Genome Resources

A Chromosome-Level Genome Assembly of the Parasitoid Wasp, Cotesia glomerata (Hymenoptera: Braconidae)

Brendan J. Pinto, Jerome J. Weis, Tony Gamble, Paul J. Ode, Ryan Paul, and Jennifer M. Zaspel

From the Department of Integrative Biology, University of Texas at Austin, Austin, TX, USA (Pinto); Milwaukee Public Museum, Milwaukee, WI, USA (Pinto, Weis, Gamble, and Zaspel); Department of Biological Sciences, Marquette University, Milwaukee, WI, USA (Gamble and Zaspel); Bell Museum of Natural History, University of Minnesota, St Paul, MN, USA (Gamble); and Colorado State University, Ft. Collins, CO, USA (Ode and Paul).

Address correspondence to B. J. Pinto at the address above, or e-mail: brendanjohnpinto@gmail.com.

Received April 7, 2021; First decision May 24, 2021; Accepted May 25, 2021.

Corresponding Editor: Beth Shapiro

Abstract

Hymenopterans make up about 20% of all animal species, but most are poorly known and lack high-quality genomic resources. One group of important, yet understudied hymenopterans are parasitoid wasps in the family Braconidae. Among this understudied group is the genus Cotesia, a clade of ~1,000 species routinely used in studies of physiology, ecology, biological control, and genetics. However, our ability to understand these organisms has been hindered by a lack of genomic resources. We helped bridge this gap by generating a high-quality genome assembly for the parasitoid wasp, Cotesia glomerata (Braconidae; Microgastrinae). We generated this assembly using multiple sequencing technologies, including Oxford Nanopore, whole-genome shotgun sequencing, and 3D chromatin contact information (HiC). Our assembly is one of the most contiguous, complete, and publicly available hymenopteran genomes, represented by 3,355 scaffolds with a scaffold N50 of ~28 Mb and a BUSCO score of ~99%. Given the genome sizes found in closely related species, our genome assembly was ~50% larger than expected, which was apparently induced by runaway amplification of 3 types of repetitive elements: simple repeats, long terminal repeats, and long interspersed nuclear elements. This assembly is another step forward for genomics across this hyperdiverse, yet understudied order of insects. The assembled genomic data and metadata files are publicly available via Figshare (https://doi.org/10.6084/m9.figshare.13010549).

Subject area: Genome resources
Key words: Braconidae, entomology, genomics, Hymenoptera, transposable elements

Introduction

The insect order Hymenoptera (comprised >150,000 described species of wasps, bees, ants, and sawflies) is one of the most species-rich animal groups on the planet (Forbes et al. 2018). Much of this diversity lies within several groups of wasps with a “parasitoid” life history, where parasitic larvae develop into free-living adults (Peters et al. 2017). However, parasitoid wasps have generally been understudied taxonomically (Jones et al. 2009). In fact, undocumented
parasitoid diversity could make the Hymenoptera the most species-rich order of animals (Forbes et al. 2018). Despite this profound species diversity and myriad of interesting biological phenomena within the Hymenoptera, high-quality genomic data for large swaths of the hymenopteran phylogeny are still lacking (Bransetter et al. 2018).

