OXFORD G3 Genes | Genomes | Genetics

G3, 2022, **12(8)**, jkac161
https://doi.org/10.1093/g3journal/jkac161
Advance Access Publication Date: 25 June 2022
Genome Report

A chromosome-level genome assembly of the orange wheat blossom midge, *Sitodiplosis mosellana* Géhin (Diptera: Cecidomyiidae) provides insights into the evolution of a detoxification system

Zhongjun Gong ,[†] Tong Li,[†] Jin Miao , Yun Duan, Yueli Jiang, Huiling Li, Pei Guo, Xueqin Wang, Jing Zhang, Yuqing Wu*

Institute of Plant Protection, Henan Academy of Agricultural Sciences, Key Laboratory of Crop Pest Control of Henan Province, Key Laboratory of Crop Integrated Pest Management of the Southern of North China, Ministry of Agriculture of the People's Republic of China, Zhengzhou 450002, P. R. China

*Corresponding author: Institute of Plant Protection, Henan Academy of Agricultural Sciences, Key Laboratory of Crop Pest Control of Henan Province, Key Laboratory of Crop Integrated Pest Management of the Southern of North China, Ministry of Agriculture of the People's Republic of China, Zhengzhou 450002, P. R. China. Email: yuqingwu36@hotmail.com
[†]These authors contributed equally to this work.

Abstract

The orange wheat blossom midge *Sitodiplosis mosellana* Géhin (Diptera: Cecidomyiidae), an economically important pest, has caused serious yield losses in most wheat-growing areas worldwide in the past half-century. A high-quality chromosome-level genome for *S. mosellana* was assembled using PacBio long read, Illumina short read, and Hi-C sequencing technologies. The final genome assembly was 180.69 Mb, with contig and scaffold N50 sizes of 998.71 kb and 44.56 Mb, respectively. Hi-C scaffolding reliably anchored 4 pseudo-chromosomes, accounting for 99.67% of the assembled genome. In total, 12,269 protein-coding genes were predicted, of which 91% were functionally annotated. Phylogenetic analysis indicated that *S. mosellana* and its close relative, the swede midge *Contarinia nasturtii*, diverged about 32.7 MYA. The *S. mosellana* genome showed high chromosomal synteny with the genome of *Drosophila melanogaster* and *Anopheles gambiae*. The key gene families involved in the detoxification of plant secondary chemistry were analyzed. The high-quality *S. mosellana* genome data will provide an invaluable resource for research in a broad range of areas, including the biology, ecology, genetics, and evolution of midges, as well as insect–plant interactions and coevolution.

Keywords: chromosome-level genome; Hi-C; *Sitodiplosis mosellana*; comparative genomics; detoxification

Introduction

Gall midges (family Cecidomyiidae) constitute one of the largest families of Diptera, with 6,651 known species and 832 genera (Gagné and Jaschhof 2021). About 75% of Cecidomyiinae species are herbivorous, and many of them induce galls in a great diversity of plants throughout the world (Yukawa and Rohfritsch 2005; Gagné and Jaschhof 2017; Dorchin *et al.* 2019). Meanwhile, most herbivorous gall midges are host specific, developing in one or a few closely related host plants, and many genera and even tribes have specialized and diversified on specific plant families (Gagné and Jaschhof 2017; Dorchin *et al.* 2019). These provide fascinating material for ecological and evolutionary studies.

The orange wheat blossom midge (OWBM), *Sitodiplosis mosellana* Géhin (Diptera: Cecidomyiidae), is an economically important pest and has caused serious yield losses in most wheat-growing areas worldwide (Berzonsky *et al.* 2002; Thomas *et al.* 2005; Bruce *et al.* 2007; Gaafar *et al.* 2011; Gong *et al.* 2013; Jacquemin *et al.* 2014). Larvae feed on young kernels, causing kernel damage, poor seed quality, and lower yield. Moderate invasion by *S. mosellana* led to a yield loss of 10–30% in China. The reduction could be as much as 30–60% when severe damage occurs (Duan *et al.* 2013). During the long course of coevolution, insects and host plants have formed intimate relationships, particularly for parasitic insect species. Insect attacks on plants cause extensive changes in gene expression in host plants. In turn, plant defense reactions to insects may cause significant changes in gene expression in insects (Mittapalli *et al.* 2005). Like other gall midges, wheat midge larvae are thought to inject saliva into developing wheat seeds, resulting in shriveled wheat kernels (Lamb *et al.* 2000). Meanwhile, adult oviposition must coincide with wheat heading, since the susceptibility of plants to wheat midge damage declines significantly after anthesis begins (Elliott and Mann 1996; Wu *et al.* 2015). The gene families involved in xenobiotic detoxification will be key in facilitating the successful exploitation of wheat.

Recent studies have focused on the mechanisms of diapause (Li *et al.* 2012), chemical communication (Gong *et al.* 2013; Cheng *et al.* 2020), long-distance migration (Miao *et al.* 2013), evaluation of midge resistance (Hao *et al.* 2019; Zhang *et al.* 2020), biological

Received: January 28, 2022. **Accepted:** June 22, 2022

control (Chavalle *et al.* 2015; Thompson and Reddy 2016; Olfert *et al.* 2020), and wheat midge–wheat interactions (Al-jbory *et al.* 2018). Because of its agricultural importance as the major pest of wheat worldwide, more knowledge about S. *mosellana* and its interaction with its host at the molecular level would be useful and benefit from comprehensive genomic analysis. A genome of S. *mosellana* was released by Agriculture and Agri-food Canada (NCBI: GCA_009176505.1). However, the genome was fragmented without anchoring on chromosomes. The quality of genome assemblies is the foundation of understanding these biological features; therefore, a more accurate S. *mosellana* genome assembly is needed.

Here, we reported a high-quality chromosome-level genome assembly for S. *mosellana* using a combination of Illumina, PacBio, and Hi-C technologies. The assembly had high completeness, providing an excellent genomic resource for subsequent research. Moreover, we described cytochrome P450 monooxygenase (P450) and glutathione S-transferase (GST) in S. *mosellana*. Our findings provide a valuable genomic resource for molecular and gene family evolutionary studies in midges (e.g. gall-forming evolution), as well as insect–host adaptive evolution, identifying genetic modifications that contribute to its insect–plant lifestyle.

Materials and methods

Insects

For genome sequencing, about 50 individuals of S. *mosellana* larvae were collected from the wheat ear in Zhumadian city, Henan province, China, in May 2019. At this time, the midge remained a mature third instar and no longer feeds. Prior to DNA preparation, S. *mosellana* larvae were starved for 15 days in pure water, and the water was replaced every 2 days to reduce the probability of contamination of the host.

Genome sequencing and assembly

High-quality DNA extracted from the larvae was used for library preparation and high-throughput sequencing. Short-insert (350 bp) paired-end libraries were prepared according to the Illumina protocol and sequenced on the Illumina NovaSeq 6000 (Illumina, Inc.) with a paired-end 150 (PE150) read layout and yield a total 28.67 Gb of sequencing data. This content corresponded to 171.49-fold genome coverage. Whole-genome sequencing was performed using the PacBio Sequel sequencer (Pacific Biosciences). The 20-kb single-molecule real-time sequencing bell libraries were constructed using the standard protocol. The PacBio Sequel platform generated 17.82 Gb of sequencing data, representing 106.59-fold coverage depth.

All raw reads from the PacBio platform were subjected to error correction according to the rate of insertions, deletions, and sequencing errors between the base pairs to obtain preassembled reads 'daligner' (Myers 2014). Then, the preassembled reads were assembled by the consensus algorithm called Overlap-Layout-Consensus to obtain contigs using FALCON (Chin *et al.* 2016). Overlapping reads and raw subreads were processed to generate consensus sequences, and the error correction of the assembly was polished using the consensus-calling algorithm Quiver (Chin *et al.* 2013). The paired-end clean reads from the Illumina platform were further corrected for any remaining errors using Pilon (Walker *et al.* 2014). Finally, the Purge Haplotigs pipeline was run to produce an improved, deduplicated assembly (Roach *et al.* 2018).

