De novo genome assembly and annotation of the saw-scaled viper *Echis carinatus*, a medically important venomous snake in the Indian Subcontinent

Avni Blotra, Anukrati Sharma, Divya Tej Sowpati, Karthikeyan Vasudevan*

CSIR- Centre for Cellular and Molecular Biology, Hyderabad, India 500007

* Corresponding author: karthik@ccmb.res.in

ABSTRACT

Echis carinatus is a widely distributed venomous snake in the Indian subcontinent. It is one of the "Big Four" snake species responsible for mortality and severe health complications caused by envenomations. Given the significance of the species for human health, we assembled and annotated the whole genome of *E. carinatus*. Using long reads for assembly and short reads for correction, we achieved a highly contiguous 1.57Gb assembly, with a contig N50 of 17.1 Mb and scaffold N50 of 23.5 Mb. We could map 96.58% of our reads back to the assembled genome, and Benchmarking Universal Single-Copy Orthologs (BUSCO) score of 97.4%, indicates near completeness. 48.59% of the genome had repeat elements. The predominant types of repeats include retrotransposons (LINEs, SINEs and LTR elements), DNA transposons and satellite DNA (SSRs). Additionally, we conducted a gene prediction analysis, identifying a total of 26,711 protein-coding genes. These genes included alternatively spliced products, leading to the prediction of 31,106 proteins, 77.6% of which were successfully annotated using *Blastp* and *InterProScan*. Out of the total predicted gene set, we identified 119 toxin-coding genes, crucial for venom toxin production in the species. The acquisition of toxin sequence-based information resulted in a valuable dataset of genes that are involved in venom production in *E. carinatus*. This dataset advances our understanding of the biological processes underlying venom production and lays the foundation for the development of new antivenom formulations.

Keywords: Snakebite, High throughput sequencing, Hybrid assembly, Toxin gene annotation, Venom evolution

INTRODUCTION

Snakes are limbless reptiles belonging to the suborder Serpentes. Despite undergoing multiple evolutionary changes that led to their widespread distribution, our understanding of snake biology remains poor. There are 3900 recognized, extant snake species in the world, out of which, roughly 600 are venomous (Uetz P. 2022). Snake venom is a mixture of proteins and peptides that act individually or in synergy to incapacitate the prey by acting on a wide range of targets in the body (Xiong & Huang, 2018). Even though all the venomous snakes have the same evolutionary origin, they show a high degree of intra and interspecific venom variation (Chippaux et al., 1991). Extensive homoplasy at the lineage level and rapid adaptations at different geographies make snake venoms an excellent model system to understand evolution of traits under constant natural selection.

Around 10% of the world's venomous snake species are found in India, and human-snake interactions are inevitable because of the high density of human population. An outcome of this interaction causes envenomation in humans. This poses a significant health challenge across various regions within the Indian subcontinent. A majority of fatal bites are caused by the "Big four", which are the Indian cobra (*Naja naja*), Russell's viper (*Daboia russelli*), saw scaled viper (*Echis carinatus*), and common krait (*Bungarus caeruleus*). Because of the knowledge gaps on the venoms and the outcomes of envenomation, snakebites are considered a neglected tropical disease (Waiddyanatha et al., 2019).

Echis carinatus has a potent hemotoxic venom that damages tissue and induces symptoms of coagulopathies (Slagboom et al., 2017). Our present understanding of *E. carinatus* venom is only based on venom proteomes, which has provided insight into the complex makeup of venom. Nonetheless, our ability to identify the venom toxins unambiguously is limited, as key isoforms of toxins responsible for diverse pathologies still remain unidentified (Bhatia & Vasudevan, 2020; Casewell et al., 2020). Genomes of venomous snakes are woefully inadequate, thereby a comprehensive library of snake venom toxins has eluded us (Suryamohan et al., 2020).

