

# Sex-based *de novo* transcriptome assemblies of the parasitoid wasp *Encarsia suzannae*, a host of the manipulative heritable symbiont *Cardinium hertigii*

Dylan L. Schultz<sup>1,2</sup>, Evelyne Selberherr<sup>3</sup>, Corinne M. Stouthamer<sup>4</sup>, Matthew R. Doremus<sup>4</sup>, Suzanne E. Kelly<sup>4</sup>, Martha S. Hunter<sup>4</sup> and Stephan Schmitz-Esser<sup>1,2,\*</sup>

- <sup>1</sup> Department of Animal Science, Iowa State University, Ames, IA 50011, USA
- <sup>2</sup> Interdepartmental Microbiology Graduate Program, Iowa State University, Ames, IA 50011, USA
- <sup>3</sup> Unit of Food Microbiology, Institute of Food Safety, Food Technology and Veterinary Public Health, Department for Farm Animals and Veterinary Public Health, University of Veterinary Medicine Vienna, 1210 Vienna, Austria
- <sup>4</sup> Department of Entomology, The University of Arizona, Tucson, AZ 85721, USA

## ABSTRACT

Parasitoid wasps in the genus *Encarsia* are commonly used as biological pest control agents of whiteflies and armored scale insects in greenhouses or the field. They are also hosts of the bacterial endosymbiont *Cardinium hertigii*, which can cause reproductive manipulation phenotypes, including parthenogenesis, feminization, and cytoplasmic incompatibility (the last is mainly studied in *Encarsia suzannae*). Despite their biological and economic importance, there are no published *Encarsia* genomes and only one public transcriptome. Here, we applied a mapping-and-removal approach to eliminate known contaminants from previously-obtained Illumina sequencing data. We generated *de novo* transcriptome assemblies for both female and male *E. suzannae* which contain 45,986 and 54,762 final coding sequences, respectively. Benchmarking Single-Copy Orthologs results indicate both assemblies are highly complete. Preliminary analyses revealed the presence of homologs of sex-determination genes characterized in other insects and putative venom proteins. Our male and female transcriptomes will be valuable tools to better understand the biology of *Encarsia* and their evolutionary relatives, particularly in studies involving insects of only one sex.

**Subjects** Genetics and Genomics, Animal Genetics, Transcriptomics

## BACKGROUND

*Encarsia suzannae* are minute parasitoid wasps within the order Hymenoptera. Our interest in this species is due to their unusual behavior and biology, their use as a biological control of the important whitefly pest *Bemisia tabaci*, their relatedness to the widespread greenhouse biological control agent *Encarsia formosa*, and because they harbor a bacterial endosymbiont capable of host reproductive manipulation, *Cardinium hertigii*. *Cardinium*, from the bacterial phylum *Bacteroidota*, shows independent evolution of reproductive manipulation from the well-known alphaproteobacterial *Wolbachia* [1]. Like other

**Submitted:** 23 June 2022  
**Accepted:** 31 August 2022  
**Published:** 31 August 2022

\* Corresponding author at Department of Animal Science, Iowa State University, 3222 NSRIC, 1029 North University Boulevard, Ames, IA 50011, USA. E-mail: [sse@iastate.edu](mailto:sse@iastate.edu)

Published by GigaScience Press.

Preprint submitted at <https://doi.org/10.1101/2022.08.05.502955>

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

*Gigabyte*, 2022, 1–13

Hymenoptera, *E. suzannae* are haplodiploid and reproduce via arrhenotoky (arrhenotokous parthenogenesis): haploid males are produced via unfertilized eggs, and females are derived from fertilized diploid eggs [2]. Most *Encarsia* species, including *E. suzannae*, are also autoparasitoids. Specifically, females develop inside and consume the nymphs of the sweet potato whitefly, *B. tabaci*; male wasps develop as hyperparasitoids, consuming the pupae of conspecific females or other aphelinid parasitoids of whiteflies. After consuming their host, both male and female *Encarsia* pupate in the whitefly cuticle and emerge as adults [3]. Thus, many *Encarsia* species are effective parasites of the whitefly species. The latter are widespread pests causing billions of dollars in crop losses yearly as they can damage plants while feeding and transmit more than 200 different plant viruses to many plant species [4, 5]. As a result, *Encarsia* species have been widely used as pest control agents to limit whitefly populations in field or greenhouse settings [6–8]. The unusual autoparasitic biology [9], sex allocation behavior, and host selection of these intriguing wasps have also been studied [10].

Like many insects, *Encarsia* may be infected with maternally-transmitted intracellular bacterial endosymbionts, such as *Wolbachia* and *Cardinium*, which influence their transmission by manipulating host reproduction [11] or oviposition behavior [12] to favor infected females. These manipulations may induce asexual reproduction via thelytokous parthenogenesis [13, 14] or a type of male reproductive sabotage called cytoplasmic incompatibility (CI) [15]. CI causes the offspring of infected males and uninfected females to die early during development; on the other hand, females infected with the same symbiont can successfully mate with infected or uninfected males. This sabotage proceeds via a two-step mechanism: the symbiont alters the male sperm with a fatal modification, then rescues the infected offspring from this fatal modification when present in the fertilized egg. Together, the modification and rescue steps of CI grant infected females a relative fitness advantage over uninfected females, driving the symbiont to high frequencies in host populations [11]. The role of endosymbionts in arthropod biology, evolution, and speciation has been the subject of intense study [16–18]. Much of this research has focused on symbiont-induced CI, given its potential role in insect speciation [19–21], its application in arthropod pest population control [22, 23], and its ability to drive desirable genetic traits through populations (e.g., the resistance to arthropod-borne diseases) [24].

The cEper1 strain of *Cardinium hertigii* is the causal agent of CI in *E. suzannae* [15]. This symbiosis between cEper1 and *E. suzannae* is the best-studied instance of *Cardinium*-induced CI, and this strain of *Cardinium* has been well-characterized by genomic and transcriptomic data [1, 3]. However, sequence information of the host, *E. suzannae*, is extremely sparse: this species currently lacks a sequenced genome and a transcriptomic profile, hampering the molecular identification of host-symbiont interactions.

