Hybrid Genome Assembly and Evidence-Based Annotation of the Egg Parasitoid and Biological Control Agent *Trichogramma brassicae*

Kim B. Ferguson,*,1 Tore Kursch-Metz,†,‡ Eveline C. Verhulst,§ and Bart A. Pannebakker*

*Wageningen University, Laboratory of Genetics, Wageningen, The Netherlands, †Technische Universität Darmstadt, Department of Biology, Darmstadt, Germany, ‡AMW Nützlinge GmbH, Pfungstadt, Germany, and §Wageningen University, Laboratory of Entomology, Wageningen, The Netherlands

ORCID IDs: 0000-0001-6483-1936 (K.B.F.); 0000-0002-4781-535X (E.C.V.); 0000-0001-8503-3896 (B.A.P.)

ABSTRACT *Trichogramma brassicae* (Bezdenko) are egg parasitoids that are used throughout the world as biological control agents and in laboratories as model species. Despite this ubiquity, few genetic resources exist beyond COI, ITS2, and RAPD markers. Aided by a *Wolbachia* infection, a wild-caught strain from Germany was reared for low heterozygosity and sequenced in a hybrid *de novo* strategy, after which several assembling strategies were evaluated. The best assembly, derived from a DBG2OLC-based pipeline, yielded a genome of 235 Mbp made up of 1,572 contigs with an N50 of 556,663 bp. Following a rigorous *ab initio*-, homology-, and evidence-based annotation, 16,905 genes were annotated and functionally described. As an example of the utility of the genome, a simple ortholog cluster analysis was performed with sister species *T. pretiosum*, revealing over 6000 shared clusters and under 400 clusters unique to each species. The genome and transcriptome presented here provides an essential resource for comparative genomics of the commercially relevant genus *Trichogramma*, but also for research into molecular evolution, ecology, and breeding of *T. brassicae*.

The chalcidoid *Trichogramma brassicae* (Bezdenko) (Hymenoptera: Trichogrammatidae) is a minute parasitoid wasp (~0.5 mm in length) that develops within the eggs of other insects (Smith 1996). For over 50 years, it has been in use world-wide as a biological control agent as many lepidopteran pests of different crops are suitable hosts (Polaszek 2009). The most common application of *T. brassicae* in Europe is against *Ostrinia nubilalis* (Hubner) (Lepidoptera: Pyralidae), the European corn borer. For example, in 2003 alone, over 11,000 hectares (ha) of maize in Germany was treated with *T. brassicae* (Zimmermann 2004). It is also released

against lepidopteran pests in spinach fields as well as in greenhouses (*e.g.*, tomato, pepper, and cucumber) (Klug and Meyhöfer 2009). With its wide application in biological control, *T. brassicae* is a well-studied species. Field trials have been conducted on several aspects, such as host location and dispersal behavior (Suverkropp *et al.* 2009, 2010), overwintering ability (Babendreier *et al.* 2003), while other biological control related studies considered issues related to low temperature storage (Lessard and Boivin 2013), reaction to insecticides (Liu and Zhang 2012; Delpuech and Delahaye 2013; Ghorbani *et al.* 2016; Jamshidnia *et al.* 2018; Thubru *et al.* 2018), or risk assessment (Kuske *et al.* 2004).

Next to its application as a biological control agent, this tiny parasitoid has been used in other research, both in genetic studies (Wajnberg 1993; Laurent *et al.* 1998; Cruaud *et al.* 2018) and ecological studies (Huigens *et al.* 2009; Fatouros and Huigens 2012; Cusumano *et al.* 2015). In addition, several initiatives investigate the infection of *T. brassicae* with *Wolbachia* bacteria (Poorjavad *et al.* 2012; Ivezić *et al.* 2018) and the consequences of such an infection (Farrokhi *et al.* 2010; Poorjavad *et al.* 2018; Rahimi-Kaldeh *et al.* 2018). As *T. brassicae* is a cryptic species with several other congenerics, misidentification and misclassification is a known issue (Polaszek 2009). In response, molecular identification of trichogrammatids is well studied

and established (Stouthamer *et al.* 1999; Sumer *et al.* 2009; Rugman-Jones and Stouthamer 2017; Ivezić *et al.* 2018). Recently, several restriction-site associated DNA sequencing (RADseq) libraries were constructed from single *T. brassicae* wasps to aide in resolving the aforementioned phylogenetic issues within *Trichogramma* (Cruaud *et al.* 2018). Otherwise, the genomics of *T. brassicae* have largely been neglected even though a well annotated genome would allow researchers and biological control practitioners access to a wealth of information and open new avenues for comparative genomics and transcriptomics for evolutionary, ecological, and applied research.

Here, we report the whole-genome sequencing and annotation of a *T. brassicae* strain infected by *Wolbachia* that had thelytokous reproduction, in which females arise from unfertilized eggs. A hybrid *de novo* sequencing strategy was chosen to address two common issues: we used long PacBio Sequel reads to bridge the large segments of repetitive sequences often found in Hymenoptera, while countering the error bias of long read technology with the accuracy of Illumina short reads. A similar strategy was recently applied to improve the *Apis mellifera* genome, where the long PacBio reads were the backbone that boosted the overall contiguity of the genome, alongside the incorporation of repetitive regions (Wallberg *et al.* 2019).

In this report, we present the hybrid *de novo* genome of *T. brassicae*. Three different assemblers were evaluated, and the most complete genome assembly was used for decontamination and *ab initio*-, homology-, and evidence-based annotation. The resulting annotation was functionally described using gene ontology analysis. Finally, a heterozygosity comparison and simple ortholog cluster analysis with the congeneric *T. pretiosum* was performed, which can be considered a starting point for future comparative genomics of the commercially important genus *Trichogramma*.

