

Supplementary Information

Telomere-to-telomere genome assembly of the Dipteran *Bactrocera dorsalis* from a single individual

Wei Liu^{1†}, Qiang Lin^{1†}, Qi Wang^{2†}, Wanfei Liu¹, Jing jia¹, Jie Zhang³, Ling Yang⁴, Yongyue Lu⁵, Peng Cui^{1*}, Guirong Wang^{1,6*}

1. Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
2. School of Forestry, Key Laboratory of Sustainable Forest Ecosystem Management-Ministry of Education, Northeast Forestry University, Harbin 150040, China.
3. Hubei Key Laboratory of Biological Resources Protection and Utilization, College of Forestry and Horticulture, HuBei Min Zu University, Enshi 445000, China
4. College of Forestry, Shanxi Agricultural University, Taigu 030801, China.
5. Department of Entomology, South China Agricultural University, Guangzhou 510640, China
6. State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 1000193, China

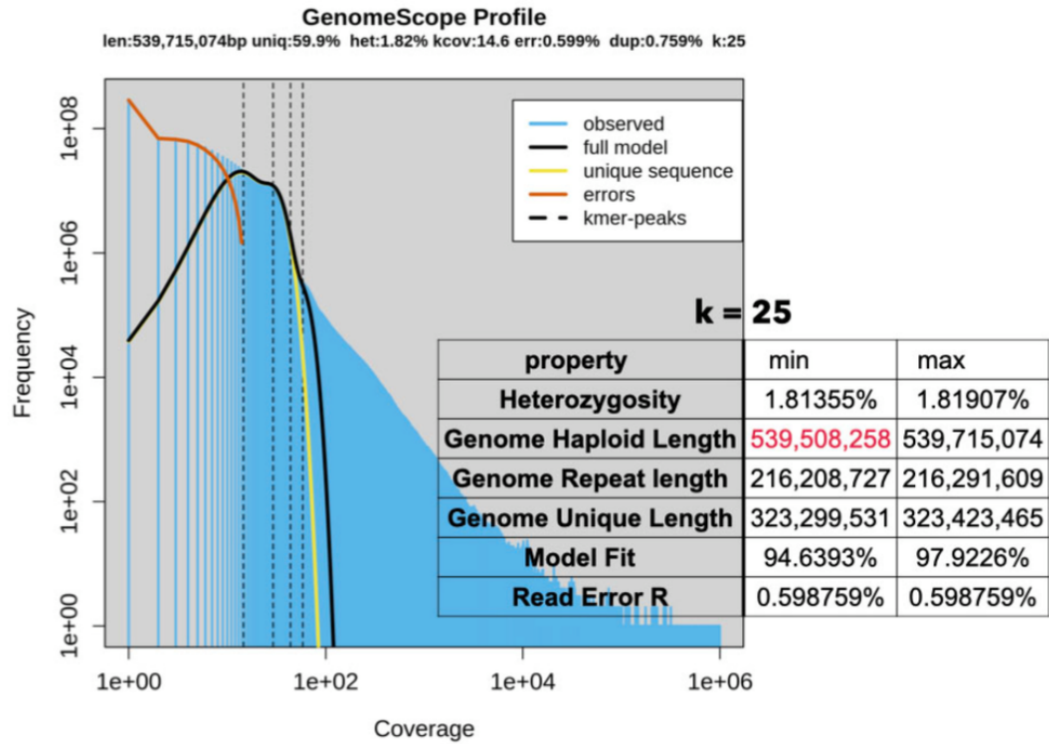
† co-authors

* These authors contributed equally

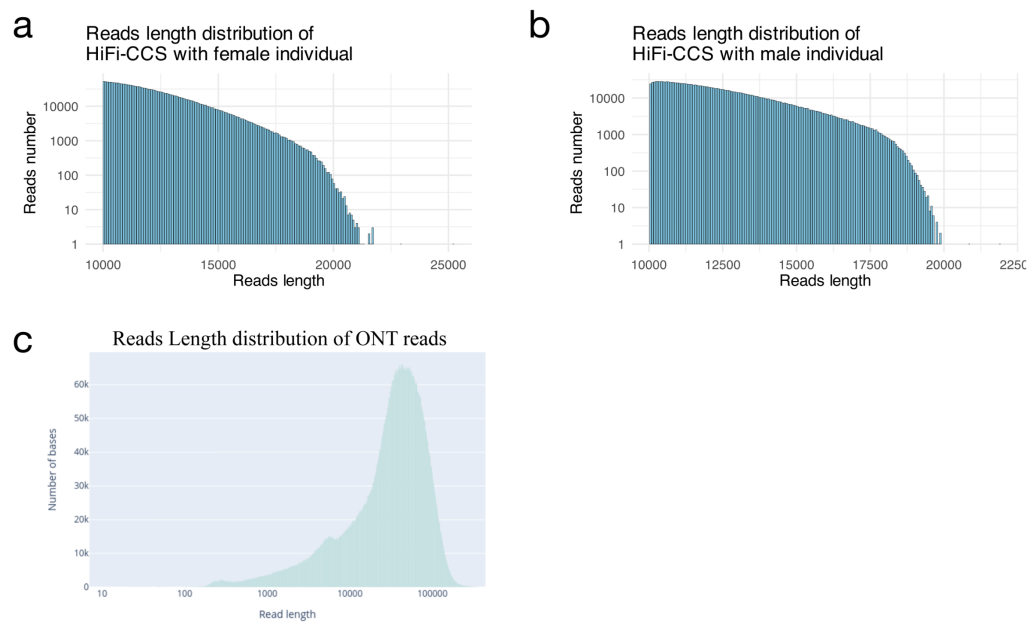
email: wangguirong@caas.cn; cuipeng@caas.cn

These authors contributed equally: Wei Liu, Qiang Lin, Qi Wang

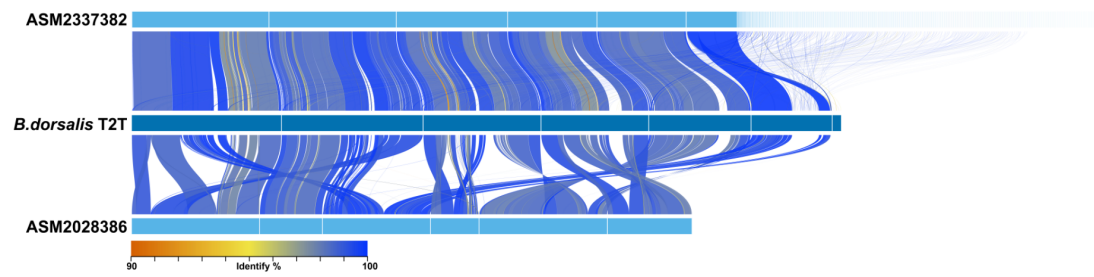
These authors jointly supervised this work: Peng Cui, Guirong Wang



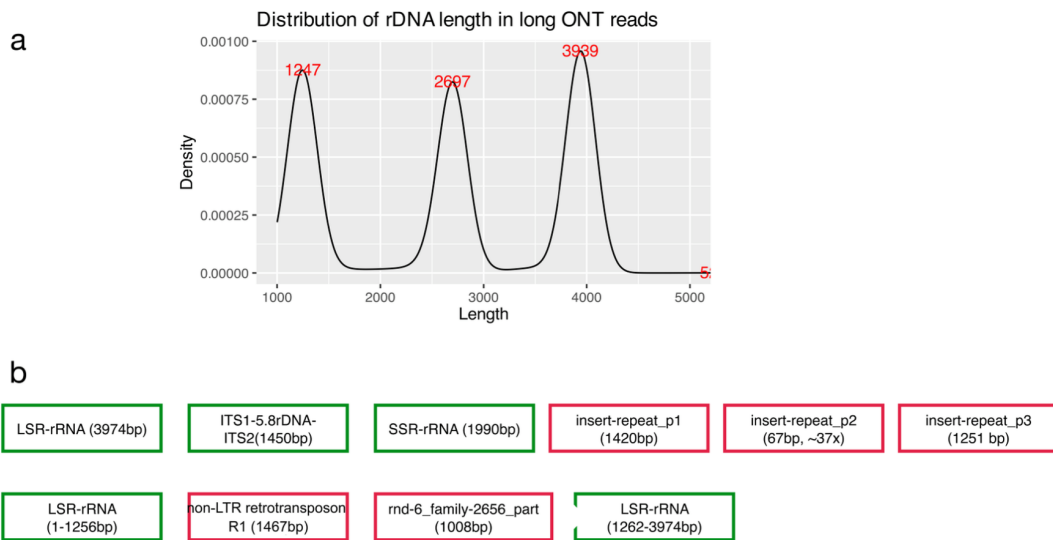
Supplementary Fig. 1. Heterozygosity of 5 individual mixed samples according to a genome survey. We determined the heterozygosity of 5 individual mixed samples via a genome survey.



