total of 24,869 functional genes were identified in *B. multicinctus* and annotated with KEGG. The majority of these genes were found to be involved in pathways related to Environmental Information Processing and Metabolism. This suggests that signal transduction-related genes play an important role in B. multicinctus. (Figure 2). In addition, the enrichment of *B. multicinctus* genes in all metabolic pathways was found in twelve metabolic pathways. Among them, the most enriched is Lipid metabolism, and the smallest metabolic pathway is Biosynthesis of other secondary metabolites.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 June 2023 doi:10.20944/preprints202306.0658.v1

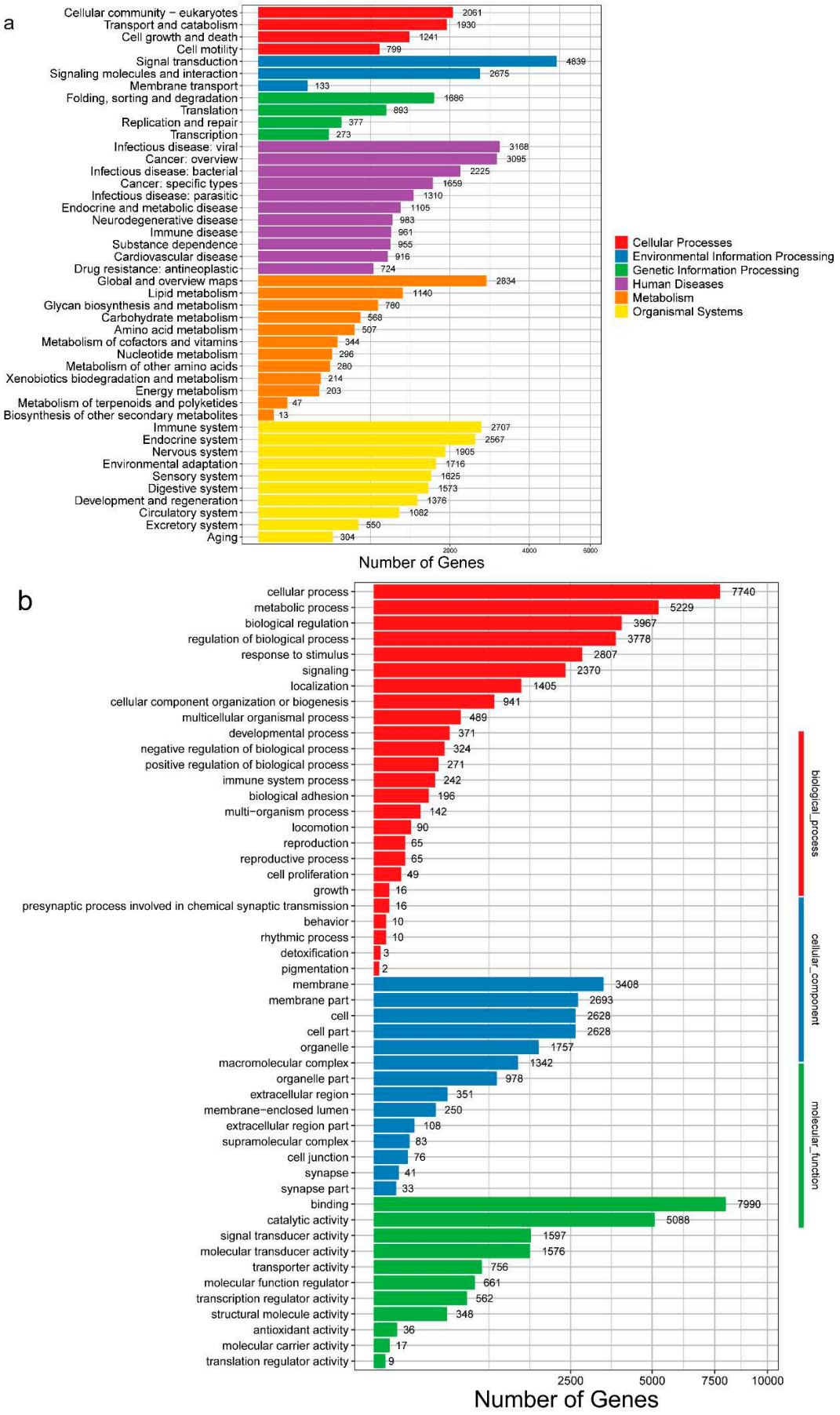

Figure 2. Gene annotation information of *B. multicinctus*. (a) KEGG enrichment of *B. multicinctus*. **(b)** GO enrichment of *B. multicinctus*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 June 2023 doi:10.20944/preprints202306.0658.v1

