De novo genome assembly 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 set out to assemble the whole genome of *E. carinatus*. Using only 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 23.4 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%, indicate near completeness. 48.59% of the genome was repeat-masked using a combination of de novo and annotated repeat databases. The predominant types of repeats include retrotransposons (LINEs, SINEs and LTR elements), DNA transposons and satellite DNA (SSRs). Our work would serve as a resource to understand the role of genes responsible for venom related adaptations, and to design and develop better 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 ~4000 extant snake species spread across the world, out of which around 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 could result in the envenomation of humans. Snake envenomation 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 about the venoms and the outcomes of envenomation, snakebites are considered a neglected tropical disease.

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 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 complete library of snake venom toxins has eluded us (Suryamohan et al., 2020).

Therapeutics for snakebite management requires 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 this novel assembly strategy, we significantly elevated the accuracy and completeness of the *E. carinatus* genome. This genomic asset has the capacity to transform snakebite therapy and lays a significant groundwork for upcoming genetic and biomedical research endeavors.

MATERIALS AND METHODS

Genome Sequencing

High molecular weight DNA was extracted from liver of male saw scaled Viper 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 were generated, giving an overall coverage of 59x based on the estimated genome size. The read N50 value was 13.9 kb. The fast5 files were basecalled into fastq format using *Guppy v6.4.6* basecaller in super accuracy mode. The fastq files were quality checked and pre-processed before analysis.

Quality check of raw reads

To generate a high-quality genome of *E. carinatus*, we employed long and short read third generation sequencing technology. The average quality of raw long reads was determined to be 15.3. The quality filtering step involved removal of adapter content and low-quality bases which was done using *Porechop v0.2.4* (Wick et al., 2017) and *Nanofit v2.8.0* (De Coster & Rademakers, 2023) respectively. For short reads *Cutadapt v4.0* (Martin, n.d.) was used to remove the poly As and Gs from the reads as those reads were overrepresented in the fastQC report.

Estimation of genome size and Coverage

The genome size was estimated using a K-mer frequency-based method. We first used *Kmergenie* (Marçais & Kingsford, 2011) to find the best K-mer size followed by counting the unique kmers using the Jellyfish 2.3.0. *GenomeScope* tool was used to determine the frequency of K-mers which assisted us in determining the genome size and 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 most recent FLO-PRO114M, R10.4.1 chemistry flowcells. The assembly was enhanced by implementing the "-scaffold" parameter, a feature within *Flye*, which strategically introduced Ns (nucleotide placeholders) to arrange contigs into coherent scaffolds. Statistical summary of assembly quality was generated by using *QUAST 5.2.0* (Gurevich et al., 2013).

Statistics to assess assembly quality

Numerical statistics like N50 are primitive and can often be misleading in identifying the best assembly. Therefore, biological information-based comparison of multiple assemblies is more a robust approach. One such widely used method is Benchmarking Universal Single-Copy Orthologs (BUSCO) that helps in quantitative assessment of annotatability of the assembled genome or transcriptome by calculating the percentage of evolutionarily conserved gene content. We have used BUSCO 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). This 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 annotation

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 done 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.

RESULTS

Quality check

Sequencing yielded a dataset of 710 million Illumina short reads with average length of 151 bases and ONT long reads with 13.9Kb of N50 and total size of 94 Gb. Initial quality assessments revealed an average quality score of 15.3 for the ONT reads, with 97.6% of the reads exceeding a quality threshold of 12. We used *porechop* and *Nanofilt* to remove adapters and filter high quality reads. The cutoff for filtering high quality reads was set at 12. The pre-processing steps enhanced the quality to 17.3 which improved the reliability of the downstream analysis.

De novo assembly

We Estimated the genome size to be 1.57Gb with the optimal K-mer size of 99 reported by *Kmergenie.* The coverage was calculated to be 60x as the available long read data size was around 94Gb. We worked with reads greater than 10 Kb in length to ensure a more contiguous assembly. This ensured a sequence coverage of 33x. The final assembly spans 1.57 Gb size, initially represented by 910 contigs. The *Flye* assembler used for scaffolding led to an assembly of 866 contigs that amplified N50 from 17.1 Mb to 23.4 Mb. *QUAST* analysis revealed the quality and completeness where 90% of assembly was covered by 70 contigs and with 99.3 % of the assembly contained in contigs over 50kb.

Assembly Evaluation

The completeness of the assembly was evaluated using BUSCO analysis. The BUSCO scores are critical in determining the quality of genome assembly. The assembly completeness score was 97.4%, indicating that it contained a substantial portion of the expected conserved gene content. Specifically, out of the total 3,327 vertebrate conserved genes screened from the vertebrate lineage database, 3239 were complete, 25 were fragmented, and 28 were missing BUSCOs (Table1).