To assist chromosome-level assembly, we used the high-throughput chromosome conformation capture (Hi-C) technique to capture genome-wide chromatin interactions (Belaghzal *et al.* 2017). For Hi-C sequencing, the chromosomal structure was fixed by formaldehyde crosslinking, and then the MboI enzyme was used to shear DNA. A Hi-C library with 350 bp insert size was constructed, which was sequenced on an Illumina NovaSeq 6000 according to the manufacturer's instructions. The Hi-C library generated 30.35 Gb sequencing data, which corresponds to approximately 181.54-fold coverage of the S. *mosellana* genome. We then used the pruning, partition, rescue, optimization, and building of the ALLHiC pipeline (Dudchenko *et al.* 2017; Zhang *et al.* 2019) to construct the chromosome-level scaffolds of the S. *mosellana* genome.

Transcriptome sequencing

To assist in genome annotation, different transcriptome profiles were generated from different sexes (male and female adults) and various developmental stages (diapause larvae, nondiapause larvae, pupae) with 2 replications (NCBI SRA: SRX 12027781–SRX 12027792). In total, 12 cDNA libraries were prepared. A total amount of 3-μg RNA per sample was used as input material for RNA sample preparation. Sequencing libraries were generated using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, USA) following the manufacturer's recommendations. Libraries were sequenced on an Illumina NovaSeq 6000 platform, and paired-end reads were generated.

Transcriptome reads assemblies were generated with Trinity (v2.1.1) for the genome annotation. To optimize the genome annotation, the RNA-Seq reads from different tissues, which were aligned to genome fasta using Hisat (v2.0.4)/TopHat (v2.0.13) with default parameters to identify exons region and splice positions. The alignment results were then used as input for Stringtie (v1.3.3)/Cufflinks (v2.1.1) with default parameters for genome-based transcript assembly. The nonredundant reference gene set was generated by merging genes predicted by 3 methods with EvidenceModeler (EVM, v1.1.1) (Haas *et al.* 2008) using program to assemble spliced alignment (PASA) (Haas *et al.* 2003) terminal exon support and including masked transposable elements (TEs) as input into gene prediction. Individual families of interest were selected for further manual curation by relevant experts.

Assessment of completeness of the assembly and gene set

To assess the accuracy of the assembled S. *mosellana* genome, a small fragment library was selected for comparison of the assembled genome using BWA software (http://bio-bwa.sourceforge.net/) (Li and Durbin 2009). To assess the completeness of the assembly and gene annotation, we performed analysis with Benchmarking Universal Single-Copy Orthologs (BUSCO, version 3.0) (http://busco.ezlab.org/) with default parameters (Waterhouse *et al.* 2018).

Genome annotation

Repeat sequences and TEs were identified using both homology-based and de novo prediction methods. For de novo predictions, LTR_FINDER (v1.0.6) (http://tlife.fudan.edu.cn/ltr_finder/), RepeatScout (v1.0.5) (http://www.repeatmasker.org/), and RepeatModeler (v2.0.1) (http://www.repeatmasker.org/RepeatModeler.html) were used to construct a de novo repeat library with default parameters, then all repeat sequences with lengths >100 bp and gap "N" less than 5% constituted the raw TE library. Prediction of tandem repeats were also searched using Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) with the following parameters: Match= 2, Mismatch= 7, Delta= 7, PM= 80, PI= 10, Minscore= 50, MaxPeriod= 2,000. For homology-based predictions, RepeatMasker (v4.1.0) (http://repeatmasker.org/) was used with Repbase library (Bao *et al.* 2015). In addition, we used

RepeatProteinMask (http://www.repeatmasker.org/) with the default parameters to identify repeat sequences at the protein level.

The tRNAs were predicted using the tRNAscan-SE program, whereas ribosomal RNAs (rRNAs) were predicted using BLASTN. Other ncRNAs, including microRNAs (miRNAs) and small nuclear RNAs (snRNAs), were identified by searching against the Rfam database with default parameters using the INFERNAL software.

For gene structure prediction, homology-based prediction, ab initio prediction, and transcriptome-based prediction were combined to predict protein-coding genes in the *S. mosellana* genome based on the repeat masked genome. For the former, protein sequences from *Mayetiola destructor*, *Belgica antarctica*, *Drosophila melanogaster*, *Aedes aegypti*, *Culex quinquefasciatus*, *Anopheles gambiae*, *Bactrocera dorsalis* were aligned to the *S. mosellana* genome using TBLASTN (*E*-value ≤ 1E-05). Then Genewise (version 2.4.1) (Birney *et al.* 2004) was used for further precise alignment and gene structure prediction. For the ab initio-based method, Augustus (v3.2.3) (Stanke *et al.* 2006), GlimmerHMM (v3.0.4) (Majoros *et al.* 2004), SNAP (v2013.11.29) (Leskovec and Sosič 2016), Geneid (v1.4) (Parra *et al.* 2000), and Genscan (v1.0) (Burge and Karlin 1997) were employed as engines to predict gene models. For RNA-seq-based prediction, transcriptome data from 12 samples were aligned to the assembled genome sequence using Hisat2 (v2.0.4) (Kim *et al.* 2015). The alignment results were then used as input for Stringtie (v1.3.3) (Pertea *et al.* 2015) with default parameters for genome-based transcript assembly. The gene prediction results derived from 3 strategies were merged using EVM (v1.1.1) (Haas *et al.* 2008) to generate a consensus gene set. Finally, PASA (Haas *et al.* 2003) was used to acquire the final gene structures after adjusting the gene models generated from EVM with the transcripts assembled by Trinity (v2.1.1) (Grabherr *et al.* 2011).

Gene functional annotation was performed based on homologue searches and the best match to the databases of Kyoto encyclopedia of genes and genomes (KEGG: http://www.genome.jp/kegg/) (Kanehisa and Goto 2000), SwissProt (http://www.uniprot.org/) (Bairoch and Apweiler 2000), nonredundant proteins (NR: http://www.ncbi.nlm.nih.gov/protein), and Pfam (http://pfam.xfam.org/) (El-Gebali *et al.* 2019). The Gene Ontology (GO: http://www.geneontology.org/) (Ashburner *et al.* 2000) analysis was executed through InterPro (https://www.ebi.ac.uk/interpro/) (Mulder and Apweiler 2007) to identify protein domains. The information from different sources of functional annotation was combined for each gene in the final integration.

Comparative genomics and phylogenetic reconstruction

Protein-coding genes from another 11 species of Diptera, as well as 1 species of Coleoptera, 1 species of Lepidoptera, 1 species of Hymenoptera, and 1 common water flea (*Daphnia pulex*), in the order of Anomopoda, were obtained from the NCBI genomes database for comparative analysis. *D. pulex* was used as an outgroup. Orthologues were identified using OrthoMCL (Li *et al.* 2003). Orthologous groups that contained only 1 gene for each species were represented by the gene encoding the longest protein sequence. Genes encoding protein sequences shorter than 50 amino acids were filtered out to exclude putative fragmented genes. All-vs-all BLASTP was applied to identify similarities among the filtered protein sequences in these species with an *E*-value cut-off of 1e^{-5}. Muscle (Edgar 2004) with default parameters was used to generate a multiple sequence alignment of the protein sequences in each single copy family. The alignments of each family were then concatenated to form a super alignment that was used for phylogenetic tree reconstruction using RAxML maximum-likelihood methods with the model LG + F+I + G4 with "-m PROTGAMMAAUTO -p 12345 -x 12345 -# 100 -f ad" (Guindon *et al.* 2010; Yang and Rannala 2012; Stamatakis 2014). Statistical support was obtained with 1,000 bootstrap replicates. The species divergence time was estimated using MCMCTree in the PAML version 4.9 package (Yang 2007) with default parameters. The calibration information for MCMCTree was extracted based on the TimeTree database (http://www.time.org/).

Gene family expansion and contraction

To further explore gene family changes under natural selection, the expansion and contraction of gene families were identified using the likelihood model originally implemented in the software package CAFE version 4.2 (De Bie *et al.* 2006) with the following parameters: "-p 0.05 -t 4 -r 10000." Gene families only present in *S. mosellana* but absent in other species were considered group specific. We used Fisher's exact test to identify overrepresented GO and KEGG pathways among the expanded and contracted genes, followed by a false discovery rate correction (FDR < 0.05).