MATERIALS AND METHODS

Species origin and description
Individuals of *Trichogramma brassicae* were acquired by AMW Nützlinge GmbH (Pfungstadt, Germany). The strain was baited in May 2013 in an apple orchard near Eberstadt, Germany. The orchard was surrounded by blackberry hedges, forest, and other orchards. For baiting, the eggs of *Sitotroga cerealella* (Olivier) (Lepidoptera: Gelechiidae) (Mega Corn Ltd., Bulgaria) were glued on paper cards (AMW Nützlinge GmbH, Germany), usually used for releasing *Trichogramma* sp. in corn fields and households. These cards were placed directly into the trees, approximately two meters above ground. After five days in the field, baiting cards were collected and incubated together at 25°. Following emergence, individuals were kept together, offered *S. cerealella* eggs, and reared in a climate chamber (27 ± 2°, L:D = 24:0h for four days, then transferred to 16 ± 2°, L:D = 0:24h until emergence).

In 2016, the offspring of twenty isolated females were transferred to Wageningen University (The Netherlands) to be reared for low heterozygosity. The resulting offspring were reared in a single general population on irradiated *Ephestia kuehniella* (Zeller) (Lepidoptera: Pyralidae) eggs as factitious hosts under laboratory conditions in a climate chamber (20 ± 5°, RH 50 ± 5%, L:D = 12:12 h). *Wolbachia* presence was determined following the PCR amplification protocol of Zhou *et al.* 1998 in a presence/absence assessment with known positive and negative control samples (Zhou *et al.* 1998). Natural *Wolbachia* infections have previously been detected in Iranian populations of *T. brassicae* (Farrokhi *et al.* 2010), but none of the Eurasian populations have been known to support this symbiosis (Stouthamer 1997; Stouthamer and Huigens 2003).

Isofemale line
Following confirmation of *Wolbachia* infection (Supplementary materials S1.1.1), a single female from the general population was isolated (generation 0, G0), and given eggs *ad libitum*. In the resulting generation (G1), unmated females were isolated and reared with eggs *ad libitum*. Offspring of the initial isolations G0 and G1 were confirmed to be entirely female, suggesting thelytokous parthenogenetic reproduction. Combined with isolating single females, this maximizes genetic similarity of the following generation (G2) of these G1 females. One of these G2 strains, S301, was reared in large population sizes for multiple generations over the period of one year. By the time of collection for sequencing, both the S301 and general population no longer harbored *Wolbachia* at detectable levels (Supplementary materials S1.1.2).

gDNA extraction
Three separate extractions were prepared in 1.5 mL safelock tubes with each several hundred *Trichogramma brassicae*. The tubes were frozen in liquid nitrogen with approximately six 1-mm glass beads and shaken for 30 s in a Silamat S6 shaker (Ivoclar Vivadent, Schaan, Liechtenstein). gDNA was then extracted using the Qiagen MagAttract Kit (Qiagen, Hilden, Germany). Following an overnight lysis step with Buffer ATL and proteinase K at 56°, extraction was performed according to the MagAttract Kit protocol. Elutions were performed in two steps with Buffer AE (Tris-EDTA) each time (first 60 μL, then 40 μL), yielding 100 μL. The two extractions yielding the largest amount of gDNA (5.49 μg and 8.24 μg) were combined for long-read sequencing, while the remaining extraction (1.67 μg) was used for short-read sequencing. gDNA concentration was measured with an Invitrogen Qubit 2.0 fluorometer using the dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, USA) while fragment length was confirmed on gel.

Library preparation and sequencing
Sequence coverage was calculated using the previously established genome size estimate for *T. brassicae* of 246 Mbp (Johnston *et al.* 2004). Library preparation and sequencing was performed by Novogene Bioinformatics Technology Co., Ltd., (Beijing, China). For Illumina sequencing, gDNA was used to construct one paired-end (PE) library according to the standard protocol for Illumina with an average insert size of 150 bp and was sequenced using an Illumina HiSeq 2000 (Illumina, San Diego, USA). For Single Molecule Real Time (SMRT) sequencing, gDNA was selected for optimal size using a Blue Pippin size selection system (Sage Science, Beverley, USA) following a standard library preparation. The library was then sequenced on a PacBio Sequel (Pacific Biosciences, Menlo Park, USA) with 16 SMRT cells.

Assembly and decontamination
Prior to assembly, Illumina reads were assessed for quality using FASTQC (Andrews *et al.* 2015), then trimmed for quality in CLC Genomics Workbench 11 using default settings (Qiagen). Trimmed Illumina reads were paired for subsequent analysis.

In order to achieve the best possible assembly, three assembly pipelines were evaluated: one for PacBio-only reads and two hybrid assemblers. The PacBio-only were assembled with Canu (v1.6) with modifications based on PacBio Sequel reads (correctedErrorRate = 0.085 corMhapSensitivity = normal alongside corMhapSensivity = normal) (Koren *et al.* 2017). This is assembly version v1.0 in the subsequent discussion.

The first hybrid assembly pipeline using both long and short sequencing read sets was SPAdes (v3.11.1) (Bankevich *et al.* 2012). The SPAdes genome toolkit supports hybrid assemblies with the hybridSPAdes algorithm (Antipov *et al.* 2016). Three iterations of the SPAdes pipeline were run with varying k-mer sizes resulting in three different assembly versions: 21, 33, 55 (default, v2.1); k-mer sizes 21, 33, 55, 77 (v2.2); and a single k-mer size of 127 (v2.3).