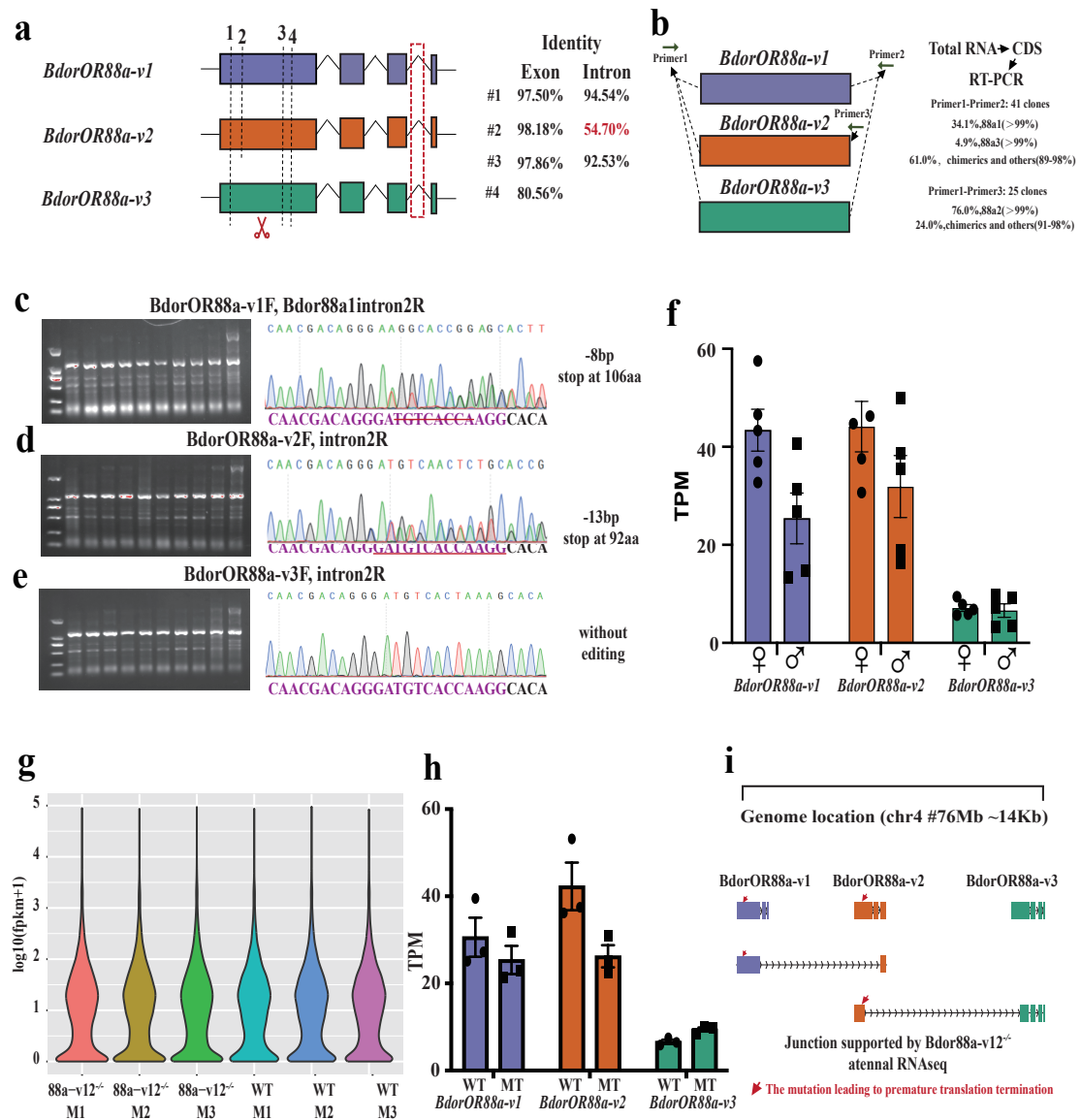
Supplementary Fig. 2. The reads length distribution profile of HiFi-CCS and ONT reads



Supplementary Fig. 3. Synteny comparison between the *B. dorsalis* T2T assembly and published genome assemblies.

