G3, 2023, **13(4)**, jkad046
https://doi.org/10.1093/g3journal/jkad046
Advance Access Publication Date: 2 March 2023
Genome Report

The genome of the soybean gall midge (*Resseliella maxima*)

Gloria Melotto,[1] Megan W. Jones,[1] Kathryn Bosley,[2] Nicole Flack,[3] Lexi E. Frank,[3] Emily Jacobson,[1] Evan J. Kipp,[3] Sally Nelson,[1] Mauricio Ramirez,[1] Carrie Walls,[2] Robert L. Koch ,[1] Amelia R.I. Lindsey ,[1] Christopher Faulk [2,*]

[1]Department of Entomology, College of Food, Agricultural and Natural Resource Sciences, University of Minnesota, Minneapolis, MN 55455, USA
[2]Department of Animal Science, College of Food, Agricultural and Natural Resource Sciences, University of Minnesota, Minneapolis, MN 55455, USA
[3]Department of Veterinary and Biomedical Sciences, College of Veterinary Medicine, University of Minnesota, Minneapolis, MN 55455, USA

*Corresponding author: Email: cfaulk@umn.edu

Abstract

The cecidomyiid fly, soybean gall midge, *Resseliella maxima* Gagné, is a recently discovered insect that feeds on soybean plants in the Midwestern United States. *R. maxima* larvae feed on soybean stems that may induce plant death and can cause considerable yield losses, making it an important agricultural pest. From three pools of 50 adults each, we used long-read nanopore sequencing to assemble a *R. maxima* reference genome. The final genome assembly is 206 Mb with 64.88× coverage, consisting of 1,009 contigs with an N50 size of 714 kb. The assembly is high quality with a Benchmarking Universal Single-Copy Ortholog (BUSCO) score of 87.8%. Genome-wide GC level is 31.60%, and DNA methylation was measured at 1.07%. The *R. maxima* genome is comprised of 21.73% repetitive DNA, which is in line with other cecidomyiids. Protein prediction annotated 14,798 coding genes with 89.9% protein BUSCO score. Mitogenome analysis indicated that *R. maxima* assembly is a single circular contig of 15,301 bp and shares highest identity to the mitogenome of the Asian rice gall midge, *Orseolia oryzae* Wood-Mason. The *R. maxima* genome has one of the highest completeness levels for a cecidomyiid and will provide a resource for research focused on the biology, genetics, and evolution of cecidomyiids, as well as plant–insect interactions in this important agricultural pest.

Keywords: soybean, gall midge, nanopore, genome assembly, DNA methylation

Introduction

The soybean gall midge, *Resseliella maxima* Gagné (Diptera: Cecidomyiidae), is a recently discovered insect pest of soybean plants (Fig. 1a) (Gagné *et al.* 2019). This insect was first described in 2019 after being associated in the prior year with dying soybean plants in the Midwestern United States (Gagné *et al.* 2019). Soybean plants become susceptible to *R. maxima* infestation during early vegetative growth stages, when natural fissures develop below the cotyledonary node (McMechan *et al.* 2021). These fissures are where the *R. maxima* females are suspected to lay their eggs (McMechan *et al.* 2021). After hatching, *R. maxima* larvae start feeding within the stem at the base of the plant (Fig. 1b). This feeding results in necrotic lesions at the base of the plant (Fig. 1c), which often results in wilting, lodging, or plant death (McMechan *et al.* 2021).

Since initial reports of *R. maxima* in 2018, the midge's presence has been confirmed in five Midwest states: Iowa, Minnesota, Nebraska, Missouri, and South Dakota (McMechan *et al.* 2021). These states are ranked as the 2nd, 3rd, 4th, 6th, and 8th most productive states, respectively, for soybean in the United States (USDA 2022). Soybean is a source of food and fuel, and it is an important commodity crop worldwide. In the United States, soybean production accounted for 20% ($48.6 billion) of US crop cash receipts in the calendar year 2021 (USDA 2022). With *R. maxima* capable of causing yield losses (McMechan *et al.* 2021; Helton *et al.* 2022), there is a growing concern over the spread and impacts of this new insect pest.

Here, we provide the first genome sequence for the genus *Resseliella*. *R. maxima* poses a threat to the US soybean industry, and its genome sequence will assist with (1) evaluating biological characteristics (e.g. overwintering ability or interactions with host plants), (2) understanding mechanisms of pesticide resistance, (3) describing cecidomyiid evolution across natural histories and host ranges, and (4) generating tools for accurate identification.

Methods

Sample collection

R. maxima adults were reared from field-collected soybean stems symptomatic of infestation with *R. maxima*. The collection of the stems occurred in summer of 2022 at one farm in Rock County, Minnesota, United States. Symptomatic soybean plants were collected by pulling the entire plants from the soil. The plants were then trimmed above the first pair of unifoliate leaves and the roots to a length of 5 cm. Each stem was wrapped with a small piece of PARAFILM® at the cut end to decelerate plant dehydration. Ten trimmed stems were set vertically into a 3-cm deep layer of potting soil (BM2 Seed Germination and Propagation Mix, Berger, Saint-Modeste, Quebec, Canada) in one emergence cage. Emergence cages consisted of plastic 5-L clear paint mixing buckets with lids (TCP Global Corporation, Lakeside, CA, USA), with a 6-cm diameter hole that had a fine-mesh (Quest Outfitters, Sarasota, Florida, USA) sleeve 30-cm-long attached to it to facilitate access to the contents of the cages. The emergence cages

