Assemblies of the genomes of parasitic wasps using meta-assembly and scaffolding with genetic linkage

Kameron T. Wittmeyer 1, Sara J. Oppenheim, 2 and Keith R. Hopper 1, *

1 USDA-ARS, Beneficial Insect Introductions Research Unit, Newark, DE 19713, USA
2 American Museum of Natural History, New York, NY 10024, USA
*Corresponding author: USDA-ARS, Beneficial Insect Introductions Research Unit, 501 South Chapel Street, Newark, DE 19713, USA. Email: keith.hopper@usda.gov

Abstract

Safe, effective biological-control introductions against invasive pests depend on narrowly host-specific natural enemies with the ability to adapt to a changing environment. As part of a project on the genetic architectures of these traits, we assembled and annotated the genomes of two aphid parasitoids, Aphelinus atriplicis and Aphelinus certus. We report here several assemblies of A. atriplicis made with Illumina and PacBio data, which we combined into a meta-assembly. We scaffolded the meta-assembly with markers from a genetic map of hybrids between A. atriplicis and A. certus. We used this genetic-linkage scaffolded (GLS) assembly of A. atriplicis to scaffold a de novo assembly of A. certus. The de novo assemblies of A. atriplicis differed in contiguity, and the meta-assembly of these assemblies was more contiguous than the best de novo assembly. Scaffolding with genetic-linkage data allowed chromosomal-level assembly of the A. atriplicis genome and scaffolding a de novo assembly of A. certus with this GLS assembly, greatly increased the contiguity of the A. certus assembly to the point where it was also at the chromosomal-level. However, completeness of the A. atriplicis assembly, as measured by percent complete, single-copy BUSCO hymenopteran genes, varied little among de novo assemblies and was not increased by meta-assembly or genetic scaffolding. Furthermore, the greater contiguity of the meta-assembly and GLS assembly had little or no effect on the numbers of genes identified, the proportions with homologs or functional annotations. Increased contiguity of the A. certus assembly provided modest improvement in assembly completeness, as measured by percent complete, single-copy BUSCO hymenopteran genes. The total genic sequence increased, and while the number of genes declined, gene length increased, which together suggest greater accuracy of gene models. More contiguous assemblies provide use other than gene annotation, for example, identifying the genes associated with quantitative trait loci and understanding of chromosomal rearrangements associated with speciation.

Keywords: genome assembly; assembly reconciliation; genetic scaffolding; comparative genomics; biological control; parasitoid; Hymenoptera

Introduction

Invasions by exotic species that become pests are an increasing problem for agriculture (Bradshaw et al. 2016). Biological control by introduction of natural enemies has proved effective at reducing the abundance and impact of such pests (Cock et al. 2016), and in principle provides a safe, cost-effective, sustainable alternative to widespread use of insecticides (Naranjo et al. 2015). Safe and effective biological-control introductions against foreign, invasive pests depend on using narrowly host-specific agents and their establishment in the new environment. Host specificity can be measured experimentally prior to introduction (Hoddle 2005), but the possibility of post-release evolution also needs to be considered. The likelihood that an introduced species will evolve to attack novel hosts depends on the genetic architecture of host use. Not much is known about the evolution or genetics of host specificity in parasitoids, but recent research suggests that the architecture is complex and will require genome-level analyses (Hopper et al. 2019, 2021). Furthermore, the establishment of introduced insects depends on their ability to adapt to the climate in a new and changing environment, and such adaptations depend on the genetic architecture of climatic adaptation.

As part of a project on the genetic architectures of host use and climatic adaptation, we assembled and annotated the genomes of Aphelinus atriplicis Walker (Hymenoptera: Aphelinidae), an important natural enemy introduced to control the Russian wheat aphid, Diuraphis noxia (Kurdjumov, Hemiptera: Aphididae), a major invasive pest of small grains in North America (Morrison and Peairs 1998), and Aphelinus certus Yasnosh (Hymenoptera: Aphelinidae), an Asian parasitoid of the soybean aphid, Aphis glycines (Hemiptera: Aphididae), that appears to have been introduced accidentally at the same time as the soybean aphid and is now a major natural enemy of this pest (Heimpel et al. 2010). We assembled these and other genomes of Aphelinus to find quantitative-trait loci and the underlying genes involved in the evolution of host specificity and climatic adaptation and to study mechanisms of reproductive isolation among the many closely related species in this genus. Ideally, one would make chromosomal-level assemblies of all the species one studied, but this would be prohibitively expensive when one is...
working with multiple species. One of the rationales for this article was to find out how much gene annotation was improved by a highly contiguous assembly made with multiple types of data and methods compared to a more fragmented assembly. Another rationale was to determine whether we could use a more contiguous assembly of one species to provide improved scaffolding of the assembly of a closely related species.