Data Validation and Quality Control

To evaluate the integrity of the assembly, we conducted BUSCO v3.1.0 (RRID:SCR_015008) assessment on the assembly [9]. The assembly captured 90.9% complete BUSCOs in the vertebrata_odb10 dataset. (Figure 3).

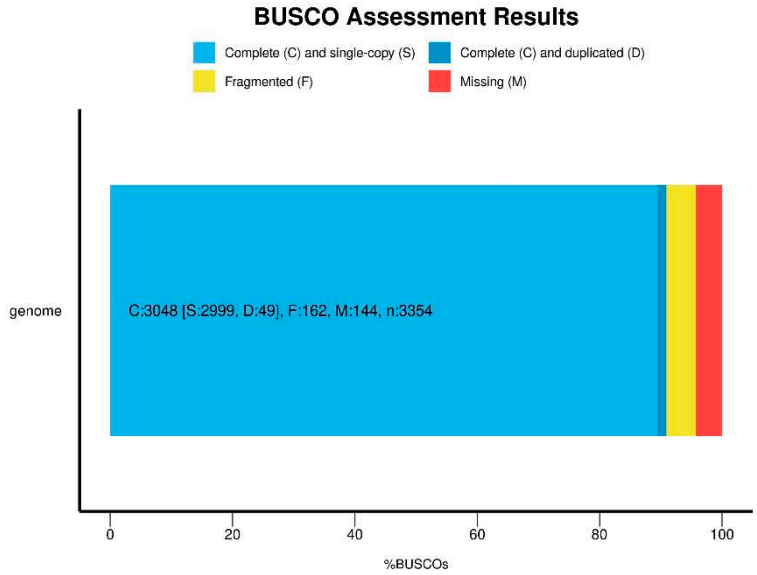

Figure 3. BUSCO Assessment result of the *B. multicinctus* genome.

To construct a phylogenetic tree, we screened closely related species, including Anolis carolinensis, Chelonia mydas, Danio rerio, Deinagkistrodon acutus, Gallus gallus, Homo sapiens, Mus musculus, Ophiophagus hannah, Python bivittatus, Xenopus tropicalis, and Alligator mississippiensis. Our data is consistent with previous studies and can be used to construct a phylogenetic tree that clusters closely related species [10]. (Figure 4)

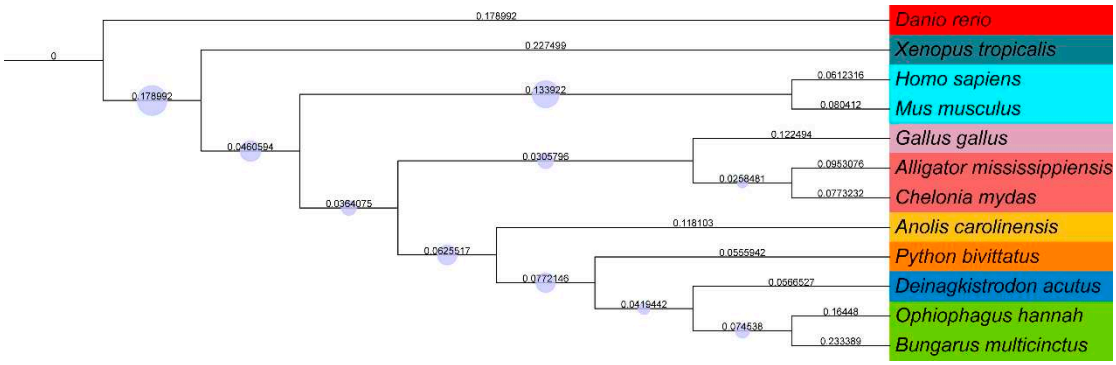

Figure 4. Phylogenetic tree reconstructed using nuclear genome single-copy genes.