Received: January 31, 2023. **Accepted:** February 16, 2023

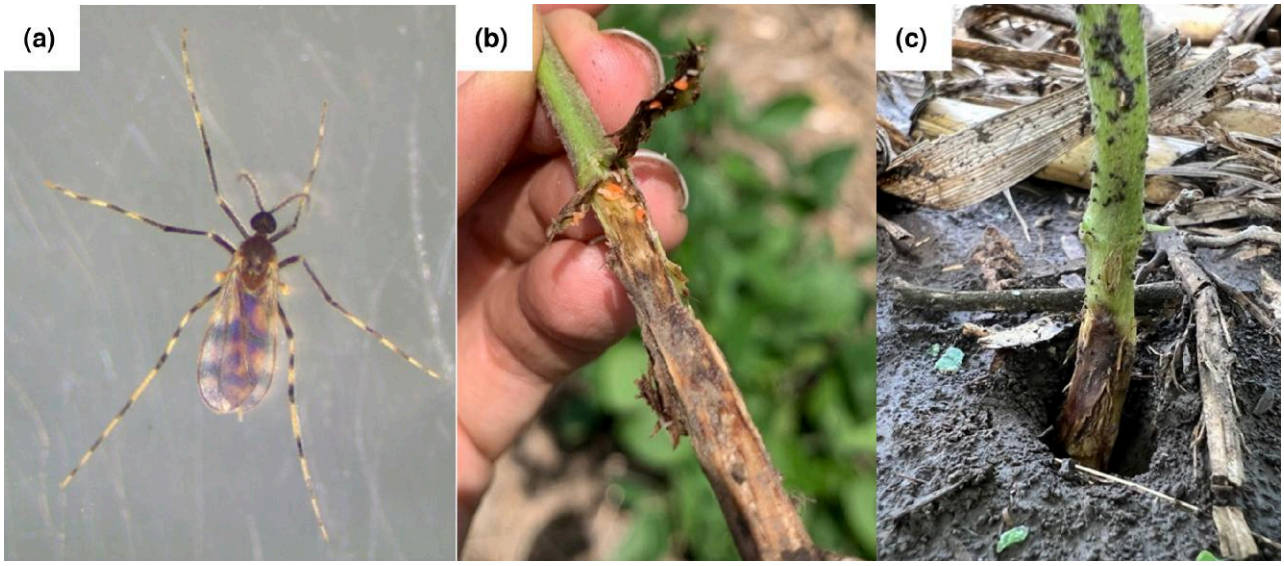

Fig. 1. Soybean gall midge biology. a) Adult female of *R. maxima*. b) *R. maxima* larvae on soybean plant lesion. c) Soybean plant showing symptoms of *R. maxima* infestation in the field.

were maintained at room temperature in 16:8 (light:dark) and watered as needed. Adult insects were collected manually into microcentrifuge tubes, freeze-killed, and morphologically confirmed to be *R. maxima* according to Gagné *et al.* (2019).

DNA extraction and sequencing

DNA was extracted using a Zymo Quick-DNA Miniprep Plus Kit (catalog number D4068, Zymo Research, Irvine, CA, USA), according to manufacturer's instructions. Due to sample timing availability and flow cell upgrade paths, we performed sequencing with two different flow cell types over three flow cells. For each of three pools, ~50 individuals were used for extraction, generating 1 µg of DNA that was loaded into the library prep kit. Libraries were prepared using the SQK-LSK110 and SQK-LSK114 ligation sequencing kits for flow cells R9.4.1 and R10.4.1, respectively. Sequencing was carried out for 24 h per flow cell. Bases were called using the guppy base caller v6.3.9 with model "dna_r10.4.1_e8.2_400bps_modbases_5mc_cg_sup.cfg."

Genome assembly and polishing

De novo assembly of the *R. maxima* nuclear genome was accomplished using Flye v2.9 (https://github.com/fenderglass/Flye) with a subsequent polishing step done using Medaka v1.6.0 (https://github.com/nanoporetech/medaka). We used Benchmarking Universal Single-Copy Ortholog (BUSCO) (v5.4.3) to assess genome completeness for the draft assembly both before and after Medaka polishing steps (Manni *et al.* 2021). Specifically, the Diptera OrthoDB v10 database, which consists of 3,285 single-copy orthologs, was chosen for scoring our assemblies. Based on these assessments, we then selected the polished assembly with the highest BUSCO score for decontamination and downstream analyses. Full commands for all steps in the bioinformatic pipeline are given in Supplementary File 7.

Decontamination

The BlobToolKit (https://blobtoolkit.genomehubs.org/blobtools2/) was used to examine contig properties by comparing GC content, contig length, coverage, and BLAST matches to the NCBI nonredundant (nr) database (Challis *et al.* 2020). When deciding cutoff values, the presence of BUSCO genes within a contig was used to determine thresholds. For instance, all contigs below 8,000 bp were removed as none below that size contained any BUSCOs. We removed contigs below 20× and above 200× coverage.

Methylation

5′ DNA methylation at cytosines in a CpG context was assessed during initial base calling by using a DNA modification aware model. Output files were converted to bed format using modbam2bed v0.6.2 (https://github.com/epi2me-labs/modbam2bed). Aggregation of DNA methylation was performed with "awk" on the command line.

Repeats

RepeatMasker v4.1.4 (https://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) with the full Dfam library v3.6 (https://www.dfam.org/home) was initially used for all cecidomyiid repeat assessments (Flynn *et al.* 2020; Storer *et al.* 2021). For *ab initio* repeat detection, RepeatModeler2 v2.0.2a (https://www.repeatmasker.org/RepeatModeler/) was used on each genome independently (Flynn *et al.* 2020). To determine shared repeat content, the *R. maxima*-specific repeat library generated from RepeatModeler2 was used as input for RepeatMasker to mask other cecidomyiid genomes.

Protein annotation

Gene Model Mapper (GeMoMa) v1.9 (http://www.jstacs.de/index.php/GeMoMa) was used for homology-based protein model prediction with both the *Drosophila melanogaster* Meigan (Diptera: Drosophilidae) (GCA_000001215.4) and *Contarinia nasturtii* Kieffer (Diptera: Cecidomyiidae) (GCF_009176525.2) transcriptomes as references (Keilwagen *et al.* 2019). BUSCO was run against the resulting GeMoMa annotation in protein mode with the Diptera odb10 database to assess quality.