The second hybrid assembly pipeline was DBG2OLC (Ye *et al.* 2016). The DBG2OLC pipeline can be readily tweaked with other programs depending on the job (Chakraborty *et al.* 2016). Following the DBG2OLC pipeline, de Bruijn graph contigs were generated using SparseAssembler using default settings and setting the expected genome size to 750 Mbp to ensure a genome size output that is unrestricted (Ye *et al.* 2012). Contigs were transformed into read overlaps using DBG2OLC with settings suggested for large genomes and PacBio Sequel data (k = 17; AdaptiveTh = 0.01; KmerCovTh = 2; MinOverlap = 20; RemoveChimera = 1), according to the DBG2OLC manual (https://github.com/yechengxi/DBG2OLC). This creates an assembly backbone of the best overlaps between the short-read de Bruijn contigs and the long reads. minimap2 (v2.9) and Racon (v1.0.2) were used for consensus calling remaining overlaps to the assembly backbone (Vaser *et al.* 2017; Li 2018). The resulting consensus assembly was polished twice using the Illumina reads with Pilon (v1.22) (Walker *et al.* 2014). This final assembly is v3.0 in subsequent discussion.

The best of the five assemblies generated was determined on the basis of N50, genome size, and completeness (Table 1). Genome statistics such as N50, number of contig, and genome size were determined using Quast (Gurevich *et al.* 2013). Assembly completeness was assessed using BUSCO (v3.0.2) with the insect_odb9 ortholog set and the fly training parameter (Simão *et al.* 2015). Based on these characteristics, the decision was made to move forward with assembly v3.0, which was then decontaminated for microbial sequences using NCBI BLASTn (v2.2.31+) against the NCBI nucleotide collection (nr).

Wolbachia contamination

Two contigs contained a large amount of *Wolbachia* content, with over 80% of the scaffold containing material with 75% or higher homology to *Wolbachia*. These contigs were assessed for homology against the NCBI nucleotide collection (nr) and removed from the assembly (Supplementary material S1.2). Post-decontamination, the assembly is referred to as v3.5.

RNA extraction, library construction, and sequencing

T. brassicae wasps from the S301 line were collected for RNAseq for evidence-based annotation. Hundreds of adult individuals (male and female) were collected and stored at -80°. For RNA extraction, samples were frozen in liquid nitrogen in a single 1.5 mL safelock tube with approximately six 1-mm glass beads and shaken for 30 s in a

Silamat S6 shaker (Ivoclar Vivadent). The RNeasy Blood and Tissue Kit (Qiagen) was used according to manufacturer's instructions, and final column elution was achieved using 60 µL sterilized water. The sample was measured for quality and RNA quantity using an Invitrogen Qubit 2.0 fluorometer and the RNA BR Assay Kit (Thermo Fisher Scientific). The RNA sample was then processed by Novogene Bioinformatics Technology Co., Ltd., (Beijing, China) using poly(A) selection followed by cDNA synthesis with random hexamers and library construction with an insert size of 300 bp. Paired-end sequencing was performed on an Illumina HiSeq 4000 according to manufacturer's instruction. Quality filtering was applied to remove adapters, reads with more than 10% undetermined bases, and reads of low quality for more than 50% of the total bases (Qscore less than or equal to 5).

Ab initio gene finding, transcriptome assembly, and annotation

For the *ab initio* gene finding, a training set was established using the reference genome of *Drosophila melanogaster* (Meigen) (Diptera: Drosophilidae) (Genbank: GCA_000001215.4; Release 6 plus ISO1 MT) (Hoskins *et al.* 2015) and the associated annotation (Adams *et al.* 2000; Dos Santos *et al.* 2015). The training parameters were used by GlimmerHMM (v3.0.1) for gene finding in the *T. brassicae* genome assembly v3.5 (Majoros *et al.* 2004). For homology-based gene prediction, GeMoMa v1.6 was used with the *D. melanogaster* reference genome alongside our RNAseq data as evidence for splice site prediction (Keilwagen *et al.* 2016). For evidence-based gene finding, the pooled RNAseq data were mapped to the to the to the *T. brassicae* genome separately with TopHat (v2.0.14) with default settings (Trapnell *et al.* 2009). After mapping, Cufflinks (v2.2.1) was used to assemble transcripts (Trapnell *et al.* 2010). CodingQuarry (v1.2) was used for gene finding in the genome using the assembled transcripts, with the strandness setting set to 'unstranded' (Testa *et al.* 2015).

The tool EVidenceModeler (EVM) (v1.1.1) was used to combine the *ab initio*, homology-based, and evidence-based information, with evidence-based weighted 1, *ab initio* weighted 2, and homology-based weighted 3 (Haas *et al.* 2008). We annotated the predicted proteins with BLASTp (v2.2.31+) on a custom database containing all SwissProt and Refseq genes of *D. melanogaster* (Boutet *et al.* 2008; Camacho *et al.* 2009; Acland *et al.* 2014), followed by an additional search in the NCBI non-redundant protein database (nr) to obtain additional homology data. The evidence-based annotation (.bam file) was compared to the final annotation (.gff file) for overlap with BEDtools coverage tool (Quinlan and Hall 2010).

GO term analysis

A list of genes was constructed for Gene Ontology (GO) term classification by deduplicating the annotated proteins and removing the non-annotated proteins. These accession IDs were converted into

■ Table 1 Statistics for five assemblies of *Trichogramma brassicae*. The first strategy was PacBio-only in Canu, while three hybrid assembly strategies were based on SPAdes and modulating k-mer sizes, and an additional hybrid assembly was based on an adapted DBG2OLC+Racon+Pilon protocol. BUSCO score is based on the insect_db09 dataset (Simão *et al.* 2015)

Assembler	Version	Size (bp)	Contigs	Longest contig (bp)	N50 (bp)	BUSCO (Complete %)
Canu	v1.0	69,522,446	3,007	126,800	27,303	18.7
SPAdes (k = 21, 33, 55)	v2.1	227,096,967	282,988	474,998	36,870	96.8
SPAdes (k = 21, 33, 55, 77)	v2.2	226,864,253	189,696	548,753	49,096	97.1
SPAdes (k = 127)	v2.3	211,402,326	73,567	537,817	63,558	96.4
DBG2OLC+ Racon+Pilon	v3.0	235,413,774	1,572	2,953,580	556,663	95.5

UniProtKB accession IDs using the UniProt ID mapping feature and deduplicated a final time (Boutet et al. 2008). These UniProtKB accession IDs were in turn used with the DAVID 6.8 Functional Annotation Tool to assign GO terms to each accession ID with the *D. melanogaster* background and generate initial functional analyses (Huang et al. 2009a, 2009b) (see supplementary S1.3 for DAVID input list).

