Chromosome-level genome assembly for the horned-gall aphid provides insights into interactions between gall-making insect and its host plant

Hongyuan Wei[1], Yu-Xuan Ye[2], Hai-Jian Huang[3], Ming-Shun Chen[4], Zi-Xiang Yang[1], Xiaoming Chen[1], and Chuan-Xi Zhang[2]

[1]Chinese Academy of Forestry Institute of Highland Forest Science
[2]Zhejiang University
[3]Ningbo University
[4]Kansas State University

March 25, 2022

Abstract

The horned gall aphid Schlechtendalia chinensis, is an economically important insect that induces galls valuable for medicinal and chemical industries. S. chinensis manipulates its host plant to form well-organized horned galls during feeding. So far, more than twenty aphid genomes have been reported; however, all of those are derived from free-living aphids. Here we generated a high-quality genome assembly of S. chinensis, representing the first genome sequence of a galling aphid. The final genome assembly was 280.43 Mb, with 97% of the assembled sequences anchored into thirteen chromosomes. S. chinensis presents the smallest aphid genome size among available aphid genomes to date. The contig and scaffold N50 values were 3.39 Mb and 20.58 Mb, respectively. The assembly included 96.4% of conserved arthropod and 97.8% of conserved Hemiptera single-copy orthologous genes based on BUSCO analysis. A total of 13,437 protein-coding genes were predicted. Phylogenomic analysis showed that S. chinensis formed a single clade between the Eriosoma lanigerum clade and the Aphidini+Macrosiphini aphid clades. In addition, salivary proteins were found to be differentially expressed when S. chinensis underwent host alternation, indicating their potential roles in gall formation and plant defense suppression. A total of 36 cytochrome P450 genes were identified in S. chinensis, considerably fewer compared to other aphids, probably due to its small host plant range. The high-quality S. chinensis genome assembly and annotation provide an essential genetic background for future studies to reveal the mechanism of gall formation and to explore the interaction between aphids and their host plants.

Chromosome-level genome assembly for the horned-gall aphid provides insights into interactions between gall-making insect and its host plant

Hong-Yuan Wei[1], Yu-Xuan Ye[2], Hai-Jian Huang[4], Ming-Shun Chen[3], Zi-Xiang Yang[1*], Xiao-Ming Chen[1*], Chuan-Xi Zhang[2,4*]

[1]Institute of Highland Forest Science, Chinese Academy of Forestry, Kunming, China

[2]Institute of Insect Sciences, Zhejiang University, Hangzhou, China

[3]Department of Entomology, Kansas State University, Manhattan, KS, USA

[4]State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products; Key Laboratory of Biotechnology in Plant Protection of MOA of China and Zhejiang Province, Institute of Plant Virology, Ningbo University, Ningbo, China

Contributed equally

*Correspondence

Zi-Xiang Yang, Institute of Highland Forest Science, Chinese Academy of Forestry, Kunming, China

E-mail: yzx1019@163.com

Xiao-Ming Chen, Institute of Highland Forest Science, Chinese Academy of Forestry, Kunming, China

E-mail: cafcxm@139.com

Chuan-Xi Zhang, Institute of Insect Sciences, Zhejiang University, Hangzhou, China

E-mail: chxzhang@zju.edu.cn

Abstract

The aphid *Schlechtendalia chinensis* is an economically important insect that can induce horned galls, which are valuable for the medicinal and chemical industries. Up to now, more than twenty aphid genomes have been reported. Most of the sequenced genomes are derived from free-living aphids. Here, we generated a high-quality genome assembly from a galling aphid. The final genome assembly is 271.52 Mb, representing one of the smallest sequenced genomes of aphids. The genome assembly is based on contig and scaffold N50 values of the genome sequence are 3.77 Mb and 20.41 Mb, respectively. Nine-seven percent of the assembled sequences were anchored onto 13 chromosomes. Based on BUSCO analysis, the assembly involved 96.9% of conserved arthropod and 98.5% of the conserved Hemiptera single-copy orthologous genes. A total of 14,089 protein-coding genes were predicted. Phylogenetic analysis revealed that, *S. chinensis* diverged from the common ancestor of*Eriosoma lanigerum* approximately 57 million years ago (MYA). In addition, 35 genes encoding salivary gland proteins showed differentially when *S. chinensis* forms a gall, suggesting they have potential roles in gall formation and plant defense suppression. Taken together, this high-quality *S. chinensis* genome assembly and annotation provide a solid genetic foundation for future research to reveal the mechanism of gall formation and to explore the interaction between aphids and their host plants.

Key words

Schlechtendalia chinensis ; PacBio sequencing; Chromosome-level genome assembly; Gall formation

1 Introduction

Numerous aphid species are economically important plant pests that feed on plant sap. Many plant-feeding aphids can also transmit plant viruses. Around 100 out of approximately 5,000 known aphid species are significant agricultural pests due to their feeding damages and/or disease transmission (Blackman & Eastop, 2020). Currently, studies on aphid genomes have mainly focused on the subfamily Aphidinae (International Aphid Genomics Consortium, 2010; Li et al., 2019; Mathers, 2020; Mathers et al., 2017; Mathers, Mugford, et al., 2020; Mathers, Wouters, et al., 2020; Nicholson et al., 2015; Thorpe et al., 2018; Wenger et al., 2016). Genome sequencing on species from other subfamilies that are distantly related to Aphidinae is relatively limited (Julca et al., 2020; Biallo et al., 2020). Unlike most free-living aphids, galling aphids can induce gall formation on their primary host plants and then live in galls. Galling aphids may are ideal models to study unique ecological and behavioral phenomena underlying insect-plant interactions and their coevolution (Moran, 1989; Wool, 2004). So far, only two galling aphids have been sequenced and assembled (*Eriosoma lanigerum* and *Hormaphis cornu*). The aphid*E. lanigerum* often causes bark deformation and cancer-like swelling on the roots, trunk or brunches of apple, and sometimes induces the formation of leaf-rosette galls on American elm (*Ulmus americana*) (Blackman and Eastop, 2020). The aphid, *H. cornu* , induces a gall on the underside of leaves of witch hazel,*Hamamelis virginiana* (Kurosu et al., 1992). The galls induced by*E. lanigerum* and *H. cornu* are quite different from the completely closed galls induced by *Schlechtendalia chinensis* , which has peculiar strategies to adapt to a closed environment that has extremely high levels of CO_2 honeydew, and other aphid metabolites (Chen et al., 2020).