Mitochondria assembly

Total reads were first blasted by mtblaster (https://github.com/nidafra92/squirrel-project/blob/master/mtblaster.py) using the *Orseolia oryzae* Wood-Mason (Diptera: Cecidomyiidae) mitogenome (KM888183) to select for reads with high identity to a cecidomyiid mitochondria sequence. Next, resulting reads were filtered by

nanofilt (https://github.com/wdecoster/nanofilt) to keep only reads above 15 kbp in length and average Q score above 30. Finally, flye was used to perform mitogenome assembly (Kolmogorov 2023). A single circular contig was recovered with 785× coverage. The assembly was polished using four rounds of racon polishing (https://github.com/lbcb-sci/racon), followed by medaka with the same parameters as the nuclear assembly.

Wolbachia detection

Wolbachia infection was determined using the references of nine *Wolbachia* genomes broadly covering the *Wolbachia* phylogeny. The strains are described in Table 1. All strain genomes were concatenated and used as the query against the set of total sequencing reads. Minimap2 was used to find high identity hits. These hits were used as input to kraken2 for species identification using the K2-Standard-16Gb database (https://benlangmead.github.io/aws-indexes/k2) version 2022-06-07 (Wood *et al.* 2019). Secondly, a PCR-based approach was used to validate the absence of *Wolbachia*. We used *Wolbachia*-specific W-Spec primers (Werren and Windsor 2000) (W-Specf (CATACCTATTCGAAGGGATAG) and W-Specr (AGCTTCGAGTGAAACCAATTC) to screen for the presence of *Wolbachia* 16S DNA in the sample. The PCR method failed to detect *Wolbachia*, affirming our computational findings.

Phylogenetic reconstruction

A whole-genome phylogeny was created with an alignment-free method, SANS, that follows a pangenomic approach to efficiently calculate a set of splits in a phylogenetic tree or network (Rempel and Wittler 2021). Default settings were used. A mitochondrial phylogeny was created using cytochrome oxidase I (COX1) sequences from (Dorchin *et al.* 2019). Amino acid sequences were aligned using MEGA11. Since many of the COX1 sequences available on NCBI are partial and vary in length, we trimmed the aligned sequences to roughly equal size in MEGA11 (Supplementary File 5). After aligning and trimming, we assembled a phylogeny in IQ-TREE 1.6.12 with the following settings: alignment: sgm_phylo_protein_align.fas, # of sequences = 47; sequence type and substitution model: amino acids, mtART; rate heterogeneity: none; state frequency: estimated by maximum likelihood; bootstrap branch support: UltraFast, # of replicates = 1,000; single branch test: none; tree search: perturbation strength = 0.5, # of unsuccessful iterations to stop = 500; and root tree: outgroups: *Rabdophaga heterobia* Loew (Diptera: Cecidomyiidae). The tree was rooted on *R. heterobia*, which is in the tribe Lasiopteridi. The remaining cecidomyiid COX1 sequences were from the tribe cecidomyiid, which includes *Resseliella* and close relatives (Dorchin *et al.* 2019). The mitogenome was visualized using GenomeVx on the web (Conant and Wolfe 2008).

Table 1. *Wolbachia* strains used for infection detection.

Strain	Host	Supergroup	Accession
*w*Mel	*Drosophila melanogaster*	A	GCF_000008025.1
*w*AlbB	*Aedes albopictus*	B	GCF_004171285.1
*w*Ov	*Onchocerca volvulus*	C	GCF_000530755.1
*w*Bm	*Brugia malayi*	D	GCF_000008385.1
*w*Fol	*Folsomia candida*	E	GCF_001931755.2
*w*Cle	*Cimex lectularius*	F	GCF_000829315.1
*w*CfeJ	*Ctenocephalides felis*	J	GCF_012277315.1
*w*Ppe	*Pratylenchus penetrans*	L	GCF_001752665.1
*w*CfeT	*Ctenocephalides felis*	T	GCF_012277295.1

Results and discussion

Assembly

Three pools of 50 adult individuals of *R. maxima* were digested for DNA extraction and sequenced over three flow cells on an Oxford Nanopore MinION sequencer (Supplementary File 1). A total of 13.7 Gb sequences was generated, with an N50 of 3,485 bp, and 76% of bases had a greater than Q20 quality score (Table 2). All bases were used for genome assembly. A draft assembly was generated containing 2,613 contigs with a length of 211 Mb. Decontamination and quality control filtering removed short contigs and those with anomalous coverage (<2× and >200×). The final assembly was 206 Mb, spread over 1,009 contigs with an N50 of 714,500 bp, and coverage of 64.88×. The genome-wide GC level is 31.60%. The assembly is available under NCBI Project number PRJNA928452 and accession number JAQOWM000000000.

BUSCO scores indicated high completeness of the assembly, with no BUSCO genes lost during generation of the final assembly (BUSCO Diptera odb10 database). The reduction of 0.1% composite score in the final assembly as compared to the draft is explained by removal of 0.2% duplicate BUSCOs and an increase of 0.1% single-copy BUSCOs during assembly polishing with medaka and removal of contaminants. The final assembly has the second highest single-copy BUSCO score of cecidomyiids, only exceeded by *Catotricha subobsoleta* (Alexander) (Diptera: Cecidomyiidae) (Fig. 2).

Our final assembly was 206 Mb, in line with other cecidomyiid genomes, such as *C. subobsoleta* at 277 Mb, swede midge; *C. nasturtii* at 186 Mb, Hessian fly; *Mayetiola destructor* Say (Diptera: Cecidomyiidae) at 186 Mb; *Porricondyla nigripennis* Meigan (Diptera: Cecidomyiidae) at 286 Mb; and wheat midge, *Sitodiplosis mosellana* Géhin (Diptera: Cecidomyiidae) at 181 Mb. Assembly contiguity was high, with an N50 of 714 kb, and the genome BUSCO score indicates our assembly has one of the highest completeness levels for a cecidomyiid. In addition to utility as a guide for genome assembly of related taxa, the *R. maxima* genome will contribute to a broader understanding of major biological traits associated with herbivory such as host adaptation, detoxification, and immunity, e.g. Grbić *et al.* (2011), Sparks *et al.* (2020), and Guan *et al.* (2021).