Therapeutics for snakebite management require a complete understanding of the venom; its composition and the mode of action. Considerable progress has been made in recent years in understanding the composition of snake venoms. Modern whole genome sequencing platforms provide an opportunity to make significant contributions in venom research by providing insights into venom evolution and identification of core toxin coding genes that might improve snakebite treatment strategies.

We used a comprehensive approach involving various genome sequencing platforms and bioinformatic tools to assemble a precise and contiguous de novo whole genome of *Echis carinatus*. The preliminary draft assembly was created using nanopore long read sequences, with subsequent iterative refinement. The errors were removed using long reads followed by polishing with quality-filtered Illumina short read sequences. We also explored the landscape of repeat elements in the genome using repeat masking tools. By implementing a novel assembly strategy, we significantly elevated the accuracy and completeness of the *E. carinatus* genome. Additionally, we conducted RNA-seq analysis of tissue samples and created a well-curated gene library, which was used to predict genes and annotate the genome. This well-annotated genome along with the transcriptome can form the basis for transforming snakebite therapy and genetic studies on snakes.

MATERIALS AND METHODS

Sample processing information

Five *E. carinatus* were sourced from different parts of the state of Telangana. Venom glands and other organs were collected after snake's post-mortem. All snakes were weighed, measured and

sexed using protocols described in the MSD Veterinary Manual (https://www.msdvetmanual.com). Genomic DNA was extracted using the Monarch® HMW DNA Extraction Kit for Tissue. To preserve the integrity of RNA, all organs were meticulously collected and stored in RNA later until the RNA extraction process, which was performed using the Qiagen RNeasy kit (Qiagen).

Genome Sequencing

High molecular weight DNA was extracted from liver of a male *E. carinatus* (EC01) using Monarch® HMW DNA Extraction Kit for Tissue (Cat#T3060L). DNA quantity, purity and integrity were assessed by Qubit fluorometer. Two types of sequencing data were generated: Illumina short-reads and nanopore long reads. For short-reads sequencing, libraries with insert size of 350 bases were prepared and sequenced on an Illumina NovaSeq 6000 platform to generate 150 bp paired-end reads. A total of 710 million short reads were generated. After quality trimming, removing adapters and eliminating the low-quality reads, 107 Gb data remained, which was used for further analysis.

Libraries for long reads were created using a ligation sequencing DNA V14 kit, SQK-LSK-114, using 1µg of high molecular weight genomic DNA as input. The nanopore libraries were sequenced on the PromethION24 sequencer. A total of 94 Gb data, with read N50 of 13.9 kb was generated, giving an overall coverage of 60x based on the estimated genome size. The fast5 files were basecalled into *fastq* format in super accuracy mode using *Guppy v6.4.6* basecaller. These raw files were quality checked and pre-processed before using it for further analyses.

RNA-sequencing

Ribonucleic acid (RNA) was extracted using the TRlzol reagent (Invitrogen, USA) followed by the Qiagen RNeasy kit (Cat #74004) from various tissues, including the venom gland, brain, heart, liver, lung, kidney, intestine, testis and ovaries. The quality, purity, and quantity of the extracted RNA were assessed using the HS RNA Qubit fluorometer and the Agilent 4200 tapestation System (Agilent, USA). Subsequently, 1 µg of RNA from each tissue was utilized for the preparation of RNA-sequencing (RNA-seq) libraries with the Illumina stranded total RNA kit. 150 bp paired end reads were generated on the NovaSeq6000 platform (Illumina). For each tissue around 50-60 million reads were generated. Following the quality check and control of the RNA-seq data, we employed the Hisat2 tool to map the data on the genome.

Quality check of raw reads

DNA Seq- To construct a high-fidelity genome, we employed both long and short-read third-generation sequencing technologies. An initial assessment revealed that ONT reads had an average q score of 15.3, with 97.6% of them exceeding the default quality threshold of 12. We utilized *Porechop v0.2.4* (Wick et al., 2017) and *Nanofit v2.8.0* (De Coster & Rademakers, 2023) to remove adapters and filter high quality reads, with a cutoff set at 12. These preprocessing

steps boosted the quality to 17.3, enhancing the reliability of subsequent analyses. For short reads, *Cutadapt* v4.0 (Martin, n.d.) was employed to eliminate poly A's and G's, which were overrepresented in the *fastQC* report. *MultiQC* report reads with quality scores <25 was removed using a *Cutadapt* quality score filter along with the adapter sequences.