Synteny analysis

Whole-genome sequence alignments between *S. mosellana*, *D. melanogaster*, and *A. gambiae* were detected and plotted using mcscan (https://github.com/tanghaibao/jcvi/wiki/MCscan-; Python-version) with default parameters (Tang *et al.* 2008). The chromosome-level genome assembly of *D. melanogaster* (Adams *et al.* 2000) was downloaded at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001215.4#/def. The chromosome-level genome assembly of *A. gambiae* (Holt *et al.* 2002) was downloaded at https://www.ncbi.nlm.nih.gov/assembly/GCF_000005575.2#/def.

Identification and phylogenetic analysis of detoxification genes in the *S. mosellana* genome

To uncover the potential detoxification genes, the *S. mosellana* predicted protein sequences were used as queries in the blast searches to the NCBI Nr database with 1×10^{-5} *E*-value threshold. In this study, the *S. mosellana*-predicted protein sequences were classified into Sm-P450 and Sm-GST sequences, as the one of the top 10 hits of them was annotated by cytochrome P450 and GST, respectively.

The protein sequences of detoxification genes were retrieved in genomes of *D. melanogaster* (assembly Release 6 plus ISO1 MT) and *Mayetiola destructor* (Zhao *et al.* 2015) to uncover the phylogenetic positions of *S. mosellana*-related genes. The sequences were aligned with MUSCLE, as implemented in MEGA 7.0 (Kumar *et al.* 2016). The phylogenetic analysis was performed using IQ-TREE 1.6.6 (Nguyen *et al.* 2015). The substitution model was selected in ModelFinder (Kalyaanamoorthy *et al.* 2017) with the Bayesian information criterion. The ultrafast bootstraps were resampled with 5,000 runs to assess the support for each node. The phylogenetic trees were visualized using the ggtree R package (Yu *et al.* 2017).

Results and discussion

Features of the assembled genome

Based on the Illumina reads, the genome size of *S. mosellana* was estimated to be 167.18 Mb, based on 17 K-mer analysis (Supplementary Fig. 1 and Supplementary Table 2). The heterozygosity rate was 1.94%. The high heterozygosity of the *S. mosellana* genome might be caused by pooling the DNA of multiple individuals for short-read sequencing. Our study demonstrates that the current methods are appropriate for high-quality de novo assembly of the genome of small highly heterozygous organism sequencing projects (Lian *et al.* 2019; Guo *et al.* 2020; Ye *et al.* 2020).

The genome of the OWBM *S. mosellana*, sequenced using both PacBio and Illumina HiSeq 2000 platforms, generated 17.8 Gb PacBio long reads and 28.7 Gb Illumina short reads, with 278.11× genome coverage. We obtained a reference *S. mosellana* genome of 180.66 Mb with a contig N50 of 988.71 kb. The GC content of the *S. mosellana* genome was 36.4% (Supplementary Fig. 2 and Supplementary Table 2). Hi-C technology was then used to improve genome assembly to the chromosomal level. A total of 30.35 Gb clean reads were generated, accounting for 181.54-fold coverage. The Hi-C scaffolding was able to anchor and order all 25 scaffolds into 4 chromosomes, with more than 91.74% of assembled bases located on the chromosomes (Fig. 1 and Supplementary Fig. 3 and Supplementary Tables 3 and 4). The length of the largest chromosome was 53.05 Mb, while the smallest chromosome was 40.43 Mb (Supplementary Table 4). The final genome assembly was approximately 180.69 Mb, with scaffold and contig N50 sizes of 44.56 Mb and 998.71 kb, respectively (Table 1). The assembled genome size was slightly higher than that obtained by K-mer estimation (167.18 Mb; Supplementary Fig. 1) and was similar to *C. nasturtii* (Mori *et al.* 2021) and *M. destructor* (Zhao *et al.* 2015). For the *S. mosellana* genome released by Agriculture and Agri-food Canada, the assembly size was 208 Mb and the scaffold N50 was 5.13 Mb, which was not a chromosome-level assembly (Table 1). The scaffold N50 of this genome was 44.56 Mb, making it a high quality, and potentially the best quality, *S. mosellana* genome available to date. These results showed that the genome reported in the current study had a high level of continuity and completeness.

For quality evaluation of the genome assembly, according to BWA software (http://bio-bwa.sourceforge.net/), a total of 91.7% of the short reads were uniquely mapped to the genome assembly and the coverage rate was 99.8%, indicating that the assembled genome was high quality (Supplementary Table 5). A BUSCO assessment showed that 93.1% of BUSCO genes were successfully detected, of which 90.7% were single copy and 2.0% were duplicated (Table 2). Compared to the Insecta databases, the results showed a high-quality assembly of *S. mosellana* above 90% of conserved genes of the database. The results of these 2 evaluations indicated that the genome assembly had a high level of completeness and was suitable for subsequent analysis.

In addition, to measure genome-wide sequencing bias, the GC content and average depth of the assembled genome were calculated and mapped using 10-kb nonoverlapping slide windows. The density points (red scatter plot) only concentrated within the 30–40% range, with the average GC content of 36.4% (Supplementary Fig. 2).

Genome annotation

Repetitive elements, including TEs, are a major sequence component of eukaryote genomes (Petersen *et al.* 2019). RepeatMasker (http://www.repeatmasker.org/) software and Repbase (http://www.girinst.org/repbase) database annotated the repeat sequences. The results of repeat prediction showed that the *S. mosellana* genome contains 21.55% repeat sequences. Repetitive sequence statistics and classification results are shown in Supplementary Table 6. Short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeats (LTRs), and DNA elements accounted for 0.02%, 0.86%, 13.24%, and 1.67% of the whole genome, respectively, and 6.57% of repeat sequences were annotated as unclassified (Fig. 1 and Supplementary Fig. 4 and Supplementary Table 7). The TEs represented 21.09% of the whole *S. mosellana* genome. Similarly, TEs occupied approximately 16% of the *M. destructor* genome (Ben Amara *et al.* 2021) and 13.9% of the *C. nasturtii* genome (Mori *et al.* 2021). TE content varies greatly among the insects and differs even between species belonging the same order. In the Diptera species, TE content ranges from less than 1% in *Belgica Antarctica* to around 50% in *Aedes aegypti*. Similar proportions were estimated in other Dipteran genomes like the Drosophilidae species whose TE content varies between 3% and 25% (Clark and Pachter 2007).

A total of 224 tRNAs were predicted by tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) (Chan and Lowe 2019). Using Blast, 18 rRNAs were identified. Using infernal software (http://infernal.janelia.org/) (Nawrocki and Eddy 2013), we also identified 21 scaRNA, 80 snRNAs, 1,406 miRNAs, and 59 other ncRNAs (Fig. 1 and Supplementary Table 8).

Gene structure prediction was performed, and 12,269 protein-coding genes were predicted, with a mean of 1,520.74 bp of coding sequence (CDS) and 5.18 exons per gene (Supplementary Table 9 and Supplementary Fig. 5). The transcript lengths of genes, CDSs, exons, and introns of *S. mosellana* are comparable to those of the genomes used for homology-based prediction (Supplementary Table 10 and Supplementary Fig. 5). The genome of *S. mosellana* is larger than the genome reported for the *Belgica antarctica* (99 Mb) (Kelley *et al.* 2014). *S. mosellana* genes tend to have much longer introns than do those of *B. antarctica*. Similarly, the genome of *Aedes aegypti*, *Culex quinquefasciatus*, and *A. gambiae* were larger than the genome of *S. mosellana*. The mosquito genes had longer introns than those of *S. mosellana*. The intron size comparison showed that a reduction in intron length also contributed to the reduced size of this genome.

Of all predicted protein-coding genes, 88.6% (10,869) had BLAST hits in the NCBI nonredundant database. Furthermore, 58.7% (7,201) were assigned GO terms, and 75.4% (9,245) were mapped to at least 1 KEGG pathway (Table 3 and Supplementary Fig. 6).