DNA methylation

Nanopore sequencing can distinguish 5′methylated cytosines (5mC) from unmethylated cytosines through base calling using a methylation-aware neural network model. Global DNA

Table 2. Assembly statistics.

Assembly	Contigs	Length (bp)	Min length	Avg length	Max length	N50
Draft	2,613	211,232,549	384	80,839	6,309,645	698,470
Final	1,009	206,036,186	8,001	204,198	6,309,645	714,500
BUSCO	Complete (%)	Single (%)	Duplicate (%)	Fragment (%)	Missing (%)	*n*
Draft	87.9	85.1	2.8	1.9	10.2	3285
Final	87.8	85.2	2.6	1.9	10.3	3285

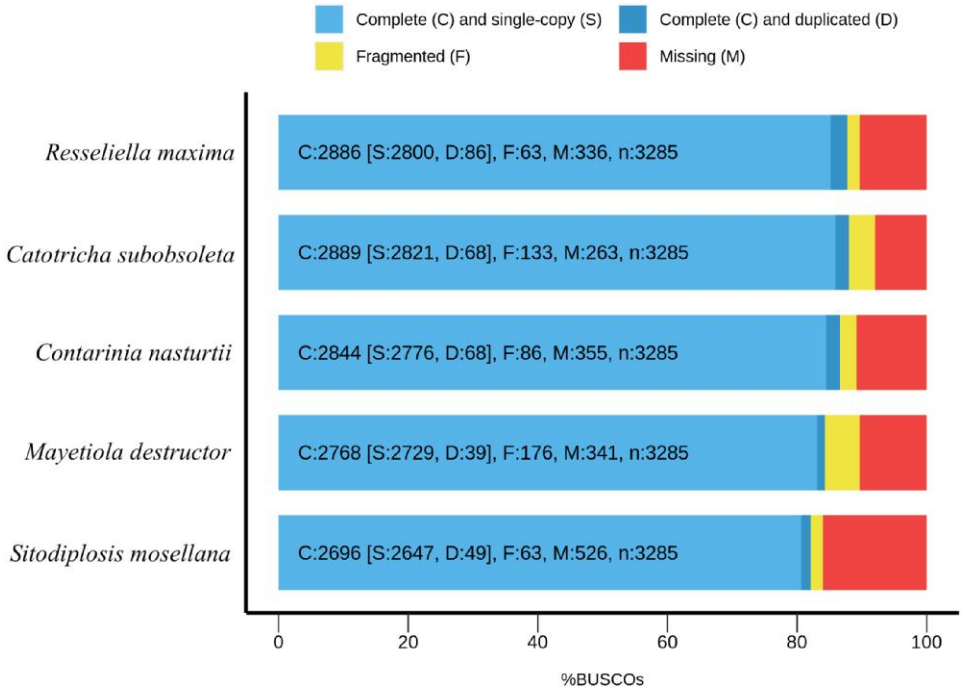

Fig. 2. BUSCO scores. *R. maxima* assembly compared to other cecidomyiid genomes available from NCBI.

methylation of cytosines in a CpG context was estimated by the fraction of 5mC divided by unmethylated cytosines, for all reads that aligned to the final assembly. Genome-wide methylation was 1.07%. The methylation was uniform across the assembly with only a single contig averaging above 2% (Supplementary File 2).

We saw extremely low levels of DNA methylation across the assembly, with most contigs averaging at the 1% level. While this could be a biologically meaningful level of methylation, other Diptera also have negligible levels of methylation, presumably due to loss of DNA methyltransferases (DNMT1 and DNMT3) in the dipteran common ancestor (Glastad *et al.* 2011). In insects with functional DNA methylation pathways, the levels are between 3% and 40% (Bewick *et al.* 2016). The level measured here could be a reflection of nonspecific background DNA damage plus uncertainty in the neural network base-calling model that detects DNA methylation especially at such low levels. It is unlikely that *R. maxima* contains a functional methylation pathway based on its evolutionary history.

Repetitive DNA

To compare repeats across species, we first masked six focal cecidomyiid genomes against the most comprehensive public repeat database from the Dfam consortium. Across all cecidomyiids, repeats are poorly annotated in the existing database, as reflected by low percentages of repeat detection (Table 3). We then used RepeatModeler2, an *ab initio* repeat finding pipeline that does not rely upon prior consensus libraries, using only the genome itself, and found the percentage of repeats detected increased from 8.15% to 21.73%. Other cecidomyiid genomes had repeat content ranging from 11.89% to 29.78%. Reasoning that some repeats may be more or less shared between species, we used the repeat library we generated from *R. maxima* by RepeatModeler2 to search the other cecidomyiid genomes. Unexpectedly, we found that other Cecidomyiidae did not share a large percentage of repeats. We found that *S. mosellana*, shared the most repeats with *R. maxima*, followed by *C. nasturtii*, *C. subobsoleta*, and *M. destructor*, in that order. Despite having a high repeat content *P. nigripennis* shared the fewest repeats with *R. maxima* (2.88%). As Cecidomyiidae is approximately 150 million years old (Dorchin *et al.* 2019), and very little is known about the evolutionary dynamics of this group, it is unclear if the differences in repeats are reflective of this divergence time or if repeats are particularly active in this group.

Overall, the genome of *R. maxima* is comprised of 15.88% interspersed repeats, 5.07% simple repeats, and 1.52% low complexity repeats (Table 4). Most interspersed repeats remain unclassified, similar to the other cecidomyiids. The full repeat complement is available in Supplementary File 3.