RNA Seq- From five *E. carinatus* - heart, kidney, brain, venom gland, intestine, lungs, liver, testis, and ovaries were collected. RNA from the tissues were isolated and sequenced. A quality assessment showed the total reads ranged from 40 to 60 million across samples. *Cutadapt* was applied to filter reads, eliminating adapters and employing a quality score cutoff of 25. Hisat2 was utilized to align RNA sequences from these tissues to the genome, with consistent mapping percentages exceeding 80% for each sample.

Estimation of genome size and coverage

The genome size was estimated using a K-mer frequency-based method. Using *Kmergenie* (Marçais & Kingsford, 2011) we arrived at the best K-mer size and counted the unique kmers using the *Jellyfish* 2.3.0. *GenomeScope* was used to determine the frequency of K-mers, the genome size and the coverage.

De novo genome assembly

To assemble the genome of *E. carinatus*, *Flye 2.9.2* (Kolmogorov et al., 2019) was used with regular ONT data "--nano-raw" that works well for the FLO-PRO114M, R10.4.1 chemistry flowcells. The assembly was enhanced by the "-scaffold" parameter, a feature within *Flye*, which introduced Ns (nucleotide placeholders) to arrange contigs into coherent scaffolds. A statistical summary of assembly quality was generated using *QUAST 5.2.0* (Gurevich et al., 2013).

Statistics to assess assembly quality

Numerical statistics, such as N50 are often misleading and could lead to erroneous assessment of the quality of assembly. Therefore, we used a robust approach using biological information-based comparison of multiple assemblies, widely referred to as Benchmarking Universal Single-Copy Orthologs (BUSCO). We used this alignment against the vertebrate lineage database to look at the coverage of the orthologs in our assembly. Furthermore, the accuracy of the assembly and the overall quality was quantified using *Inspector v1.0.2* (Chen et al., 2021). It identified errors after mapping of the long reads over the contigs. Subsequently, a polishing procedure using *Pilon* v1.24 (Walker et al., 2014) with high quality illumina data was carried out.

Repeat element identification

The simple repeat sequences (SSRs) and all tandem repeat elements in the whole genome were annotated using *RepeatModeler v2.0.4* (https://github.com/Dfam-consortium/RepeatModeler). The *de novo* predictions were made using a custom Dfam library - an open database of transposable elements families, *TRF*- tandem repeat finder and *RepeatScout* 1.0.6 - De novo

repeat finder. The constructed library was given as an input for *RepeatMasker v4.0.7* (http://www.repeatmasker.org) to predict the repeats, which were masked using a combination of de novo and annotated repeat databases for downstream analyses

Gene prediction and annotation

We used the *BRAKER v3.0.7* annotation pipeline for comprehensive gene annotation. It employs GeneMark-ETP and AUGUSTUS for accurate gene training and prediction and integrates de novo prediction, homology-based searches, and transcriptome/proteome-assisted methods. Additionally, a custom database subset of Sauropsida in the Vertebrata database was incorporated to enhance annotation efficiency. For a comprehensive functional annotation of genes, we employed the *Blastp* function using *Diamond v2.0.9.147*. Hits were then filtered using the following parameters, E-value threshold of 1.0×10^{-4}, a query coverage > 70% and identity >70%. We then aligned the protein sequences from the finalized gene set to the GenBank *Non-Redundant Protein Database (NRDB)* to reveal sequence similarities and potential functional associations. We incorporated *InterProScan* to further enrich our annotation. This application aided in predicting essential motifs and domains, and facilitated the retrieval of Gene Ontology (GO) terms, to create a nuanced and informative characterization of the genes. To investigate the presence of potential toxin-coding genes within our predicted set, we utilized the Animal Toxin Annotation Project (Tox-Prot) database. We applied two distinct filtering criteria: Relaxed criteria, with query coverage >50%, identity > 50% and an E-value $<1 \times 10^{-4}$, and more stringent criteria, identity > 80% and query coverage >80%.