Evolutionary analysis

Dipteran diversity was traditionally partitioned into 2 principal suborders: the Nematocera and the Brachycera (Wiegmann *et al.* 2011). The gall midges, along with marsh flies, gnats, and other midges, made up the nematoceran infraorder, Bibionomorpha. Protein sequences from the 1,024 single-copy gene families were used for phylogenetic tree reconstruction, and the estimation of divergence time was performed (Fig. 2a) with the MCMC tree program implemented in the PAML. The results showed that *S. mosellana* and 10 other flies were clustered together (Fig. 2a). Our analysis showed that the OWBM *S. mosellana*, Swede midge *C. nasturtii*, and Hessian fly *M. destructor* formed a sister lineage to Cecidomyiidae, while *D. melanogaster*, *D. mojavensis*, and *B. dorsalis* were in another sister lineage. Therefore, the placement of *S. mosellana*, *C. nasturtii*, and *M. destructor* with the Drosophilids (Brachycera) confirmed the Nematocera to be a paraphyletic group, consistent with previous analyses placing the Bibionomorpha as a sister group to the Brachycera (Wiegmann *et al.* 2011; Zhao *et al.* 2015). The genes used for gene family clustering in each species are shown in Supplementary Table 11. In total, 1,024 single-copy gene families are common to all 15 species. The distributions of single-copy orthologs, multiple-copy orthologs, genes unique to *S. mosellana*, and other orthologs in different species are shown in Fig. 2a and Supplementary Table 12. Overall, 321 gene families (564 genes included) were unique to *S. mosellana*.

The family Cecidomyiinae usually was divided into 2 supertribes: the Lasiopteridi and the Cecidomyiidi. The genera Mayetiola was in the former, while Contarinia, and Sitodiplosis are in the latter (Hall *et al.* 2012; Dorchin *et al.* 2019). A total of

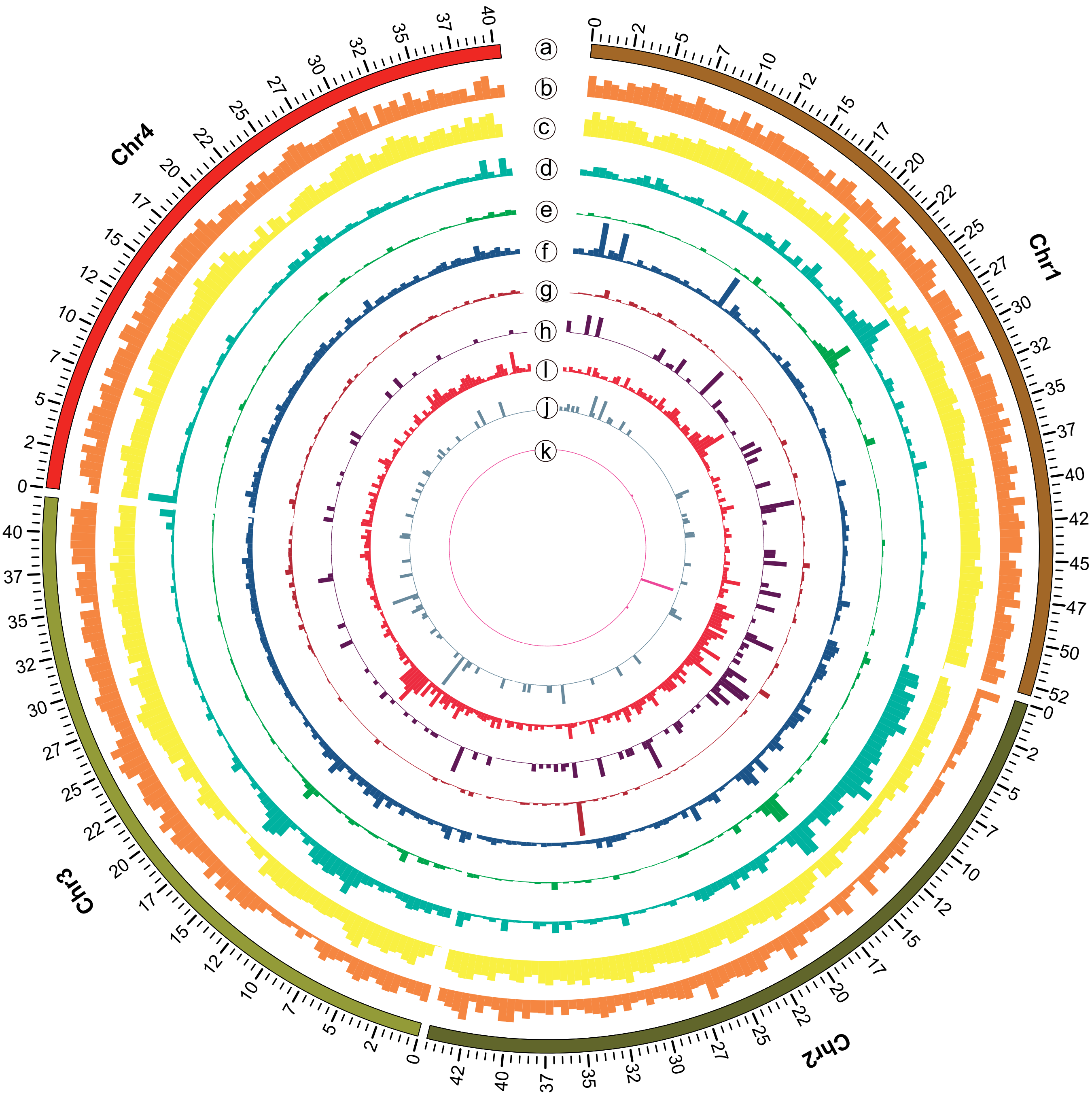

Fig. 1. The genome characteristics of OWBM, *S. mosellana*. Circos plot showing the genomic features. Units on the circumference are megabase values of pseudomolecules. From outermost to innermost circles: Track a: 4 chromosomes of the genome; Track b: gene distribution on each chromosomes; Track c: GC content distribution on each chromosomes; Track d: LTR distribution on each chromosomes; Track e: LINE distribution on each chromosomes; Track f: DNA distribution on each chromosomes; Track g: SINE distribution on each chromosomes; Track h: tRNA located on chromosomes; Track i: miRNA located on chromosomes; Track j: snRNA located on chromosomes; Track k: rRNA located on chromosomes.

6,795 homologous gene families were shared by the 3 species. *S. mosellana* shared 8,191 gene families with *C. nasturtii* and more than 7,101 with *M. destructor* (Fig. 2b), which showed more homology between *S. mosellana* and *C. nasturtii*.

Estimated divergence times of *S. mosellana* and other species (calculated using MCMCTREE) suggest that *S. mosellana* diverged from the common ancestor of *C. nasturtii* 32.7 MYA, and from the ancestor of *M. destructor* 62.7 MYA. Thus, the divergence of *S. mosellana* postdated that of *M. destructor*, a plant parasitic gall midge and a pest of wheat (*Triticum* spp.). The split of the Neodiptera lineage from other Diptera clusters was inferred to be around 211.9 MYA. All 11 Diptera insects diverged from the sister lineage *B. mori* about 233.7 MYA (Fig. 2a). The fly phylogenetic relationships were consistent with previous studies (Wiegmann *et al.* 2011; Vicoso and Bachtrog 2015).

Expansion and contraction of gene families in *S. mosellana*

Of the 22,953 gene families in the most recent common ancestor (MRCA) of all 15 species, 3 were expanded and 33 were contracted

Table 1. Comparison of *S. mosellana* genome assemblies from this and a previous study.

Assembly	ASM2101890v1 (this study)	AAFC_SMos_1.0 (from Agriculture and Agri-food Canada)
Bioproject	PRJNA720212	PRJNA563698
DNA resource	Third-instar larvae	Single pupa
Assembly approach	Falcon	Supernova
Sequencing platform	NovaSeq/PacBio	Illumina HiSeq
Assembly level	Chromosomes	Scaffolds
Number of contigs	381	11,287
Contig N50 (bp)	988,708	62,752
Number of Scaffolds	25	7,269
Scaffold N50 (bp)	44,562,869	5,125,045
Total gap length (bp)	35,600	13,573,270
Total sequence length	180,693,642	208,800,104
Ungapped bases (bp)	180,658,042	195,226,834

Table 2. Statistics of the completeness of the assembled *S. mosellana* genome by BUSCO.