Heterozygosity estimates

The heterozygosity of the S301 line was assessed using sequence reads and k-mer counting, and compared to the congeneric *Trichogramma pretiosum* (Riley) (Hymenoptera: Trichogrammatidae), for which sequence data exists for both a thelytokous (asexual) *Wolbachia*-infected strain as well as an inbred arrhenotokous (sexual) line (Lindsey et al. 2018). Using jellyfish (v2.3.0) to count k-mers, the same trimmed and paired Illumina reads used for assembly were assessed using the default k-mer size of 21 (m = 21), with results exported to a histogram (Marçais and Kingsford 2011). This histogram file was then used with GenomeScope (v1.0) to estimate heterozygosity of the reads based on a statistical model, where a Poisson distribution is expected for a homozygous sample while a bimodal distribution is expected for a heterozygous distribution (Vurture et al. 2017). This genome profiling gives a reliable estimate for heterozygosity as well as estimates of repetitive content. The same jellyfish and GenomeScope analyses were performed on *T. pretiosum* short-read sequence data for the thelytokous strain (NCBI SRA database, SRR1191749) and the arrhenotokous line (SRR6447489), with adaptions for reported insert sizes (Lindsey et al. 2018).

Ortholog cluster analysis

The complete gene set of *T. brassicae* was compared to that of *T. pretiosum* (Lindsey et al. 2018), which was retrieved from the i5K Workspace (Poelchau et al. 2016). An ortholog cluster analysis was performed on both gene sets via OrthoVenn2 with the default settings of E-values of 1e-5 and an inflation value of 1.5 (Xu et al. 2019). For *T. brassicae* protein set, see supplementary materials S1.5. For full results from the cluster analysis, see supplementary materials S1.6.

Data availability

All sequence data are available at the EMBL-ENA database under BioProject PRJEB35413, including assembly (CADCXV010000000.1). An overview of supplementary material is available on figshare (https://doi.org10.6084/m9.figshare.12794771.v1), with additional material found on the DANS EASY Repository (https://doi.org/10.17026/dans-23w-a9tn), such as gel images, the *Wolbachia* contaminated contigs, input gene list for DAVID, GenomeScope images, and complete protein set. Contained within these supplementary materials are an additional GFF file that can be found on figshare (https://doi.org/10.6084/m9.figshare.12073833.v1) along with the OrthoVenn2 outputs (https://doi.org/10.6084/m9.figshare.12624629.v1).

RESULTS AND DISCUSSION

Sequencing, assembly, and decontamination

Sequencing of the Illumina 150 bp paired-end library yielded 80,489,816 reads. After quality filtering and trimming, 80,483,128 paired-end reads were retained. Sequencing the PacBio Sequel library yielded 2,500,204 subreads with an average length of 6377 bp. The genome size estimate for *T. brassicae* is 246 Mbp (Johnston et al. 2004) indicating that short-read coverage was 98x while long-read coverage was 64x, resulting in a total coverage of 162x. Three assembly pipelines were used, resulting in five potential assemblies

where one, v3.0, was eventually selected for further use. Results of these assemblies are detailed in Table 1.

The first draft assembly generated with Canu with the altered settings for PacBio Sequel data resulted in an assembly of approximately 70 Mbp in size, drastically smaller than the 246 Mbp expected, and contained a total of 3,007 contigs with an N50 of 27,303. The longest contig was 126,800 bp in size.

The second assembly strategy relied on hybrid assembly pipelines, and SPAdes was used with the default k-mer settings, which resulted in an assembly of approximately 227 Mbp in size with an N50 of 36,870 and a BUSCO completeness of 96.8%. Three different assembly runs were done with differing k-mer sizes: the default k-mer sizes of 21, 33, 55 (v2.1); default k-mer sizes plus 77 (v2.2); or the highest possible k-mer size of 127 (v2.3). Increasing the k-mer size only improved N50 scores to a point, along with decreasing the number of contigs, and stable BUSCO scores, however, the assembled genome size drops dramatically with the third attempt shrinking down to 211 Mbp. Based on BUSCO scores and N50 alone, the second SPAdes attempt, v2.2, would be the best of the three, though all three are similar in most measures.

The third assembly strategy used the DGB2OLC+Racon+Pilon pipeline, which resulted in assembly v3.0. Here, there is a large difference compared to the previous SPAdes assemblies. Particularly, the number of contigs is reduced dramatically from the 70,000 to 280,000 range of the SPAdes output down to a mere 1,572. Meanwhile, the assembled genome size is now 235 Mbp and with an N50 of 556,663 and a BUSCO score of 95.5%. The full completeness score for this assembly, using the 1658 BUSCO groups within the insect_od09 BUSCO set, returned 1531 (92.3%) complete and single-copy BUSCOs, 53 (3.2%) complete and duplicated BUSCOs, 22 (1.3%) fragmented BUSCOs, and 52 (3.2%) missing BUSCOs (Simão et al. 2015).

While the PacBio-only assembly in Canu could have been improved using different settings or additional tools, we decided to focus on using the additional sequence information of the Illumina reads in the subsequent hybrid assembly strategies. The SPAdes assemblies (v2.1-3) were already decent but could have been further improved using Pilon, a tool that improves assemblies at the base pair level using high quality Illumina data. However, the v3.0 assembly was by far the best assembly based on assembled genome size, N50, and BUSCO scores and therefore we chose this strategy for our *T. brassicae* genome assembly.