RESULTS

De novo genome assembly

The genome size of *E. carinatus* was estimated to be 1.57Gb with an optimal K-mer size of 99 as reported by *Kmergenie*. The coverage was calculated to be 60x, as the available long read data size was approximately 94Gb. To generate a contiguous assembly, we filtered out the small and fragmented sequences. We used the reads that were greater than 10 Kb in length and it resulted in a reduction of sequence coverage to 33x. However, this enhanced the N50 metric. The final assembly was 1.57 Gb in size with 910 contigs. After refinement the assembly had 711 contigs that improved the N50 from 17.1 Mb to 23.5 Mb. 90% of the assembly had 70 contigs (L90) and 99.3 % of the assembly contained contigs over 50kb.

Assembly Evaluation and correction

The assembly completeness score (BUSCO) was 97.4%. Out of the total 3,327 conserved genes screened in the vertebrate lineage database, 3,239 were complete, 25 were fragmented, and 62 were missing BUSCOs (Table1). *Inspector* tool was used to map long reads onto assembly contigs to identify errors. This step achieved a mapping rate of 96.58% and initially reported a

Quality Value (QV) of 33.2. Additionally, it detected 254 minor assembly errors per million base pairs and introduced 113 Ns during scaffold-level assembly. The approach of correcting the assembly using long reads enhanced the QV to 36.7 and reduced the number of Ns to 33. The minor assembly errors were brought down to 36 per Mbp. After further polishing using the short illumina reads, the assembly metrics had 711 contigs with N50 of 23.5 Mb, and L90 of 69.

Repeat Annotation

We were able to mask 6.74% of repeats in the genome with the Viperidae repeat database. We could construct a custom library with Dfam and Viperidae database and mask 48.59% of repeats (Table 2; Fig. 3). 3.4 % SSRs were masked using our in-house tool DiviSSR v0.1.2 (Avvaru et al., 2021).

Gene prediction and annotation

We used *Braker* pipeline to annotate the genome with gene expression data from 8 different tissues. We predicted 26,711 protein-coding genes, with 74.8% supported by either RNA-Seq data or protein databases. This gene set encoded a total of 31,106 predicted proteins including the alternatively spliced products, each with corresponding orthologs in NCBI's non-redundant database. *Blastp* analysis facilitated the successful annotation of 22,219 transcripts (71.4%), corresponding to 20,113 distinct genes. Subsequently, integrating *InterProScan* resulted in the successful annotation of 24,159 transcripts that improved the annotation from 71.4% to 77.6%. Notably, 29,331 transcripts (94.2%) had canonical start and stop codons.

The genome assembly consisted of 711 fragments, out of which, 678 were contigs and 33 were scaffolds. We could map 54.6 % of functionally annotated genes on 32 scaffolds and the remaining genes on 186 contigs. We initially identified 4,407 toxin coding transcripts that met the relaxed criteria out of the 7,839 protein sequences in Tox-Prot. By employing a stringent criteria, we identified 119 toxin-coding genes, crucial for venom toxin production in the species. The completeness of the annotation was 97.6% with the Vertebrate database and 95% with the Sauropsida database (Table 3).

Future Perspective

The Indian subcontinent faces a huge burden of snakebite deaths (Suraweera et al., 2020). Out of the common four venomous snake species, only *Naja naja* has whole genome and transcriptome data. The lack of genomic and transcriptomic data has impeded our understanding of venom toxin evolution and the mechanisms that regulate toxin-gene expression. With the *E. carinatus* genome and transcriptome, we will perform differential expression analysis and refine the venom proteome using a species-specific protein database. This would provide insights into the molecular basis of venom production. More importantly, an in-depth understanding of the toxin

gene sequences will lay the foundation for developing targeted therapies for snake bite envenomations.