Correction

For the identification of errors in the assembly mapping long reads over the assembly contigs is one of the steps of *Inspector* that showed 95.58% of mapping and reported an initial Quality Value (QV) of 33.2, alongside 254 minor assembly errors per Mbp, and the introduction of 113 Ns during scaffolding. The approach of correcting the assembly using long reads increased the QV to 36.7 and reduced the number of Ns to 33 where small-scale errors were only 36 per Mbp. Further polishing was done using the short illumina reads with *pilon*. Finally, the assembly metrics were reported using *Quast* that had 711 contigs with N50 of 23.5 Mb and L90 of 69.

Repeat Annotation

RepeatMasker masked 6.74% of repeats in the genome with the Viperidae repeat database. *RepeatModeler* was used to construct a custom library with Dfam database that was used along with Viperidae database in *RepeatMasker* and could mask 48.59% of repeats (Fig. 3). We masked 3.4 % SSRs using inhouse tool DiviSSR v0.1.2 that reported the locations of short tandem repeats in the genome.

Future Perspective

The Indian subcontinent is a global hotspot of snakebite deaths (Suraweera et al., 2020). With the exception of *Naja naja*, other venomous snakes of the region lack whole genome data. This has impeded our understanding of venom toxin evolution and the mechanisms that regulate toxin-gene expression. With this data, we will combine the venom gland and multi-tissue transcriptome of the species to gain a comprehensive understanding of the evolutionary pressure that shaped venom production and delivery systems. More importantly, an in-depth understanding of the toxin gene sequences will lay the foundation for developing targeted therapies for envenomations.

Data availability

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

REFERENCES

1. Bhatia, S., & Vasudevan, K. (2020). Comparative proteomics of geographically distinct saw-scaled viper (Echis carinatus) venoms from India. *Toxicon: X*, *7*. https://doi.org/10.1016/j.toxcx.2020.100048
2. Casewell, N. R., Jackson, T. N. W., Laustsen, A. H., & Sunagar, K. (2020). Causes and Consequences of Snake Venom Variation. In *Trends in Pharmacological Sciences* (Vol. 41, Issue 8, pp. 570–581). Elsevier Ltd. https://doi.org/10.1016/j.tips.2020.05.006
3. 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). https://doi.org/10.1186/s13059-021-02527-4
4. Chippaux, J.-P., Williams2, V., White', J., Chippaux, J. .-P., & Williabis, V. (1991). Snake venom variability: methods of study, results and interpretation. In *Adelaide Childrens Hospital* (Vol. 29, Issue 11).
5. 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
6. 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
7. 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
8. 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
9. 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
10. Slagboom, J., Kool, J., Harrison, R. A., & Casewell, N. R. (2017). Haemotoxic snake venoms: their functional activity, impact on snakebite victims and pharmaceutical promise. In *British Journal of Haematology* (Vol. 177, Issue 6, pp. 947–959). Blackwell Publishing Ltd. https://doi.org/10.1111/bjh.14591
11. 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*, 1–37. https://doi.org/10.7554/eLife.54076

12. 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
13. Uetz, P. (2022). The reptile database. Zoological Museum Hamburg. http://reptile-database.org.
14. 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). https://doi.org/10.1371/journal.pone.0112963
15. 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
16. 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 *E. 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%**