Braconidae is one of the largest hymenopteran families with >17,000 described species (Yu et al. 2005) and diverged from their closest relatives (Ichneumonidae) in the early-to-mid Jurassic (~175 million years ago) (Peters et al. 2017). The group includes many introduced species important for biological control of insect pests in field and greenhouse agriculture (Heimpel and Lundgren 2000; Van Driesche 2008; Acbes and Messing 2013). Perhaps most notably, the release of Eurasian parasitoid species, Cotesia glomerata (Braconidae; Microgastrinae), in North America, in the 1880s to control populations of imported cabbageworm (Pieris rapae) was one of the first documented cases of introductions for biological control in the world (Heimpel and Cock 2017). Cotesia glomerata has also been a valuable model in a wide range of topics including parasitoid–microbe interactions (Casamano et al. 2018; Zhu et al. 2018), multitrophic interactions (Bukovinský et al. 2009; Poelman et al. 2011, 2012), competition and life-history trade-offs (Geervliet et al. 2000; Gu and Dorn 2003; Vyas et al. 2019), nectar nutrition (Wäckers 2001; Wanner et al. 2006), neurobiology (Van Vlugt et al. 2015), sex determination (Zhou et al. 2006), and host foraging (Horikoshi et al. 1997; Shoji et al. 2006; De Rijk et al. 2016). Continued exploration into these aspects, and others, of the biology and evolution of braconid wasps is currently hindered by the lack of available high-quality genomic data (Bransetter et al. 2018). To date, multiple genomic resources have been developed in braconid wasps. However, these resources are generally lacking in contiguity and phylogenetic breadth. To address this gap, we present a high-quality genome assembly for the parasitoid wasp, C. glomerata, generated using long-read, short-read, and long-range contact (HiC) sequencing data. The resulting C. glomerata genome assembly reported herein is among the most contiguous and complete assemblies currently available across all hymenopterans.

Methods

Data Generation

We obtained C. glomerata specimens from a colony maintained at Colorado State University. Field-collected founders of this colony were initially collected from cabbage fields at the Colorado State Agricultural Research, Development, and Education Center near Wellington, CO. Cotesia glomerata are haplodiploid with 2n females having 20 chromosomes and 1n males having 19 chromosomes (Zhou et al. 2006). We extracted high-molecular-weight DNA (HMW DNA) from 1 male [CgM1] and 1 female [CgF2] C. glomerata specimens using a custom low-input DNA extraction protocol designed for small vertebrates and adapted for single insect input (Supplementary Methods). For each individual, we generated a sequencing library for the Oxford Nanopore Technologies (ONT) platform (SQK-RAD004) and conducted long-read DNA sequencing using a MinION (2 FLO-MINSP6 flowcells). We quality-filtered male and female ONT reads using NanoFilt [v2.5.0] (De Coster et al. 2018) by discarding any sequence with an average quality of <7 and a read length <1000 bp. We then concatenated male and female ONT reads to produce our final ONT dataset, which contained 1 864 940 reads (9 440 627 459 bp) with a read N50 of 8222 bp (~32x coverage).

To maximize the quality of the final assembly with these data, we generated and compared several polished draft genome assemblies using 4 methods: Flye [v2.8.1] (Kolmogorov et al. 2019), wtdbg [v2.5] (Ruan and Li, 2020), Canu [v2.0] (Koren et al. 2017), and Miniasm [v0.3] [Racon [v1.4.13] (Li, 2016; Vaser et al. 2017). We then polished each assembly using ONT and Illumina reads with NextPolish [v1.3.1] (Hu et al. 2020) and assessed the contiguity and completeness of each step using the program Benchmarking Using Single-Copy Orthologs [v3.0] (BUSCO) with the arthropod ortholog database (odb9) (Simão et al. 2015), via the gVolante web server [v1.2.1] (Nishimura et al. 2017) (Table 1). At this point, we chose the highest quality genome assembly (the Flye assembly; Table 1) to continue through HiC scaffolding and genome annotation. We broke and rescaffolded the polished Flye assembly using 2 iterations of 3D-DNA [v201008] (Dudchenko et al. 2017), which yielded 11 chromosome scale scaffolds with no apparent large-scale misassemblies. We visualized the final HiC contact map for misassemblies using Juicebox [v1.11] (Durand et al. 2016). However, we identified 2 arms of a single chromosome that were not scaffolded together and manually connected the scaffolds using Juicebox. Thus, postcuration, this assembly contained 10 chromosome-level scaffolds.