Methods

Detailed stepwise protocols are gathered together in a protocols.io collection [dx.doi.org/10.17504/protocols.io.5jyl8j6e9g2w/v2], and can be summarised as follows.

Sample Collection and Sequencing

B. multicinctus specimens were collected from Beiliu Longgukeng, Guangxi, and immediately transferred to dry ice for quick freezing. The samples were then stored at -80 °C until further use. High-molecular-weight DNA was isolated using the protocol described by Wang et al. [11], and a

 doi:10.20944/preprints202306.0658.v1

stLFR co-barcoding DNA library was constructed using the MGIEasy stLFR Library Prep Kit (MGI, China). The libraries were sequenced using a BGISEQ-500 sequencer. In addition, genomic DNA was isolated using the AxyPrep genomic DNA kit (AxyPrep, USA) for whole-genome sequencing.

We extracted total RNA using TRlzol reagent (Invitrogen, USA) following the manufacturer's protocol. The quality, purity, and quantity of RNA were assessed using the Qubit 3.0 fluorometer (Life Technologies, USA) and the Agilent 2100 Bioanalyzer System (Agilent, USA). The cDNA libraries were generated by reverse-transcribing RNA fragments of 200-400 bp. All experimental procedures were approved by the Institutional Animal Care and Use Committee of Northeast Forestry University.

Genome Assembly, Annotation and Assessment

The stLFR sequencing data obtained from the manybanded krait were subjected to assembly using Supernova software (v2.1.1) [12]. To improve the quality of the assembly, GapCloser (v1.12-r6) and redundans (v0.14a) were utilized for gap filling and redundancy removal, respectively, by incorporating the WGS data.

To identify known repeat elements in the genome of the Many-banded Krait, Repeat Finder (TRF) [13] (v4.09), LTR_FINDER [14], and RepeatModeler[15] (v1.0.8) were utilized, and RepeatMasker [16] (v3.3.0) and RepeatProteinMask [17] (v3.3.0) were employed for repeat element annotation. Protein-coding genes were predicted using de novo, homology-based, and transcript mapping approaches. De novo gene prediction was performed using Augustus [18] (v3.0.3). RNA-seq data were filtered using Trimmomatic [19] (v0.30), and transcripts were assembled based on clean RNA-seq data using Trinity [20] (v2.13.2) for RNA-seq-based prediction. The Program to Assemble Spliced Alignments (PASA) [21] (v2.0.2) was utilized to align transcripts against the Many-banded Krait genome to obtain gene structures. Homology-based prediction was performed by mapping protein sequences of UniProt database (release-2020_05), *Pseudonaja textilis*, *Crotalus tigris*, *Thamnophis elegans*, and *Notechis scutatus* to the *B. multicinctus* genome using Blastall [22] (v2.2.26). Gene models were predicted by analyzing the alignment results using GeneWise [23] (v2.4.1). Finally, the MAKER pipeline [24] (v3.01.03) was employed to generate the final gene set, which represented RNA-seq, homology, and de novo predicted genes.

To perform functional annotation, a BLAST search was conducted against several databases, including SwissProt [25], TrEMBL [25], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [26], with an E-value cut-off of 1e-5. Furthermore, InterProScan [27] (v5.52-86.0) was applied to predict motifs and domains, as well as Gene Ontology (GO) terms.

The genome completeness was evaluated through the analysis of sets of Benchmarking Universal Single-Copy Orthologs (BUSCO v5.2.2) using genome mode and lineage data from vertebrata_odb10 [28], following standard scientific methodology.

To reconstruct the phylogenetic tree, OrthoFinder (v2.3.7) (RRID: SCR_017118) [29] to used to search for single-copy orthologs among the protein sequences of *Anolis carolinensis* (GCA_000090745.2), *Chelonia mydas* (GCA_015237465.2), *Danio rerio* (GCA_000002035.4), *Deinagkistrodon acutus* (http://gigadb.org/dataset/100196), *Gallus gallus* (GCA_016699485.1), *Homo sapiens* (GCA_000001405.29), *Mus musculus* (GCA_000001635.9), *Ophiophagus hannah* (GCA_000516915.1), *Python bivittatus* (GCA_000186305.2), *Xenopus tropicalis* (GCA_000004195.4) and *Alligator mississippiensis* (GCA_000281125.4). The number of orthogroups with all species were 7788.