The horned gall aphid,*S. chinensis* (Hemiptera: Aphididae: Eriosomatinae: Fordini), is one of the most economically valuable insects. Gallnuts induced by the aphids are valuable for medicinal purposes and in chemical industries. The components in gallnuts, such as tannins, are important gradients for producing inks, wine, food, cosmetic antioxidants, and animal feed. High levels of tannins (50- 70%) have been found in horned galls (Zhang, Tang, & Cheng, 2008). The annual yield of gallnuts in China is 8,000-10,000 tons, accounting for over 90% of the total yield worldwide (Zhang, Tang, & Cheng, 2008).

S. chinensis has a complex life cycle involving both sexual and asexual reproduction stages with a host alternation between the Chinese sumac (*Rhus chinensis* , Anacardiaceae) and mosses of the genus (*Plagiomnium* spp.,Mniaceae). In this holocyclic life cycle, a fundatrix produced by a mated female crawls along the trunk and feeds on a new leaf, where it induces the formation of a horned gall. The fundatrix can produce wingless fundatrigeniae in galls via parthenogenesis. In autumn, wingless fundatrigeniae will produce winged fundatrigeniae named autumn migrants. When galls become mature and burst open, the alate autumn migrants will fly to nearby mosses and produce nymphs for overwintering. In the following spring, nymphs on mosses will develop into spring winged migrants, which then fly back to the primary host, *R. chinensis*and produce both female and male offspring called sexuales. After mating, each female reproduces only one fundatrix, starting the cycle again (Figure 1) (Zhang, Qiao, Zhong & Zhang, 1999; Blackman and Eastop, 2020). This representing an unusual life cycle with comprising various morphologically distinct aphid forms at different stages, and its evolution was likely driven by the adaptation to different environmental conditions. Unlike most free-living aphids from the Aphidinae taxon, galling aphids exhibit diverse biological characteristics. For example, most galling aphid species do not seriously affect the health of their host plants. In some cases, the galls are thought to be beneficial to host plants (Chen et al., 2020).

For *S. chinensis* , the complexities in its developmental process and the structure of its induced galls imply that it may possess unique gene sets that regulate its development and manipulate its host plants (Takeda et al., 2019; Hirano et al., 2020). The molecular mechanisms underlying its complex life cycle remain largely unknown. Galls are produced through the insect-driven dramatic reprogramming of plant cell biology. Previous studies have shown that gall induction is highly species-specific, and that different galling insects deliver unique sets of effectors into plant tissues, resulting in gall formation (Zhao et al., 2015; Aljbory et al., 2018). The underlying mechanisms of the parasitic ability of galling aphids on host plants via apparently harmless galls remain unknown so far. To understand the genetic basis of the complex lifestyle, a high-quality chromosome-level genome assembly of *S. chinensis* accomplished, representing the first genome sequence of aphids that induces the formation of completely closed galls. Phylogenetic relationship between *S. chinensis* and closely related species was analyzed to better understand the unique biological characteristics of *S. chinensis* .

2 Materials and Methods

2.1 Sample collection

S. chinensis samples were collected from fresh mature galls on *R. chinensis* , in Wufeng county (30°10' N, 110deg52' E, 960 m above sea level), Hubei Province China, on October, 2019. A colony was established through artificial cultivation for further genetic studies. Briefly, autumn migrants of*S. chinensis* from mature galls, transferred to a nursery of the moss *Plagiomnium maximoviczii* , and maintained in a greenhouse. In the following year, nymphs and spring migrants (sexuparae) were harvested from mosses and cultivated in laboratory. Male and female produced by spring migrants were collected in laboratory. After fundatrix emergence, aphids were transferred to host trees for gall induction. Aphid samples were collected separately at different stages, including fundatrix, fundatrigeniae, autumn migrants, overwinter nymphs, spring migrants, male and female sexuales. Fundatrigeniae (females) contained in a gall were transferred to a petri dish after dissecting the gall. Impurities like waxes were removed manually. All aphids within a gall were presumed to be the clonal offspring of a single fundatrix, since all the *S. chinensis* galls contained only one single fundatrix that produced offspring in the gall via parthenogenesis. All aphid samples were immediately frozen in liquid nitrogen for two hours and subsequently stored at -80degC until further analysis.

2.2 Genomic and transcriptomic sequencing

Genomic DNA (gDNA) was isolated from 200 individual female and male using the DNeasy Blood & Tissue Extraction Kit (Qiagen Inc., Valencia, CA, USA), following the manufacturer's instructions. After quality and quantity measurements, the gDNA was used to construct a 150-bp paired-end sequencing library for Illumina platform. A 20 kb long-read sequencing library was constructed by gDNA isolated from 200 fundatrigeniae for PacBio Sequel II platform. For Hi-C analysis, 200 fundatrigeniae were soaked in 1% formaldehyde for 10 min at room temperature and in a 2.5 M-glycine solution to terminate the isolation and cross linking of aphid cells. The Hi-C assays and the sequencing procedures were performed via a commercial contract with Annoroad Gene Technology Co., Ltd. (Beijing, China) (Rao et al., 2014).

Transcriptomes were generated from RNA samples extracted from different stages including fundatrix, fundatrigeniae, autumn migrants, nymphs, spring migrants (sexuparae), male and female sexuales, separately. RNA quantity, purity and integrity were determined on a NanoPhotometer and an Agilent 2100 Bioanalyzer. cDNA libraries were subsequently constructed following the chain specific method. The libraries were initially quantified by the qubit 2.0 fluorometer and diluted to 1.5 ng/ul. Later, different libraries were pooled according to the requirements of effective concentration and target data volume for Illumina sequencing. Low-quality bases in the RNA-Seq raw reads were filtered using Trimmomatic (version 0.36) (Bolger, Lohse, & Usadel, 2014). Clean reads were mapped to the genome assembly using Hisat2 (version 2.1.0.5) (Kim et al., 2015), so as to obtain the putative transcripts. Transcript levels were analyzed using cufflinks (version 2.2.1) (Ghosh, & Chan, 2016).

2.3 Genome assembly

The Illumina paired end reads were used for k-mer analysis to estimate the genome size and heterozygosity with a k-mer length of 17 bases. Specifically, the k-mer number and distribution were calculated based on Jellyfish (version 1.1.10, parameters set to -C, -m 17, -s 10G, -t 80), whereas the genomic information was counted and visualized using GenomeScope (version 2.0, parameters set to 12, 150) (Ranallo-Benavidez, Jaron, & Schatz, 2020, Marcais & Kingsford, 2011). Pacbio sequencing data were used to assemble the draft genome using Wtdbg2 (version 2.5, parameters set to -t 8, -p 21, -S 4, -s 0.05, -g 274m, -L 5000) (Ruan & Li, 2020). Potential sequences from bacteria, fungi and other microorganisms were removed by aligning the genome sequences to the Nt database. Both long and short reads were utilized to correct base errors in the draft genome using NextPolish (Hu, Fang, Su, & Liu, 2019). HaploMerger2 (with default parameters) and purge_haplotigs (parameters set to -m 4G; -t 60; -l value1, -m value2, -h value3; -t 60, -a 70) were adopted to remove the heterozygous regions in the genome (Huang, Kang, & Xu, 2017, Roach, Schmidt, & Borneman, 2018).