Decontamination of this assembly (v3.0) resulted in the removal of two contigs as the homology analysis using BLASTn with the NCBI nr database indicated that both contigs were confirmed to be largely composed of *Wolbachia* genomic content. Contig "Backbone_1176" is 9,448 bp in length and two areas of the contig, representing over 80% of its length, showed high homology to *Wolbachia*. Similarly, contig "Backbone_1392" is 17,350 bp and three separate areas representing over 80% showed similar levels of homology to *Wolbachia* After decontamination this final assembly (v3.5) was used for annotation.

Ab initio gene finding, transcriptome assembly, and annotation

In our RNA sequencing experiment, we generated 26,479,830 150bp paired-end cDNA reads. Filtering the reads for quality retained 99.3% of these reads to be used for evidence-based gene finding via transcriptome assembly.

The annotations from the evidence-based gene finding were used alongside homology-based findings and *ab initio* annotations in a weighted model, resulting in a complete annotation for the assembly. In 865 mRNA tracks, representing approximately 5.1% of the official

	Heterozygosity (%)	Repetitive content (%)	Source of sequence data
T. brassicae, thelytokous S301 line	0.0332	0.608	This publication
T. pretiosum, thelytokous *Wolbachia* line	0.0289	0.482	Lindsey *et al.*, 2018
T. pretiosum, arrhenotokous inbred line	0.863	2.64	Lindsey *et al.*, 2018

gene set, a gene model could not be annotated via the SwissProt database, and these tracks are named "No_blast_hit." The majority of tracks are annotated with reference to SwissProt or GenBank accession number of the top BLASTp hit.

Transcriptome assembly and mapping resulted in 45,876,158 mapped transcripts (48,327,134 total). CodingQuarry predicted 45,454 evidence-based genes from these mapped transcripts, while *ab initio* gene finding using GlimmerHMM resulted in 16,877 genes and homology-based gene finding with GeMoMa resulted in 6,675 genes. The final complete gene set was created using EVidenceModeler, where a weighted model using all three inputs resulted in a complete gene set of 16,905 genes. 38.96% of the annotation is supported by RNAseq based on coverage comparison to the mapped transcripts.

GO term analysis

The complete gene set of 16,905 genes was deduplicated and genes with no correlating BLASTp hit were removed from this analysis. The remaining 9,373 genes were subjected to UniProtKB ID mapping, resulting in 8,247 genes with a matching ID after another round of deduplication (828 duplicates found). The remaining 755 accession IDs were not able to be matched, half of which are obsolete proteins within the UniParc database (377).

The DAVID Functional Annotation Tool used 6,585 genes for the analysis and showed that 80.8% (5,320) contribute to 530 biological processes, 77.5% (5,104) contribute to 115 different cellular component categories, and 74.2% (4,889) contribute to 93 molecular functions (genes can code to multiple GO terms). The remaining 1,662 genes are uncategorized.

Heterozygosity estimates

Using short-read data and k-mer counting, heterozygosity was estimated for our isofemale S301 line and compared to both a parthenogenesis inducing *Wolbachia*-infected strain and an arrhenotokous line of *T. pretiosum* (Lindsey *et al.* 2018). The average estimated heterozygosity for our S301 *T. brassicae* line is 0.0332% with approximately 0.608% repetitive content (for full details, see Table 2). This is similar to the thelytokous *T. pretiosum* line, which has a slightly lower estimated heterozygosity (0.0289%) and a lower amount of repetitive content (0.482%). Both have a very distinct Poisson distribution, indicating a low heterozygosity (Figure S1.4.1-2). The arrhenotokous *T. pretiosum* showed a higher estimated heterozygosity (0.863%), a larger amount of repetitive content (2.64%), and a slightly bimodal distribution (Figure S1.4.3). The fact that both thelytokous *Trichogramma* species have a similar low level

of heterozygosity when compared to the arrhenotokous *T. pretiosum* suggests that in both cases *Wolbachia* infection had a severe effect on genetic diversity. As the canonical mechanism of parthenogenesis-induction in other *Wolbachia* infected thelytokous *Trichogramma* species is gamete duplication (Stouthamer and Kazmer 1994; Pannebakker *et al.* 2004), in which unfertilized eggs are diploidized and results in fully homozygous progeny in a single generation, the low genomic heterozygosity rate suggests a similar mechanism for *Wolbachia*-induced parthenogenesis in *T. brassicae*. However, the involvement of *Wolbachia* in causing all-female offspring in this *T. brassicae* strain and the presence and mechanisms of *Wolbachia* in other thelytokous *T. brassicae* strains (Farrokhi *et al.* 2010; Poorjavad *et al.* 2012, 2018) does require further investigation.

Ortholog cluster analysis

The complete gene set of *T. brassicae* was compared to that of *T. pretiosum* using OrthoVenn2 (full output in Table 3). Both species show a different absolute number of proteins (16,905 in *T. brassicae* and 13,200 in *T. pretiosum*) that form a similar number of clusters (6,562 in *T. brassicae* and 6,507 in *T. pretiosum*). The two species share 6,178 clusters (of 16,858 proteins), while *T. brassicae* has 384 unique clusters (1,804 proteins) and *T. pretiosum* has 329 unique clusters (998 proteins), as shown in Figure 1. These unique clusters account for approximately 5% of the entire cluster set for both species, and may both indicate true areas of differentiation, or result from differences in the annotation strategies. There is a similar amount of singleton clusters (proteins that do not cluster with others) in *T. brassicae* (5,268) and *T. pretiosum* (5,177). Finally, between the two species, there are 5,828 single copy-gene clusters.

Both the unique clusters and the singleton genes could be novel proteins, regions of contamination, evidence of unique horizontal gene transfer, or pseudogenes. While the total number of genes shows a difference of over 3,000 genes, the cluster analysis shows both a similar number of unique clusters for both species, as well as a similar number of singletons. The BUSCO analysis would indicate that gene inflation due to assembly error is unlikely, as only 3.2% of the BUSCOs are duplicated and 1.3% are fragmented. In addition to the possibility of actual gene duplication in *T. brassicae*, the difference in the number of genes between the two species could also be due to the different annotation tools and methods used between the two projects. More investigation into these protein clusters in addition to a more comprehensive manual annotation of *T. brassicae* should shed some light on the differences between these closely related yet geographically distinct parasitoid wasps.

