Chromosome-Level Genome Assembly of *Callitettix versicolor* (Rice Spittlebug)

Hong Chen[1,2], Gexia Qiao[1,2,*], and Aiping Liang[1,2,3,*]

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[2]College of Life Sciences, University of Chinese Academy of Sciences (UCAS), Beijing, China

[3]College of Life Sciences, Tianjin Normal University, Tianjin, China

*Corresponding authors: E-mails: qiaogx@ioz.ac.cn (G.Q.), liangap@ioz.ac.cn (A.L.).

Accepted: 09 August 2022

Abstract

The spittlebug family Cercopidae (Hemiptera: Auchenorrhyncha: Cicadomorpha: Cercopoidea) is distributed worldwide. Some Cercopidae species are agricultural pests that are responsible for substantial economic damage. However, the genomics of spittlebugs has rarely been studied and their complete genome assembly is yet to be reported. Here, we present the draft reference genome of *Callitettix versicolor* Fabricius (Hemiptera: Cercopidae) at the chromosome level. The assembled draft genome was 974.99 Mb with a contig N50 of 5.63 Mb, and the longest contig being 24.54 Mb. Hi-C technology was used to obtain an approximately 958.71 Mb chromosome-level genome on 10 pseudochromosomes, which covered 98.33% of the assembly. Repeat sequences accounted for 38.88% of the genomic sequences. A total of 21,937 protein-coding genes were detected in the reference genome, 89.97% of which were annotated in public databases. The high-quality reference genome of *C. versicolor* reported in this study will provide a valuable genomic resource for future ecological and evolutionary studies of spittlebugs.

Key words: *Callitettix versicolor*, genome, gene annotation, nanopore sequencing.

Significance

Callitettix versicolor belongs to the Cercopidae family and is a common rice pest in southern China. In this study, we sequenced and assembled a high-quality draft reference genome of *C. versicolor* to gain a genome insight into its genetic basis. Our results represent a valuable resource for further study on the evolutionary biology of Cercopidae.

Introduction

Cercopidae is a xylem-feeding insect group which forms the largest family of Cercopoidea (Dietrich 2009). Cercopoidea are commonly known as spittlebugs or froghoppers as their nmyphs secrete a foam that protects them against harmful radiation and higher temperatures (Biedermann 2003; Chen et al. 2018). Several species are known to cause economic loss to crops (Paladini et al. 2018). Notable examples include *Mahanarva fimbriolata* Stål and *Mahanarva andigena* Jacobi, which feeds on sugarcane (Madaleno et al. 2008; Chaves et al. 2014), *Philaenus spumarius* Linnaeus, a pest in Italy (Avosani et al. 2020), and *Callitettix versicolor*, a rice pest that is widespread throughout southern China (fig. 1*a*). Both the adults and nymphs of *C. versicolor* can damage crops (Tang and Gao 1995; Li et al. 2001; Wang et al. 2014).

With a change in farming systems and warming of the climate, *C. versicolor* has spread to northern China (Chen and Liang 2012). Previous studies have been made into the morphology, anatomy, and phylogeography of *C. versicolor* so as to provide a theoretical foundation for minimizing their economic damage (Chen and Liang

2012; Yang et al. 2016). Here, we report a high-quality draft genome of *C. versicolor*, including nanopore long reads, Illumina short reads, and Hi-C sequencing data at the chromosome level. This is the first reference genome of *C. versicolor* and the first chromosome-level genome assembly of spittlebugs. The genome will provide useful resource for future ecological, evolutionary, and conservation studies of *C. versicolor*.

Results and Discussion

Genome Assembly

To acquire a high-quality genome assembly, we generated and filtered three types of clean data for the genome, including 105.34 Gb Oxford Nanopore Technologies (ONT) clean data, 114.77 Gb Illumina clean data, and 100.05 Gb Hi-C clean data. The final assembly had a length of 974.99 Mb, with a scaffold/contig N50 size of 98.12/5.63 Mb and a GC content of 33.34% (fig. 1*b*), which is close to the k-mer-based genome size estimate (962.21 Mb/31.98%; Supplementary fig. 1, Supplementary Material online). Our Hi-C analyses scaffolded for *C. versicolor*, anchoring 98.33% of the genome assembly in ten pseudochromosomes (fig. 1c, *d*; supplementary table 1, Supplementary Material online).