To construct the chromosome-level genome assembly, Hi-C sequences were aligned to the haploid genome assembly using Juicer (version 1.5, with default parameters). An initial chromosome-level assembly was generated via the 3D de novo assembly (3D-DNA) (version 180114) analysis with the parameter "-r 3" (Dudchenko et al., 2017). The final chromosome-level assembly was reviewed using Juicebox Assembly Tools (JBAT, version 1.11.0, with default parameters) (Dudchenko et al., 2018). The completeness of genome assembly was assessed using BUSCO (v5.1.3) (Waterhouse et al., 2018) to scan the universal single-copy orthologous genes selected from Eukaryota, Arthropoda, Insecta and Hemiptera datasets (odb_10). The final assembly was validated based on the Illumina reads and RNA sequencing (RNA-seq) reads via bowtie2 (Table S1).

2.4 Localization of the sex chromosomes and autosomes

The mapped reads per million (MRPM) of each chromosome for female and male Illumina reads were calculated to locate the sex chromosomes and autosomes (Ye et al., 2021). The normalized read counts of the X chromosome are approximately twice higher in females than those in males, because males have only one copy of the X chromosome, while female have two copies. Both males and females have two copies in the autosomes, and the ratio of males and females is expected to approach 1 (Pal & Vicoso, 2015). Male and female DNA reads were mapped separately to the genomic scaffolds using Bowtie2 with default

parameters (Langmead & Salzberg, 2012). The resulting alignments were later filtered to remove the low-quality mapped reads via SAMtools view (-b -q 30). The read counts of each chromosome were calculated using SAMtools idxstats (Li et al., 2009). The sex chromosomes were then verified by comparison with other species. Syntenic blocks of genes were identified between the chromosome-level genome assemblies of*S. chinensis* , *Acyrthosiphon pisum* , *Rhopalosiphum maidis* , *E. lanigerum* by adopting MCSCANX and visualization via Dual Systeny Plotter for MCSCANX of the synteny visualization of TBtools (version 1.09, Chen et al., 2020) (Table S1).

2.5 Gene annotation

To predict the repetitive regions, RepeatMasker (version 4.1.1) (Tarailo-Graovac & Chen, 2009) was employed to screen the *S. chinensis* genome against the Repbase library (Bao, Kojima, & Kurtz, 2015), and the parameter was set to RepeatMasker -pa 4 -e ncbi -species Hemiptera ch -dir. Further, an aphid- specific database was generated using RepeatModeler (version 2.0.1, with default parameters), so as to predict the transposons and repetitive regions (Flynn et al., 2020). Statistical results of RepeatMasker and Repeatmodeler analyses were combined.

Gene structures were predicted using GETA pipeline (version 2.4.2, https://github.com/chenlianfu/geta) to merge the results of the RNA-seq assisted, homology-based and ab initio methods. Briefly, In the RNA-seq assisted method, RNA-seq data generated from Illumina were aligned to the assembled *S. chinensis* genome using Hisat2 (version 2.1.0.5) (Kim et al., 2015). In the homology-based method, genes were predicted based on homology to map protein sequences using GeneWise (version 2.4.1) (Birney, Michele, & Durbin, 2004). Augustus (version 2.5.5) (Stanke et al., 2006) was used to generate ab initio gene prediction (Stanke et al., 2006; Blanco, Parra & Guigo, 2007). Gene prediction results were then pooled and screened against the PFAM database.

To assign functions to the newly annotated genes in the *S. chinensis* genome, these genes were aligned to sequences in databases including NCBI Non-Redundant Protein Sequence (Nr), Non-Redundant Nucleotide Sequence Database (Nt), SwissProt, Cluster of Orthologous Groups for eukaryotic complete genomes (KOG), Integrated Resource of Protein Domains and Functional Sites (InterPro), Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes, Orthology database (KEGG), and evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG). A localBlast2GO database was also built for GO annotation, which was later processed via Blast2GO (version 2.5). The KAAS of KEGG databases were utilized to annotate the *S. chinensis* genome sequence, and then BBH pattern was chosen.

2.6 Non-coding RNA identification

Transfer RNAs (tRNAs) were identified using the tRNAscan-SE program (version 1.3.1, with default parameters for eukaryotes) (Chan & Lowe, 2019). RNAmmer (version 1.2, parameters set to "-s euk -m tsu, ssu, lsu") was used to identify 5S/ 8S, 16S/ 18S and 23S/ 28S ribosomal RNAs (rRNAs) (Karin et al., 2007). rRNAs, microRNAs (miRNAs) and small nuclear RNAs (snRNAs) were identified based on the Rfam database (version 12.2) using BLASTN (E-value [?]1x10^{-5}) (Kalvari et al., 2017).

2.7 Phylogenetic analysis

Phylogenetic trees for *S. chinensis* and eight other aphid species including *Daktulosphaira vitifoliae* , *Sipha flava* , *Aphis glycines* , *R. maidis* , *A. pisum* , *Myzus persicae* , *Diuraphis noxia* ,*E. lanigerum* were reconstructed (International Aphid Genomics Consortium, 2010; Li et al., 2019; Mathers, 2020; Mathers et al., 2017; Mathers, Mugford, et al., 2020; Mathers, Wouters, et al., 2020; Nicholson et al., 2015; Thorpe et al., 2018; Wenger et al., 2016). The whitefly, *Bemisia tabaci* was used as the outgroup. The aphid genome sequence and gene structure annotation files were downloaded from the NCBI genome database, genes containing mRNA information were retained, and the CDS was modified. The longest isoform was selected as the representative sequence of the gene. Predicted proteins encoded by all putative genes were obtained. Orthologous groups were assigned by OrthMCL (v2.0.9) (Li, Stoeckert & Roos, 2003) based on the all-versus-all BLASTP results (E-value [?]1x10^{-5}). Single copy orthologous groups were extracted