Here, we generated separate *de novo* assembled transcriptomes for male and female *E. suzannae* using previously obtained RNA-seq data that was generated to characterize the *Cardinium hertigii* transcriptome [3]. To our knowledge, there is only one other publicly available *Encarsia* transcriptome: that of the widely used greenhouse whitefly biocontrol agent *Encarsia formosa*, which was published as part of a phylogenetic characterization of Chalcidoidea parasitoid wasps [25, 26]. However, based on the morphology and lifestyle differences between *E. suzannae* and *E. formosa*, as well as their phylogenetic relationship, the two species are distantly related within the diverse *Encarsia* genus [27–30]. Our dataset will be a valuable asset for an ecologically important lineage within the chalcidoid wasps

(Aphelinidae) that is sorely lacking sequencing data. We also provide the first molecular characterization of the host in the model *Cardinium* CI system.

## METHODS

### Sample information and sequencing

We used the transcriptome data obtained by Mann *et al.* [3]. Whereas they focused on *Cardinium* data, here, we focused on the host (non-*Cardinium*) reads of the same dataset. The data was collected as described in the original manuscript [3]. In brief, the initial *E. suzannae* (NCBI:txid1892410) culture was obtained in 2006 in Weslaco, TX, from whitefly (*B. tabaci*) hosts. Male and female wasps were reared separately in a laboratory culture as described previously [3]. For females, mated *E. suzannae* were introduced to cages bearing whitefly nymphs on cowpea (*Vigna unguiculata*) plants. For males, unmated *E. suzannae* were provided with *Eretmocerus* sp. nr. *emiratus* larvae or pupae developing within whitefly nymphs. The total RNA from 6 groups of 350–500 male or female 1- to 3-day old *E. suzannae* wasps was extracted using TRIzol (Invitrogen). Next, the digestion of genomic DNA was done with the Turbo DNA-free kit (Ambion). The quality of the extracted RNA was assessed with a 2100 bioanalyzer (Agilent Technologies, RRID:SCR\_018043). Three libraries for each sex were generated with the NEBNext Ultra RNA Library Prep Kit (Illumina) combined with the Ribo-Zero Magnetic Gold Kit (Epicentre Biotechnologies) for rRNA depletion. Samples were sequenced on an HiSeq 2500 platform (Illumina, RRID:SCR\_016383) at the Vienna BioCenter Core Facilities (VBCF) NGS unit [31], producing a range of 127 to 162 million 50 bp paired-end reads per sample [3].

### Read preparation and assembly

Raw read files were processed with BBDuk (RRID:SCR\_016969) from the BBTools software suite (v37.36, RRID:SCR\_016968) [32] to remove the Illumina adapter sequences, trim and/or filter out whole reads with a quality score less than 15, and remove reads shorter than 36 bp after trimming using the following options: “ref=adapters.fa ktrim=r ordered k=23 hdist=1 mink=11 tpe tbo maq=15 qtrim=rl trimq=15 minlen=36”. We utilized FastQC (v0.11.9, RRID:SCR\_014583) to visualize the sequence quality of each sample before and after trimming and to confirm the successful removal of adapter sequences [33]. Due to the complex biology of this species and its host insects, sequence contamination from a variety of organisms throughout the rearing system is inevitable, including *Cardinium* cEper1, the different insect hosts of male and female *E. suzannae*, and the endosymbionts of those insect hosts. Thus, we employed a mapping-and-removal approach to enrich for *E. suzannae* reads prior to assembly and limit the generation of contaminating transcripts. For this approach, BBDuk (RRID:SCR\_016965) from BBTools was used to initially map the quality-trimmed reads to the genomes of *Cardinium hertigii* cEper1 and the endosymbionts of *Bemisia tabaci* MEAM1, with which *E. suzannae* females and males have direct or indirect contact (i.e., *Hamiltonella defensa*, *Portiera aleyrodidarum*, and *Rickettsia* sp. MEAM1 [34, 35]). It was also determined that the *E. sp. nr. emiratus* hosts of *E. suzannae* males contain *Wolbachia* [36]; thus, the *Wolbachia* wPip genome was added and mapped to the male samples. Reads that did not map to any of these bacterial genomes with a greater than 94% identity were retained (to allow for a difference of 3 nucleotides between sequenced transcripts and reference endosymbiont genomes). These reads were then subsequently mapped to the *B. tabaci* MEAM1 genome with a more stringent 97% identity

Table 1. Pre-assembly contaminant read mapping and removal of *Encarsia suzannae* transcriptome sequencing data.

Organism	Reason for removal	Proportion of mapped trimmed reads	GenBank accession number
<i>Cardinium hertigii</i> cEper1	CI-causing secondary <i>E. suzannae</i> endosymbiont	Female: 1.183% Male: 0.991%	GCA_000304455.1
<i>Portiera aleyrodidarum</i> MEAM1	Primary endosymbiont of <i>B. tabaci</i>	Female: 0.043% Male: 0.035%	GCA_002285875.1
<i>Rickettsia</i> sp. MEAM1	Secondary endosymbiont of <i>B. tabaci</i>	Female: 0.058% Male: 0.065%	GCA_002285905.1
<i>Hamiltonella defensa</i> MEAM1	Secondary endosymbiont of <i>B. tabaci</i>	Female: 0.040% Male: 0.037%	GCA_002285855.1
<i>Bemisia tabaci</i> MEAM1	Parasitized by female <i>E. suzannae</i> offspring and <i>E. sp. nr. emiratus</i>	Female: 5.343% Male: 5.289%	GCA_001854935.1
<i>Wolbachia pipientis</i> wPip (male only)	Secondary endosymbiont of <i>E. sp. nr. emiratus</i> , which is parasitized by male <i>E. suzannae</i> offspring	Female: N/A Male: 0.050%	GCA_000073005.1

List of the organisms whose reads were removed prior to assembly with Trinity. Quality-controlled reads were mapped to the genomes of the listed organisms, and the reads mapped to any of the references were removed.

threshold using BBMap to avoid mapping *E. suzannae* reads from genes highly conserved in both *Encarsia* and *Bemisia* (see Table 1 for mapping and removal details). Again, only unmapped reads were retained for assembly, as these final reads were expected to be mainly attributed to *E. suzannae*.