Type	BUSCO groups	Percentage (%)
Complete BUSCOs	907	92.7
Complete and single-copy BUSCOs	887	90.7
Complete duplicated BUSCOs	20	2.0
Fragmented BUSCOs	4	0.4
Missing BUSCOs	67	6.9
Total BUSCO groups searched	978	100

Table 3. Statistics of gene function annotation of *S. mosellana*.

Database	Number	Percentage (%)
Total	12,269	–
Swissprot	9,292	75.70
Nr	10,869	88.60
KEGG	9,245	75.40
InterPro	10,235	83.40
GO	7,201	58.70
Pfam	9,121	74.30
Annotated	11,169	91.00
Unannotated	1,100	9.00

in *S. mosellana* compared with gene families of the common ancestor of *S. mosellana* and *C. nasturtii* (Fig. 3). In contrast, *C. nasturtii* had 91 expanded and 7 contracted gene families. The common ancestor of Cecidomyiidae species showed 27 expanded and 8 contracted gene families compared to that of the common ancestor of Drosophilidae species and Tephritidae.

GO enrichment analysis reveal that the *S. mosellana*-contracted gene families are enriched in carbohydrate metabolic process (GO:0005975, 4 genes, $P = 0.002629$, Adjusted *P*-value), oxidation–reduction process (GO:0055114, 5 genes, $P = 0.005992$), sensory perception of smell (GO:0007608, 2 genes, $P = 0.006189$) (Supplementary Table 13).

As *S. mosellana* adults did not feed and larvae had no capability for host selection, a reduced role for sensory perception was consistent with the general loss of chemoreceptors, the same way as in *M. destructor* (Zhao *et al.*, 2015).

Chromosome synteny

Synteny referred to genes that reside on the same chromosome. Conserved synteny indicated that homologous genes were syntenic between species, regardless of gene order (Ehrlich *et al.* 1997). Syntenic relationships between *S. mosellana*, *D. melanogaster*, and *A. gambiae* showed a high level of collinearity among the 3 chromosome-level genomes, and a relatively low frequency of fragment rearrangements was observed (Fig. 2c). We defined a syntenic block as including at least 3 orthologous genes. In total, 48 syntenic blocks were found between *S. mosellana* and *D. melanogaster*, and the gene number in these blocks ranged from 4 to 12, with a mean of 5.31. Eighty-one blocks were found between *S. mosellana* and *A. gambiae*, with the same gene number range of 4–15 and a mean of 6.20. The most conserved pairs of chromosomal arms were SmChr2/Dm2L and SmChr2/Ag3R, with 75% and 80% of synteny blocks in SmChr2 mapping to Dm2L and Ag3R, respectively. The remaining blocks represented exchanges with other arms. Other relationships were 70% and 46% of synteny blocks in SmChr1 mapping to Dm3R and Ag 2R, respectively. In our analysis, *S. mosellana* showed slightly higher synteny with *A. gambiae* than *D. melanogaster*, despite the closer phylogenetic relationship of *S. mosellana* and *D. melanogaster*.

Gene families were commonly found in genomes and were thought to evolve by gene duplication and neofunctionalization. The Osiris gene family was a large conserved family first described in *D. melanogaster* (Dorer *et al.* 2003). Twenty-three Osiris genes were originally found in the *D. melanogaster* genome, with 20 of them located on chromosome 3R in a cluster. The Osiris gene family was also present in the mosquito *A. gambia* genome (Dorer *et al.* 2003) and *S. mosellana* genome. The families maintained a remarkable degree of synteny displays remarkable synteny and sequence conservation with the *Drosophila* cluster (Shah *et al.* 2012).

Evolution of detoxification gene families in *S. mosellana*

Herbivorous insects have developed detoxification enzymes to metabolize otherwise deleterious plant secondary metabolites (Ramsey 2010; Simon *et al.* 2015). As a strict specialist, *S. mosellana* likely had adaptations that allowed it to detoxify these chemicals (Smith *et al.* 2004, 2007). The vast array of GST and CYP450 genes in insects represents the largest repertoire of detoxification enzyme genes known (Hazzouri *et al.* 2020). There were 4 large clades of insect P450 genes: the CYP2 clade, the CYP3 clade, the CYP4 clade, and the mitochondrial clade (Feyereisen 2006). With homology searching, 95 P450 genes were annotated and grouped into the 4 major clades (Supplementary Fig. 7). CYP3 ranked as the largest clade, consisting of 51 members, and strong gene expansion was observed (36 for *D. melanogaster* and 38 for *M. destructor*). The CYP4 clade included 23 P450 members. The remainder belonged to the mitochondrial (10 ones) and CYP2 (11 ones) clades. CYP6 and CYP313 were the most expanded gene families, with a more specific expansion of subfamilies CYP6D (25 genes) (Supplementary Fig. 7). Other examples of several such blooms in a diversity of species are the 17 CYP6AS genes in honeybee (Claudianos *et al.* 2006), 12 CYP6A genes in the fruit fly (Tijet *et al.* 2001), and 13 CYP6BQ genes in *T. castaneum* (Zhu *et al.* 2013). Although few of the CYP6 enzymes have been characterized and in many (but not all) studies, they are shown to metabolize xenobiotics and plant natural compounds (Li *et al.* 2002; Feyereisen 2012; Edi *et al.* 2014). Greater expression levels of P450 genes were found in *M. destructor* and *Aphis glycine* feeding on resistant plants (Bansal *et al.* 2014; Chen *et al.* 2016). In fact, the induction of xenobiotic response genes by plant secondary metabolite exposure was thought to be the first step leading to eventual detoxification and virulence adaptation (Bansal *et al.* 2014).