Gene annotation

Putative coding regions were predicted using GeMoMa against the repeat masked version of the assembly. GeMoMa uses annotations of related species as hints to detect coding regions. We used *D. melanogaster* and *C. nasturtii* as source annotations. GeMoMa predicted 14,798 proteins with a protein BUSCO score of 89.9%, similar to the swede midge's score of 92.0%. The full set of proteins and associated data are available in Supplementary File 4.

Gene annotation is only available for a single cecidomyiid reference genome, *C. nasturtii*. Fortunately, that species is relatively closely related to *R. maxima* and provides good hints to the GeMoMa annotation pipeline used here. Our count of 14,798 protein coding genes is typical of an animal genome. The *C. nasturtii* assembly was created using Illumina short-read sequencing with 57× coverage, while our *R. maxima* relied solely on long-read nanopore reads. Although nanopore reads are inherently less accurate than Illumina reads, we were still able to generate an accurate consensus, and its BUSCO protein score of 89.9% indicates good resolution of the assembly. Other nanopore-only assemblies have shown overall quality scores of Q45 (1 in 50,000 base error rate) (Flack *et al.* 2022), considered as a platinum status genome by the Vertebrate Genomes Project (Morin *et al.* 2020).

Table 3. Repetitive content of cecidomyiid genomes.

Species	Assembly	Common name	vs Dfam (%)	vs Self (%)	vs R. maxima (%)	Size (Mb)
Resseliella maxima	Resmax_1	soybean gall midge	8.15	21.73	NA	206
Contarinia nasturtii	AAFC_CNas_1.1	swede midge	2.30	13.25	6.04	186
Porricondyla nigripennis	ASM2654663v1	NA	2.26	18.95	2.88	285
Sitodiplosis mosellana	ASM2101890v1	wheat midge	3.28	17.26	7.01	181
Mayetiola destructor	Mdes_1.0	Hessian fly	3.47	11.89	4.46	186
Catotricha subobsoleta	ASM1163474v2	NA	5.51	29.78	5.57	277

Table 4. Interspersed repeats in *R. maxima*.

Name	Number	Length (bp)	Percent (%)
Retroelements	12,207	8,108,723	3.94
Penelope class	117	29,589	0.01
LINE class	8858	5,531,322	2.68
L2/CR1/Rex	5902	3,795,681	1.84
RTE/Bov-B	2250	1,427,140	0.69
LTR class	3349	2,577,401	1.25
BEL/Pao	902	975,078	0.47
Ty1/copia	645	559,287	0.27
Gypsy/DIRS1	1802	1,043,036	0.51
DNA transposons	12,991	5,092,926	2.47
hobo-Activator	1428	588,862	0.29
Tc1-IS630-Pogo	8110	3,313,329	1.61
MULE-MuDR	61	27,086	0.01
Other	55	25,938	0.01
Rolling circles	283	137,313	0.07
Unclassified	82,923	17,722,456	8.6
Total interspersed		30,924,105	15.01
Small RNA	453	142,243	0.07
Simple repeats	252,445	10,436,422	5.07
Low complexity	57,017	3,135,087	1.52
Bases masked		44,775,170	21.73

Whole-genome phylogeny

First, we created a phylogeny based on whole-genome comparisons and found *R. maxima* most closely related to *S. mosellana* and *C. nasturtii*. We created our assembly using an alignment-free whole-genome–based reconstruction method yielding the tree in Fig. 3. We used genomes of the same six cecidomyiids as for repeat detection (plus *D. melanogaster* as an outgroup) as these are the only full genomes publicly available from Cecidomyiidae. Here, *R. maxima* is sister to *S. mosellana* and *C. nasturtii*, which matches the repeat content analysis where these three share the most repeats. In contrast, the multilocus phylogeny of Dorchin *et al.* (2019) places at least one other *Resseliella* species more closely to *C. nasturtii* before *S. mosellana*. Importantly, our phylogeny was limited to seven species for which whole genomes exist.

Mitochondrial structure and phylogeny

We extracted the *R. maxima* mitogenome from a subset of the total reads that were identified by BLAST as homologous to the mitogenome of the Asian rice gall midge, *O. oryzae*. The assembly is a single circular contig of 15,301 bp and matches 79.79% identity over 93% of its length to the *O. oryzae* mitogenome (KM888183.1). Gene annotation indicates some errors in assembly, likely due to polymorphisms within the pooled population used for sequencing (Fig. 4). The mitogenome has been deposited under accession number OQ342780.

The *R. maxima* mitogenome contains 22 tRNA genes. $tRNA^{Leu}$ and $tRNA^{Ser}$ have been duplicated as in *O. oryzae* and other gall midges. We were unable to annotate $tRNA^{Glu}$ in *R. maxima*, despite multiple attempts in MITOS2. This is potentially due to the unusually truncated tRNA genes observed in other cecidomyiid midges (Mori *et al.* 2021). Our mitogenome likely contains a few errors as assembly was particularly difficult in light of pooled sampling; however, others have shown nanopore-only mitogenomes are useful and reliable despite imperfections (De Vivo *et al.* 2022).

We compared *R. maxima* mitochondrial gene order to the closest relative for which a complete mitogenome is available, *O. oryzae*. Gene order in *R. maxima* varies significantly from *O. oryzae*, with some conserved elements. Nad4 and nad5 have been inverted in both *O. oryzae* and *R. maxima* (Supplementary File 5). A region containing COII, $tRNA^{Lys}$, atp8, atp6, and COIII is coded on the positive strand in *R. maxima*, while this entire region is inverted in *O. oryzae*. Additionally, $tRNA^{Leu}$ and $tRNA^{Asp}$ are present within this contiguous region in *R. maxima*, but not in *O. oryzae*. Another contiguous region on the positive strand containing $tRNA^{Thr}$, nad6, cytB, and $tRNA^{Ser}$ is present in both midges. With the exception of these relatively conserved regions, gene order is not well conserved between *R. maxima* and *O. oryzae*.