■ Table 3 Output of OrthoVenn2 ortholog cluster analysis of *Trichogramma brassicae* and *Trichogramma pretiosum*.

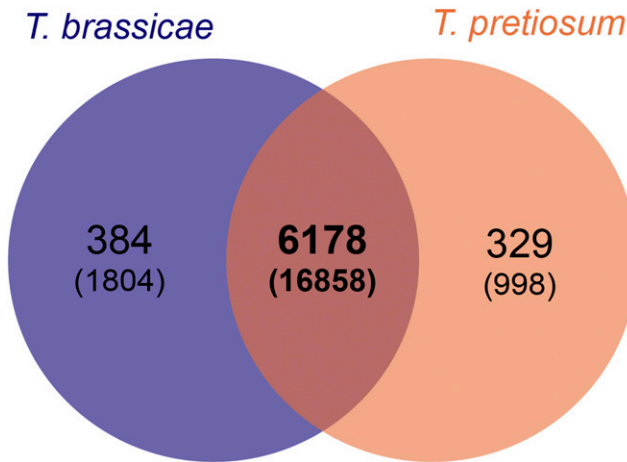| Species	Proteins	Clusters	Singletons	Source of gene set
T. brassicae	16,905	6,562	5,268	This work (S1.5)
T. pretiosum	13,200	6,507	5,177	Lindsey *et al.*, 2018; Poelchau *et al.*, 2015

Figure 1 Ortholog cluster analysis between *Trichogramma brassicae* and *Trichogramma pretiosum* using OrthoVenn2 (Xu *et al.* 2019). The number of clusters shared between the two organisms is in bold, with the number of proteins within each cluster grouping underneath in parentheses.

CONCLUSIONS AND PERSPECTIVES

Here, we present the genome of biological control agent *Trichogramma brassicae*, a chalcidoid wasp used throughout the world for augmentative biological control as well as genetic and ecological research. This unique strain hosted a parthenogenesis-inducing *Wolbachia* infection and is the first European *Trichogramma* genome to be published, allowing for comparative analyses with other *Trichogramma* genomes, as we have shown. Our genomic data also illuminates the possible mechanism of parthenogenesis-induction by *Wolbachia* in this strain. Furthermore, the variety of genomic and transcriptomic data generated for this genome provide much-needed resources to bring *T. brassicae* into the -omics era of biological research.

A hybrid approach was used, resulting in a highly contiguous assembly of 1,572 contigs and 16,905 genes based on *ab initio*, homology-based, and evidence-based annotation, for a total assembly size of 235 Mbp. Two scaffolds were identified that were of *Wolbachia* origin and removed. Ortholog cluster analysis with a sister species showed 384 unique protein clusters containing 1,804 proteins. Future studies are needed to show whether these clusters are truly unique in addition to manual annotation that would shed light on possible gene duplication events. This genome and annotation provides the basis for future, more in-depth comparative studies into the genetics, evolution, ecology, and biological control use of *Trichogramma* species.

LITERATURE CITED

Acland, A., R. Agarwala, T. Barrett, J. Beck, D. A. Benson *et al.*, 2014 Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 42: 8–13. https://doi.org/10.1093/nar/gkt1146

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of Drosophila melanogaster. Science 287: 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Andrews, S., F. Krueger, A. Seconds-Pichon, F. Biggins, and S. Wingett, 2015 FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics. Babraham Inst. 1: 1.

Antipov, D., A. Korobeynikov, J. S. McLean, and P. A. Pevzner, 2016 HybridSPAdes: An algorithm for hybrid assembly of short and long reads. Bioinformatics 32: 1009–1015. https://doi.org/10.1093/bioinformatics/btv688

Babendreier, D., M. Rostas, M. C. J. Höfte, S. Kuske, and F. Bigler, 2003 Effects of mass releases of Trichogramma brassicae on predatory insects in maize. Entomol. Exp. Appl. 108: 115–124. https://doi.org/10.1046/j.1570-7458.2003.00075.x

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin *et al.*, 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19: 455–477. https://doi.org/10.1089/cmb.2012.0021

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, 2008 UniProtKB/Swiss-Prot: The manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Methods Mol. Biol. 406: 89–112. https://doi.org/10.1007/978-1-4939-3167-5_2

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: Architecture and applications. BMC Bioinformatics 10: 421. https://doi.org/10.1186/1471-2105-10-421

Chakraborty, M., J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson, 2016 Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44: 1–12. https://doi.org/10.1093/nar/gkw654

Cruaud, A., G. Groussier, G. Genson, L. Sauné, A. Polaszek *et al.*, 2018 Pushing the limits of whole genome amplification: successful sequencing of RADseq library from a single microhymenopteran (Chalcidoidea, Trichogramma). PeerJ 6: e5640. https://doi.org/10.7717/peerj.5640

Cusumano, A., B. T. Weldegergis, S. Colazza, M. Dicke, and N. E. Fatouros, 2015 Attraction of egg-killing parasitoids toward induced plant volatiles in a multi-herbivore context. Oecologia 179: 163–174. https://doi.org/10.1007/s00442-015-3325-3

Delpuech, J. M., and M. Delahaye, 2013 The sublethal effects of deltamethrin on Trichogramma behaviors during the exploitation of host patches. Sci. Total Environ. 447: 274–279. https://doi.org/10.1016/j.scitotenv.2012.12.096

dos Santos, G., A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby *et al.*, 2015 FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 43: D690–D697. https://doi.org/10.1093/nar/gku1099

Farrokhi, S., A. Ashouri, J. Shirazi, H. Allahvari, and M. E. E. Huigens, 2010 A comparative study on the functional response of Wolbachia-infected and uninfected forms of the parasitoid wasp Trichogramma brassicae. J. Insect Sci. 10: 167. https://doi.org/10.1673/031.010.14127

Fatouros, N. E., and M. E. Huigens, 2012 Phoresy in the field: natural occurrence of Trichogramma egg parasitoids on butterflies and moths. BioControl 57: 493–502. https://doi.org/10.1007/s10526-011-9427-x

Ghorbani, M., M. Saber, M. Bagheri, and N. Vaez, 2016 Effects of diazinon and fipronil on different developmental stages of Trichogramma brassicae Bezdenko (Hym.; Trichogrammatidae). J. Agric. Sci. Technol. 18: 1267–1278.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler, 2013   QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen *et al.*, 2008   Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 19: 1–22. https://doi.org/10.1186/gb-2008-9-1-r7.