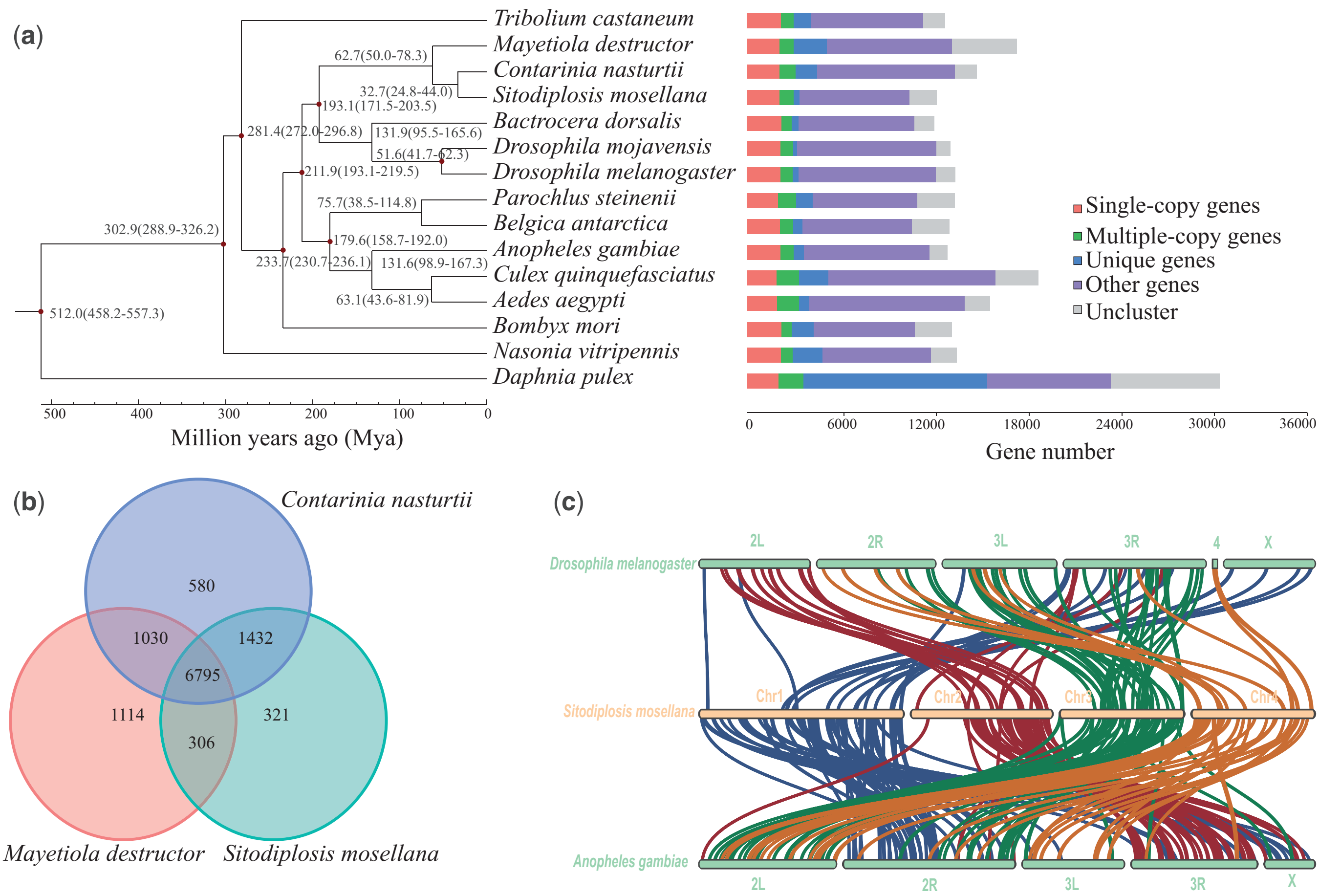

Fig. 2. Phylogenetic tree, gene orthology, and synteny blocks. a) The phylogenetic tree was constructed based on 1,024 single-copy gene families with 14 insects and 1 noninsect species, using RAxML maximum-likelihood methods. Bootstrap values are 100 in all nodes based on 100 replicates. The numbers near each node correspond to the estimated divergence time of these species. The colored bars to the right are subdivided to represent different types of orthology. "Single-copy genes" indicates single copy orhologous genes in common gene families; "Multiple-copy genes" indicates mutiple copy orthologous genes in common gene families; "Unique genes" indicates genes from unique gene family from each species; "Other genes" indicates genes that do not belong to any above-mentioned ortholog categories; "Uncluster" indicates genes that do not cluster to any families. b) Venn diagram of the orthologous gene families from 3 gall midges: *S. mosellana*, *C. nasturtii*, and *M. destructor*. c) Synteny blocks between *S. mosellana*, *D. melanogaster*, and *A. gambiae*.

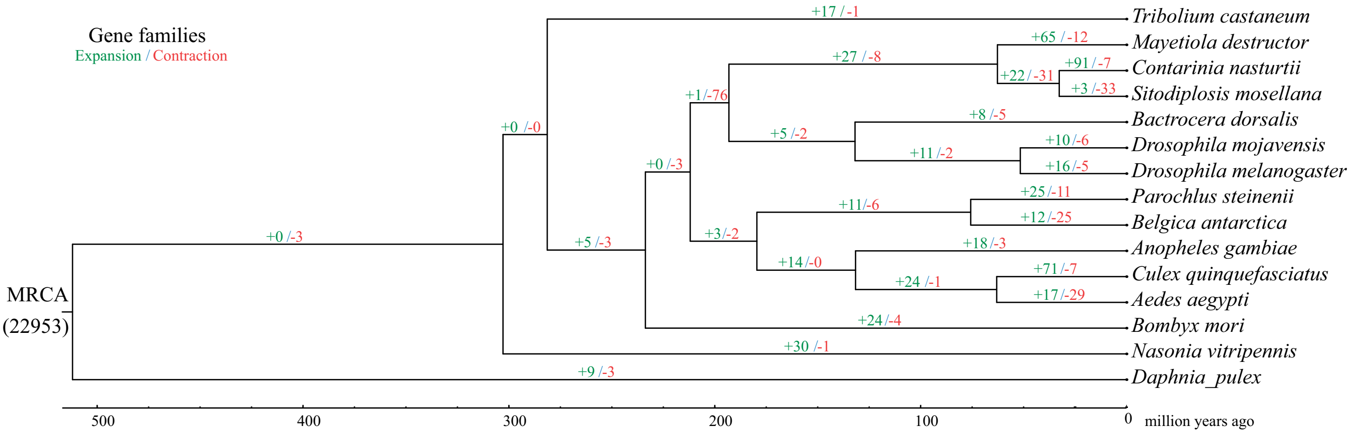

Fig. 3. Gene family evolution between genomes of *S. mosellana* and 14 other arthropod species. Left number indicates gene family expansions and right number indicates gene family contractions. The length of branch indicate the divergence time. MRCA: most recent common ancestor.

Another group of detoxification enzymes was GSTs. GSTs are involved in many cellular physiological activities, such as detoxification of endogenous and xenobiotic compounds, intracellular transport, biosynthesis of hormones, and protection against oxidative stress (Enayati *et al.* 2005; Shi *et al.* 2012). In *S. mosellana*, 26 GST genes were identified. GST genes were grouped into 5 GST classes, with 5 Delta GSTs, 14 Epsilon GSTs, 1 Omega GST, 5 Theta GSTs, and 1 Sigma GST. Genes belonging to the Zeta class were not found. Contractions of GST genes were derived from Delta classes (Supplementary Fig. 8). The similar numbers of GST genes in *S. mosellana* and *M. destructor* (22 genes) were in contrast with the significantly higher number of GST genes in *D. melanogaster* (36 genes). This might correspond to midges' narrow host range. *S. mosellana* and *M. destructor* had similar diets (consisting mainly of wheat) and habitats (primarily wheat dominated). The great number of orthologous GST gene groups in *S. mosellana* and *M. destructor* species suggested that the radiation event or independent expansion of the GST gene family in these species may have occurred relatively recently, and this was consistent with previous studies in 3 planthoppers (Zhou *et al.* 2013). As in *Drosophila* and *Ceratitis capitata*, many of the insect-specific genes of the Delta and Epsilon subclasses are putatively involved in insect responses to environmental conditions, as well as in xenobiotic and insecticide resistance (Enayati *et al.* 2005; Li *et al.* 2007; Papanicolaou *et al.* 2016).

Conclusion

In summary, we successfully assembled a genome for the wheat pest *S. mosellana*, providing the first chromosome-level genome for a species from the family Cecidomyiidae of Diptera insects using Illumina and PacBio sequencing platforms with Hi-C technology. The availability of the genome sequence will facilitate the future evaluation of unique biological characteristics of *S. mosellana*, such as olfactory reception, prolonged diapauses, and insect–host interactions.

Data availability

The raw reads and assembled genome produced in this study were deposited in the National Center for Biotechnology Information (NCBI) with the BioProject accession number PRJNA720212. The whole genome sequence reported in this article was deposited in the Genome Warehouse in the Beijing Institute of Genomics Data Center (https://bigd.big.ac.cn/) under accession number GWHBEIQ00000000.

Supplemental material is available at *G3* online.

Acknowledgments

We thank Prof. Zengrong Zhu and Wenwu Zhou, Zhejiang University, for helpful comments. Genome sequencing and assembly were conducted by Novogene Bioinformatics Institute, Beijing, China. The authors thank Qun He, Depo Liu, Suchan Ji, Ruijie Lu, and Dongmei Kang for their help with sample collection.

Funding

This work was financially supported by the China Agriculture Research System of MOF and MARA (CARS-03), National Key Research and Development Program of China (2017YFD0301104).

Conflicts of interest

None declared.

Literature cited

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.* The genome sequence of *Drosophila melanogaster*. Science. 2000;287(5461): 2185–2195.

Al-jbory Z, Anderson KM, Harris MO, Mittapalli O, Whitworth RJ, Chen MS. Transcriptomic analyses of secreted proteins from the salivary glands of wheat midge larvae. J Insect Sci. 2018;18(1):17.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–29.

Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000; 28(1):45–48.