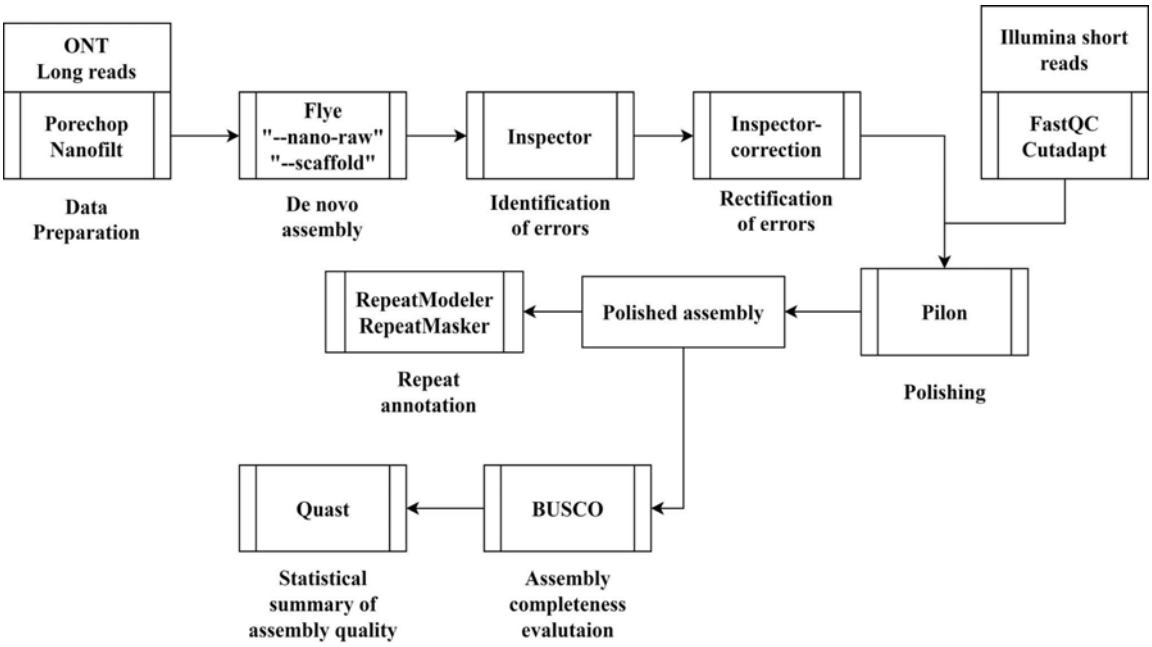

Fig.1: Overview of the de novo assembly pipeline representing the steps and the tools that were used for genome assembly of *E. carinatus*

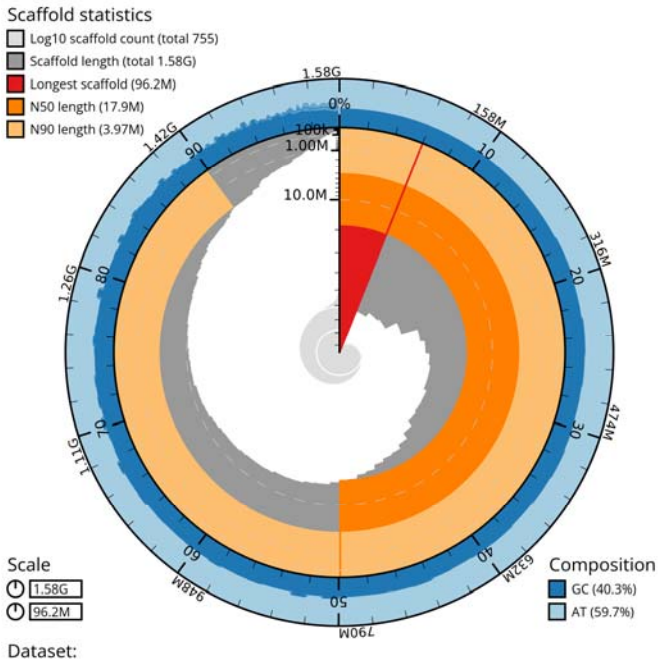

Fig. 2: Snail plot visualization summarizing metrics of the *E. carinatus* genome including the length of the longest contig (96.2 Mb; red line), N50 (17.9 Mb; dark orange), N90 (3.9 Mb; light orange), and base composition.

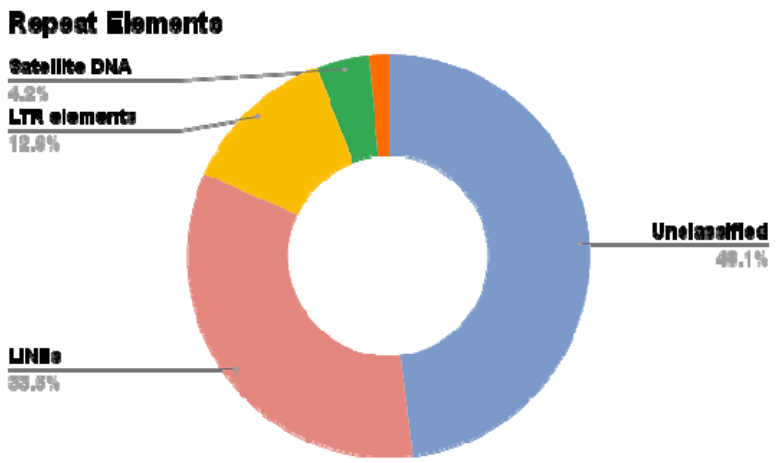

Fig 3: Plot displaying the predominant types of repeat elements present in *E. carinatus* genome

Acknowledgements

We are grateful to the Telangana Forest Department for granting us permission 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, 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. Aaron Comeault and Axel Barlow provided bioinformatics training for the whole genome analysis.