To assess the quality of the assembled genome, Illumina reads were mapped on the reference genome and 92.39% of the pair-end clean reads were correctly mapped. In addition, the completeness of universal single-copy orthologs and conserved core eukaryotic genes (CEGs) was examined using Benchmarking Universal Single-Copy Ortholog assessment (BUSCO) and Core Eukaryotic Genes Mapping Approach (CEGMA), respectively. As a result, the BUSCO completeness of the *C. versicolor* assembly reached 94.69% including 1,531 single-copy BUSCOs, 39 duplicated BUSCOs, 9 fragmented BUSCOs, and 79 missing BUSCOs (fig. 1*b*). The reference genome also contained 242 of the 248 (97.58%) highly complete CEGs and 455 of the 458 (99.34%) complete CEGs. These results suggest a high completeness of the assembled genome.

Genome Annotation

Approximately 379 Mb of repeat sequences were identified, accounting for 38.88% of the genomic sequence. DNA transposons and retroelements accounted for 21.84% and 17.03% of the genome, respectively. Long interspersed nuclear element and long terminal repeats (LTRs) constituted a higher proportion of the retroelements of *C. versicolor* genome, at 8.67% and 17.5%, respectively (Supplementary table 2, Supplementary Material online). A total of 237 pseudogenes and noncoding RNAs were identified in the draft genome, including 61,076 transfer RNAs (tRNAs), 201 ribosomal RNAs (rRNAs), 46 micro-RNAs (miRNAs), 173 small nuclear RNAs (snRNAs), and 28 small nucleolar RNAs.

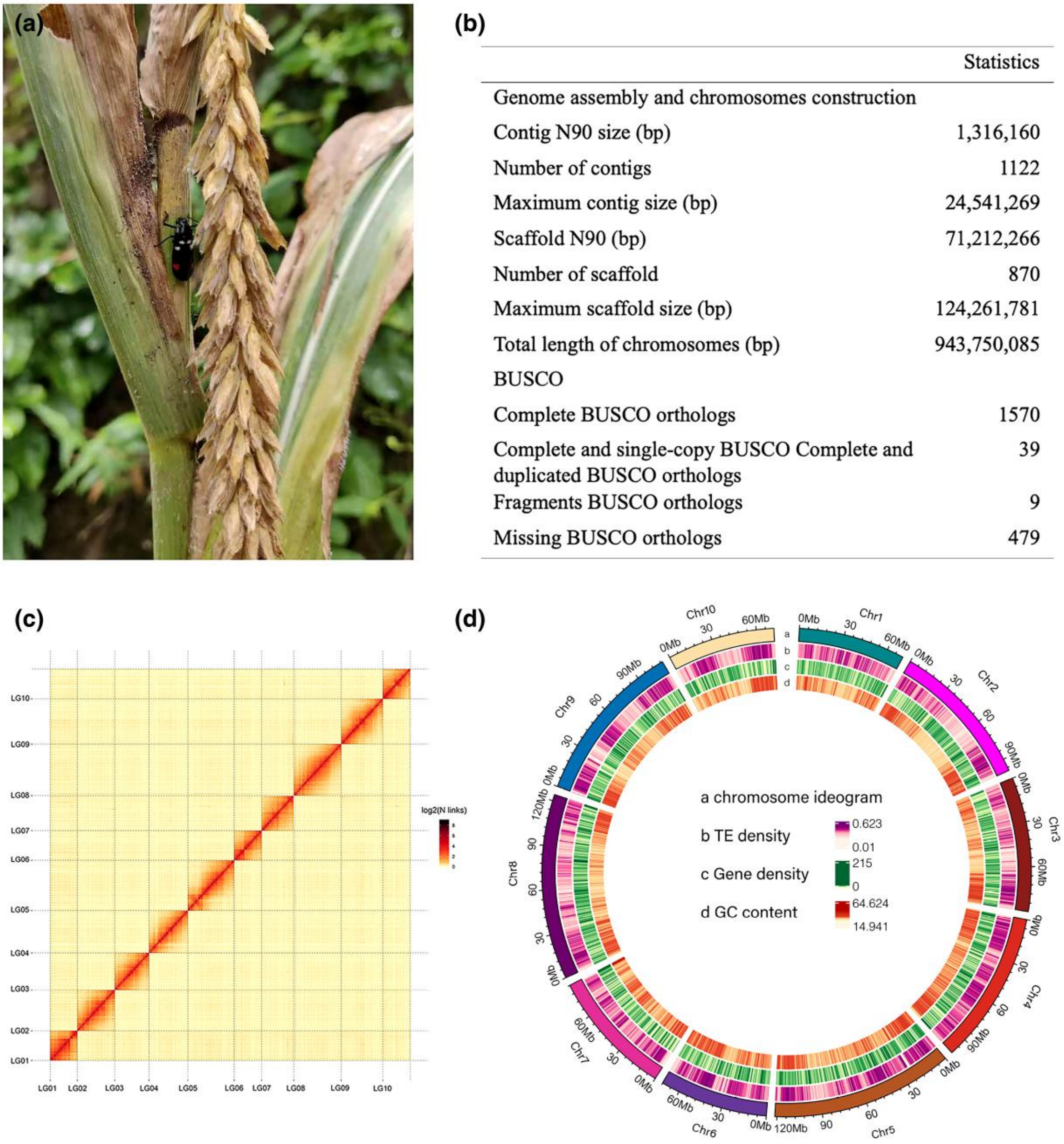
The results from three gene prediction strategies were integrated, and a total of 21,937 protein-coding genes with an average length of 14,237 bp was annotated for the genome (Supplementary table 3, Supplementary Material online). The BUSCO completeness for protein sequences reached 97% ($n = 1{,}367$), 1,306 single-copy, 20 duplicated, 2 fragmented, and 39 missing BUSCOs were identified. Together, these indicate a high level of quality for the predictions. A comparison of the predicted genes was made against public genome databases, including evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG), Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Non-Redundant Protein Sequence Database (NR), Pfam, Swiss-Prot, and Translated (Tr) EMBL. The comparison annotated a total of 19,736 genes in at least one of the databases, representing 89.97% of the *C. versicolor* genome (Supplementary table 4, Supplementary Material online).