Bansal R, Mian MAR, Mittapalli O, Michel AP. RNA-Seq reveals a xenobiotic stress response in the soybean aphid, *Aphis glycines*, when fed aphid-resistant soybean. BMC Genomics. 2014;15(1): 972.

Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6(1):11.

Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: an optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods. 2017;123:56–65.

Ben Amara W, Quesneville H, Khemakhem MM. A genomic survey of Mayetiola destructor mobilome provides new insights into the evolutionary history of transposable elements in the cecidomyiid midges. PLoS One. 2021;16(10):e0257996.

Berzonsky WA, Ding H, Haley SD, Harris MO, Lamb RJ, McKenzie RIH, Ohm H, Patterson FL, Pearis FB, Porter DR, *et al.* Breeding wheat for resistance to insects. In: Janick J, editor. Plant Breeding Reviews. New York (NY): John Wiley & Sons, Inc. 2002;22: 221–296.

Birney E, Clamp M, Durbin R. GeneWise and genomewise. Genome Res. 2004;14(5):988–995.

Bruce TJA, Hooper AM, Ireland L, Jones OT, Martin JL, Smart LE, Oakley J, Wadhams LJ. Development of a pheromone trap monitoring system for orange wheat blossom midge, *Sitodiplosis mosellana*, in the UK. Pest Manag Sci. 2007;63(1):49–56.

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.

Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA Genes in Genomic Sequences. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York (NY): Springer New York; 2019. p. 1–14.

Chavalle S, Buhl PN, Censier F, De Proft M. Comparative emergence phenology of the orange wheat blossom midge, *Sitodiplosis mosellana* (Géhin) (Diptera: Cecidomyiidae) and its parasitoids (Hymenoptera: Pteromalidae and Platygastridae) under controlled conditions. Crop Protect. 2015;76:114–120.

Chen MS, Liu S, Wang H, Cheng X, El Bouhssini M, Whitworth RJ. Genes expressed differentially in Hessian fly larvae feeding in resistant and susceptible plants. Int J Mol Sci. 2016;17(8):1324.

Cheng WN, Zhang YD, Yu JL, Liu W, Zhu-Salzman K. Functional analysis of odorant-binding proteins 12 and 17 from wheat blossom midge *Sitodiplosis mosellana* Géhin (Diptera: Cecidomyiidae). Insects. 2020;11(12):891.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–569.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–1054.

Clark AG, Pachter L. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450(7167):203–218.

Claudianos C, Ranson H, Johnson RM, Biswas S, Schuler MA, Berenbaum MR, Feyereisen R, Oakeshott JG. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. Insect Mol Biol. 2006;15(5):615–636.

De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006; 22(10):1269–1271.

Dorchin N, Harris KM, Stireman JO. Phylogeny of the gall midges (Diptera, Cecidomyiidae, Cecidomyiinae): systematics, evolution of feeding modes and diversification rates. Mol Phylogenet Evol. 2019;140:106602.

Dorer DR, Rudnick JA, Moriyama EN, Christensen AC. A family of genes clustered at the Triplo-lethal locus of *Drosophila melanogaster* has an unusual evolutionary history and significant synteny with *Anopheles gambiae*. Genetics. 2003;165(2):613–621.

Duan Y, Wu YQ, Luo LZ, Miao J, Gong ZJ, Jiang YL, Li T. Genetic diversity and population structure of *Sitodiplosis mosellana* in Northern China. PLoS One. 2013;8(11):e78415.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–95.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–1797.

Edi CV, Djogbénou L, Jenkins AM, Regna K, Muskavitch MAT, Poupardin R, Jones CM, Essandoh J, Kétoh GK, Paine MJI, *et al.* CYP6 P450 enzymes and ACE-1 duplication produce extreme and multiple insecticide resistance in the malaria mosquito *Anopheles gambiae*. PLoS Genet. 2014;10(3):e1004236.

Ehrlich J, Sankoff D, Nadeau JH. Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics. 1997;147(1):289–296.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, *et al.* The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1): D427–D432.

Elliott RH, Mann LWJCE. Susceptibility of red spring wheat, *Triticum aestivum* L. cv. Katepwa, during heading and anthesis to damage by wheat midge, *Sitodiplosis mosellana* (Géhin) (Diptera: Cecidomyiidae). Can Entomol. 1996;128(3):367–375.

Enayati AA, Ranson H, Hemingway J. Insect glutathione transferases and insecticide resistance. Insect Mol Biol. 2005;14(1):3–8.

Feyereisen R. Evolution of insect P450. Biochem Soc Trans. 2006; 34(Pt 6):1252–1255.

Feyereisen R. 8—insect CYP genes and P450 enzymes. In: Gilbert LI, editor. Insect Molecular Biology and Biochemistry. San Diego (CA): Academic Press; 2012. p. 236–316.

Gaafar N, Volkmar C, Cöster H, Spilke J. Susceptibility of winter wheat cultivars to wheat ear insects in Central Germany. Gesunde Pflanzen. 2011;62(3–4):107–115.

Gagné RJ, Jaschhof M. 2017. A Catalog of the Cecidomyiidae (Diptera) of the World. 4th ed. Washington, DC: United States Department of Agriculture. Digital. 762p.

Gagné RJ, Jaschhof M. 2021. A Catalog of the Cecidomyiidae (Diptera) of the World. 5th ed. Washington, DC: United States Department of Agriculture. Digital. 816.p

Gong ZJ, Wu YQ, Miao J, Duan Y, Jiang YL, Li T. Global transcriptome analysis of orange wheat blossom midge, *Sitodiplosis mosellana* (Gehin) (Diptera: Cecidomyiidae) to identify candidate transcripts regulating diapause. PLoS One. 2013;8(8):e71564.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–652.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–321.

Guo S-K, Cao L-J, Song W, Shi P, Gao Y-F, Gong Y-J, Chen J-C, Hoffmann AA, Wei S-J. Chromosome-level assembly of the melon thrips genome yields insights into evolution of a sap-sucking lifestyle and pesticide resistance. Mol Ecol Resour. 2020;20(4): 1110–1125.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):R7–22.

Hao Z, Geng M, Hao Y, Zhang Y, Zhang L, Wen S, Wang R, Liu G. Screening for differential expression of genes for resistance to *Sitodiplosis mosellana* in bread wheat via BSR-seq analysis. Theor Appl Genet. 2019;132(11):3201–3221.

Hall DR, Amarawardana L, Cross JV, Francke W, Boddum T, Hillbur Y. The chemical ecology of Cecidomyiid midges (Diptera: Cecidomyiidae). J Chem Ecol. 2012;38(1):2–22.

Hazzouri KM, Sudalaimuthuasari N, Kundu B, Nelson D, Al-Deeb MA, Le Mansour A, Spencer JJ, Desplan C, Amiri KMA. The genome of pest *Rhynchophorus ferrugineus* reveals gene families important at the plant-beetle interface. Commun Biol. 2020;3(1): 323.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JC, Wides R, *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. Science. 2002;298(5591):129–149.

Jacquemin G, Chavalle S, De Proft M. Forecasting the emergence of the adult orange wheat blossom midge, *Sitodiplosis mosellana* (Géhin) (Diptera: Cecidomyiidae) in Belgium. Crop Protect. 2014; 58:6–13.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–589.

Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.

Kelley JL, Peyton JT, Fiston-Lavier A-S, Teets NM, Yee M-C, Johnston JS, Bustamante CD, Lee RE, Denlinger DL. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nat Commun. 2014;5(4611):4611.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4): 357–360.

Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016; 33(7):1870–1874.

Lamb RJ, Tucker JR, Wise IL, Smith MAH. Trophic Interaction between *Sitodiplosis Mosellana* (Diptera: Cecidomyiidae) and spring wheat: implications for yield and seed quality. Can Entomol. 2000;132(5):607–625.

Leskovec J, Sosič R. SNAP: a general-purpose network analysis and graph-mining library. ACM Trans Intell Syst Technol. 2016;8(1):1–20.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–1760.

Li L, Stoeckert CJ, Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–2189.

Li X, Berenbaum MR, Schuler MA. Plant allelochemicals differentially regulate *Helicoverpa zea* cytochrome P450 genes. Insect Mol Biol. 2002;11(4):343–351.