Genome Annotation

Prior to annotation, we trimmed the minimal scaffold length to 1000 bp (Table 1). To functionally annotate the genome assembly, we used the Funannotate pipeline [v1.5.0] (Palmer 2018) in an isolated docker computing environment (Merkel 2014). Briefly, the funannotate pipeline provides a simple pipeline to train and predict gene models using both curated databases (Simão et al. 2015) and RNAseq data (Haas et al. 2008; Keller et al. 2011; Hoff et al. 2016). To compile a posteriori annotation evidence, we fetched available paired-end RNAseq reads from previously conducted RNAseq experiments from Van Vlugt et al. (2015) (Bioproject PRJNA289655; SRR2105801, SRR2105795, SRR2105791, SRR2105785). We quality and adapter trimmed reads before filtering PCR duplicates (as described above) of the concatenated RNAseq reads files. As one line of evidence, we mapped these reads directly to the genome.
assembly using HiSat2 [v2.1] (Kim et al. 2019). As another line of evidence, we assembled a de novo transcriptome using the *De novo* RNAseq Assembly Pipeline (DRAP) [v1.91] (Cabau et al. 2017) to assist in genome annotation. An in-depth description of the utility of DRAP in the production of high-quality transcriptome assemblies can be found elsewhere (Cabau et al. 2017; Pinto et al. 2019). This transcriptome, along with the hymenopteran BUSCO database (odb9), was used to train gene prediction models during the annotation process. We conducted a systematic assessment of other, freely available hymenopteran genome assemblies using the same method to compare with our *C. glomerata* assembly using the same metrics as Table 1 (see Table 2).

### Table 1. Tracking quality and contiguity of the genome assemblies across versions using 5 common metrics: Contig/scaffold N50, size of the smallest scaffold comprising the largest 50% of the assembly; Scaffold L50, number of scaffolds comprising the largest 50% of the genome; Scaffolds, total number of scaffolds comprising the analyzed assembly; Size, the number of base pairs in the assembly; and BUSCO score using the arthropod database (odb9) (Nishimura et al. 2017)

<table>
<thead>
<tr>
<th>Version</th>
<th>Step</th>
<th>N50</th>
<th>L50</th>
<th>Scaffolds</th>
<th>Size (Mb)</th>
<th>BUSCO (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flye Assembly v0.1</td>
<td>Flye</td>
<td>1 105 129</td>
<td>52</td>
<td>4767</td>
<td>287</td>
<td>79.0</td>
</tr>
<tr>
<td>Flye Assembly v0.2</td>
<td>NextPolish</td>
<td>1 122 554</td>
<td>52</td>
<td>4767</td>
<td>290</td>
<td>99.0</td>
</tr>
<tr>
<td>Flye Assembly v0.3</td>
<td>3D-DNA</td>
<td>27 795 534</td>
<td>5</td>
<td>3973</td>
<td>291</td>
<td>98.8</td>
</tr>
<tr>
<td>C. glomerata MPM v2.0</td>
<td>Funannotate</td>
<td>27 795 534</td>
<td>5</td>
<td>3355</td>
<td>290</td>
<td>98.9</td>
</tr>
<tr>
<td>Transcript</td>
<td>Funannotate</td>
<td>Genome annotations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peptide</td>
<td>Funannotate</td>
<td>Genome annotations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. Representative genome assemblies for publicly available hymenopteran species compared with the new *C. glomerata* assembly (bold) across 5 common metrics: Scaffold N50, size of the smallest scaffold comprising the largest 50% of the assembly; Scaffold L50, number of scaffolds comprising the largest 50% of the genome; Scaffolds, total number of scaffolds comprising the analyzed assembly; Size, the number of base pairs in the assembly; and BUSCO score using the arthropod database