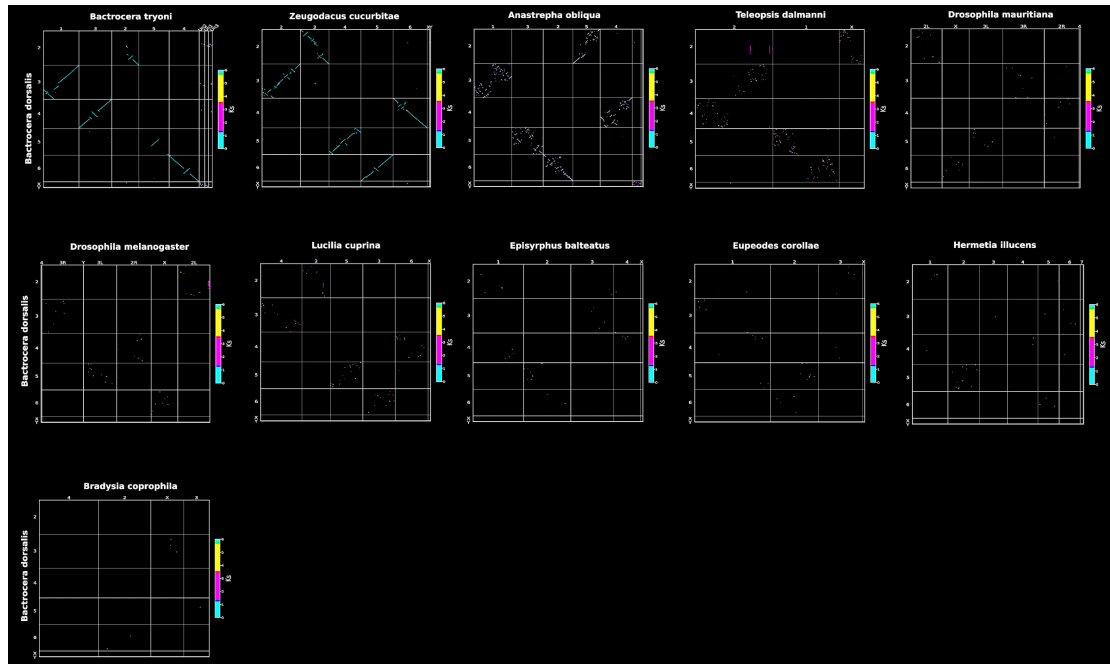


Supplementary Fig. 5. Distribution of rDNA block length in ONT reads (>80 kb). (a) All large subunit rDNA (LSU-rDNA) related ONT long reads (>80 kb) were identified using srfutls.js ‘reads enlong’ and aligned with minimap2 using the parameters (-c -N1000000 -f1000 -r100,100). Aligned LSU-rDNA blocks were filtered with a minimum coverage and identity threshold of 0.2. If a read contains insertions within LSU-rDNA element, the aligned block was fragmented into two regions of different lengths. In the raw ONT long reads, the LSU-rDNA homologous regions displayed three distinct length peaks at 1,247 bp, 2,697 bp, and 3,939 bp (the complete unit), indicated that approximately 48% of LSU-rDNA units were fragmented into two parts rather than existing as intact units (b) Discontinuous rDNA structure identified in ONT long reads. Specific repeat sequences were found downstream of the entire rDNA large subunit and following the rDNA repeat unit.



Supplementary Fig. 7. Gene structure, verification of chimeric transcripts, expression patterns of BdorOR88as, and establishment of knockout strains.