We created a larger phylogeny using a database of COX1 amino acid sequences and found that *R. maxima* grouped with several other *Resseliella* but that the genus appears to be polyphyletic overall. We used 46 COX1 sequences to reconstruct Cecidomyiidae relationships (Supplementary File 5). The resulting tree placed *R. maxima* as sister to *Resseliella oleisuga* Targioni-Tozzetti (Diptera: Cecidomyiidae) with 88% bootstrap support. *R. maxima* and *R. oleisuga* are grouped together with other *Resseliella* species in a polyphyletic clade that includes other gall midges in the subtribe Cecidomyiini, as well as gall midges in Aphidoletini, Lopesiini, and Lestodiplosini. These polyphyletic clades are likely a result of (1) a single-locus phylogenetic reconstruction, (2) the significant changes we see in mitochondrial genomes across Cecidomyiidae, and/or (3) the poor coverage of available cecidomyiid mitogenomes. However, this could potentially indicate a need for reexamination of cecidomyiid phylogeny.

Absence of *Wolbachia* infection

Some members of Cecidomyiidae host *Wolbachia* as intracellular bacteria (Behura *et al.* 2001; Bel Mokhtar *et al.* 2020). To test whether *R. maxima* was infected, we examined the sequencing reads for the presence of *Wolbachia*. Minimap2 was used to identify putative candidates in the total set of sequenced reads by matching to the full genomes of a set of nine strains that cover a wide range of *Wolbachia* diversity. There were 4,929 sequences (0.000036%) that matched these reference *Wolbachia* genomes. Of these 92% ($n = 4{,}480$) were identified as human origin, and the remaining 29 hits matched to a variety of non-*Wolbachia* bacteria (Supplementary File 6). This evidence indicates that this population of *R. maxima* is not hosting *Wolbachia* infection. As validation, we performed PCR on genomic isolates and found no amplification using general *Wolbachia* primers.

Summary

Here, we provide the first whole-genome assembly of a *Resseliella* species (206 Mb with 64.88× coverage) that has one of the highest

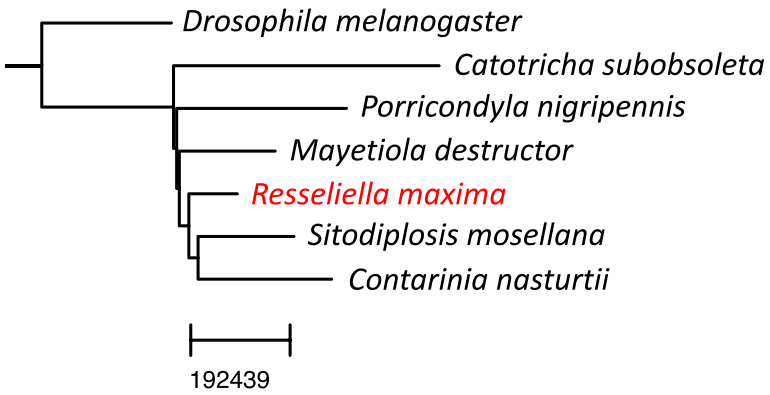

Fig. 3. Whole-genome phylogeny. A reference-free alignment method was used to create this whole-genome phylogeny.