We assembled separate transcriptomes for male and female *E. suzannae* whole adult wasps with the remaining unmapped reads using Trinity (v2.6.6, RRID:SCR\_013048) and its default settings [37]. Transcript abundance was then estimated for each transcriptome with kallisto (RRID:SCR\_016582) using the “align\_and\_estimate\_abundance.pl” command bundled with Trinity [38]. Transcripts with an estimated abundance below 0.5 transcripts per million were removed from both assemblies as these may be lowly expressed isoforms of other transcripts, poorly assembled or chimeric transcripts, or simply contaminants and, thus, not from *Encarsia* [39, 40]. Next, TransDecoder (v5.5.0, RRID:SCR\_017647) [41] was used to predict coding sequences within the remaining transcripts in each assembly and translate those coding sequences into predicted protein sequences with a minimum amino acid length of 67. Similar protein-coding sequences were then clustered using CD-HIT (v4.6.8, RRID:SCR\_007105) [42, 43] with a 95% amino acid identity threshold, and the longest protein isoform of each cluster was selected as the representative sequence for that cluster. The final assemblies are presented as the nucleotide sequences of the representative proteins of each cluster. For a comprehensive list of the number of reads or transcripts at each step in the pipeline, see Table 2.

Quality control and data validation

Along with our mapping-and-removal approach to limit contaminations while enriching for *Encarsia* reads prior to assembly, we also utilized additional methods to improve the quality of our assemblies. First, to comply with the National Center for Biotechnology Information (NCBI)’s Transcriptome Shotgun Assembly (TSA) database requirements, we removed all coding sequences below 200 bp. Furthermore, we used blastn (RRID:SCR\_001598) with the remaining sequences against the NCBI’s vector database to identify contaminating sequences and synthetic RNA spike-in controls; hits with a 100% nucleotide identity to vector sequences were removed from each assembly [44]. Prior to submission, any remaining coding sequences flagged by NCBI’s contamination check as sequencing vectors



**Table 2.** *E. suzannae* transcriptome read and transcript statistics.

	<i>E. suzannae</i> female	<i>E. suzannae</i> male	<i>E. formosa</i>
Total number of reads	439,763,386	449,368,298	14,341,314
Reads after trimming and mapping	401,213,202	412,945,938	N/A
Initial transcripts	146,798	211,544	48,232
Final transcripts	122,465	136,359	47,852*
Coding sequences	45,986*	54,762*	27,161
Average length of final sequences (bp)	697.74	692.03	772.51
Assembly N50	1,275	1,200	1,237
Average% GC	44.94	44.88	37.5
% Annotated	58.27	65.34	0
Assembly software	Trinity v2.6.6	Trinity 2.6.6	SOAPdenovo-Trans-31kmer (v1.01, RRID:SCR_013268)
Reference	This study	This study	[25]

Assembly and annotation statistics at each step in the pipeline for our *E. suzannae* transcriptomes compared to the previously-published *E. formosa* transcriptome assembly. The \* highlight the number and type of final sequences in the public version of each assembly.

**Table 3.** Prediction of *E. suzannae* transcriptome assembly completeness using BUSCO.

	Male <i>E. suzannae</i>		Female <i>E. suzannae</i>	
BUSCO results	BUSCOs present	Percent of total	BUSCOs present	Percent of total
Complete BUSCOs	4,953	82.6%	4,915	82.1%
Complete single-copy BUSCOs	3,591	59.9%	4,492	75.0%
Complete duplicated BUSCOs	1,362	22.7%	423	7.1%
Fragmented BUSCOs	279	4.7%	280	4.7%
Missing BUSCOs	759	12.7%	796	13.2%
Total BUSCO groups searched	5,991	100%	5,991	100%

Assessment of assembly completeness using BUSCO v5.3.2 to search the assembled proteins against a database of proteins identified as Hymenopteran BUSCOs. All BUSCO groups searched were determined to be present in a single copy in >90% of the Hymenopteran species tested; therefore, a high number of complete single-copy BUSCOs indicates a comprehensive and non-redundant assembly [45].

or contaminants were also removed. In total, 71 and 109 contaminating sequences were removed from the female and male assemblies, respectively.

The final assemblies were then assessed for completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO) (v5.3.2, RRID:SCR\_015008) in protein mode against the hymenoptera\_odb10 reference lineage (v2020-08-05) [45, 46]. The female and male assemblies were found to possess, respectively, 82.1% and 82.6% of the 5,991 complete orthologs identified as single-copy and nearly universal within the order Hymenoptera (present in >90% of the tested species). This indicates a high level of completeness for both *E. suzannae* transcriptomes, although with varying degrees of duplication (shown in Table 3).

One issue we could not rectify with the currently available sequencing data was the presence of *E. sp. nr. emiratus* transcripts within the male *E. suzannae* assembly. As mentioned above, haploid male *E. suzannae* eggs are laid into *Eretmocer* pupae. Since this host does not have a sequenced genome (in contrast to *B. tabaci*), we could not apply the same mapping-and-removal approach to *E. sp. nr. emiratus*. This fact may at least partly explain the high number of total sequences and duplicated BUSCOs in the *E. suzannae* male assembly compared to the female one (see Tables 2 and 3). However, there are likely other contributing factors. Due to the relatedness of *Encarsia* and *Eretmocer*, we could not differentiate sequences originating from either of these organisms at the read or assembled

transcript level without their reference genomes. However, we are confident that the abundance of the *Eretmocerus* transcripts in the male assembly is low, and many may have been removed from the assembly during the transcript abundance filtering step. This is evidenced by the very low *Eretmocerus* biomass in/on fully emerged adult *E. suzannae* (larval *Encarsia suzannae* void their guts before pupation [47]). Additionally, using the average abundance of the *B. tabaci* reads as a proxy for the *E. sp. nr. emiratus* reads suggests an abundance of around 5% for *Eretmocerus* in either assembly (Table 1).

## Annotation

The male and female *E. suzannae* assemblies are available as unannotated coding sequences in the NCBI's TSA database under the accession numbers [GJLB00000000](#) and [GJLI00000000](#), respectively. Here, we also provide the annotation information for both assemblies from multiple sources.