Materials and Methods

Sample Collection and Sequencing

All *C. versicolor* samples used for genome sequencing are second generation inbreeding lines. The inbred laboratory strain was derived from a field population collected in Ji'shou, Hunan province, China. The spittlebugs were reared on rice and wheat plants at a relative humidity of 70%. The spittlebugs light/dark regime consisted of a photoperiod of 16:8 h and a temperature split of 28 °C/26 °C. Approximately 62.5 μg genomic DNA was extracted from ten male adults. Sodium dodecyl sulfate (SDS) lysis buffer was applied for DNA extraction, and the whole extraction steps are following as SDS DNA extraction methods from Chen et al. (2010). The concentration and purity of DNA were detected by NanoDrop and Qubit, respectively. The integrity of DNA was detected using pulsed field electrophoresis.

For nanopore DNA sequencing, the genomic DNA was prepared using the NEB Next FFPE DNA Repair Mix kit, and then a 12-kb insert size library was constructed using the ONT Template prep kit (SQK-LSK109). The library was sequenced on the ONT PromethION platform with the R9 cell and ONT-sequencing reagent kit (EXP-FLP001.PRO.6) in accordance with the manufacturer's instructions. The raw signal data in nanopore sequencing are stored in FAST5, and Guppy v0.5.1 which is supplied by Oxford Nanopore was used for PHRED standard base calling. For Illumina DNA sequencing, three short libraries with an insert size of 350 bp were constructed using the NEB DNA Library Rapid Prep Kit. The products were quantified using Bioanalyzer 2100 (Agilent Technologies). The high-throughput/resolution chromosome conformation capture-based (Hi-C) library sequencing used one lane of Illumina NovaSeq



	Statistics
Genome assembly and chromosomes construction	
Contig N90 size (bp)	1,316,160
Number of contigs	1122
Maximum contig size (bp)	24,541,269
Scaffold N90 (bp)	71,212,266
Number of scaffold	870
Maximum scaffold size (bp)	124,261,781
Total length of chromosomes (bp)	943,750,085
BUSCO	
Complete BUSCO orthologs	1570
Complete and single-copy BUSCO Complete and duplicated BUSCO orthologs	39
Fragments BUSCO orthologs	9
Missing BUSCO orthologs	479

Fig. 1.—(*a*) *Callitettix versicolor*. (*b*) Genome assembly and completeness assessment statistics features of *C. versicolor*. (*c*) Hi-C interaction heatmap of ten linkage groups in *C. versicolor* genome. The depth of color indicated the interaction between fragments. (*d*) Genomic landscape of the ten assembled pseudochromosomes.

6000 (150 bp PE reads) to construct Hi-C fragment libraries with insert sizes of 300–700 bp by using Biomarker Hi-C Library Prep Kit for Illumina (Rao et al. 2014).

Genome Assembly

For genomic contig assembly, Canu v1.3 was used for the error correction of the long reads with default parameters (Koren et al. 2017). The longest supported range of error-corrected reads was obtained and then assembled using Wtdbg2 with default parameters (Ruan and Li 2020). Next, the ONT-sequencing data were mapped to assembly using Minimap2 (Li 2018), and Racon (Vaser et al. 2017) was used for polishing with the default parameters. After the polishing of Racon, the Illumina paired-end reads mapped to the polished assembly genome with BWA v 0.7.17 (Li and Durbin 2009), and Pilon (Walker et al. 2014) was used to polish the second round.