Data availability

Raw and assembled sequence data will be available via the NCBI GEO

Acknowledgements

We are grateful to the Telangana Forest Department for permitting us to collect saw scaled viper and its venom under order no – Rc. No.26803/2012/WL-2 dated 08.04.2021. A. Srinivas and N. Tulasi from the CCMB's NGS facility generated the sequencing data. Sampath Kumar, Alka Sahu, Gopi Krishnan, Omkar Yadav, K. Rajyalakshmi and Harika Katakam extended help for this work. Anita Malhotra facilitated AB's research through Bangor University, Taith Research Mobility Scheme for student exchange grant. Aaron Comeault and Axel Barlow provided bioinformatics training for the whole genome analysis.

REFERENCES

1. Avvaru, A.K., Mishra, R. K., Sowpati, D. T. (2021). DiviSSR: Simple arithmetic for efficient identification of tandem repeats; bioRxiv 2021.10.05.462997, doi: https://doi.org/10.1101/2021.10.05.462997
2. Bhatia, S., & Vasudevan, K. (2020). Comparative proteomics of geographically distinct saw-scaled viper (Echis carinatus) venoms from India. Toxicon: X, 7, 100048. https://doi.org/10.1016/j.toxcx.2020.100048
3. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. NAR Genomics and Bioinformatics 3(1), Lqaa108.
4. Bruna, T., Lomsadze, A., & Borodovsky, M. (2023). GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistence with Extrinsic Data. BioRxiv, 2023-01.
5. Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods, 18(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x
6. Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59–60. https://doi.org/10.1038/nmeth.3176
7. Casewell, N. R., Jackson, T. N. W., Laustsen, A. H., & Sunagar, K. (2020). Causes and Consequences of Snake Venom Variation. Trends in Pharmacological Sciences, 41(8), 570–581. https://doi.org/10.1016/j.tips.2020.05.006
8. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., & Chong, Z. (2021). Accurate long-read de novo assembly evaluation with Inspector. Genome Biology, 22(1), 312. https://doi.org/10.1186/s13059-021-02527-4
9. Chippaux, J.-P., Williams, V., & White, J. (1991). Snake venom variability: methods of study, results and interpretation. Toxicon, 29(11), 1279–1303. https://doi.org/10.1016/0041-0101(91)90116-9
10. de Coster, W., & Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics, 39(5). https://doi.org/10.1093/bioinformatics/btad311
11. Gotoh, O. (2008). A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Research, 36(8), 2630–2638. https://doi.org/10.1093/nar/gkn105
12. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a

reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883
13. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086
14. Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with BRAKER. Gene Prediction (Pp. 65-95). Humana, New York, NY.
15. Iwata, H., & Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Research, 40(20), e161–e161. https://doi.org/10.1093/nar/gks708
16. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics (Oxford, England), 30(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
17. Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4
18. Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology, 37(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8
19. Kovaka, S., Zimin, A. v., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology, 20(1), 278. https://doi.org/10.1186/s13059-019-1910-1
20. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
21. Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k -mers. Bioinformatics, 27(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011
22. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10. https://doi.org/10.14806/ej.17.1.200
23. Merck & Co., Inc. , & Rahway, N. (n.d.). MSD Veterinary Manual (MVM). Https://Www.Msdvetmanual.Com/.
24. Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Research, 9, 304. https://doi.org/10.12688/f1000research.23297.2
25. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 33(3), 290–295. https://doi.org/10.1038/nbt.3122
26. Slagboom, J., Kool, J., Harrison, R. A., & Casewell, N. R. (2017). Haemotoxic snake venoms: their functional activity, impact on snakebite victims and pharmaceutical promise. British Journal of Haematology, 177(6), 947–959. https://doi.org/10.1111/bjh.14591

27. Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008a). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013
28. Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008b). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013
29. Stanke. M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62.
30. Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics, 7(1), 62. https://doi.org/10.1186/1471-2105-7-62
31. Suraweera, W., Warrell, D., Whitaker, R., Menon, G., Rodrigues, R., Fu, S. H., Begum, R., Sati, P., Piyasena, K., Bhatia, M., Brown, P., & Jha, P. (2020). Trends in snakebite deaths in India from 2000 to 2019 in a nationally representative mortality study. ELife, 9. https://doi.org/10.7554/eLife.54076
32. Suryamohan, K., Krishnankutty, S. P., Guillory, J., Jevit, M., Schröder, M. S., Wu, M., Kuriakose, B., Mathew, O. K., Perumal, R. C., Koludarov, I., Goldstein, L. D., Senger, K., Dixon, M. D., Velayutham, D., Vargas, D., Chaudhuri, S., Muraleedharan, M., Goel, R., Chen, Y.-J. J., … Seshagiri, S. (2020). The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. Nature Genetics, 52(1), 106–117. https://doi.org/10.1038/s41588-019-0559-8
33. The reptile database. Zoological Museum Hamburg. http://reptile-database.org. (2022). The reptile database. Zoological Museum Hamburg. http://reptile-database.org.
34. Waiddyanatha, S., Silva, A., Siribaddana, S., & Isbister, G. (2019). Long-term Effects of Snake Envenoming. Toxins, 11(4), 193. https://doi.org/10.3390/toxins11040193
35. Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963
36. Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics, 3(10). https://doi.org/10.1099/mgen.0.000132
37. Xiong, S., & Huang, C. (2018). Synergistic strategies of predominant toxins in snake venoms. Toxicology Letters, 287, 142–154. https://doi.org/10.1016/j.toxlet.2018.02.004

Tables and Figures

Table 1: Representing the BUSCO completeness statistics of the *Echis carinatus* genome assembly using vertebrate lineage database

BUSCO Statistics	
Complete BUSCO	3267 (97.4%)
Complete and single copy BUSCOs	3239 (96.6%)
Complete and duplicated BUSCOs	28 (0.8%)
Fragmented BUSCOs	25 (0.7%)
Missing BUSCOs	62 (1.9%)
Total BUSCO groups searched	3354

Table 2: Repeat element distribution in *Echis carinatus* genome

Repeat element distribution	
Retrotransposons	**21.88%**
LINEs	15.9%
L2/CR1/Rex	6.9%
R2/R4/NeSL	0.66%
RTE/Bov-B	4.1%
L1/CIN4	2.8%
LTR elements	5.98%
BEL/Pao	0.06%
Ty1/Copia	0.65%
Gypsy/DIRS1	4.98%
Retroviral	0.29%
DNA transposons	**0.79%**
hobo-Activator	0.13%
Tc1-IS630-Pogo	0.65%
Unclassified	**23.63%**
Satellites	**2%**

Table 3: Representing the BUSCO completeness statistics of the *Echis carinatus* transcriptome using two lineage databases

	Vertebrate database	**Sauropsida database**
Complete BUSCO	3274 (97.6%)	7105 (95%)
Complete and single copy BUSCOs	2877 (85.8%)	5907 (79%)
Complete and duplicated BUSCOs	397 (11.8%)	1198 (16%)
Fragmented BUSCOs	25 (0.7%)	42 (0.6%)
Missing BUSCOs	55 (1.7%)	42 (4.4%)
Total BUSCO groups searched	3354	7480