The final clustered proteins were annotated through the eggNOG-mapper (v2, [RRID:SCR\\_021165](#)) web-based pipeline using default settings to assign taxonomy information to the sequences and generate an annotation report including Gene Ontology terms, Pfam domains, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, and other relevant information [48, 49]. Additionally, the final proteins were searched against the NCBI's non-redundant (nr) protein database (release 242.0) using DIAMOND ([RRID:SCR\\_016071](#)) and the “-very-sensitive” option [50]. The final proteins were also searched using blastp ([RRID:SCR\\_001010](#)) [51, 52] and an *e*-value cutoff of  $10^{-5}$  against a targeted database of well-annotated predicted insect proteomes: *Nasonia vitripennis* Nvit\_psr\_1\_1 (Genbank accession: [GCA\\_009193385.2](#)), *Trichogramma pretiosum* Tpre\_2\_0 (Genbank accession: [GCA\\_000599845.3](#)), and *Bemisia tabaci* MEAM1 (Genbank accession: [GCA\\_001854935.1](#)). Although not closely related to *Encarsia*, *Bemisia* was included in the targeted insect database. Its thorough annotation and presence as an outgroup may be useful in annotating proteins retained in *Encarsia* that *Nasonia* or *Trichogramma* may have lost. This database was also found to generate fewer hits labeled as “hypothetical” or “uncharacterized” when compared to a search against the nr protein database. The annotation results from each reference for both assemblies were pooled into a single Microsoft Excel spreadsheet (Additional File 1). We also provide a .fasta file for each assembly containing the final nucleotide sequences and the sequence headers containing the annotations from blastp against the targeted insect database (female: Additional file 2; male: Additional file 3).

Approximately 58% and 65% of the female and male assembled proteins were annotated using one of the listed methods, with the characterization against the NCBI's nr database annotating the highest number of proteins (26,155 female and 35,073 male), followed closely by the targeted insect database (24,478 female and 33,353 male). Some transcripts of note that were annotated in both the male and female assemblies are putative homologs to an array of insect sex-determination genes characterized in *Drosophila*. These homologs included *sex lethal* (*sxl*), the master regulator of the *Drosophila* sex-determination cascade, and some genes it regulates, including *transformer* (*tra*), *doublesex* (*dsx*), and *fruitless* (*fru*). *Sex lethal* controls the splicing of *tra*, which is involved in the sex-specific splicing of *dsx* and *fru* [53] and results in either male isoforms of *dsx* and *fru* or a female isoform of *dsx* and a truncated and untranslated female *fru* isoform. The different *dsx* isoforms are crucial for male and female somatic sexual development, while *fru* appears to be key in the male

courtship behavior of *Drosophila* [54, 55]. We also searched the assemblies for homologs of *wasp overruler of masculinization (wom)* [56] but found none. This gene was identified in *N. vitripennis* as the instructor of sex determination via the activation of *tra* expression and autoregulation, which results in female development. However, we cannot rule out the presence of *wom* in *E. suzannae* as this gene in *N. vitripennis* is mainly transcribed in diploid (female) embryos prior to 7 h post oviposition and is not expressed in adults, which we sampled for our transcriptome assemblies. We also did not find homologs of *complementary sex determiner (csd)*, the instructor of sex determination in *Apis mellifera*.

Sex determination in the Chalcidoidea has been a matter of some speculation [57]. However, the presence or absence of these transcripts provides insights into the nature of sex determination and development in *E. suzannae* and lays the foundation for understanding how the mechanisms of sexual development in *Encarsia* may interface with the reproductive manipulation of *Cardinium*. Particularly applicable are cases of symbiont-induced parthenogenesis, in which unfertilized eggs are diploidized by the endosymbiont and biological females are produced [13, 58].

Furthermore, the identification of many transcripts harboring coding sequences annotated as putative venom proteins in both the male and female *E. suzannae* transcriptomes is notable as these are believed to be important mechanisms used by female parasitoid wasps to enhance the survivability of their offspring. Venom proteins are diverse and predicted to have a variety of impacts on the host undergoing parasitism, including immune system suppression, developmental arrest, lipid accumulation, and apoptosis [59]. In the case of *E. suzannae*, parasitism causes the whitefly host to undergo developmental arrest during a late nymphal stage. As arrest occurs regardless of wasp larva survival, it is possible that it is induced by venom injected into the whitefly during oviposition [15]. The presence of predicted proteins annotated as venom proteins in the male *E. suzannae* assembly is intriguing since only female wasps host feed and lay eggs into their host, while adult males seemingly have no need to express venom genes. It is unclear whether these putative proteins are actually venom genes expressed in male *E. suzannae* or if they were annotated as such due to the presence of domains similar to those found in venom proteins. Regardless, detecting putative venom proteins in *E. suzannae* provides more insight into how these wasps effectively parasitize their hosts. However, it should be noted that reliable identifications of venom proteins require additional experimental verifications.

## Transcriptome comparisons

As stated above, the only other publicly available transcriptome of an *Encarsia* species belongs to *E. formosa* [26]; thus, limited comparisons can be made within this genus. An overview of all currently known *Encarsia* transcriptomes is shown in Table 2. Compared to the *E. formosa* transcriptome assembly, the male and female *E. suzannae* assemblies were generated from more initial reads and produced more pre-filtering transcripts, meaning they could be subject to more stringent transcript filtering than the *E. formosa* assembly. While the *E. formosa* assembly underwent limited post-assembly contaminant filtering, the *E. suzannae* assemblies utilized additional measures to (1) limit potential nonsense, low-abundance, and redundant transcripts through post-assembly filtering and processing, and (2) eliminate as many contaminants as possible prior to the assembly via mapping-and-removal. Furthermore, the publicly available *E. formosa* assembly consists of full-length mRNA transcripts instead of coding sequences, as seen in the *E. suzannae*