The completeness of the assembled genome was evaluated in terms of three aspects. Firstly, Illumina reads were mapped on the reference genome using BWA. Next, BUSCO v3 was run using the data from the INSECTA database (OrthoDB v9), which contains 1,658 conserved insect genes (Waterhouse et al. 2018). Finally, CEGMA v2.5, which contains 458 conserved eukaryotic core genes and 248 highly conserved CEGs, was used to assess the completeness of the assembly genome using the default parameters (Parra et al. 2007). CEGMA relies on some highly conserved proteins are encoded in essentially all eukaryotic genomes, which is based on euKaryotic clusters of Orthologous Groups, resulting in a set of CEGs.

The clean Hi-C reads of the *C. versicolor* genome were first truncated at the putative Hi-C junctions and the trimmed reads were aligned with assembly results using BWA. Only unique read pairs that can be aligned with a mapping quality of over 20 were included for further analysis. HiC-Pro v2.10.0 was used to filter out invalid reads such as dangling-end and self-cycle, religation, and dumped products (Servant et al. 2015). LACHESIS was used for Hi-C scaffold correction and assembly (Burton et al. 2013). Manual inspection was conducted for any two segments that showed an inconsistent connection with information from the raw scaffolds. The corrected scaffolds were assembled with the following parameters: CLUSTER_MIN_RE_SITES = 58; CLUSTER_MAX_LINK_DENSITY = 2; ORDER_MIN_N_RES_IN_TRUNK = 56; ORDER_MIN_N_RES_IN_SHREDS = 56.

RNA Isolation and Sequencing

For genome annotation, RNA was extracted from living samples of *C. versicolor*, which including one male adult and one three-instar nymph. RNAprep Pure Tissue Kit (Tiangen, China) was used for extracted from adult and nymph, respectively. The library was constructed using the NEBNext® Ultra™ RNA Library Prep Kit (NEB, UK) according to the manufacturer's instructions. The RNA was isolated by using mRNA Capture Beads, and then fragmented by using first strand synthesis reaction buffer and random primers. After the second strand synthesis, the double-strand cDNA was purified by adding VAHTS™ DNA Clean Beads, and then repair the terminal and add poly-A by using NEBNext End Prep (End Repair Reaction Buffer and End Prep Enzyme Mix). Then, added the adaptors and selected the target fragment. Finally, polymerase chain reaction was performed. The library with an insert size of 250–350 bp was constructed, and then sequenced on Illumina NovaSeq 6000 platform. At last, 12 Gb RNA data were obtained.

Genome Annotation

We customized a de novo repeat library of the genome using RepeatModeler, which automatically execute two de novo repeat finding programs, including RECON v1.08 and RepeatScout (Bao and Eddy 2002; Price et al. 2005; Flynn et al. 2020). RepeatClassifier was used to classify the predicted results in the following three databases: Rebase v19.06, Rexdb v3.0, and Dfam v3.2 (Jurka et al. 2005; Wheeler et al. 2013; Neumann et al. 2019). The de novo repeats library of *C. versicolor* was analyzed using Repeatmasker v1.331 to identify repetitive sequences and transposable elements. Software that employ LTR, LTRharvest v1.5.9 and LTR_FINDER v1.1, was used to predict specific repeats (Xu and Wang 2007; Ellinghaus et al. 2008).

Three approaches, including de novo prediction, homology-based prediction, and transcriptome-based prediction, were adopted to predict gene structures. For ab initio annotation, Augustus v2.4 and SNAP were applied (Korf 2004; Stanke et al. 2008). For homology-based prediction, GEMOMA v1.6.1 (Keilwagen et al. 2016) with default parameters was used to predict homology-based species. The genome data of four insect species, including the model *Drosophila melanogaster* (GCA_000001215.4), and three Hemiptera species: *Halyomorpha halys* (GCA_000696795.2), *Cimex lectularius* (GCA_00064865.3), and *Nilaparvata lugens* (GCA_014356525.1) were downloaded form GenBank for gene annotation (supplementary table 5, Supplementary Material online). For transcript-based prediction, RNA-sequencing data were mapped to the reference genome using HISAT2 and assembled by Stringtie v1.2.3 (Pertea et al. 2015; Kim et al. 2019). GeneMarkS-T v5.1 was used to predict genes in the assembled transcripts, and the program PASA was used to align spliced transcripts and annotate candidate genes (Tang et al. 2015). Finally, genes predicted from the three models were merged by EVidenceModeler (Haas et al. 2008). The functions of protein-coding genes were predicted using eggNOG, GO, KEGG, NR, Pfam, Swiss-Prot, and TrEMBL.

