

Deciphering the structural organization of sex chromosomes in Siberian silk moth, *Dendrolimus sibiricus*

Shipova A.A.¹, Ershov N.I.¹, Yakimova M.E.^{2, 3}, Martemyanov V.V.^{2*}

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia

³ Department of Information Biology, Novosibirsk State University, Novosibirsk, Russia

* martemyanov79@yahoo.com

Key words: sex determination; whole-genome sequencing; chromosome evolution; W chromosome

Motivation and Aim: The W chromosome of *Lepidoptera* is typically saturated by genomic repeats, gene-poor and completely heterochromatinized. These properties as well as haploid state of sex chromosomes in contrast to diploid state of the rest genome significantly hamper the accurate assembly of W chromosome from heterogametic individuals. To avoid these problem, many genome sequencing projects prefer to study homogametic individuals. We explored the possibility to circumvent these problems at low cost using an approach of contig classification and targeted reassembly of available long-read genomic data obtained from female *D. sibiricus*.

Methods and Algorithms: Two short-read genomic libraries of female and male *D. sibiricus* were sequenced at low depth and mapped on the primary assembly of *D. sibiricus* female genome. The ratios of coverage by these libraries were compared to the theoretically expected values according to the ZW/ZZ sex determination scheme and used to classify the contigs on W, Z and autosomal. From the raw HiFi data, the portion mapped on W-derived contigs was extracted and assembled using the hifiasm with parameters configured for haploid assembly.

RepeatMasker and RepeatModeler software were used to predict and identify genomic repeats in the assembly and to perform Kimura 2-parameter divergence analysis. The analysis of syntenic regions between the Z and W chromosomes included their pairwise alignment by minimap2 and aggregation into syntenic blocks by Asynt R-package. Non-partitioned phylogenetic analysis of W-associated genes was performed using IQ-TREE software with 10,000 bootstrap iterations.

Results: The primary diploid assembly of *D. sibiricus* contained a significant portion of fragments that were represented in one haplovariant but not linked to the Z chromosome, i.e. potentially related to the W chromosome. Using two additional genomic short-read libraries of female and male *D. sibiricus*, we built a classifier to partition all contigs into autosomes, the Z chromosome, and the W chromosome according to the expected ZW/ZZ sex determination scheme.

The initially used hifiasm assembler is designed to assemble diploid genomes, thus it may misassemble unpaired sex chromosomes. We performed a targeted reassembly of the W chromosome, limited to only reads mapped to W-classified unitigs. Nevertheless, the assembly retained some fragmentation, accounting 10 contigs with a total length of 22.46 Mbp. Unresolved assembly fragmentation of the W chromosome is explained by its exceptional saturation with genomic repeats together with the fact that the W chromosome contained several large and extremely homologous duplications.

Accounting them, the actual length of W is expected to be 27.4 Mbp, while we kept only one copy of each in the final assembly.

Interestingly, we discovered two large fragments of the symbiotic *Wolbachia pipientis* genome as part of the W chromosome. The inserted fragments both are characterized by relative preservation of the coding ORFs, assuming that these insertions occurred quite recently.

The W chromosome sequences are also distinguished by a clear expansion of LTR transposons. The proportion of these repeats was more than 3 times higher compared to the whole genome. The total proportion of repeats in the W chromosome was about 80 %. The analysis of interspersed repeat landscape based on Kimura 2D distances revealed significant differences between the W chromosome and the rest nuclear genome, suggesting that the evolutionary age of the large-scale expansion of the LTR transposons along the W chromosome is very recent. Moreover, a series of repeats were strongly overrepresented or even specific to the W chromosome. Among them was dispersed telomere-associated non-LTR transposon accounting for almost 3 % of its sequences.

Analysis of synteny between Z and W chromosomes revealed significant regions of their collinearity at both distal subtelomeric regions. Almost all protein-coding genes predicted on the W chromosome were concentrated in these regions of synteny, having a close paralog on Z. A comparison of the Z and W paralogues indicated that most W chromosome genes are pseudogenes with a damaged ORF. The phylogenetic analysis of the preserved genes confirmed that at least part of the W chromosome was formed from a copy of Z of *D. sibiricus*.

Conclusion: The abundance of young repeats and young insertions of *Wolbachia* genomic fragments indicate the recent formation of the W chromosome, which echoes the theory of multiple non-canonical emergence of W chromosomes in the evolution of *Lepidoptera*, including its formation from B chromosomes [1]. At the same time, the W exhibits significant sequence similarity to the Z, which confirms its evolutionary origin from the Z chromosome.

References

1. Fraïsse C., Picard M.A.L., Vicoso B. The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. *Nat Commun.* 2017;8(1):1486. doi 10.1038/s41467-017-01663-5