**Table 4.** Software and version specifications.

Software	Usage	Version	Reference(s)
BBTools	BBduk for read trimming; BBDuk for read mapping	37.36	[32]
FastQC	Visualization of sequence quality	0.11.9	[33]
SAMtools (RRID:SCR_005227)	.bam file manipulation	1.10	[62]
Trinity	<i>De novo</i> transcriptome assembly	2.6.6	[37]
kallisto	Transcript abundance estimation	0.46.2	[38]
TransDecoder	Prediction of coding sequences	5.5.0	[41]
CD-HIT	Clustering similar protein sequences	4.6.8	[42, 43]
BUSCO	Assessing assembly completeness	5.3.2	[45]
eggNOG-mapper	Annotation of assembled proteins	2.1.6	[48, 49]
Blast+	Annotation of assembled proteins	2.11.0	[51]
Diamond	Annotation of assembled proteins	2.0.4	[50]
OrthoVenn2	Orthologous protein group clustering and visualization	N/A	[60]

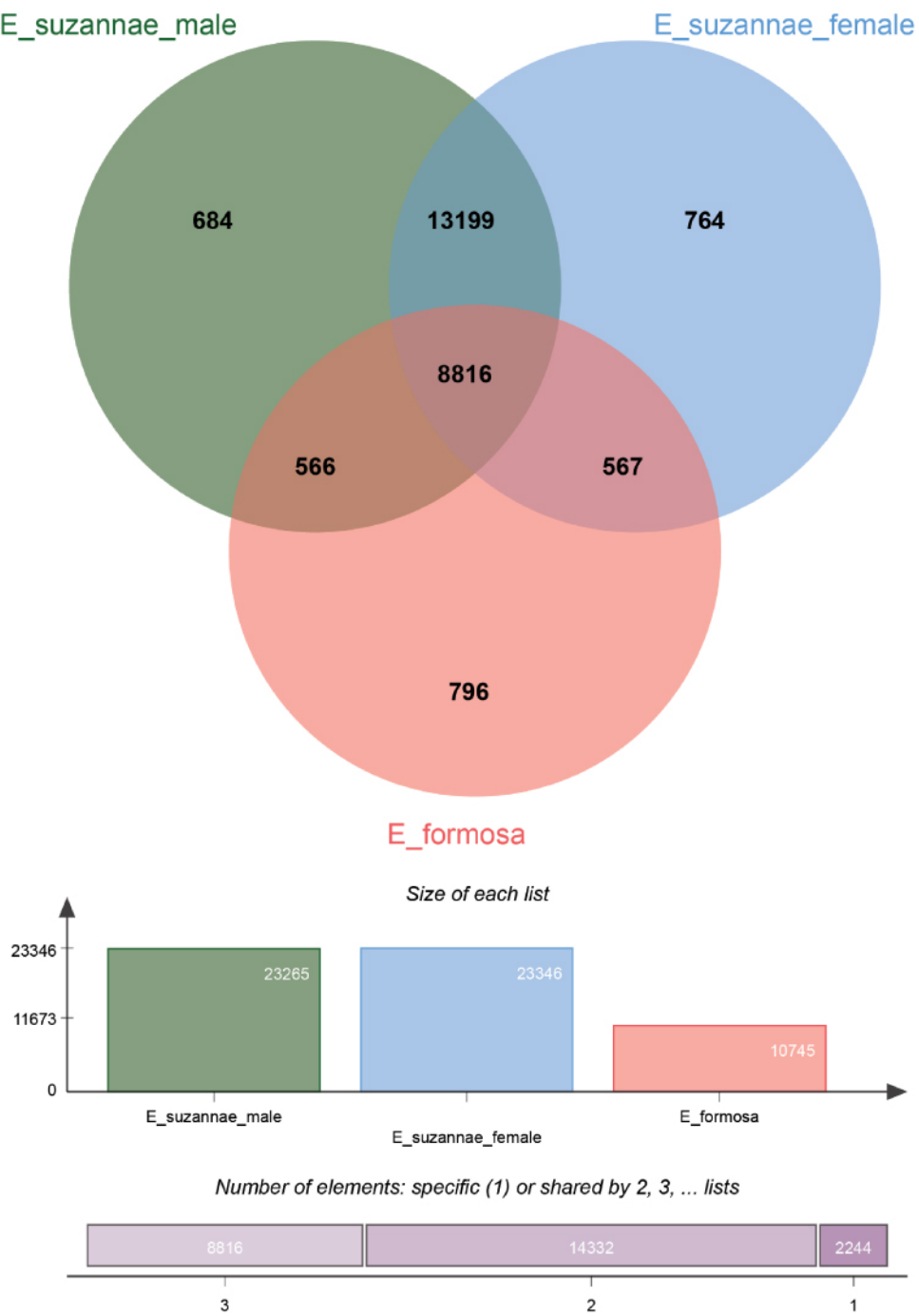
assemblies [25]. After running TransDecoder on the *E. formosa* transcripts, only 27,161 coding sequences were predicted using a minimum length of 50 amino acids. This indicates that the female (45,986) and male (54,762) *E. suzannae* assemblies contain twice or nearly twice as many coding sequences compared to the *E. formosa* assembly, even though the *E. formosa* coding sequences were predicted with a shorter minimum protein size than *E. suzannae*.

Finally, OrthoVenn2 (RRID:SCR\_022504) was used to determine the orthologous groups between the predicted proteins in both the *E. suzannae* assemblies presented in this paper and the *E. formosa* assembly published elsewhere [26, 60]. Using the default settings and an e-value cutoff of  $1 \times 10^{-5}$ , 8816 orthologs were found to be shared across all three transcriptomes, and a total of 22,015 orthologous groups were shared between male and female *E. suzannae* out of the total of 23,265 and 23,346 clusters, respectively (see Figure 1). These results indicate a high degree of similarity between the different sex assemblies while showing the presence of over one thousand sex-specific protein clusters. It is also striking that the female and male *E. suzannae* transcriptomes are equally similar to the *E. formosa* transcriptome, although *E. formosa* exists as an asexual species consisting of nearly all females (due to the presence of parthenogenesis-inducing *Wolbachia*), and its transcriptome therefore only reflects female individuals [61].