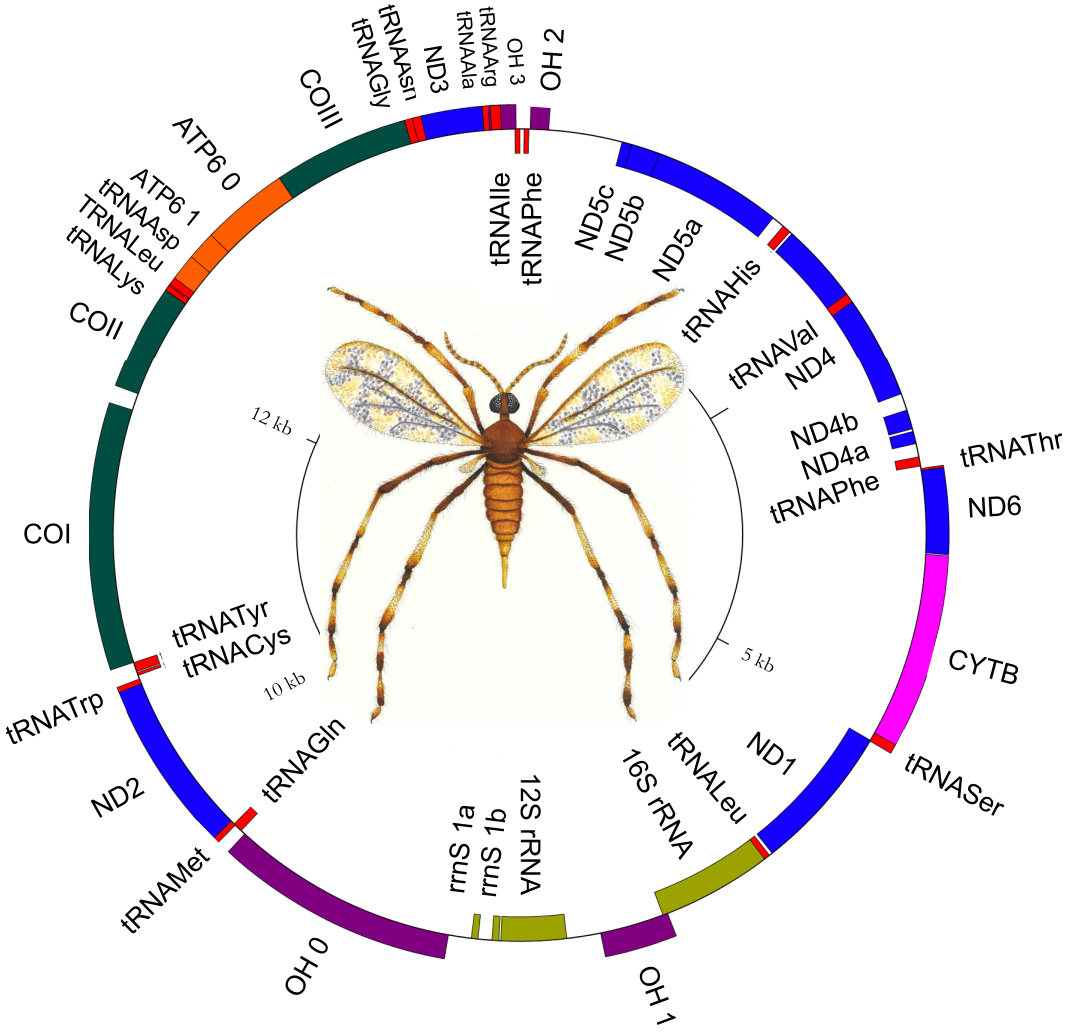

Fig. 4. Mitochondrial structure. The mitogenome of *R. maxima*.

completeness levels (BUSCO score of 87.8%) for a cecidomyiid. The *R. maxima* genome will provide a resource for research focused on the biology, genetics, and evolution of cecidomyiids, as well as development of management tactics.

Genomes of insects and other invertebrates can be difficult to assemble for a range of factors including (1) high heterozygosity when inbreeding is unrealistic, (2) small sizes that present very little DNA, requiring pooling individuals to make libraries, (3) lack of high-

quality genome assemblies of close relatives to assist with assembly due to high diversity across arthropods, (4) the need to optimize DNA preparations for new insect species, and (5) the fact that some arthropods have large genomes (Richards and Murali 2015). However, long-read sequencing technologies, e.g. Pacific Biosciences (PacBio) and Oxford Nanopore, are contributing to improvement in quality of genome assembly, producing assemblies ~48× more contiguous than short-read–based approaches (Li *et al.* 2019; Hotaling *et al.* 2021). Other pest insect genomes have recently been sequenced using nanopore-only approaches such as the black carpenter ant, *Camponotus pennsylvanicus* De Geer (Hymenoptera: Formicidae), and the coconut rhinoceros beetle, *Oryctes rhinoceros* [L.] (Coleoptera: Scarabaeidae) (Faulk 2022; Filipović *et al.* 2022). Here, we not only assembled a high-quality genome from long reads but were able to do so from pooled samples.

Knowledge of arthropod pest genomes can reveal important evolutionary innovations. For example, the genome of the spider mite, *Tetranychus urticae* Koch (Trombidiformes: Tetranychidae), showed the evolutionary innovation of silk production and signatures of polyphagy and detoxification (Grbić *et al.* 2011). Guan *et al.* (2021) used whole-genome sequencing to detect insecticide resistance mutations in fall armyworm, *Spodoptera frugiperda* (JE Smith) (Lepidoptera: Noctuidae). Genomic analysis of the brown marmorated stink bug, *Halyomorpha halys* (Stål) (Hemiptera: Pentatomidae), revealed genetic elements associated with immunity and detoxification that have potential for biomolecular pesticide applications (Sparks *et al.* 2020).

Even though the family Cecidomyiidae has more than 6,600 described species (Dorchin *et al.* 2019), there are only five genome assemblies from this family and none from the genus *Resseliella* available on GenBank. Not surprisingly, most of the genomes belong to agricultural insect pests (i.e. *M. destructor*, *C. nasturtii*, and *S. mosellana*). Availability of the *R. maxima* genome will facilitate population genetics-based understandings of the origin of *R. maxima* and its spread to new areas and provide possibilities for future work on developing alternative pest control methods.

Data availability

Assembly and BioSample information is available at NCBI Project number PRJNA928452 and accession number JAQOWM000000000. The mitochondrial assembly is available separately at accession number OQ342780. Supplementary material is available at figshare: https://doi.org/10.25387/g3.21984575.

Acknowledgments

We thank Bruce Potter for his field assistance.

Funding

This work was supported by the National Institutes of Health Office of the Director T32OD010993 (NF), NIH R21AG071908 (CF), Impetus Grant (Norn Foundation) (CF), USDA-NIFA MIN-16-129 (CF), and the Minnesota Rapid Agricultural Response Fund (RLK and ARIL).

Conflicts of interest statement

The authors declare no conflict of interest.

Author contributions

This work was a collaborative effort by the members of graduate course ANSC 8509 taught by CF in the Department of Animal Science at the University of Minnesota in Fall 2022. Students were GM, MWJ, KB, NF, LEF, EJ, EJK, SN, MR, and CW. Students performed analyses and provided text for the manuscript. ARIL and RLK provided the midge samples, analytical guidance, and edited the manuscript. CF performed analyses and edited the manuscript. All authors have read and approved the manuscript prior to submission.

Literature cited

Behura SK, Sahu SC, Mohan M, Nair S. *Wolbachia* in the Asian rice gall midge, *Orseolia oryzae* (Wood-Mason): correlation between host mitotypes and infection status: *Wolbachia* infection in the rice gall midge. Insect Mol Biol. 2001;10(2):163–171. doi:10.1046/j.1365-2583.2001.00251.x.

Bel Mokhtar N, Maurady A, Britel MR, El Bouhssini M, Batargias C, Stathopoulou P, Asimakis E, Tsiamis G. Detection of *Wolbachia* infections in natural and laboratory populations of the Moroccan Hessian fly, *Mayetiola destructor* (Say). Insects. 2020;11(6):340. doi:10.3390/insects11060340.

Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. Evolution of DNA methylation across insects. Mol Biol Evol. 2016;34(3):654–665. doi:10.1093/molbev/msw264.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. Blobtoolkit—interactive quality assessment of genome assemblies. G3 (Bethesda). 2020;10(4):1361–1374. doi:doi:10.1534/g3.119.400908.