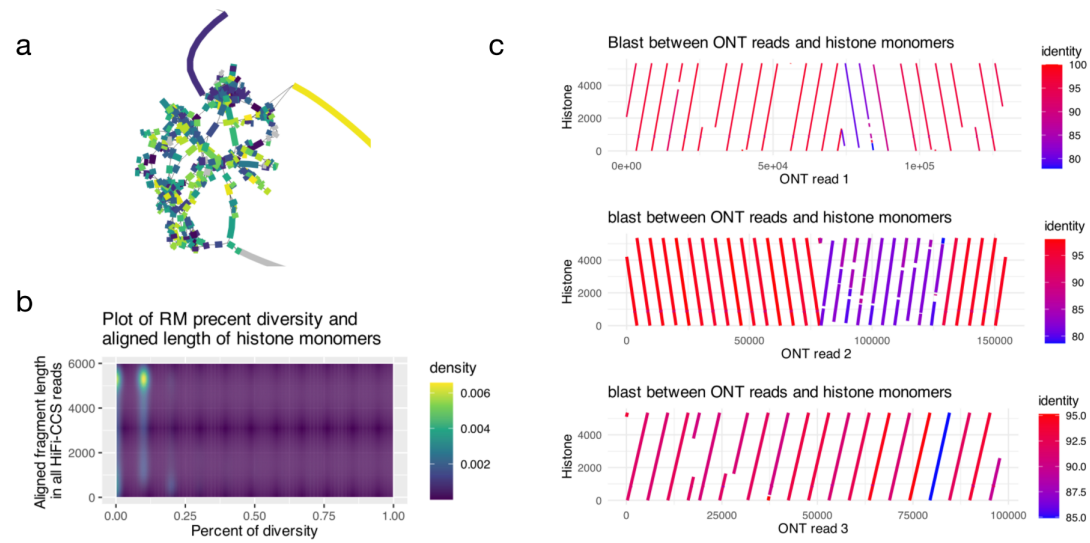
(a) Gene structure and exon-intron identities across three copies. (b) Verification of chimeric transcripts formed by exons from different copies. (c-e) Genotyping of the BdorOR12^{-/-} knockout mutant. (f) Expression profiles of BdorOR1-3s in male and female adults. (g) Distribution of gene expression from the transcriptome of male antennae in wildtype and BdorOR12^{-/-} knockout mutants. (h) Expression profiles of BdorOR1-3s in male antennae between wildtype and BdorOR12^{-/-} knockout mutants. (i) Chimeric transcripts of BdorOR88as in BdorOR12^{-/-} knockout mutants, with red arrows indicating the mutations that lead to premature translation termination.



Supplementary Fig. 8. Ks dot plot for syntenic blocks between *B. dorsalis* and other species.