## CONCLUSION AND RE-USE POTENTIAL

We are confident that our assemblies are among the purest possible transcriptome representations of *E. suzannae* that can be obtained with the currently available data and tools for assembly and filtering (for a list of all software names and versions utilized in this study, see Table 4). This study is also one of the first to present sex-specific transcriptome assemblies of a single insect species. In an organism such as *E. suzannae* – where males and females develop within different hosts, are impacted differently by endosymbiotic bacteria, and exhibit distinct behaviors – it is highly valuable to have the availability of a reference database for both sexes to ensure more accurate studies when wasps of only one sex are used. Furthermore, these assemblies greatly expand our host knowledge of the *Cardinium* cEper1 CI system and pave the way for future studies exploring how this endosymbiont interacts with its *E. suzannae* host in causing CI. We also believe that these data will be a valuable reference when studying the diverse members of the ecologically important genus *Encarsia* and other chalcidoid parasitic wasps, many of which have interesting biology and potential as pest biological control agents.





**Figure 1.** Orthologous groups between the *E. formosa* females and the male and female *E. suzannae* transcriptomes. The above figure shows an OrthoVenn2 diagram of the orthologous groups between the *E. formosa* females and the male and female *E. suzannae* ( $e\text{-value} = 1 \times 10^{-5}$ ) [60]. TransDecoder, using a minimum amino acid length of 50, was run on the *E. formosa* assemblies to obtain the coding sequences. The resulting peptide sequence output (27,161 sequences) was tested against the predicted proteins from the male and female *E. suzannae* transcriptomes. The top Venn diagram depicts the number of orthologous protein clusters shared between the three transcriptomes. The middle bar graph depicts the total number of orthologous clusters present for each transcriptome. Lastly, the bottom graph shows (left to right) the number of clusters that were shared by all three transcriptomes, by any two transcriptomes, or were unique to one of the three assemblies.

## DATA AVAILABILITY

All the raw sequencing data and the final assemblies from this study are publicly available. The *E. suzannae* female and male raw read data and unannotated assemblies were submitted to the NCBI's Sequence Read Archive (SRA) and Transcriptome Shotgun Assembly (TSA) databases under the BioProjects [PRJNA737477](#) for male *E. suzannae* and [PRJNA737478](#) for female *E. suzannae*. Detailed annotation information from multiple sources is provided alongside the annotated female and male assemblies in FASTA format in the GigaDB repository [63].

## LIST OF ABBREVIATIONS

BUSCO: Benchmarking Universal Single-Copy Orthologs; CI: Cytoplasmic incompatibility; KEGG: Kyoto Encyclopedia of Genes and Genomes; NCBI: National Center for Biotechnology Information; nr: non-redundant; SRA: Sequence Read Archive; TSA: Transcriptome Shotgun Assembly; VBCF: Vienna BioCenter Core Facilities.

## COMPETING INTERESTS

The authors declare no competing interests.

## FUNDING SOURCES

This work was supported by the National Science Foundation grant no. IOS-2002987 and IOS-202934 to SSE, MSH, and Manuel Kleiner (North Carolina State University).

## AUTHORS' CONTRIBUTIONS

SSE and MSH conceived the experiments and provided supervision. DLS and SSE developed the analysis pipeline. ES, CS, and SEK performed the experiments. DLS analyzed and visualized the data and wrote the draft manuscript. All authors wrote and edited the manuscript. SSE and MSH obtained funding.

## ETHICS APPROVAL

The authors declare that ethical approval was not required for this type of research.

## ACKNOWLEDGEMENTS

We would like to thank Manuel Kleiner for his constructive comments and advice regarding the filtering and assembly pipeline.

## REFERENCES

- 1 Penz T, Schmitz-Esser S, Kelly SE et al. Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium hertigii*. *PLoS Genet.*, 2012; 8(10): e1003012. doi:[10.1371/journal.pgen.1003012](#).
- 2 Heimpel GE, de Boer JG. Sex determination in the hymenoptera. *Annu. Rev. Entomol.*, 2008; 53: 209–230. doi:[10.1146/annurev.ento.53.103106.093441](#).
- 3 Mann E, Stouthamer CM, Kelly SE et al. Transcriptome sequencing reveals novel candidate genes for *Cardinium hertigii*-caused cytoplasmic incompatibility and host-cell interaction. *mSystems*, 2017; 2(6): e00141-17. doi:[10.1128/mSystems.00141-17](#).
- 4 Sani I, Ismail SI, Abdullah S et al. A review of the biology and control of whitefly, *Bemisia tabaci* (Hemiptera: Aleyrodidae), with special reference to biological control using entomopathogenic fungi. *Insects*, 2020; 11(9): 619. doi:[10.3390/insects11090619](#).