The rationale for the work reported here was to determine the most effective way to allocate sequencing expenditures in a project involving a set of closely related species. Would generating a highly contiguous, chromosomal-level assembly using multiple data types and methods improve the accuracy and completeness of gene annotation in comparison with a fragmented assembly? And, could a highly contiguous assembly be used to improve scaffolding of the assembly of the genome a closely related species for which only short-read data were available?

We present assemblies of *A. atriplicis* using different data types (Illumina vs PacBio), processing (raw vs error-corrected), and different *de novo* assemblers. The minimal case would be to create assemblies using only one data type, which would be cheaper than using combined data. (Within the single-data assemblies, Illumina-only is currently about 1/10 the cost of PacBio-only). We made a meta-assembly with selected *de novo* assemblies and scaffolded this meta-assembly with a genetic map. We evaluate the output from different approaches and describe the pipeline for the procedure that produced the overall best assembly (Figure 1).

Meta-assembly, also known as assembly reconciliation, is an increasingly popular approach to providing higher quality genome assemblies (Wences and Schatz 2015; Babb et al. 2017; Torresen et al. 2017; Patel et al. 2018; Song et al. 2019; Tang et al. 2020). This approach takes advantage of differences in the sequence information captured by assemblers that differ in how well they deal with coverage level, repetitive or duplicated sequences, and sequencing errors. Because of these differences, one assembler may cover gaps between contigs produced by another assembler, and the combination of multiple assemblies can improve contiguity substantially.

We scaffolded the meta-assembly of *A. atriplicis* using markers from a genetic map of hybrids between *A. atriplicis* and *A. certus*. We compared the completeness and contiguity of the best *de novo* assembly, the meta-assembly made using 11 *de novo* assemblies, and the genetic linkage-scaffolded (GLS) assembly of this meta-assembly. For *A. certus*, we used the GLS assembly of *A. atriplicis* to scaffold an Illumina-only *de novo* assembly. Finally, we compared the gene annotations of these assemblies. The pipeline for our approach is given in Figure 1, and details of the programs and versions we used are in Supplementary Table S1, and parameters, commands, and scripts are in Supplementary Document S1.

The *de novo* assemblies differed in several metrics of contiguity with no one assembly being best across all metrics. For *A. atriplicis*, the meta-assembly metrics were superior to *de novo* assemblies. Scaffolding the meta-assembly with genetic linkage data improved contiguity more than six-fold and produced a chromosomal-level assembly of the *A. atriplicis* genome. For *A. certus*, scaffolding the *de novo* assembly with the GLS assembly of *A. atriplicis* increased its contiguity more than 2000-fold and produced a chromosomal-level assembly. Assembly completeness, as measured by percent complete, single-copy BUSCO hymenopteran genes, showed far less improvement and generally varied little among assemblies. Furthermore, the total numbers of genes identified in assemblies, the proportion of genes with homologs on GenBank, and the proportion of genes with Gene Ontology or InterProScan annotations, showed little improvement between assemblies. Nonetheless, increased contiguity of assemblies is important regardless of impact on gene annotation, for example, finding genes associated with quantitative trait loci (QTL) and understanding chromosomal rearrangement associated with speciation both require highly contiguous assemblies.

**Materials and methods**

**Study system**
The genus *Aphelinus* (Hymenoptera: Aphelinidae) comprises at least 94 species (Noyes 2019). From published host records, all