<table>
<thead>
<tr>
<th>Species</th>
<th>Family</th>
<th>N50 (kb)</th>
<th>L50 (kb)</th>
<th>Total scaffolds</th>
<th>Size (Mb)</th>
<th>BUSCO (%)</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ceratosolen solmsi</em></td>
<td>Agoniidae</td>
<td>9558</td>
<td>10</td>
<td>2457</td>
<td>277</td>
<td>99.2</td>
<td>Xiao et al. (2013)</td>
</tr>
<tr>
<td><em>Apis mellifera</em></td>
<td>Apidae</td>
<td>13 219</td>
<td>8</td>
<td>5320</td>
<td>250</td>
<td>99.5</td>
<td>Elsik et al. (2014)</td>
</tr>
<tr>
<td><em>Bombus terrestris</em></td>
<td>Apidae</td>
<td>12 868</td>
<td>9</td>
<td>5609</td>
<td>249</td>
<td>99.0</td>
<td>Sadd et al. (2015)</td>
</tr>
<tr>
<td><em>Cotesia congregata</em></td>
<td>Braconidae</td>
<td>20 027</td>
<td>5</td>
<td>2039</td>
<td>199</td>
<td>97.5</td>
<td>Gauthier et al. (2021)</td>
</tr>
<tr>
<td><em>Cotesia glomerata</em></td>
<td>Braconidae</td>
<td>27 796</td>
<td>5</td>
<td>3355</td>
<td>290</td>
<td>98.9</td>
<td>This work</td>
</tr>
<tr>
<td><em>Cotesia vestalis</em></td>
<td>Braconidae</td>
<td>51</td>
<td>818</td>
<td>39 682</td>
<td>172</td>
<td>97.3</td>
<td>Shi et al. (2019)</td>
</tr>
<tr>
<td><em>Diacrasma allorum</em></td>
<td>Braconidae</td>
<td>645</td>
<td>128</td>
<td>3968</td>
<td>389</td>
<td>98.6</td>
<td>Tvedte et al. (2019)</td>
</tr>
<tr>
<td><em>Fopius arisanus</em></td>
<td>Braconidae</td>
<td>978</td>
<td>49</td>
<td>1042</td>
<td>154</td>
<td>97.8</td>
<td>Burke, Simmonds, et al. (2018)</td>
</tr>
<tr>
<td><em>Macrocentrus cingulata</em></td>
<td>Braconidae</td>
<td>192</td>
<td>179</td>
<td>5696</td>
<td>132</td>
<td>98.8</td>
<td>Yin et al. (2018)</td>
</tr>
<tr>
<td><em>Copidosoma floridanum</em></td>
<td>Encyrtidae</td>
<td>1037</td>
<td>153</td>
<td>5445</td>
<td>555</td>
<td>97.1</td>
<td>i5k:GCA_000648653</td>
</tr>
<tr>
<td><em>Atta cephalotes</em></td>
<td>Formicidae</td>
<td>5154</td>
<td>21</td>
<td>2835</td>
<td>318</td>
<td>97.3</td>
<td>Suen et al. (2011)</td>
</tr>
<tr>
<td><em>Monomorium pharaonis</em></td>
<td>Formicidae</td>
<td>15 646</td>
<td>7</td>
<td>9622</td>
<td>259</td>
<td>97.7</td>
<td>Warner et al. (2019)</td>
</tr>
<tr>
<td><em>Ocerea lati</em></td>
<td>Formicidae</td>
<td>16 888</td>
<td>6</td>
<td>139</td>
<td>224</td>
<td>99.3</td>
<td>Oxley et al. (2014)</td>
</tr>
<tr>
<td><em>Temnothorax longispinosus</em></td>
<td>Formicidae</td>
<td>514</td>
<td>137</td>
<td>3987</td>
<td>261</td>
<td>98.1</td>
<td>Kaur et al. (2019)</td>
</tr>
<tr>
<td><em>Diadromus collaris</em></td>
<td>Ichneumonidae</td>
<td>1030</td>
<td>107</td>
<td>63 392</td>
<td>399</td>
<td>99.6</td>
<td>Shi et al. (2019)</td>
</tr>
<tr>
<td><em>Orussus albietinus</em></td>
<td>Orussidae</td>
<td>612</td>
<td>63</td>
<td>1789</td>
<td>186</td>
<td>98.5</td>
<td>i5k:GCA_000612105</td>
</tr>
<tr>
<td><em>Nasonia giralda</em></td>
<td>Peromelidae</td>
<td>759</td>
<td>64</td>
<td>4912</td>
<td>284</td>
<td>97.1</td>
<td>Werren et al. (2010)</td>
</tr>
<tr>
<td><em>Nasonia longicornis</em></td>
<td>Peromelidae</td>
<td>758</td>
<td>65</td>
<td>5214</td>
<td>286</td>
<td>97.2</td>
<td>Werren et al. (2010)</td>
</tr>
<tr>
<td><em>Nasonia vitripennis</em></td>
<td>Peromelidae</td>
<td>897</td>
<td>21</td>
<td>6098</td>
<td>296</td>
<td>97.1</td>
<td>Werren et al. (2010)</td>
</tr>
<tr>
<td><em>Trichogramma pretiosum</em></td>
<td>Trichogrammatidae</td>
<td>3706</td>
<td>19</td>
<td>357</td>
<td>195</td>
<td>99.3</td>
<td>Lindsey et al. (2018)</td>
</tr>
</tbody>
</table>
Results and Discussion