- 5 Oliveira MRV, Henneberry TJ, Anderson P. History, current status, and collaborative research projects for *Bemisia tabaci*. *Crop Prot.*, 2001; 20(9): 709–723. doi:10.1016/S0261-2194(01)00108-9.
- 6 Gerling D, Alomar Ò, Arnó J. Biological control of *Bemisia tabaci* using predators and parasitoids. *Crop Prot.*, 2001; 20(9): 779–799. doi:10.1016/S0261-2194(01)00111-9.
- 7 Hoddle MS, Van Driesche RG, Sanderson JP. Biology and use of the whitefly parasitoid *Encarsia formosa*. *Annu. Rev. Entomol.*, 1998; 43: 645–669. doi:10.1146/annurev.ento.43.1.645.
- 8 Pickett CH, Ball JC, Casanave KC et al. Establishment of the ash whitefly parasitoid *Encarsia inaron* (Walker) and its economic benefit to ornamental street trees in California. *Biol. Control*, 1996; 6(2): 260–272. doi:10.1006/bcon.1996.0033.
- 9 Hunter MS, Woolley JB. Evolution and behavioral ecology of heteronomous aphelinid parasitoids. *Annu. Rev. Entomol.*, 2001; 46: 251–290. doi:10.1146/annurev.ento.46.1.251.
- 10 Heraty JM, Polaszek A, Schauff ME. Systematics and biology of *Encarsia*. In: Classical biological control of *Bemisia tabaci* in the United States—A review of interagency research and implementation. Dordrecht: Springer, 2008; pp. 71–87. doi:10.1007/978-1-4020-6740-2\_4.
- 11 Doremus MR, Hunter MS. The saboteur's tools: Common mechanistic themes across manipulative symbioses. *Adv. Insect Physiol.*, 2020; 58: 317–353. doi:10.1016/bs.aiip.2020.03.003.
- 12 Kenyon SG, Hunter MS. Manipulation of oviposition choice of the parasitoid wasp, *Encarsia pergandiella*, by the endosymbiotic bacterium *Cardinium*. *J. Evol. Biol.*, 2007; 20(2): 707–716. doi:10.1111/j.1420-9101.2006.01238.x.
- 13 Zchori-Fein E, Gottlieb Y, Kelly SE et al. A newly discovered bacterium associated with parthenogenesis and a change in host selection behavior in parasitoid wasps. *Proc. Natl. Acad. Sci. USA*, 2001; 98(22): 12555–12560. doi:10.1073/pnas.221467498.
- 14 Zchori-Fein E, Perlman SJ, Kelly SE et al. Characterization of a 'Bacteroidetes' symbiont in *Encarsia* wasps (Hymenoptera: Aphelinidae): proposal of 'Candidatus *Cardinium hertigii*'. *Int. J. Syst. Evol. Microbiol.*, 2004; 54(3): 961–968. doi:10.1099/ijs.0.02957-0.
- 15 Hunter MS, Perlman SJ, Kelly SE. A bacterial symbiont in the *Bacteroidetes* induces cytoplasmic incompatibility in the parasitoid wasp *Encarsia pergandiella*. *Proc. Biol. Sci.*, 2003; 270(1529): 2185–2190. doi:10.1098/rspb.2003.2475.
- 16 Ma WJ, Vavre F, Beukeboom LW. Manipulation of arthropod sex determination by endosymbionts: diversity and molecular mechanisms. *Sex Dev.*, 2014; 8(1–3): 59–73. doi:10.1159/000357024.
- 17 Irina G, Boris A. Reproductive parasitism in insects. The interaction of host and bacteria. *Biol. Commun.*, 2021; 66(1): 17–27. doi:10.21638/spbu03.2021.103.
- 18 Shropshire JD, Bordenstein SR. Speciation by symbiosis: the microbiome and behavior. *mBio*, 2016; 7(2): e01785. doi:10.1128/mBio.01785-15.
- 19 Gebiola M, Kelly SE, Hammerstein P et al. "Darwin's corollary" and cytoplasmic incompatibility induced by *Cardinium* may contribute to speciation in *Encarsia* wasps (Hymenoptera: Aphelinidae). *Evolution*, 2016; 70(11): 2447–2458. doi:10.1111/evo.13037.
- 20 Shoemaker DD, Katju V, Jaenike J. *Wolbachia* and the evolution of reproductive isolation between *Drosophila recens* and *Drosophila subquinaria*. *Evolution*, 1999; 53(4): 1157–1164. doi:10.1111/j.1558-5646.1999.tb04529.x.
- 21 Bordenstein SR, O'Hara FP, Werren JH. *Wolbachia*-induced incompatibility precedes other hybrid incompatibilities in *Nasonia*. *Nature*, 2001; 409(6821): 707–710. doi:10.1038/35055543.
- 22 Zheng X, Zhang D, Li Y et al. Incompatible and sterile insect techniques combined eliminate mosquitoes. *Nature*, 2019; 572(7767): 56–61. doi:10.1038/s41586-019-1407-9.
- 23 Li TP, Zhou CY, Zha SS et al. Stable establishment of *Cardinium spp.* in the brown planthopper *Nilaparvata lugens* despite decreased host fitness. *Appl. Environ. Microbiol.*, 2020; 86(4): e02509-19. doi:10.1128/AEM.02509-19.
- 24 Ryan PA, Turley AP, Wilson G et al. Establishment of wMel *Wolbachia* in *Aedes aegypti* mosquitoes and reduction of local dengue transmission in Cairns and surrounding locations in northern Queensland, Australia. *Gates Open Res.*, 2020; 3: 1547. doi:10.12688/gatesopenres.13061.2.
- 25 Peters RS, Niehuis O, Gunkel S et al. Transcriptome sequence-based phylogeny of chalcidoid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. *Mol. Phylogenet. Evol.*, 2018; 120: 286–296. doi:10.1016/j.ympev.2017.12.005.
- 26 NCBI Transcriptome Shotgun Assembly. 2017; <https://identifiers.org/nucleotide:GBVN000000000.1>.