Reuse Potential

Venomous animals have fascinated and influenced humans since ancient times, and the venom gland is a special evolutionary mechanism that snakes have developed to adapt to their ecological environment [30]. In recent years, ecosystems have undergone changes due to climate variations, and toxic species pose a threat not only to humans but also to native species and livestock [31,32]. Therefore, it is crucial to collect genomic resources of venomous snakes and explore the formation mechanism of venom glands and venom production.

Genome assembly of reptiles, including snakes, has always been a challenging task. However, we have observed that Xu et al. recently published an article on the origin of neurotoxins in the Elapidae family, based on a high-quality genome assembly of the Banded Krait [33]. Using third-generation sequencing and HIC, Xu et al. assembled the Banded Krait genome to the chromosome level, achieving a BUSCO score of 94.6% and a scaffold N50 of 149.80 Mbp. Our assembly resulted in a BUSCO score of only 90.9%. Although our assembly did not achieve the same level of genome continuity as Xu et al., we obtained a relatively complete genome of the Banded Krait using stLRF second-generation sequencing data. This provides a genomic resource for our future research exploring the evolution and origin of reptilian species, including snakes.

Our data can be combined with already published and new venomous snake genome data to construct the evolutionary history of venomous snakes and other reptiles. The genome data can also be used in venomics research for exploring the toxic gland genes and the mechanism of toxic gland production.

Author Contributions: Song Huang, Shuhui Yang and Yanchun Xu designed and initiated the project. Zhangwen Deng, Diancheng Yang and Yanan Gong collected the samples. Boyang Liu, Liangyu Cui and Yue Ma performed the DNA extraction and data analysis. Boyang Liu, Liangyu Cui and Zhangwen Deng wrote the manuscript. All authors read and approved the final manuscript.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study have been deposited into CNGB Sequence Archive (CNSA) [34] of China National GeneBank DataBase (CNGBdb) [35] with accession number CNP0004003. The data are also hosted in public databases with accession number PRJNA934116. Additional data is available in the GigaScience GigaDB repository [cite when ready].

Acknowledgments: This work was supported by the the Fundamental Research Funds for the Central Universities (No. 2572020DY02) and the Guangdong Provincial Key Laboratory of Genome Read and Write (grant No. 2017B030301011). This work was also supported by China National GeneBank (CNGB).

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

References

1. Uetz P, Koo M, Aguilar R, Brings E, Catenazzi A, Chang A, et al. A Quarter Century of Reptile and Amphibian Databases. Herpetological Review. 2021;52:246-55.
2. Wiens JJ, Hutter CR, Mulcahy DG, Noonan BP, Townsend TM, Sites JW, et al. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biology Letters. 2012;8 6:1043-6. doi:10.1098/rsbl.2012.0703.
3. Rao W-q, Kalogeropoulos K, Allentoft ME, Gopalakrishnan S, Zhao W-n, Workman CT, et al. The rise of genomics in snake venom research: recent advances and future perspectives. GigaScience. 2022;11:giac024. doi:10.1093/gigascience/giac024.
4. Liang Q, Huynh TM, Ng YZ, Isbister GK and Hodgson WC. In Vitro Neurotoxicity of Chinese Krait (Bungarus multicinctus) Venom and Neutralization by Antivenoms. Toxins. 2021. doi:10.3390/toxins13010049.
5. Tian Y, Liu Z, Ma L, Yu Y, Shi Q, Zhao S, et al. Forensic identification of a fatal snakebite from Bungarus multicinctus (Chinese krait) by pathological and toxicological findings: a case report. Forensic Science, Medicine and Pathology. 2022;18 4:497-502. doi:10.1007/s12024-022-00517-x.
6. <ARMC2.pdf>. doi:10.1016/j.
7. Castoe TA, de Koning APJ, Hall KT, Card DC, Schield DR, Fujita MK, et al. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proceedings of the National Academy of Sciences. 2013;110 51:20645-50. doi:10.1073/pnas.1314475110.
8. Sotero-Caio CG, Platt RN, II, Suh A and Ray DA. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biology and Evolution. 2017;9 1:161-77. doi:10.1093/gbe/evw264.
9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31 19:3210-2. doi:10.1093/bioinformatics/btv351.
10. Vidal N and Hedges SB. The molecular evolutionary tree of lizards, snakes, and amphisbaenians. Comptes Rendus Biologies. 2009;332 2:129-39. doi:https://doi.org/10.1016/j.crvi.2008.07.010.