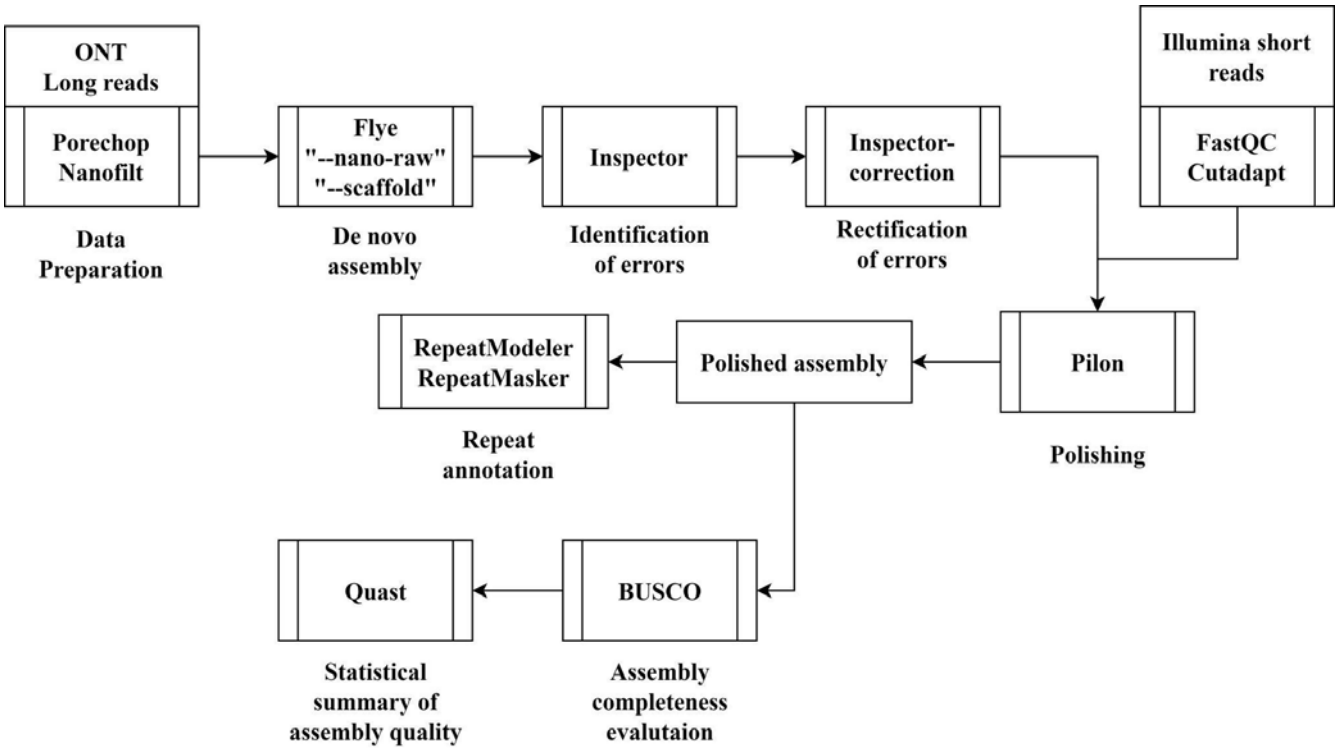

Fig.1: Overview of the de novo assembly pipeline representing the steps and the tools involved