- 27 Manzari S, Polaszek A, Belshaw R et al. Morphometric and molecular analysis of the *Encarsia inaron* species-group (Hymenoptera: Aphelinidae), parasitoids of whiteflies (Hemiptera: Aleyrodidae). *Bull. Entomol. Res.*, 2002; 92(2): 165–176. doi:10.1079/BER2001144.
- 28 Schmidt S, Driver F, De Barro P. The phylogenetic characteristics of three different 28S rRNA gene regions in *Encarsia* (Insecta, Hymenoptera, Aphelinidae). *Org. Divers. Evol.*, 2006; 6(2): 127–139. doi:10.1016/j.ode.2005.07.002.
- 29 Babcock CS, Heraty JM, De Barro PJ et al. Preliminary phylogeny of *Encarsia* Forster (Hymenoptera: Aphelinidae) based on morphology and 28S rDNA. *Mol. Phylogenet. Evol.*, 2001; 18(2): 306–323. doi:10.1006/mpev.2000.0875.
- 30 Gebiola M, Monti MM, Johnson RC et al. A revision of the *Encarsia pergandiella* species complex (Hymenoptera: Aphelinidae) shows cryptic diversity in parasitoids of whitefly pests. *Syst. Entomol.*, 2017; 42(1): 31–59. doi:10.1111/syen.12187.
- 31 Vienna BioCenter Core Facilities. <https://www.viennabiocenter.org/vbcbf/>. Accessed 5 April 2022.
- 32 Bushnell B. BBMap. <https://sourceforge.net/projects/bbmap/>. Accessed 5 April 2022.
- 33 Andrews S. FastQC: a quality control tool for high throughput sequence data. Version 0.11.9. 2010; <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 5 April 2022.
- 34 Andreason SA, Shelby EA, Moss JB et al. Whitefly endosymbionts: Biology, evolution, and plant virus interactions. *Insects*, 2020; 11(11): 775. doi:10.3390/insects11110775.
- 35 Himler AG, Adachi-Hagimori T, Bergen JE et al. Rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias. *Science*, 2011; 332(6026): 254–256. doi:10.1126/science.1199410.
- 36 Chiel E, Kelly SE, Harris AM et al. Characteristics, phenotype, and transmission of *Wolbachia* in the sweet potato whitefly, *Bemisia tabaci* (Hemiptera: Aleyrodidae), and its parasitoid *Eretmocerus sp. nr. emiratus* (Hymenoptera: Aphelinidae). *Environ. Entomol.*, 2014; 43(2): 353–362. doi:10.1603/EN13286.
- 37 Grabherr MG, Haas BJ, Yassour M et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 2011; 29(7): 644–652. doi:10.1038/nbt.1883.
- 38 Bray NL, Pimentel H, Melsted P et al. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 2016; 34(5): 525–527. doi:10.1038/nbt.3519.
- 39 Kerkvliet J, de Fouchier A, van Wijk M et al. The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras. *Ecol. Evol.*, 2019; 9(18): 10513–10521. doi:10.1002/ece3.5571.
- 40 Sim SB, Calla B, Hall B et al. Reconstructing a comprehensive transcriptome assembly of a white-pupal translocated strain of the pest fruit fly *Bactrocera cucurbitae*. *GigaScience*, 2015; 4(1): 14. doi:10.1186/s13742-015-0053-x.
- 41 Haas BJ. TransDecoder. Version 5.5.0. <https://github.com/TransDecoder/TransDecoder>. Accessed 5 April 2022.
- 42 Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 2012; 28(23): 3150–3152. doi:10.1093/bioinformatics/bts565.
- 43 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 2006; 22(13): 1658–1659. doi:10.1093/bioinformatics/btl158.
- 44 Jiang L, Schlesinger F, Davis CA et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, 2011; 21(9): 1543–1551. doi:10.1101/gr.121095.111.
- 45 Manni M, Berkeley MR, Seppey M et al. 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.
- 46 BUSCO Index of /v4/data/lineages/. <https://busco-data.ezlab.org/v4/data/lineages/>. Accessed 5 April 2022.
- 47 Gerling D. Studies with whitefly parasites of Southern California: I. *Encarsia pergandiella* Howard (Hymenoptera: Aphelinidae). *Can. Entomol.*, 1966; 98(7): 707–724. doi:10.4039/Ent98707-7.
- 48 Huerta-Cepas J, Szklarczyk D, Heller D et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, 2019; 47(D1): D309–D314. doi:10.1093/nar/gky1085.
- 49 Cantalapiedra CP, Hernández-Plaza A, Letunic I et al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.*, 2021; 38(12): 5825–5829. doi:10.1093/molbev/msab293.



- 50 Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, 2021; **18**(4): 366–368. doi:10.1038/s41592-021-01101-x.
- 51 Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. *BMC Bioinform.*, 2009; **10**: 421. doi:10.1186/1471-2105-10-421.
- 52 Altschul SF, Madden TL, Schaffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 1997; **25**(17): 3389–3402. doi:10.1093/nar/25.17.3389.
- 53 Penalva LO, Sanchez L. RNA binding protein sex-lethal (Sxl) and control of *Drosophila* sex determination and dosage compensation. *Microbiol. Mol. Biol. Rev.*, 2003; **67**(3): 343–359. table of contents. doi:10.1128/MMBR.67.3.343-359.2003.
- 54 Rideout EJ, Dornan AJ, Neville MC et al. Control of sexual differentiation and behavior by the doublesex gene in *Drosophila melanogaster*. *Nat. Neurosci.*, 2010; **13**(4): 458–466. doi:10.1038/nn.2515.
- 55 Yamamoto D, Kohatsu S. What does the fruitless gene tell us about nature vs. nurture in the sex life of *Drosophila*? *Fly (Austin)*, 2017; **11**(2): 139–147. doi:10.1080/19336934.2016.1263778.
- 56 Zou Y, Geuverink E, Beukeboom LW et al. A chimeric gene paternally instructs female sex determination in the haplodiploid wasp *Nasonia*. *Science*, 2020; **370**(6520): 1115–1118. doi:10.1126/science.abb8949.
- 57 Beukeboom LW, van de Zande L. Genetics of sex determination in the haplodiploid wasp *Nasonia vitripennis* (Hymenoptera: Chalcidoidea). *J. Genet.*, 2010; **89**(3): 333–339. doi:10.1007/s12041-010-0045-7.
- 58 Giorgini M, Monti MM, Caprio E et al. Feminization and the collapse of haplodiploidy in an asexual parasitoid wasp harboring the bacterial symbiont *Cardinium*. *Heredity (Edinb)*, 2009; **102**(4): 365–371. doi:10.1038/hdy.2008.135.
- 59 Danneels EL, Rivers DB, de Graaf DC. Venom proteins of the parasitoid wasp *Nasonia vitripennis*: recent discovery of an untapped pharmacopee. *Toxins (Basel)*, 2010; **2**(4): 494–516. doi:10.3390/toxins2040494.
- 60 Xu L, Dong Z, Fang L et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.*, 2019; **47**(W1): W52–W58. doi:10.1093/nar/gkz333.
- 61 Zchori-Fein E, Roush RT, Hunter MS. Male production induced by antibiotic treatment in *Encarsia formosa* (Hymenoptera: Aphelinidae), an asexual species. *Experientia*, 1992; **48**(1): 102–105. doi:10.1007/BF01923619.
- 62 Danecek P, Bonfield JK, Liddle J et al. Twelve years of SAMtools and BCFtools. *GigaScience*, 2021; **10**(2): giab008. doi:10.1093/gigascience/giab008.
- 63 Schultz DL, Selberherr E, Stouthamer CM et al. Supporting data for “Sex-based de novo transcriptome assemblies of the parasitoid wasp *Encarsia suzannae*, a host of the manipulative heritable symbiont *Cardinium hertigii*”. *GigaScience Database*, 2022; <http://dx.doi.org/10.5524/102258>.