Hoskins, R. A., J. W. Carlson, K. H. Wan, S. Park, I. Mendez *et al.*, 2015   The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 25: 445–458. https://doi.org/10.1101/gr.185579.114

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009a   Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37: 1–13. https://doi.org/10.1093/nar/gkn923

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009b   Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4: 44–57. https://doi.org/10.1038/nprot.2008.211

Huigens, M. E., F. G. Pashalidou, M.-H. Qian, T. Bukovinszky, H. M. Smid *et al.*, 2009   Hitch-hiking parasitic wasp learns to exploit butterfly antiaphrodisiac. Proc. Natl. Acad. Sci. USA 106: 820–825. https://doi.org/10.1073/pnas.0812277106

Ivezić, A., P. Rugman-Jones, R. Stouthamer, and A. Ignjatović-Ćupina, 2018   Molecular identification of Trichogramma egg parasitoids of Ostrinia nubilalis in northeastern Serbia. Arch. Biol. Sci. 70: 425–432. https://doi.org/10.2298/ABS171103002I

Jamshidnia, A., S. Abdoli, S. Farrokhi, and R. Sadeghi, 2018   Efficiency of spinosad, Bacillus thuringiensis and Trichogramma brassicae against the tomato leafminer in greenhouse. BioControl 63: 619–627. https://doi.org/10.1007/s10526-018-9893-5

Johnston, J. S., L. D. Ross, L. Beani, D. P. Hughes, and J. Kathirithamby, 2004   Tiny genomes and endoreduplication in Strepsiptera. Insect Mol. Biol. 13: 581–585. https://doi.org/10.1111/j.0962-1075.2004.00514.x

Keilwagen, J., M. Wenk, J. L. Erickson, M. H. Schattat, J. Grau *et al.*, 2016   Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 44: e89. https://doi.org/10.1093/nar/gkw092

Klug, T., and R. Meyhöfer, 2009   Performance of two Trichogramma brassicae strains under greenhouse and field conditions for biocontrol of the silver Y moth in spinach cultures. J. Pest Sci. 82: 73–79. https://doi.org/10.1007/s10340-008-0224-y

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017   Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27: 722–736. https://doi.org/10.1101/gr.215087.116

Kuske, S., D. Babendreier, P. J. Edwards, T. C. J. Turlings, and F. Bigler, 2004   Parasitism of non-target lepidoptera by mass released Trichogramma brassicae and its implication for the larval parasitoid Lydella thompsoni. BioControl 49: 1–19. https://doi.org/10.1023/B:BICO.0000009379.13685.47

Laurent, V., E. Wajnberg, B. Mangin, T. Schiex, C. Gaspin *et al.*, 1998   A composite genetic map of the parasitoid wasp Trichogramma brassicae based on RAPD markers. Genetics 150: 275–282.

Lessard, E., and G. Boivin, 2013   Effect of low temperature on emergence, fecundity, longevity and host-feeding by Trichogramma brassicae. BioControl 58: 319–329. https://doi.org/10.1007/s10526-012-9493-8

Li, H., 2018   Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lindsey, A. R. I., Y. D. Kelkar, X. Wu, D. Sun, E. O. Martinson *et al.*, 2018   Comparative genomics of the miniature wasp and pest control agent Trichogramma pretiosum. BMC Biol. 16: 54. https://doi.org/10.1186/s12915-018-0520-9

Liu, T. X., and Y. Zhang, 2012   Side effects of two reduceδ-risk insecticides, indoxacarb and spinosad, on two species of Trichogramma (Hymenoptera: Trichogrammatidae) on cabbage. Ecotoxicology 21: 2254–2263. https://doi.org/10.1007/s10646-012-0981-5

Majoros, W. H., M. Pertea, and S. L. Salzberg, 2004   TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics 20: 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Marçais, G., and C. Kingsford, 2011   A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770. https://doi.org/10.1093/bioinformatics/btr011

Pannebakker, B. A., L. P. Pijnacker, B. J. Zwaan, and L. W. Beukeboom, 2004   Cytology of *Wolbachia*-induced parthenogenesis in *Leptopilina clavipes* (Hymenoptera: Figitidae). Genome 303: 299–303. https://doi.org/10.1139/g03-137

Poelchau, M., C. Childers, G. Moore, V. Tsavatapalli, J. Evans *et al.*, 2015   The i5k Workspace@NAL-enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res. 43: D714–D719. https://doi.org/10.1093/nar/gku983

Poelchau, M. F., B. S. Coates, C. P. Childers, A. A. Peréz De León, J. D. Evans *et al.*, 2016   Agricultural applications of insect ecological genomics. Curr. Opin. Insect Sci. 13: 61–69. https://doi.org/10.1016/j.cois.2015.12.002

Polaszek, A., 2009   Species Diversity and Host Associations of Trichogramma in Eurasia, pp. 237–266 in *Egg Parasitoids in Agroecosystems with Emphasis on Trichogramma*, edited by Consoli, F. L., J. R. P. Parra, and R. A. Zucchi. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-1-4020-9110-0_9

Poorjavad, N., S. H. Goldansaz, and T. van Leeuwen, 2018   Fertility life table parameters, coi sequences and wolbachia infection in populations of trichogramma brassicae collected from chilo suppressalis. Bull. Insectol. 71: 89–96.