Data Description

Across individuals and sequencing platforms, we generated a data set of ~130× coverage for genome assembly, ~32× in Oxford Nanopore (ONT) and ~100× PE Illumina reads. We also generated a dataset of 140× Illumina reads derived from 3D chromatin capture experiment (HiC). Our initial assembly depended on the ONT long reads, Illumina reads for downstream polishing, and HiC for scaffolding. Data are summarized further in Supplementary Table 1.

Genome Assembly

By almost every metric, Flye produced the best draft genome assembly. It was the most contiguous (N50) and had the highest per-base quality of all 4 assemblies (represented by a higher BUSCO score, given that each assembly was derived from the same dataset). In contrast, Canu provided the lowest-quality assembly by all metrics including an order of magnitude increase in runtime, while Miniasm/Racon and wtdbg generated intermediate assemblies. Of note, assemblers varied in speed by orders of magnitude with wtdbg being the fastest from reads to consensus (~2 h), then Miniasm and Racon (~9 h), Flye (~8 h), and finally Canu (>96 h). However, after polishing, only the wtdbg assembly approached the quality of the Flye assembly (summarized further in Table 1). Thus, we chose the Flye assembly for additional scaffolding using the HiC data.

Our *C. glomerata* genome assembly achieved “chromosome-level” status with 90% of the genome attributed to the largest 10
scaffolds; representative of each of *C. glomerata*’s 10 chromosomes (1× = 10, sensu Zhou et al. 2006). Additionally, 95.6% of the final assembly is represented in 46 scaffolds that are longer than 50 kb. Quality-control read-mapping of Illumina back to the final assembly produced extremely high mapping rates for both CgM1 (98.9% reads mapped; 96.1% properly paired) and CgM2 (98.4% reads mapped; 95.4% properly paired). The base composition across the genome assembly was relatively homogenous with an overall GC content of 31% (±6%). The BUSCO statistics for the final assembly, relative to a set of 1066 conserved arthropod orthologs, included 97.3% complete single-copy genes with 1.5% complete and duplicated gene versions, 0.5% fragmented gene copies, and 0.7% missing genes. Genome annotation, using RNAseq data derived from experiments of expression in the heads of multiple female *C. glomerata*, successfully passed 29 455 annotations, including 24 170 genes. Interestingly, the final genome assembly approached 300 Mb in size, which far exceeded our initial predictions for its genome size given the available data from their closest relatives (Figure 1).

Genome sizes across Hymenoptera can vary greatly between taxa, such as the 3-fold difference between *Macrocentrus cingulum* and *Diadromus collaris* within the superfamily Ichneumonoidea (Table 2). Previous genome assemblies within *Cotesia*, *C. congregata* (Gauthier et al. 2021) and *C. vestalis* (Shi et al. 2019), had haploid genome sizes that were far closer to ~200 Mb (Table 2). We set out to compare across the genomes of these taxa to shed light on this apparent genome size discrepancy. However, drastic differences in sequencing technologies and genome quality across these species prevented direct comparisons between all 3 species. To close this gap, we dropped *C. vestalis* from downstream analyses and focused solely on comparing the *C. glomerata* and *C. congregata* [v2.0] (Gauthier et al. 2021) genomes.