In addition, noncoding RNAs, including tRNA, miRNA, rRNA, and snRNA, were identified. rRNA, miRNA, and snRNA were annotated by mapping the transcripts against the Rfam database, whereas tRNA was predicted using tRNAscan-SE v2.0 with eukaryote parameters (Chan et al. 2021).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31872279 and 2070470) and the Biodiversity Survey and Assessment Project of the Ministry of Ecology and Environment, China (Grant No. 2019HJ2096001006).

Data Availability

The genomic assembled sequences, ONT raw reads, and Illumina sequencing data have been deposited in the NCBI database, under the BioProject accession number PRJNA772103. The GenBank assembly accession number is GCA_022606455.1.

Literature Cited

Avosani S, et al. 2020. Vibrational communication and mating behavior of the meadow spittlebug *Philaenus spumarius*. Entomol Gen. 40:307–321.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12: 1269–1276.

Biedermann R. 2003. Aggregation and survival of *Neophilaenus albipennis* (Hemiptera: Cercopidae) spittlebug nymphs. Eur J Entomol. 100:493–500.

Burton JN, et al. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 31:1119–1125.

Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 49:9077–9096.

Chaves VDV, et al. 2014. Biology and preferred oviposition site of the *Mahanarva indentata* froghopper (Hemiptera: Cercopidae) on sugarcane. Fla Entomol. 97:73–79.

Chen X, Liang AP. 2012. Laboratory rearing of *Callitettix versicolor* (Hemiptera: Cicadomorpha: Cercopidae), with descriptions of the immature stages. Ann Entomol Soc Am. 105:664–670.

Chen X, Meyer-Rochow VB, Fereres A, Morente M, Liang AP. 2018. The role of biofoam in shielding spittlebug nymphs (Insecta, Hemiptera. Cercopidae) against bright light. Ecol Entomol. 43: 273–281.

Chen H, Rangasamy M, Tan SY, Wang H, Siegfried BD. 2010. Evaluation of five methods for total DNA extraction from western corn rootworm beetles. PLoS One 5:e11963.

Dietrich CH. 2009. Chapter 15—Auchenorrhyncha: (cicadas, spittlebugs, leafhoppers, treehoppers, and planthoppers. In: Resh VH, Cardé RT, editors. Encyclopedia of insects. 2nd ed. San Diego: Academic Press. p. 56–64.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinform. 9:1–14.

Flynn JM, et al. 2020. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 117: 9451–9457.

Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 9(1):R7.

Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Keilwagen J, et al. 2016. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 44:e89.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37:907–915.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinform. 5:1–9.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li JQ, Zhao ZM, Wu SY, Ming K, Hou LN. 2001. Biology and ecology of rice spittle bug (*Callitettix versicolor*). J Southwest Agric Univ. 23:156–159.

Madaleno LL, et al. 2008. Influence of *Mahanarva fimbriolata* (Stål) (Hemiptera: Cercopidae) injury on the quality of cane juice. Neotrop Entomol. 37:68–73.

Neumann P, Novák P, Hoštáková N, Macas J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA 10:1–17.

Paladini A, Takiya DM, Urban JM, Cryan JR. 2018. New world spittlebugs (Hemiptera: Cercopidae: Ischnorhininae): dated molecular phylogeny, classification, and evolution of aposematic coloration. Mol Phylogenet Evol. 120:321–334.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067.

Pertea M, et al. 2015. Stringtie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33:290–295.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21:i351–i358.

Rao SS, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159:1665–1680.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 17:155–158.

Servant N, et al. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16:1–11.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644.

Tang CS, Gao JZ. 1995. Studies on the spatial distribution pattern and sampling techniques on the adult stage of rice spittlebug in paddy field. J Hunan Agric Univ. 21:421–426.

Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 43:e78.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.

Wang CP, Peng DC, Wu NJ, Liu HM, Huang XY. 2014. Occurrence status and control countermeasures of Callitettix versicolor Fabricius in Xiangxi autonomous prefecture. J Hunan Agric Sci. 9:49–51.

Waterhouse RM, et al. 2018. BUSCO Applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 35: 543–548.

Wheeler TJ, et al. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 41:D70–D82.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35: W265–W268.

Yang H, Lin CP, Liang AP. 2016. Phylogeography of the rice spittle bug (*Callitettix versicolor*) implies two long-term mountain barriers in South China. Zool Sci. 33:592–602.

Associate editor: Christopher Wheat