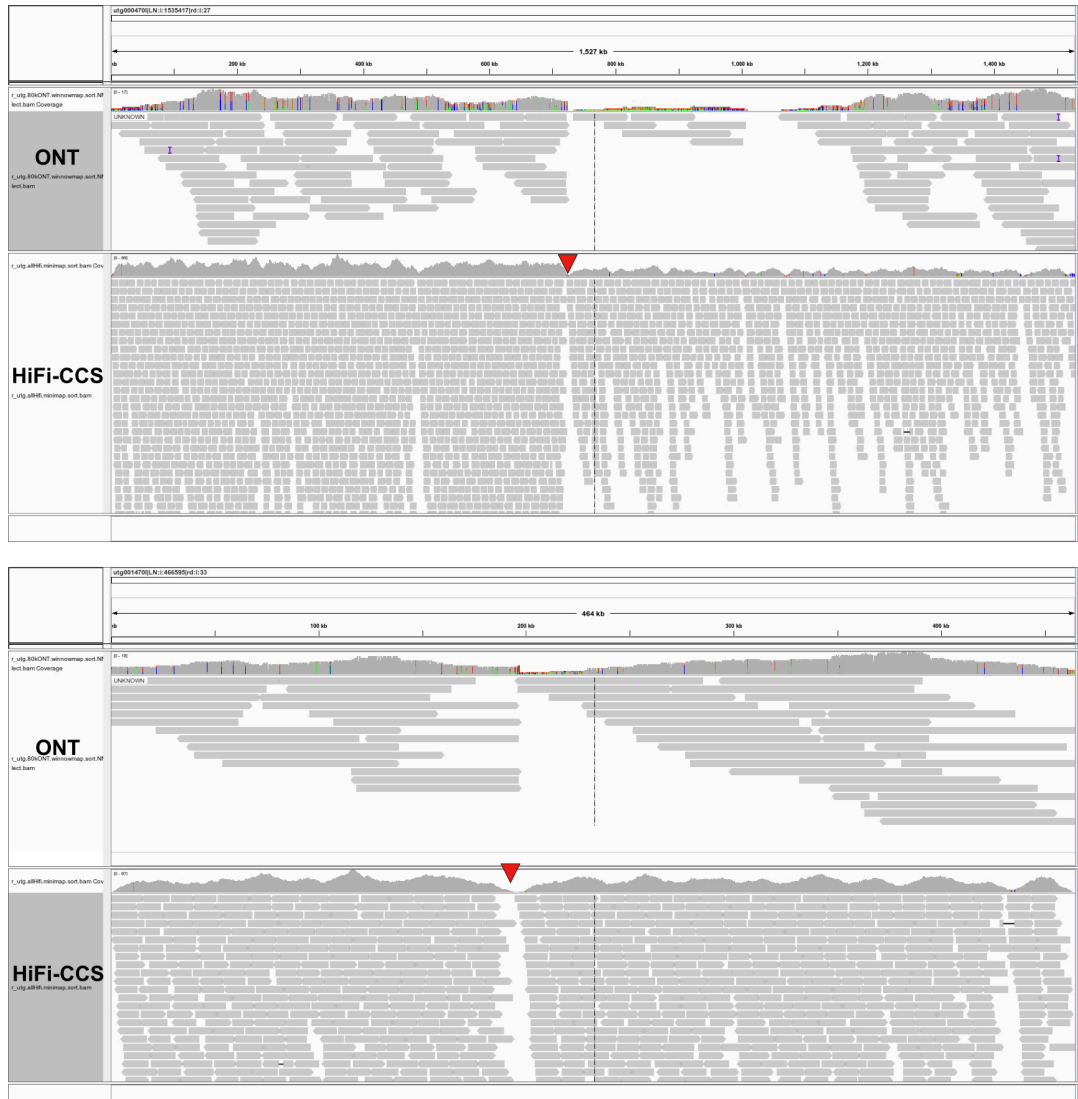
Conserved Muller element structures are observed only in closely related species (row 1), where synteny with the *B. dorsalis* genome is largely preserved. In more distantly related species (rows 2 and 3), extensive genomic inversions have disrupted gene-based synteny patterns. However, upon closer inspection, homologous genes still tend to cluster within corresponding chromosomal structures, reflecting a degree of structural conservation.

Blocks are colored according to Ks values.



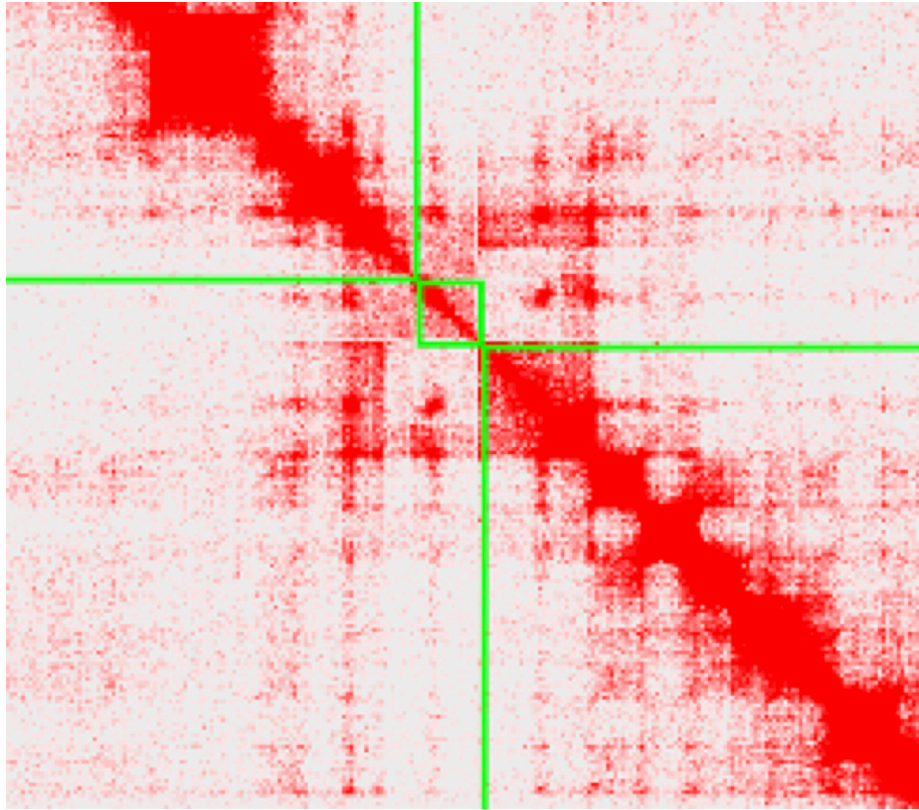
Supplementary Fig. 9. Sequence conservation of histone array