from OrthoMCL results where single copy genes covered at least 50% of all species. And if the shortest sequence of the single copy ortholog group is longer than 6000 bp, the single copy ortholog group is filtered out to avoid too long sequences that may affect the accuracy of tree. Multi-sequence alignments of single copy orthologous genes were performed using MAFFT (version 7.221, Katoh, Misawa, Kuma, & Miyata, 2002; Katoh & Standley, 2013) and the conserved amino-acid sites were identified using Gblocks (version 0.91, Clore, 2014). RAxML (version 8.1.24) (Stamatakis 2014) was employed to construct the phylogenetic tree under the GTRGAMMA model with 1000 bootstrapping replicates (Castresana, 2000). The branch length of homologous genes was analyzed with PAML (Yang, 2007), and compared with the standard tree to eliminate abnormal genes. Then, the tree was rebuilt using RAxML again (Stamatakis, 2014). By providing the root number and multiple sequence alignment results with calibration point information, the species divergence time was calculated using MCMCtree of PAML software (version 14.9). Divergence time within the evolutionary tree was obtained with 95% confidence interval (CI) (Yang, 2007). Meanwhile, divergence time and age of fossil records were derived from TimeTree (http://www.timetree.org/) and applied as the calibration points. According to the divergence times from TimeTree, the nodal dates of *Ac. pisum* and *Ap. glycines* were 28-61 million years ago (MYA), those of *D. vitifoliae* and *S. flava* were 87-162 MYA and those of *B. tabaci* and *D. vitifoliae* were 245- 351 MYA (Johnson et al., 2018).

2.8 Gene family expansion and contraction

CAFE (version 3.1) (Hahn et al., 2007) was used to analyze gene family expansion and contraction by comparing the *S. chinensis* genome with those from eight other aphid species (namely *D. vitifoliae* ,*S. flava, E. lanigerum, Ap. glycines, R. maidis, Ac. pisum, D. noxia* and *M. persicae*). Briefly, the quantitative information of gene families of 10 insects was obtained based on the OrthoMCL results. The number of gene families in each species and the trees with divergence time were used as the input information of CAFE (parameters set to "lambda -s, -t"). The best rates for gene birth and death were decided using CAFE, and all branches had the same rates of gene birth and death. Expansion and contraction of gene families were identified using CAFE (Hahn, Demuth & Han, 2007). GO and KEGG enrichment analyses were conducted using Omicshare CloudTools under default instructions settings (http://www.omicshare.com/).

2.9 Identification of genes potentially involved in gall formation and host manipulation

One hundred forty-one proteins were identified from the saliva of*S. chinensis* in a previous study (Yang et al., 2018). These identified proteins were used to identify genes potentially involved in gall formation and host manipulation. tBLASTN was used to search the corresponding genes in the *S. chinensis* genome with the 141 salivary proteins as queries (E-value [?]1x10^{-5}, identify [?] 50). The expression levels of salivary protein-encoding genes were quantified in three stages based on the RNA-seq data. Up-regulated genes in fundatrix were subject to GO and KEGG enrichment analyses using Omicshare CloudTools with default parameters (http://www.omicshare.com/).

3 Results and Discussion

3.1 Genome sequencing and *de novo* assembly

The k-mer (K=17) analysis indicated that the heterozygosity of *S. chinensis* was 0.786% and the estimated genome size was 274,512,001 bp (Figure S3). The sequencing of the fundatrigenia genome (using the PacBio Sequel II platform) generated 130 Gb raw data with an N50 length of 21,033 bp. The raw contig-level assembly was composed of 304,774,269 bases with 1,409 contigs and the N50 length of 2,961,835 bp (Table 1). After removing the heterozygosity, the length of final contig-level assembly was 271,416,320 bp with 378 contigs, and N50 length of 4,333,385 bp (Table 1).

The chromosome-level genome was generated via Hi-C data (Table S1) with a total length of 271,524,833 bp, with a scaffold of N50 20,405,002 (Table 1). More than 97% of the total genome bases were successfully anchored to the 13 chromosomes (Figure 2). The remaining 2.8% sequences was comprised 341 small scaffolds (Table 1). Chromosome lengths ranged from 14,859,000 bp to 10,104,278 bp. As revealed by BUSCO analyses against the Eukaryota, Arthropoda, Insecta and Hemiptera datasets, the*S. chinensis* genome as-

sembly contained a higher number of conserved single-copy Arthropoda genes than any other published aphid genome, suggesting the completeness and high quality of our genome assembly (Figure 4A). The genomic short reads were mapped to the assembled genome sequences, resulting in a 97.79% mapping rate and 60 Gb average sequence depth (Table S2). RNA-seq isolated from seven samples including fundatrix, fundatrigeniae, autumn migrants, nymphs, spring migrants (sexuparae), male and female sexuales, a total of 124.22 Gb raw data were generated using the Illumina platform, and more than 86% of the assembled RNA-seq transcripts were mapped to the genome (Table S3). Altogether 260,508 transcripts (280,520,495 bp in total) were produced by Trinity (Table S4).

3.2 Sex chromosomes and autosomes

Male and female Illumina paired-end reads were mapped separately to genomic scaffolds to estimate MRPM. The MRPM value of female reads for chrX1, chrX2 and chrX3 were 1,439,092, 1,333,387 and 1,051,602, whereas those for the corresponding male reads were 781,901, 726,210 and 576,946 respectively. As expected, MRPM values of female reads were roughly twice as high as those of male reads in chrX1, chrX2 and chrX3. For the other 10 chromosomes, no significant difference in total reads was observed between females and males, with the female-to-male ratio ranging from 0.90 to 1.00 (Table S5).

It has been shown that the X chromosome is conserved in aphids while chromosomal rearrangements are common for autosomes (Li et al. 2021, Mathers et al. 2021). The syntenic blocks were compared between the*S. chinensis* assembly and that of *Ac. pisum* from Macrosiphini (Li, et al., 2020), *R. maidis* from Aphidini (Chen et al. 2019), and *E. lanigerum* from Eriosomatinae (Figure 3B). The comparisons revealed high levels of genome rearrangements between autosomes. The three *S. chinensis* chromosomes were mapped to the conserved X chromosome of Macrosiphini and Aphidini, and two X chromosomes of *E. lanigerum* . The observed multiple X chromosomes were consistent with previous reports (Biello et al., 2020), which were speculated to result from the fragmentation of the X chromosome in*S. chinensis* and *E. lanigerum* or from the ancient fusion event of the large X chromosome in Aphidinae (Macrosiphini + Aphidini). This observation strongly supports that chrX1, chrX2 and chrX3 are the sex chromosomes and the karyotype of *S. chinensis* is XX+X (Yuan et al., 2021).