We investigated this genome size discrepancy by interrogating the repeat content using Illumina reads to account for differences in sequencing technologies between assemblies—along with a estimating total repeat content from the assemblies. First, we confirmed this was the haploid genome size by estimating the *C. glomerata* (CgM2) and *C. congregata* (ENA: ERR3829581) genome size using kmers with Jellyfish [v2.2.10] (Marçais and Kingsford 2011). The estimated genome size of *C. glomerata* roughly approximated the assembled genome size at 272 Mb, whereas the estimated genome size of *C. congregata* was 142 Mb (Figure 2). Next, using RepeatModeler [v1.0.11] (Smit and Hubley 2008), we calculated percent repeat content for each genome 54.75% (159.28 Mb) of the *C. glomerata* assembly and 35.04% (69.72 Mb) of the *C. congregata* assembly. This >100% increase in repeat content is sufficient to account for the large increase in genome size in *C. glomerata*, relative to *C. congregata*. Lastly, we approximated the categories of genomic repeats responsible for the discrepancies in genome size between these 2 *Cotesia* species using dnaPipeTE [v1.3.1] (Goubert et al. 2015). We identified 3 primary categories of repetitive elements that increased proportionally between *C. glomerata* relative to *C. congregata*. A large proportion of this amplified repetitive content appears to be made up of simple repeats, long terminal repeats, and long interspersed nuclear elements (Figure 1).

Here, we present a chromosome-level genome assembly for *C. glomerata*. This assembly is currently the most complete and, in part due to its exceptionally large genome size (larger chromosomes), is also one of the most contiguous hymenopteran genomes to date, i.e., having the highest N50 (Table 2). Interestingly, the genome size was much larger than initially hypothesized based on information from its closest relatives, and this discrepancy appears to be due to runaway amplification of repetitive elements in the *C. glomerata* lineage (Figure 1). As more high-quality braconid genomes continue to be generated, fruitful areas of research may include diving deeper into the unanswered questions, such as the impact of repeat content

---

**Figure 2.** Comparative genome TE content between *Cotesia glomerata* and *C. congregata*. “Assembly Size” is the total assembled genome size for each species; “Estimated Size” is the genome size estimation calculated using kmer counting with Jellyfish; and “Repeat Content” is the total repeat content estimated using RepeatModeler. Categories of genomic repeat content and pie charts generated by dnaPipeTE, unspecified % pie slices represent <1% estimated content. It should also be noted that the color palette for each wheel begins at the top of both the key and the wheels (with LTR) and continues clockwise down the palette in the order listed in the key.
on the parasitoid lifestyle and the evolution of sex determination in Hymenoptera—among others. Indeed, this assembly is another step forward for genomics across this hyperdiverse, yet understudied, order of insects.

Supplementary Material
Supplementary data are available at Journal of Heredity online.

Acknowledgments
The authors thank J. Colby (MPM) and C. Smith (UT) for assistance organizing and genome assembly-related discussions, respectively, and m4n4b for computational their efficiency over the course of this project. Also, we would like to thank members of the Kirkpatrick Lab for feedback on an earlier version of the manuscript. This work was funded by the John J. Brander and Christine E. Rundbald Research Fellowship to J.J.W. and American Genetic Association EECG award to B.J.P.

Author Contributions
B.J.P., T.G., P.O., J.Z., and J.W. conceptualized the study. R.P. and P.O. conducted fieldwork. B.J.P. and J.W. extracted DNA and B.J.P. generated sequencing libraries, conducted ONT sequencing, analyzed data, and drafted the manuscript. All authors edited, read, and approved the final manuscript.

Data Availability
Sequence data is submitted to the NCBI Sequence Read Archive (SRA) under BioProject PRJNA715233. Assembled data is publicly available at https://doi.org/10.6084/m9.figshare.13010549.

References