(a) The histone array regions were highly complex histone array regions of in the initial assembly graph. (b) We assessed the histone-associated segments in the original HiFi CCS reads annotated by RepeatMasker. Most annotated regions showed high sequence conservation. (c) Three longer histone-associated reads from the ONT dataset were selected to illustrate the structure and sequence identification of histone units.



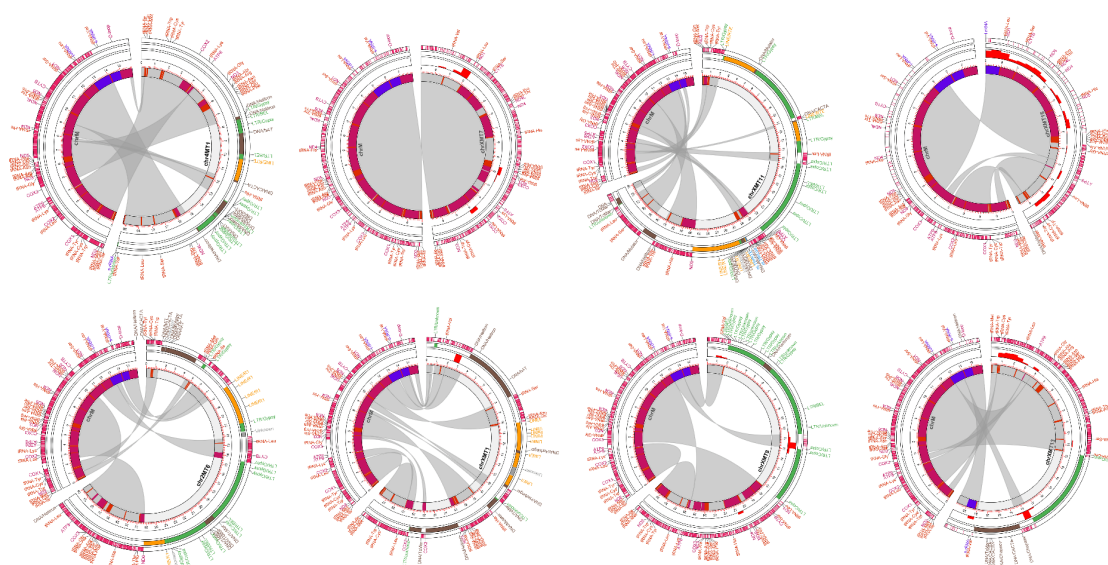
Supplementary Fig. 10. Coverage breaks in HiFi-CCS read mappings indicate candidate mis-joins.

We present two loci of potential mis-assembled in Hifiasm unitigs, as revealed by abrupt coverage drops in HiFi-CCS read alignments. These regions are further supported by discordant ONT read distribution.



Supplementary Fig. 11. Alternate contigs identified and removed using Hi-C contact maps.

Hi-C interaction density plots were used to identify and remove haplotypic duplications from the assembly. Specifically, they allowed us to detect redundant alternate contigs located in high-heterozygosity or repetitive regions, often appearing as bubbles in the p_utg assembly graph.



Supplementary Fig. 12. NUMTs in nuclear genome