3.3 Genome annotation

A total of 79,136,004 bp repetitive sequences were obtained in the*S. chinensis* genome, yielding a repeat percentage of 29% (Table S6). A total of 14,089 (15,987 transcripts) genes were predicted to encode proteins. There were 97.37% of the annotated genes located on the 13 chromosome-level scaffolds (Figure 2B). The average CDS length, exon number per gene, exon length and intron length were 1,536 bp, 73, 212 bp and 910 bp, respectively, similar to those in most of the reported aphid species (Table S7, Figure S2). According to our results, 96.9%, 97.7%, 97.8% and 96.7% of BUSCO genome/gene sets were identified in the *S. chinensis* genome in comparison with Eukaryota, Arthropod, Hemiptera and Insecta datasets, respectively, demonstrating the completeness of the gene set (Figure 4B). The percentage of RNA-Seq reads assigned to a gene set was up to 80% (Table S3). Among the 14,078 predicted genes, 12,584 (89.32%) were functionally annotated, including 9,272 (65.81%) genes found via GO database and 7,285 (51.71%) genes via KEGG database (Table 2). Non-coding RNAs (ncRNAs) were also identified in the *S. chinensis*genome, including 130 tRNAs, 29 rRNAs, 29 miRNAs, and 72 snRNAs (Table S8).

3.4 Phylogenetic analysis

Protein sequences of *S. chinensis* and eight other closely related species were retrieved from public databases along, *B. tabaci* as an outgroup. A total of 3479 single copy orthologous groups extracted by OrthoMCL were incorporated to construct the phylogenetic tree. The results showed that *S. chinensis* was a sister taxon to the wooly apple aphid *E. lanigerum* . The two Eriosomatinae species diverged from their common ancestor at approximately 57 million years ago (MYA) (Figure 5). Eriosomatinae and Aphidinae (including *Ap. glycines* , *R. maidis* , *Ac. pisum* , *M. persicae* or *D. noxia*) diverged from their common ancestor at about 63 MYA, similar to the previous study (Mather et al., 2020). Compared with the subfamily Chaitophorinae (including*S. flava*) in the family Aphididae, the subfamily Eriosomatinae has a closer relationship with

the subfamily Aphidinae. Significant expansion and contraction of gene families is usually related to the adaptive divergence of species. To elucidate the key genomic changes associated with adaptation, expansion and contraction of gene families were analyzed in all the nine aphids and *B. tabaci* . Eriosomatinae showed 40 expanded and 986 contracted gene families compared with those of the common ancestor of Aphidinae and Eriosomatinae (Figure S4A). KEGG and GO enrichment analyses suggested that most of the expanded genes were involved in the detoxification of natural xenobiotics from plants (Figure S4B, S4C). *S. chinensis* genome displayed 235 expanded and 1,037 contracted gene families compared with of the common ancestor. KEGG pathway enrichment analysis suggested that most of the expanded gene families were involved in IL-17 signaling pathway, arachidonic acid metabolism, NF-kappa B signaling pathway, ovarian steroidogenesis, VEGF signaling pathway, necroptosis, regulation of lipolysis in adipocyte, TNF signaling pathway, and c-type lectin receptor signaling pathway (Figure S4E). Similarly GO annotation analysis revealed that most of the expanded gene families were involved in prostaglandin-endoperoxide synthase activity, arachidonate 15-lipoxygenase activity, nucleosomes, ovarian cumulus expansion, intrinsic apoptotic signaling pathway in response to osmotic stress, regulation of fever generation, regulation of platelet-derived growth factor production, response to lead ion, and chromatin assembly or disassembly (Figure S4D, Table S9). The expanded gene families of the*S. chinensis* genome were enriched not only in detoxification but also in the immune system.

3.5 Salivary protein-encoding genes and other gall formation associated genes

S. chinensis can induce the formation of closed galls on host plants. Previous studies have reported that gall induction is highly species-specific, and that galling insects deliver effectors into plant tissues, resulting in gall formation (Yang et al., 2018). The gall midge*Mayetiola destructor* can inject effector proteins into tissues via its saliva during feeding, leading to the conversion of a whole wheat seedling into a gall (Wang et al., 2018; Aljbory et al., 2020). A novel family of insect secreted proteins named BICYCLE have been identified in *Hormaphis cornu* , which induce gall formation on the leaves of witch hazel, *Hamamelis virginiana* (Korgaonkar et al., 2021). *BICYCLE* may regulate numerous aspects of gall development, due to their abundant expression in salivary glands specifically in gall aphids. *S. chinensis* feeds on host leaves where it presumably injects saliva into host leaf cells, resulting in gall formation. A total of 141 proteins have been identified from its salivary glands by LC-MS/MS analysis (Yang et al., 2018). In comparison with salivary proteins from 10 other free-living Hemipterans, the presence of a high proportion of proteins with binding activity is noticeable, including DNA-, protein-, ATP-, and iron-binding proteins. These proteins may be involved in gall formation. In this study, we did not identify any BICYCLE protein in the salivary glands of*S. chinensis* , suggesting the different mechanisms of gall induction between *S. chinensis* and *H. cornu* . As demonstrated by RNA-Seq analysis, transcripts corresponding to 35 genes (Sc.chr03.1184- Sc.chr10.506) that encoded salivary gland proteins exhibited high expression levels in the gall forming fundatrix stage (Figure S5). These salivary proteins were potentially related to the interaction between insects and host plants. According to their predicted functions, these genes can be divided into several categories, including detoxification, signal transduction, secreted protein metabolism, energy metabolism, basic biological processes and movement (Table S10). The largest number of genes related to detoxification may be related to defense inhibition in host plants. On the other hand, gene belonging to movement and energy metabolism categories may be associated with the contraction of salivary gland muscle and the supply of energy for salivation.

4. Conclusions

A high-quality chromosome-level genome assembly of the galling aphid*S. chinensis* was established in this study. Phylogenetic analysis indicated that *S. chinensis* diverged from *E. lanigerum* at approximately 57 million years ago (MYA). Transcriptome analysis showed that 35 genes that encoded salivary gland proteins were highly expressed in the gall forming fundatrix stage. Some of these salivary proteins might be involved in gall formation. Our results will benefit future research to study the molecular mechanisms underlying the unique biology associated with galling aphids, their gall induction ability, and molecular interactions between insects and plants.

Data accessibility statement

All data mentioned in this paper have been deposited in the National Center for Biotechnology Information with the BioProject accession number PRJNA700780 (genomic sequencing) and PRJNA702264 (transcriptome sequencing). The final DNA sequence assembly has been deposited at DDBJ/ENA/GenBank under the accession JAFHKX000000000. The genome assembly and annotation, orthogroup clustering results and salivary gland genes are available for download from Zenodo (10.5281/zenodo.3797131).