Li X, Schuler MA, Berenbaum MR. Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu Rev Entomol. 2007;52:231–253.

Li YP, Wu JX, Cheng WN, Song WW, Yuan XQ. Comparison of silk glands of diapause and non-diapause larval *Sitodiplosis mosellana*. J Insect Sci. 2012;12(81):1–9.

Lian Y, Wei H, Wang J, Lei C, Li H, Li J, Wu Y, Wang S, Zhang H, Wang T, *et al*. Chromosome-level reference genome of X12, a highly virulent race of the soybean cyst nematode *Heterodera glycines*. Mol Ecol Resour. 2019;19(6):1637–1646.

Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–2879.

Miao J, Wu YQ, Gong ZJ, He YZ, Duan Y, Jiang YL. Long-distance wind-borne dispersal of *Sitodiplosis mosellana* Géhin (Diptera: Cecidomyiidae) in northern China. J Insect Behav. 2013;26(1):120–129.

Mittapalli O, Neal JJ, Shukle RH. Differential expression of two cytochrome P450 genes in compatible and incompatible Hessian fly/wheat interactions. Insect Biochem Mol Biol. 2005;35(9):981–989.

Mori BA, Coutu C, Chen YH, Campbell EO, Dupuis JR, Erlandson MA, Hegedus DD. De novo whole-genome assembly of the swede midge (*Contarinia nasturtii*), a specialist of Brassicaceae, using linked-read sequencing. Genome Biol Evol. 2021;13(3):evab036.

Mulder N, Apweiler R. InterPro and InterProScan. In: Comparative Genomics. In: Bergman HH, editor. Totowa, NJ: Humana Press; 2007. p. 59–70.

Myers G. Efficient local alignment discovery amongst noisy long reads. In: Brown D, Morgenstern B, editors. Algorithms in Bioinformatics. WABI 2014. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2014;8701:52–67.

Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–2935.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–274.

Olfert O, Weiss RM, Vankosky M, Hartley S, Doane JF. Modelling the tri-trophic population dynamics of a host crop (*Triticum aestivum*; Poaceae), a major pest insect (*Sitodiplosis mosellana*; Diptera: Cecidomyiidae), and a parasitoid of the pest species (*Macroglenes penetrans*; Hymenoptera: Pteromalidae): a cohort-based approach incorporating the effects of weather. Can Entomol. 2020;152(3):311–329.

Papanicolaou A, Schetelig MF, Arensburger P, Atkinson PW, Benoit JB, Bourtzis K, Castañera P, Cavanaugh JP, Chao H, Childers C, *et al*. The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. Genome Biol. 2016;17(1):192.

Parra G, Blanco E, Guigo R. GeneID in *Drosophila*. Genome Res. 2000;10(4):511–515.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–295.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol Biol. 2019;19(1):1–15.

Ramsey JS, Rider DS, Walsh TK, De Vos M, Gordon KHJ, Ponnala L, Macmil SL, Roe BA, Jander G. Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*. Insect Mol Biol. 2010;19(S2):155–164.

Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):1–10.

Shah N, Dorer DR, Moriyama EN, Christensen AC. Evolution of a large, conserved, and syntenic gene family in insects. G3 (Bethesda). 2012;2(2):313–319.

Shi H, Pei L, Gu S, Zhu S, Wang Y, Zhang Y, Li B. Glutathione S-transferase (GST) genes in the red flour beetle, *Tribolium castaneum*, and comparative analysis with five additional insects. Genomics. 2012;100(5):327–335.

Simon J-C, d'Alençon E, Guy E, Jacquin-Joly E, Jaquiéry J, Nouhaud P, Peccoud J, Sugio A, Streiff R. Genomics of adaptation to host-plants in herbivorous insects. Brief Funct Genomics. 2015;14(6):413–423.

Smith MAH, Lamb RJ, Wise IL, Olfert OO. An interspersed refuge for *Sitodiplosis mosellana* (Diptera: Cecidomyiidae) and a biocontrol agent *Macroglenes penetrans* (Hymenoptera: Pteromalidae) to manage crop resistance in wheat. Bull Entomol Res. 2004;94(2):179–188.

Smith MAH, Wise IL, Lamb RJ. Survival of *Sitodiplosis mosellana* (Diptera: Cecidomyiidae) on wheat (Poaceae) with antibiosis resistance: implication for the evolution of virulence. Can Entomol. 2007;139(1):133–140.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–1313.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web Server issue):W435–W439.

Tang HB, Bowers JE, Wang XY, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. Science. 2008;320(5875):486–488.

Thomas J, Fineberg N, Penner G, McCartney C, Aung T, Wise I, McCallum B. Chromosome location and markers of Sm1: a gene of wheat that conditions antibiotic resistance to orange wheat blossom midge. Mol Breeding. 2005;15(2):183–192.

Thompson BM, Reddy GVPJCP. Status of *Sitodiplosis mosellana* (Diptera: Cecidomyiidae) and its parasitoid, *Macroglenes penetrans* (Hymenoptera: Pteromalidae), in Montana. Crop Protect. 2016;84:125–131.

Tijet N, Helvig C, Feyereisen R. The cytochrome P450 gene superfamily in *Drosophila melanogaster*: annotation, intron-exon organization and phylogeny. Gene. 2001;262(1–2):189–198.

Vicoso B, Bachtrog D. Numerous transitions of sex chromosomes in Diptera. PLoS Biol. 2015;13(4):e1002078.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35(3):543–548.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, *et al.* Episodic radiations in the fly tree of life. Proc Natl Acad Sci USA. 2011;108(14):5690–5695.

Wu YQ, Duan Y, Zhang ZQ, Liu CY, Liu ST, Miao J, Gong Z, Duan Y, Jiang Y, Li T. The synchronization of ear emerging stages of winter wheat with occurrent periods of the orange wheat blossom midge, *Sitodiplosismosellana* (Gehin) (Diptera: Cecidomyiidae) adults and its damaged level. Acta Ecol Sin. 2015;35(11):3548–3554.

Yang ZH. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–1591.

Yang ZH, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012;13(5):303–314.

Ye X, Yan Z, Yang Y, Xiao S, Chen L, Wang J, Wang F, Xiong S, Mei Y, Wang F, *et al.* A chromosome-level genome assembly of the parasitoid wasp *Pteromalus puparum*. Mol Ecol Resour. 2020;20(5):1384–1402.

Yu GC, Smith DK, Zhu HC, Guan Y, Lam TT-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.

Yukawa J, Rohfritsch O. Biology and ecology of gall-inducing Cecidomyiidae (Diptera). In: Raman A, Schaefer CW, Withers, TM, editors. Biology, Ecology and Evolution of Gall-Inducing Arthropods. Enfield(NH), USA, Plymouth, UK: Science Publishers, Inc.; 2005. p. 273–304.

Zhang L, Geng M, Zhang Z, Zhang Y, Yan G, Wen S, Liu G, Wang R. Molecular mapping of major QTL conferring resistance to orange wheat blossom midge (*Sitodiplosis mosellana*) in Chinese wheat varieties with selective populations. Theor Appl Genet. 2020;133(2):491–502.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019;5(8):833–845.

Zhao C, Escalante LN, Chen H, Benatti TR, Qu J, Chellapilla S, Waterhouse RM, Wheeler D, Andersson MN, Bao R, *et al.* A massive expansion of effector genes underlies gall-formation in the wheat pest *Mayetiola destructor*. Curr Biol. 2015;25(5):613–620.

Zhou W-W, Liang Q-M, Xu Y, Gurr GM, Bao Y-Y, Zhou X-P, Zhang C-X, Cheng J, Zhu Z-R. Genomic insights into the glutathione S-transferase gene family of two rice planthoppers, *Nilaparvata lugens* (Stål) and *Sogatella furcifera* (Horváth) (Hemiptera: Delphacidae). PLoS One. 2013;8(2):e56604.

Zhu F, Moural TW, Shah K, Palli SR. Integrated analysis of cytochrome P450 gene superfamily in the red flour beetle, *Tribolium castaneum*. BMC Genomics. 2013;14:174–174.

Communicating editor: A. Sethuraman