Poorjavad, N., S. H. Goldansaz, T. Machtelinckx, L. Tirry, R. Stouthamer *et al.*, 2012   Iranian Trichogramma: ITS2 DNA characterization and natural Wolbachia infection. BioControl 57: 361–374. https://doi.org/10.1007/s10526-011-9397-z

Quinlan, A. R., and I. M. Hall, 2010   BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rahimi-Kaldeh, S., A. Ashouri, and A. Bandani, 2018   Does Wolbachia Infection Change the Overwintering Ability of Trichogramma brassicae (Hymenoptera: Trichogrammatidae)? Neotrop. Entomol. 47: 583–590. https://doi.org/10.1007/s13744-017-0549-4

Rugman-Jones, P. F., and R. Stouthamer, 2017   High-resolution melt analysis without DNA extraction affords rapid genotype resolution and species identification. Mol. Ecol. Resour. 17: 598–607. https://doi.org/10.1111/1755-0998.12599

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015   BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smith, S. M., 1996   Biological control with Trichogramma: advances, successes, and potential of their use. Annu. Rev. Entomol. 41: 375–406. https://doi.org/10.1146/annurev.en.41.010196.002111

Stouthamer, R., 1997   Wolbachia-induced parthenogenesis, pp. 102–124 in *Influential passengers: inherited microorganisms and arthropod reproduction*, edited by O'Neill, S. L., A. A. Hoffmann, and J. H. Werren. Oxford University Press, Oxford, UK.

Stouthamer, R., J. Hu, F. J. P. M. P. M. Van Kan, G. R. Platner, and J. D. Pinto, 1999   The utility of internally transcribed spacer 2 DNA sequences of the nuclear ribosomal gene for distinguishing sibling species of Trichogramma. BioControl 43: 421–440. https://doi.org/10.1023/A:1009937108715

Stouthamer, R., and M. Huigens, 2003   Parthenogenesis Associated With Wolbachia, pp. 247–266 in *Insect Symbiosis*, edited by Bourtzis, K., and T. A. Miller. CRC Press, Boca Raton. https://doi.org/10.1201/9780203009918.ch15

Stouthamer, R., and D. J. Kazmer, 1994   Cytogenetics of microbe-associated parthenogenesis and its consequences for gene flow in Trichogramma wasps. Heredity 73: 317–327. https://doi.org/10.1038/hdy.1994.139

Sumer, F., A. S. Tuncbilek, S. Oztemiz, B. Pintureau, P. Rugman-Jones *et al.*, 2009   A molecular key to the common species of Trichogramma of the Mediterranean region. BioControl 54: 617–624. https://doi.org/10.1007/s10526-009-9219-8

Suverkropp, B. P., F. Bigler, and J. C. van Lenteren, 2009   Dispersal behaviour of Trichogramma brassicae in maize fields. Bull. Insectol. 62: 113–120.

Suverkropp, B. P., F. Bigler, and J. C. van Lenteren, 2010   Movement and host finding of Trichogramma brassicae on maize plants. Bull. Insectol. 63: 115–127.

Testa, A. C., J. K. Hane, S. R. Ellwood, and R. P. Oliver, 2015   CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16: 170. https://doi.org/10.1186/s12864-015-1344-4

Thubru, D. P., D. M. Firake, and G. T. Behere, 2018   Assessing risks of pesticides targeting lepidopteran pests in cruciferous ecosystems to eggs parasitoid, Trichogramma brassicae (Bezdenko). Saudi J. Biol. Sci. 25: 680–688. https://doi.org/10.1016/j.sjbs.2016.04.007

Trapnell, C., L. Pachter, and S. L. Salzberg, 2009   TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan et al., 2010   Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–515. https://doi.org/10.1038/nbt.1621

Vaser, R., I. Sovic, N. Nagarajan, and M. Sikic, 2017   Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27: 737–746. https://doi.org/10.1101/gr.214270.116

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, M. C. Schatz, J. Gurtowski et al., 2017   GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33: 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wajnberg, E., 1993   Genetic variation in sex allocation in a parasitic wasp: variation in sex pattern within sequences of oviposition. Entomol. Exp. Appl. 69: 221–229. https://doi.org/10.1111/j.1570-7458.1993.tb01745.x

Walker, B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel et al., 2014   Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963. https://doi.org/10.1371/journal.pone.0112963

Wallberg, A., I. Bunikis, O. V. Pettersson, M.-B. Mosbech, A. K. Childers et al., 2019   A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics 20: 275. https://doi.org/10.1186/s12864-019-5642-0

Xu, L., Z. Dong, L. Fang, Y. Luo, Z. Wei et al., 2019   OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 47: W52–W58. https://doi.org/10.1093/nar/gkz333

Ye, C., C. M. Hill, S. Wu, J. Ruan, and Z. Ma, 2016   DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci. Rep. 6: 1–9. https://doi.org/10.1038/srep31900

Ye, C., Z. S. Ma, C. H. Cannon, M. Pop, and D. W. Yu, 2012   Exploiting sparseness in de novo genome assembly. BMC Bioinformatics 13: S1. https://doi.org/10.1186/1471-2105-13-S6-S1

Zhou, W., F. Rousset, and S. O'Neill, 1998   Phylogeny and PCR-based classification of Wolbachia strains using wsp gene sequences. Proc. Biol. Sci. 265: 509–515. https://doi.org/10.1098/rspb.1998.0324

Zimmermann, O., 2004   Der Einsatz von Trichogramma-Schlupfwespen in Deutschland. Gesunde Pflanzen 56: 157–166. https://doi.org/10.1007/s10343-004-0031-1

Communicating editor: S. Celniker