Author contributions

Z.C.X., Y.Z.X. and C.X.M conceived and designed the study. Z.C.X., Y.Z.X and W.H.Y. collected samples. W.H.Y. and Y.Y.X. performed the genome assembly, gene model prediction, gene annotation and comparative analyses. H.H.J. performed the chromosome analyses. W.H.Y., Y.Z.X. performed the transcriptome analyses. W.H.Y., Y.Z.X. and Y.Y.X. wrote the manuscript with input from all authors. Z.C.X., Y.Z.X. and C.M.S. analyzed the data and discussed the results. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank to Prof. Kirst King-Jones (University of Alberta), and Dr. Yi-Yuan Li (University of Texas at Austin), for their helpful comments.

Funding information

The national natural science foundation of China (31872305); the basic research program of Yunnan Province (202001AT070016); the grant for innovative team of Yunnan Province (202005AE160011).

Figure 1 Life cycle diagram of *Schlechtendalia chinensis*

A typical life cycle of the horned gall aphid in Hubei, China. A fundatrix (1) finds a suitable tender leaf on the primary host*Rhus chinensis* , to feed and initialize gall formation, and feeds inside the induced gall by the end of April or the beginning of May. Afterwards, the fundatrix and the wingless daughters (called fundatrigeniae) (2) reproduce for generations viviparously and parthenogenetically within the gall from May to October. The gall size increases gradually along with the growth of the aphid population in it. At the end of October, winged autumn migrants (3) emerge from the gall and fly away after the gall opened. The migrants find the moss*Plagiomnium maximoviczii* nearby where they produce nymphal offspring. (4) The nymphs feed on the moss and secrete waxes to wrap themselves up for overwintering. Winged spring migrants (5) emerge by the end of March, then fly back to the primary host and reproduce sexual females (6) and males (7) in the trunk crevices. After mating, the female reproduces a fundatrix to begin the next life cycle. * Graphs not in scale. Stippled sector indicating the in-gall stages.

Figure 2 Chromosomal analysis of *Schlechtendalia chinensis*

(a) Contact maps of Hi-C interactions among chromosomes in the *S. chinensis* genome. The heatmap was generated by Juicebox software using in situ Hi-C data (the resolution is 300 kb). (b) From the outside towards the inside, the first circle shows the 13 chromosomes. The second circle shows GC contents. The third circle represents repeat density across the genome. The fourth circle displays gene density across the genome. The fifth to eleventh circles show autumn migrant, fundatrix, fundatrigenia, nymph, spring migrant, male and female.

Figure 3 Identification of the X chromosome through syntenic blocks of chromosomal regions

Pairwise synteny relationships are shown between *S. chinensi* s and the chromosome-scale genome assemblies of three aphids. (a)*Acyrthosiphon pisum* , ApX is X chromosome. (b)*Eriosoma lanigerum* , EI5 and EI6 is X chromosome. (c)*Rhopalosiphum maidis* , Rm3 is X chromosome.

Figure 4 Assessments of BUSCO completeness

(a) The genome completeness values of *S. chinensis* , *Aphis glycines* , *Rhopalosiphum maidis* , *Acyrthosiphon pisum* ,*Myzus persicae* , *Diuraphis noxia* and *Eriosoma lanigerum* assessed by the recovery of universal single-copy genes (BUSCOs) using the Arthropoda gene set (odb_10 and odb_9). (b) The gene set completeness of the predicted gene model of *S. chinensis* . The genome completeness and gene set completeness were calculated using BUSCO against Eukaryota, Arthropoda, Insecta and Hemiptera. C: complete BUSCOs, S: complete and single-copy BUSCOs, D: complete and duplicated BUSCOs, F: fragmented BUSCOs, M: missing BUSCOs.

Figure 5 Timing of inferred divergence of *Schlechtendalia chinensis* and other nine insects

References

Aljbory, Z., El-Bouhssini, M., & Chen, M.S. (2018) Conserved and unique putative effectors expressed in the salivary glands of three related gall midges species. *Journal of Insect Science 18*(5): 15. https://doi.org/10.1093/jisesa/iey094 Aljibory, Z., Micheal, J. A., Park, Y., Reeck, G. R., & Chen M. S. (2020). Differential localization of Hessian fly candidate effectors in resistant and susceptible wheat plants. *Plant direct, 00*: 1-15. https://doi.org/10.1002/pld3.246 Bao, W. D., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6. https://doi.org/10.1186/s13100-015-0041-9 Biello, R., Singh, A., Godfrey, C. J., Fernandez, F. F., Mugford, S. T., & Powell, & G., Hogenhout, S. A., Mathers, T. C. (2021). A chromosome level genome assembly of the woolly apple aphid, *Eriosoma lanigerum* Hausmann (Hemiptera: Aphididae). *Molecular Ecology Resources, 21*(1), 316-326. https://doi.org/10.1111/1755-0998.13258 Birney, E., Clamp, M., & Durbin, Rd. (2004). GeneWise and Genomewise.*Genome Research, 14*(5), 988-995. https://doi.org/10.1101/gr.1865504 Blackman, R. L., & Eastop, V. F. (2020). *Aphids on the world's plants: An online identification and information guide.* Chichester, UK, John Wiley & Sons Ltd. http://www.aphidsonworldsplants.info/. Blanco, E., Parra, G., & Guigo, R. (2002). *Using geneid to Identify Genes.* Chichester, UK, John Wiley & Sons Ltd. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,30*(15), 2114-2120. https://doi.org/ 10.1093/bioinformatics/btu170 Carolan, J. C., D Caragea, Reardon, K. T., Mutti, N. S., & Edwards, O. R. (2011). Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): a dual transcriptomic/proteomic approach. *Journal of Proteome Research*, *10*(4), 1505-18. https://doi.org/10.1021/pr100881q Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology Evolution, 17*(4), 540-552.*https://doi.org/10.1093/oxfordjournals.molbev.a026334* Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods in Molecular Biology, 1962*:1-14. In book: Gene Prediction.*https://doi.org/10.1007/978-1-4939-9173-0_1* Chen, C., Chen, H., Y Zhang, Thomas, H. R., Frank, M., H., He, Y., & Xia, R. (2020). Tbtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant,13*(8).*https://doi.org/10.1016/j.molp.2020.06.009* Chen, W., Shakir, S., Bigham, M., Fei, Z., & Jander, G. (2019). Genome sequence of the corn leaf aphid (*Rhopalosiphum maidis* Fitch).*GigaScience*, *8*(4). https://doi.org/ 10.1093/gigascience/giz033 Chen, X. M., Yang, Z. X., Chen, H., Qi, Q., Liu, J., & Wang, C., Shao, S. X., Lu, Q., Li, Y., Wu, H. X., King-Jones, K., & Chen, M. S. (2020). A complex nutrient exchange between a gall-forming aphid and its plant host. *Frontiers in Plant Science, 11*, 811. https://doi.org/10.3389/fpls.2020.00811 Clore, A. (2014). gBlocks gene fragments for gene construction and more. *Journal of Immunological Methods, 188*(1), 165-167. https://doi.org/10.1016/0022-1759(95)00229-4 Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., P., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds.*Science, 356*, 92-95.*https://doi.org/10.1126/science.aal3327*Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C. & Aiden, E. L. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *BioRxiv.* https://doi.org/10.1101/254797 Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS,*