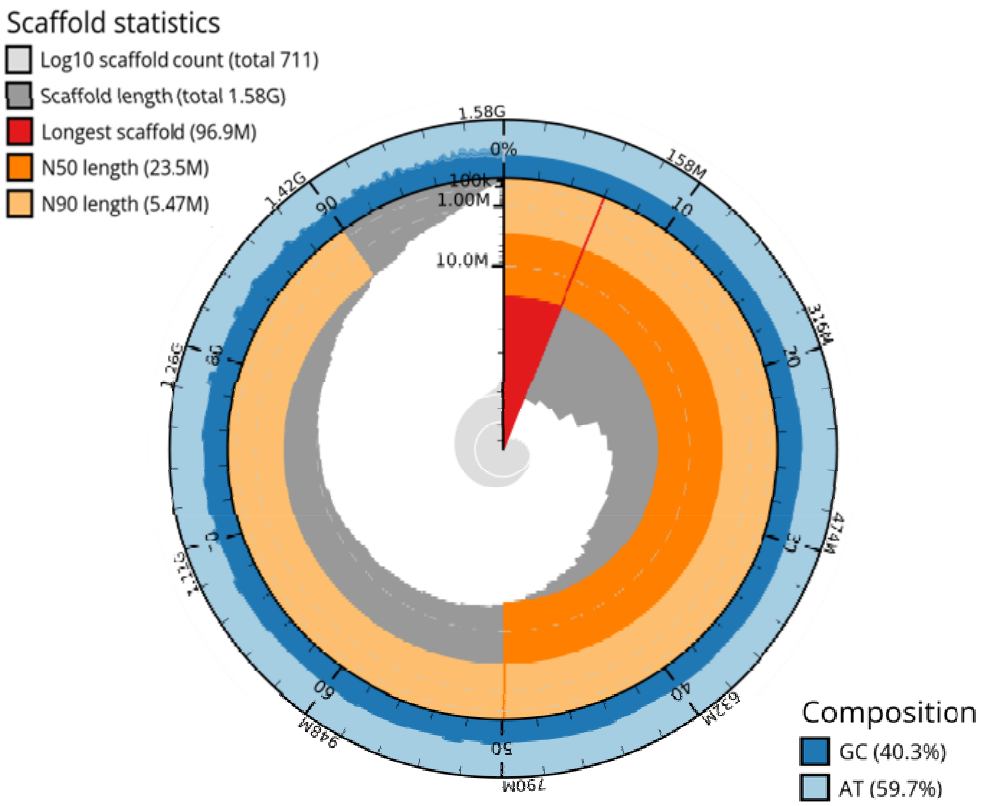

Fig. 2: Snail plot summarizing metrics of the *E. carinatus* genome including the length of the longest contig (96.9 Mb; red line), N50 (23.5 Mb; dark orange), N90 (5.47 Mb; light orange), and base composition.

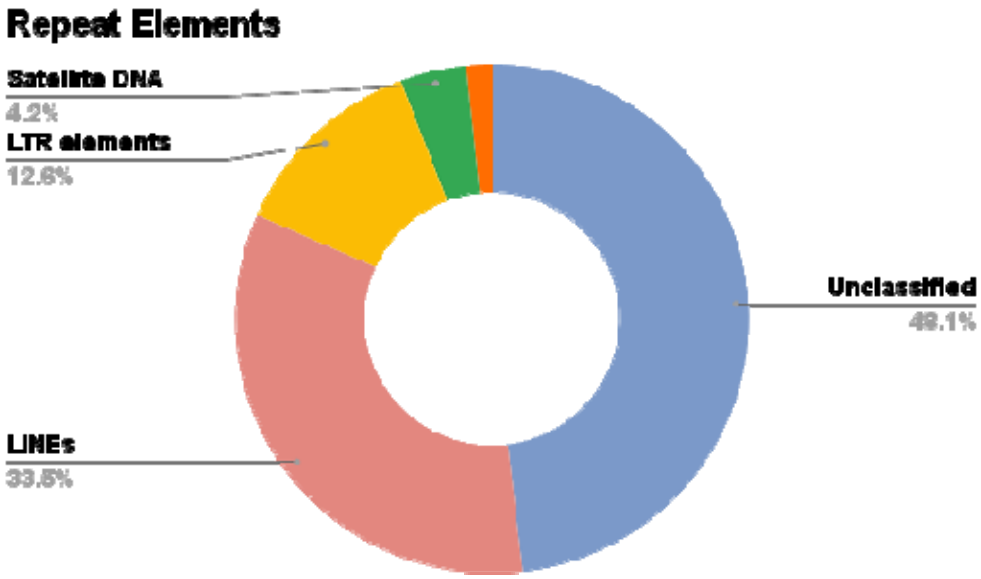

Fig 3: Plot displaying the predominant types of repeat elements present in *Echis carinatus* genome