Conant GC, Wolfe KH. Genomevx: simple web-based creation of editable circular chromosome maps. Bioinformatics. 2008;24(6):861–862. doi:10.1093/bioinformatics/btm598.

De Vivo M, Lee H-H, Huang Y-S, Dreyer N, Fong C-L, de Mattos FMG, Jain D, Wen Y-HV, Mwihaki JK, Wang T-Y, *et al.* Utilisation of Oxford nanopore sequencing to generate six complete gastropod mitochondrial genomes as part of a biodiversity curriculum. Sci Rep. 2022;12(1):9973. doi:10.1038/s41598-022-14121-0.

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. doi:10.1016/j.ympev.2019.106602.

Faulk C. De novo sequencing, diploid assembly, and annotation of the black carpenter ant, *Camponotus pennsylvanicus*, and its symbionts by one person for $1000, using nanopore sequencing. Nucleic Acids Res. 2022;51(1):17–28. doi:10.1093/nar/gkac510.

Filipović I, Rašić G, Hereward J, Gharuka M, Devine GJ, Furlong MJ, Etebari K. A high-quality de novo genome assembly based on nanopore sequencing of a wild-caught coconut rhinoceros beetle (*Oryctes rhinoceros*). BMC Genomics. 2022;23(1):426. doi:10.1186/s12864-022-08628-z.

Flack N, Drown M, Walls C, Pratte J, McLain A, Faulk C. Chromosome-level, nanopore-only genome and allele-specific DNA methylation of Pallas's Cat, Otocolobus manul. 2022. doi:10.1101/2022.11.30.518596.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117-(17):9451–9457. doi:10.1073/pnas.1921046117.

Gagné RJ, Yukawa J, Elsayed AK, McMechan AJ. A new pest species of *Resseliella* (Diptera: Cecidomyiidae) on soybean (Fabaceae) in North America, with a description of the genus. Proc Entomol Soc Wash. 2019;121(2):168–177. doi:10.4289/0013-8797.121.2.168.

Glastad KM, Hunt BG, Yi SV, Goodisman MAD. DNA Methylation in insects: on the brink of the epigenomic era. Insect Mol Biol. 2011;20(5):553–565. doi:10.1111/j.1365-2583.2011.01092.x.

Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Thi Ngoc PC, Ortego F, *et al*. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature. 2011;479(7374):487–492. doi:10.1038/nature10640.

Guan F, Zhang J, Shen H, Wang X, Padovan A, Walsh TK, Tay WT, Gordon KHJ, James W, Czepak C, *et al*. Whole-genome sequencing to detect mutations associated with resistance to insecticides and Bt proteins in *Spodoptera frugiperda*. Insect Sci. 2021;28(3):627–638. doi:10.1111/1744-7917.12838.

Helton ML, Tinsley NA, McMechan AJ, Hodgson EW. Developing an injury severity to yield loss relationship for soybean gall midge (Diptera: Cecidomyiidae). J Econ Entomol. 2022;115(3):767–772. doi:10.1093/jee/toac038.

Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB. Long reads are revolutionizing 20 years of insect genome sequencing. Genome Biol Evol. 2021;13(8):evab138. doi:10.1093/gbe/evab138.

Keilwagen N, Hartung F, Grau J. Gene prediction: methods and protocols. In: Kollmar M, editor. Methods in Molecular Biology. New York, NY: Springer New York; 2019.

Kolmogorov M. Flye assembler; 2023. https://github.com/fenderglass/Flye.

Li F, Zhao X, Li M, He K, Huang C, Zhou Y, Li Z, Walters JR. Insect genomes: progress and challenges. Insect Mol Biol. 2019;28(6):739–758. doi:10.1111/imb.12599.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–4654. doi:10.1093/molbev/msab199.

McMechan AJ, Hodgson EW, Varenhorst AJ, Hunt T, Wright R, Potter B. Soybean gall midge (Diptera: Cecidomyiidae), a new species causing injury to soybean in the United States. J Integr Pest Manag. 2021;12(1):8. doi:10.1093/jipm/pmab001.

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. doi:10.1093/gbe/evab036.

Morin PA, Alexander A, Blaxter M, Caballero S, Fedrigo O, Fontaine MC, Foote AD, Kuraku S, Maloney B, McCarthy ML, *et al*. Building genomic infrastructure: sequencing platinum-standard reference-quality genomes of all cetacean species. Mar Mammal Sci. 2020;36(4):1356–1366. doi:10.1111/mms.12721.

Rempel A, Wittler R. SANS Serif: alignment-free, whole-genome-based phylogenetic reconstruction. Bioinformatics. 2021;37(24):4868–4870. doi:10.1093/bioinformatics/btab444.

Richards S, Murali SC. Best practices in insect genome sequencing: what works and what doesn't. Curr Opin Insect Sci. 2015;7:1–7. doi:10.1016/j.cois.2015.02.013.

Sparks ME, Bansal R, Benoit JB, Blackburn MB, Chao H, Chen M, Cheng S, Childers C, Dinh H, Doddapaneni HV, *et al*. Brown marmorated stink bug, *Halyomorpha halys* (Stål), genome: putative underpinnings of polyphagy, insecticide resistance potential and biology of a top worldwide pest. BMC Genomics. 2020;21(1):227. doi:10.1186/s12864-020-6510-7.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):2. doi:10.1186/s13100-020-00230-y.

USDA. Crop production 2021 summary 01/12/2022. Crop Prod. 119; 2022.

Werren JH, Windsor DM. *Wolbachia* infection frequencies in insects: evidence of a global equilibrium? Proc Biol Sci. 2000;267(1450):1277–1285. doi:10.1098/rspb.2000.1139.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. doi:10.1186/s13059-019-1891-0.

Editor: K. Vogel