117(17), 9451-9457.*https://doi.org/10.1073/pnas.1921046117*Ghosh, S., & Chan, C. K. (2016). Analysis of rna-seq data using tophat and cufflinks. *Methods in Molecular Biology*, 1374, 339-61. https://doi.org/10.1007/978-1-4939-3167-5_18 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., & Orvis, J., White, O., Buell, C. R., & Wortman J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology, 9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7 Hahn, M. W., Demuth, J. P., & Han, S. G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics, 177*(3). https://doi.org/10.1534/genetics.107.080077 Hirano, T., Kimura, S., Sakamoto, T., Okamoto, A., Nakayama, T., Matsuura, T., Ikeda, Y., Takeda, S., Suzuki, Y., OhshimaI., & Sato, M. H. (2020). Reprogramming of the developmental program of *Rhus javanica* during initial stage of gall induction by*Schlechtendalia chinensis*. *Frontiers in Plant Science*,*11*, 471.*https://doi.org/10.3389/fpls.2020.00471*Hu, J., Fang, J. P., Su, Z. Y., & Liu, S. L. (2019). NextPolish: a fast and efficient genome polishing tool for long-read assembly.*Bioinformatics, 36*(7), 2253-2255. https://doi.org/10.1093/bioinformatics/btz891 Huang, S. F., Kang, M. J., & Xu, A. L. (2017). HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics 16*, 2577.*https://doi.org/10.1093/bioinformatics/btx220*International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum*. *Plos Biology, 8*(2), 1-25. https://doi.org/10.1371/journal.pbio.1000313 Julca, I., Marcet-Houben, M., Cruz, F., Vargas-Chavez, C., Johnston, J. S., Gomez-Garrido, J., Frias, L., Corvelo, A., Loska, D., Camara, F., Gut, M., Alioto, T., Latorre, A., & Gabaldon, T. (2020). Phylogenomics identifies an ancestral burst of gene duplications predating the diversification of aphidomorpha. *Molecular Biology and Evolution*,*37*(3), 730-756. https://doi.org/10.1093/molbe v/msz261 Johnson, K. P., Dietrich, C. H., Friedrich, F., Beutel, R. G., Wipfler, B., & Peters, R. S., et al. (2018). Phylogenomics and the evolution of hemipteroid insects. *PNAS*, *115*(50). https://doi.org/10.1073/pnas.1815820115 Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*(14), 3059-3066. https://doi.org/10.1093/nar/gkf436 Katoh, K., & Standley, D. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability.*Molecular Biology and Evolution, 30*(4), 772-780. https://doi.org/10.1093/molbev/mst010 Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., & Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, *46* (Database issue), D335-D342. http://dx.doi.org/10.1093/nar/gkx1038 Karin, L., Peter, H., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research, 35*(9), 3100-3108.*https://doi.org/10.1093/nar/gkm160*Korgaonkar, A., Han, C., Lemire, A. L., Siwanowicz, I., & Stern, D. L. (2021). A novel family of secreted insect proteins linked to plant gall development. *Current Biology*, *31*(9):2038. https://doi.org/10.1016/j.cub.2021.03.001. Kim, D., Landmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*,*12*(4), 357-360. https://doi.org/ 10.1038/nmeth.3317 Kurosu, U., & Aoki, S. (1992). Gall cleaning by the aphid *Hormaphis betulae*. *Journal of Ethology*, *9*, 51-55. https://doi.org/10.1007/BF02350191 Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., & Walters, J. R. (2019). Insect genomes: progress and challenges. *Insect Molecular Biology, 28*(6), 739-758. https://doi.org/10.1111/imb.12599 Li, L., Stoeckert, C. J., & Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research, 13*(9), 2178-2189. https://doi.org/10.1101/gr.1224503 Liu, P., Yang, Z. X., Chen, X. M., & Foottit, R. G. (2014). The Effect of the gall-forming aphid *Schlechtendalia chinensis* (Hemiptera: Aphididae) on leaf wing ontogenesis in *Rhus chinensis*(Sapindales: Anacardiaceae). *Annals of the Entomological Society of America, 107*(1), 242-250. http://www.bioone.org/doi/full/10.1603/AN13118 Li, Y., Park, H., Smith, T. E., & Moran, N. A. (2019). Gene family evolution in the pea aphid based on chromosome-level genome assembly.*Molecular Biology and Evolution, 36*(10), 2143-2156.*https://doi.org/10.1093/molbev/msz138*Li, Y., Zhang, B., & Moran, N. A. (2020). The aphid X chromosome is a dangerous place for functionally important genes: diverse evolution of Hemipteran genomes based on chromosome-level assemblies. *Molecular Biology and Evolution, 37*(8), 2357-2368. https://doi.org/ 10.1093/molbev/msaa095 Mathers, T. C. (2020). Improved genome assembly and annotation of the soybean aphid (*Aphis glycines* Matsumura). *G3: Genes, Genomes, Genetics, 10*(3),