11. Wang O, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res. 2019;29 5:798-808. doi:10.1101/gr.245126.118.
12. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27 5:757-67. doi:10.1101/gr.214874.116.
13. Benson and G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 27 2:573-80.
14. Zhao X and Hao W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research. 2007; suppl_2:suppl_2.
15. Smit A, Hubley R and Green P. RepeatModeler Open-1.0. 2008–2015. Seattle, USA: Institute for Systems Biology Available from: httpwww repeatmasker org, Last Accessed May. 2015;1:2018.
16. Tarailo-Graovac M and Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2009;Chapter 4 Unit 4:Unit 4.10.
17. Tempel S. Using and understanding RepeatMasker. Mobile Genetic Elements. Springer; 2012. p. 29-51.
18. Stanke M, Steinkamp R, Waack S and Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research. 2004;32 suppl_2:W309-W12. doi:10.1093/nar/gkh379.
19. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30 15:2114-20. doi:10.1093/bioinformatics/btu170.
20. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8 8:1494-512. doi:10.1038/nprot.2013.084.
21. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology. 2008;9 1:R7. doi:10.1186/gb-2008-9-1-r7.
22. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). CSH protocols. 2007;2007:pdb.top17. doi:10.1101/pdb.top17.
23. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. Genome Research. 2004;14 5:988-95. doi:10.1101/gr.1865504.
24. Campbell MS, Holt C, Moore B and Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. Current Protocols in Bioinformatics. 2014;48 1:4.11.1-4..39. doi:https://doi.org/10.1002/0471250953.bi0411s48.
25. Amos B and Rolf A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research. 2000; 1:45.
26. Pitk E. KEGG database. Novartis Foundation Symposium. 2006;247:91-103.
27. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30 9:1236-40. doi:10.1093/bioinformatics/btu031.
28. Wick RR and Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research. 2019;8.
29. Emms DM and Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology. 2015;16 1:157. doi:10.1186/s13059-015-0721-2.
30. von Reumont BM, Anderluh G, Antunes A, Ayvazyan N, Beis D, Caliskan F, et al. Modern venomics—Current insights, novel methods, and future perspectives in biological and applied animal venom research. GigaScience. 2022;11:giac048. doi:10.1093/gigascience/giac048.
31. Linardich C, Brookson CB and Green SJ. Trait-based vulnerability reveals hotspots of potential impact for a global marine invader. Global Change Biology. 2021;27 18:4322-38. doi:https://doi.org/10.1111/gcb.15732.
32. Giallongo G, Douek J, Harbuzov Z, Galil BS and Rinkevich B. Long-term changes in population genetic features of a rapidly expanding marine invader: implication for invasion success. Biological Invasions. 2021;23 8:2541-52. doi:10.1007/s10530-021-02521-8.
33. Xu J, Guo S, Yin X, Li M, Su H, Liao X, et al. Genomic, transcriptomic, and epigenomic analysis of a medicinal snake, Bungarus multicinctus, to provides insights into the origin of Elapidae neurotoxins. Acta Pharmaceutica Sinica B. 2023;13 5:2234-49. doi:https://doi.org/10.1016/j.apsb.2022.11.015.

34. Guo X, Chen F, Gao F, Li L, Liu K, You L, et al. CNSA: a data repository for archiving omics data. Database. 2020;2020 2020:baaa055.
35. Feng ZC, Li JY, Fan Y, Li NW and Xiao FW. CNGBdb: China National GeneBank DataBase. Hereditas. 2020;42 8:799-809.