Snake W Sex Chromosome: The Shadow of Ancestral Amniote Super-Sex Chromosome

Worapong Singchat, Syed Farhan Ahmad, Nararat Laopichienpong, Aorarat Suntronpong, Thitipong Panthum, Darren K Griffin and Kornsorn Srikulnath

Supplementary Note 1: Gene ontology enrichment analysis and comparative repeatomic landscaping of Indian cobra Z and W chromosomes

(This note describes the methodology of the analysis of Indian cobra sex chromosomes as performed in the present study)

A recently published analysis of the Indian cobra genome [1] reported for the first time a high-quality assembly of complete W chromosomes for any snake species. This provides an excellent reference resource and enabled us to perform the first comprehensive comparison of the repeatome between the snake Z and W chromosomes and report the relative abundance of repeats on heteromorphic chromosomes. We retrieved the unmasked assembled genome sequences of W and Z chromosomes from male (Biosample ID: SAMN11155092) and female (Biosample ID: SAMN15493627) assemblies from the National Center for Biotechnology Information (NCBI) database with accession ID GCA_009733555.1 (BioProject: PRJNA527614). Both assembled chromosomes were subjected to repeat annotation using RepeatMasker v4.1 [2]. We customized RepeatMasker with the "hmmer", "slow", "align", "gff", and "metazoa" parameters. We created a summary of the RepeatMasker output results and applied a Perl script (buildSummary.pl), available as a utility package of the RepeatMasker software. The generated results files from RepeatMasker were then utilized as input data for further Perl scripts (createrepeatlandscape.pl and calcdivergencefromalign.pl) to determine the Kimura divergence values and visualize the transposable elements landscape plots (Figure 3).

In addition to repeatomics, we performed gene ontology (GO) enrichment analysis of the W sex chromosome to conduct a survey of the biological functions of W-linked genes and to test our hypothesis that the W chromosome might carry a diverse set of genes associated with multiple functions.

We first downloaded a complete set of reference genes from Ensembl [3], and performed BLAST alignments of the complete W sex chromosome against this reference. The reference genes consisted of the super-set of all transcript-coding sequences of the green anole *Anolis carolinensis* genome, resulting from Ensembl gene predictions. We chose anole as the reference set because it represents highly accurate annotation and is a phylogenetically close reptilian species available in Ensembl (last accessed August, 2020). W-linked annotated genes with the highest confidence hits (with alignment size of at least 200 bp and E-value < 0.01) were selected for GO enrichment analysis. The Ensembl IDs of these W-linked genes were extracted using the BioMart package (http://www.ensembl.org/biomart/) from Ensembl and two separate lists of IDs were generated. The list corresponding to W-linked genes was considered the selection, whereas the list corresponding to whole-genome reference genes was considered the background to perform GO analysis using the R package ViSEAGO ("Visualization, Semantic Similarity and Enrichment Analysis of Gene Ontology") [4]. Using ViSEAGO, a functional analysis was conducted with the following steps: (1) in R, the W-linked genes IDs file was read as the input list of genes of interest and the whole-genome genes IDs file was associated with a reference gene set (i.e., gene background); (2) we loaded the last current GO terms' annotations from the selected "anole lizard" database; (3) we then performed functional enrichment (Fisher exact) tests; and (4) the semantic similarity was computed and clusters of GO terms were visualized (Figure 2).

Reference

1. Suryamohan, K.; Krishnankutty, S.P.; Guillory, J.; Jevit, M.; Schröder, M.S.; Wu, M.; Kuriakose, B.; Mathew, O.K.; Perumal, R.C.; Koludarov, I.; et al. The Indian cobra reference genome and transcriptome enables

comprehensive identification of venom toxins. *Nat. Genet.* **2020**, *52*, 106–117, doi:10.1038/s41588-019-0559-8.

2. Smit, A.; Hubley, R.; Grenn, P. RepeatMasker Open-4.0. **2015**, http://www.repeatmasker.org.
3. Cunningham, F.; Achuthan, P.; Akanni, W.; Allen, J.; Amode, M.R.; Armean, I.M.; Bennett, R.; Bhai, J.; Billis, K.; Boddu, S.; et al. Ensembl 2019. *Nucleic Acids Res.* **2019**, *47*, 745–751, doi:10.1093/nar/gky1113.
4. Brionne, A.; Juanchich, A.; Hennequet-Antier, C. ViSEAGO: A Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. *BioData Min.* **2019**, *12*, 16, doi:10.1186/s13040-019-0204-1