g3.400954.2019. https://doi.org/10.1534/ g3.119.400954 Mathers, T. C., Chen, Y., Kaithakottil, G., Legeai, F., Mugford, S. T., Baa-Puyoulet, P., Bretaudeau, A., Clavijo, B., Colella, S., Collin, O., Dalmay, T., Derrien, T., Feng, H., Gabaldon, T., Jordan, A., Julca, I., Kettles, G. J., Kowitwanich, K., Lavenier, D., ... Hogenhout, S. A. (2017). Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biology, 18*(1), 27. https://doi.org/10.1186/s13059-016-1145-3 Mathers, T. C., Mugford, S. T., Hogenhout, S. A. T., & Tripathi, L. (2020). Genome sequence of the banana aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and its symbionts. *G3: Genes, Genomes, Genetics, 10*(12), 4315-4321. https://doi.org/10.1534/g3.120.401358 Mathers, T. C., Wouters, R. H. M., Mugford, S. T., Swarbreck, D., Van Oosterhout, C., & Hogenhout, S. A. (2020). Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome. *Molecular Biology and Evolution, 38*(3):856-875. https://doi.org/10.1093/molbev/msaa246 Marcais, Guillaume, Kingsford, & Carl. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*, 764-770. https://doi.org/10.1093/bioinformatics/btr011 Moran, N. A. (1989). A 48-million-year-old aphid-host plant association and complex life cycle: biogeographic evidence. *Science, 245*(4914), 173-175. https://doi.org/10.1126/science.245.4914.173 Nicholson, S. J., Nickerson, M. L., Dean, M., Song, Y., Hoyt, P. R., Rhee, H., Kim, C., & Puterka, G. J. (2015). The genome of *Diuraphis noxia*, a global aphid pest of small grains. *BMC Genomics, 16*(1), 1-16. https://doi.org/10.1186/s12864-015-1525-1 Pal, A. & Vicoso, B. (2015). The X chromosome of hemipteran insects: conservation, dosage compensation and sex-biased expression. *Genome Biology and Evolution, 7*(12), 3259-3268.https://doi.org/10.1093/gbe/evv215 Quan, Q. M., Hu, X., Pan, B. H., Zeng, B. S., Wu, N. N., Fang, G. Q., Cao, Y. H., Chen, X. Y., Li, X., Huang, Y. P., & Zhan, S. (2019). Draft genome of the cotton aphid *Aphis gossypii*. *Insect Biochemistry and Molecular Biology, 105*, 25-32. https://doi.org/10.1016/j.ibmb.2018.12.007 Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., & Lander, E. S. (2014). A 3D map of the human genome at Kilobase resolution reveals principles of chromatin looping. *Cell, 158*, 1-6. https://doi.org/10.1016/j.cell.2014.11.021 Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications, 11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3 Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods, 17*(Supp l 6), 1-4.*https://doi.org/10.1101/530972*Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics, 19*(1). https://doi.org/10.1186/s12859-018-2485-7 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033 Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research, 34* (Web Server issue), W435-439. https://doi.org/10.1093/nar/gkl200 Thorpe, P., Escudero-Martinez, C. M., Cock, P. J. A., Eves-van den Akker, S., & Bos, J. I. B. (2018). Shared transcriptional control and disparate gain and loss of aphid parasitism genes. *Genome Biology and Evolution, 10*(10), 2716-2733. https://doi.org/10.1093/gbe/evy183 Takeda, S., Yoza, M., Amano, T., Ohshima, I., Hirano, T., Sato, M. H., Sakamoto, T., & Seisuke Kimura, S. (2019) Comparative transcriptome analysis of galls from four different host plants suggests the molecular mechanism of gall development. *PLoS One, 14*(10), e0223686.*https://doi.org/10.1371/journal.pone.0223686*Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, Chapter 4, Unit 4.10.*https://doi.org/10.1002/0471250953.bi0410s25*Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q. D., 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), e112963.*https://doi.org/10.1371/journal.pone.0112963*Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., . . . Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543-548. https://doi.org/10.1093/molbev/msx319 Wang, Z., Ge, J., Chen, H., Cheng, X., Yang, Y., Li, J., Whitworth, R. J., & Chen M. C. S. (2018). An insect nucleoside diphosphate kinase (NDK) functions as an effector protein in wheat - Hessian fly interactions. *Insect*

Biochemistry and Molecular Biology, 100, 30-38. https://doi.org/10.1016/j.ibmb.2018.06.003 Wenger, J. A., Cassone, B. J., Legeai, F., Johnston, J. S., Bansal, R., Yates, A. D., Coates, B. S., Pavinato, V. A. C., & Michel, A. (2016). Whole genome sequence of the soybean aphid, *Aphis glycines.Insect Biochemistry and Molecular Biology*, *123*, 102917. https://doi.org/10.1016/j.ibmb.2017.01.005 Wool, D. Galling aphids: specialization, biological complexity, and variation. (2004). *Annual Review of Entomology, 49*(1), 175. https://doi.org/10.1146/annurev.ento.49.061802.123236 Yang, Z. H. (2007). PAML 4: phylogenetic analysis by maximum likelihood.*Molecular Biology and Evolution*, *24*(8), 1586-1591. https://doi.org/10.1093/molbev/msm088 Yang, Z. X., Ma, L., Francis, F., Yang, Y., Chen, H., Wu, H. X., & Chen, X. M. (2018). Proteins identified from saliva and salivary glands of the Chinese gall aphid*Schlechtendalia chinensis*. *Proteomics, 18*, 1700378. https://doi.org/10.1002/pmic.201700378 Ye, Y. X., Zhang, H. H., Li, D. T., Zhou, J. C., Shen, Y., Hu, Q. L., & Zhang, C. X. (2021). Chromosome level assembly of the brown planthopper genome with a characterized Y chromosome. *Molecular Ecology Resources*, *21*:1287-1298.https://doi.org/10.1111/1755-0998.13328 Yuan, H., Huang, Y., Mao, Y., Zhang, N., & Mao, S. (2021). The evolutionary patterns of genome size in ensifera (insecta: orthoptera). *Frontiers in Genetics*, *12*, 693541. https://doi.org/10.3389/fgene.2021.693541 Zhang, C. X., Tang, X. D., & Cheng, J. A. (2008). The utilization and industrialization of insect resources in China. *Entomological research, 38*, S38-S47. https://doi.org/10.1111/j.1748-5967.2008.00173.x Zhang, G. X., Qiao, G. X., Zhong, T. S., & Zhang, W. Y. (1999). *Fauna Sinica, Insecta* Vol. 14 Homoptera, Mindaridae and Pemphigidae. Science Press, Beijing.

Zhao, C. Y., Escalante, L.N., Chen, H., Benatti, T. R., Qu, J. X., Chellapilla, S., Waterhouse, R. M., Wheeler, D., Andersson, M. N., Bao, R., Batterton, M., Behura, S. K., Blankenburg, K. P., Caragea, D., Carolan, J. C., Coyle, M., El-Bouhssini M., Francisco L., ... Richards S. (2015). A massive expansion of effector genes underlies gall-formation in the wheat pest *Mayetiola destructor* .*Current Biology, 25* (5): 613-620. https://doi.org/10.1016/j.cub.2014.12.057