---

**Figure 1** Pipeline for assembly of the genome of *Aphelinus atriplicis*. Orange boxes are inputs and outputs; blue boxes are processing steps. The step numbers reference text in Materials and Methods and Supplementary Material.
Aphelinus species are internal parasitoids of aphids, and many Aphelinus species are used in biological control of pest aphids (Hopper et al. 1998, 2017a, 2017b). Aphelinus atriplicis was collected from D. noxia in the Caucasus region of southern Russia and introduced into the United States in 1991 to control this pest (Hopper et al. 1998). Aphelinus certus appeared in North America soon after the soybean aphid invaded and has spread throughout the region where soybean aphid is a pest (Heimpel et al. 2010). In the laboratory, A. atriplicis and A. certus are partially compatible: the species will mate with one another, and crosses produce viable, fertile F1 females (Heraty et al. 2007). When F1 females from crosses between A. certus females and A. atriplicis males are backcrossed to the parental species, they produce as many progeny as either parental species on their common host, Aphis glycines, but when F1 females from the reciprocal cross (A. atriplicis females × A. certus males) are backcrossed to the parental species, they produce half as many progeny as either parental species (Wu et al. 2004). In spite of this partial compatibility, A. atriplicis and A. certus are phylogenetically distinct and differ in courtship behavior, morphology, and host specificity (Heraty et al. 2007; Hopper et al. 2017a, 2017b). The populations of A. atriplicis and A. certus we studied differ by ~4 million single-nucleotide polymorphisms, which is four times greater than the intraspecific variation among 17 populations of A. certus (Hopper et al. 2019). Natural populations of A. atriplicis and A. certus are separated by 5000 km of deserts and mountains, including some of the highest mountain ranges in the world, thus their partial compatibility likely results from a lack of selection against hybridization (Heraty et al. 2007).

**Insect strains and rearing**

The parasitoids and aphids used in the research reported here were from cultures maintained at the USDA-ARS, Beneficial Insect Introductions Research Unit, Newark, Delaware. The culture of A. atriplicis was started in 2000 with 101 adults from parasitized D. noxia collected in fields of Triticum sp. (wheat) near Tbilisi, Republic of Georgia (permit P526P-20-03892). The A. certus culture was started in 2001 with 14 adults from parasitized Aphis glycines (soybean aphid) collected in fields of Glycine max (soybean) near Tsukuba City, Japan (permit P526F-18-04656). To help maintain genetic variation, each parasitoid culture was divided into 4–6 subcultures, and each subculture is maintained with an adult population size >200 (Roush and Hopper 1995). The culture of A. glycines was started in 2004 with aphids collected from a soybean field near Newark, Delaware, while that of D. noxia was started in 1998 with aphids from a culture at the University of Wyoming, which was started with aphids collected in southeastern Wyoming in 1992, 1994, and 1995. The aphid cultures are maintained as 18–36 subcultures with thousands of aphids per subculture. All cultures are kept in plant-growth chambers (model AR66L2, Percival Scientific, Perry, IA, USA) at 20°C, 40–70% relative humidity, and 16:8 h (L:D) photoperiod. Aphelinus atriplicis is reared on D. noxia on Hordeum vulgare (barley) and A. certus on A. glycines on Glycine max (soybean). Vouchers of insect species are maintained at −20°C in molecular grade ethanol at the Beneficial Insect Introduction Research Unit, Newark, Delaware.

**Genomic DNA libraries**

For assemblies of the genome of A. atriplicis, we used a combination of an Illumina paired-end (2 × 150 nt) library and PacBio long-read libraries, prepared with standard kits and protocols from Illumina (Illumina, San Diego, CA, USA) and PacBio (Pacific Biosciences, Menlo Park, CA, USA). To make an Illumina library of A. atriplicis, we extracted genomic DNA from an individual male, using a DNaseasy Blood and Tissue Kit, that we then whole-genome amplified (WGA) the DNA with a REPLI-g Kit (Qiagen, Valencia, CA, USA). Because Hymenoptera are haplo-diploid with males being hemizygous, using a single male ensured that this library would have no heterozygosity. To make the PacBio libraries of A. atriplicis, however, we had to extract genomic DNA from pooled adults because, at the time these libraries were made and sequenced, several micrograms of DNA were needed for the PacBio technology (Aphelinus species are small, yielding ~100 ng DNA per adult). We used the Genomic Tip 20/G Kit (Qiagen) to extract DNA from a pool of 275 individuals for one library and 700 for the other. The libraries were prepared and sequenced at the Delaware Biotechnology Institute, Newark, Delaware, using an Illumina HiSeq 2500 and a PacBio RS. After size selection, the average PacBio-library fragment size was 35,290 bp (see Supplementary Figure S1).

After quality trimming with Trimmomatic (Bolger et al. 2014), analysis with FastQC showed that the Illumina library for A. atriplicis yielded 97 million reads with a length of 150 nt and a total of 15 billion nt of sequence. After subread and quality filtering, the PacBio libraries for A. atriplicis yielded ~1 million reads with an N50 for subread length of 22 kb and a total of 19 billion nucleotides of sequence. Given the 353-Mb size of the A. atriplicis genome estimated from flow cytometry (Gokhman et al. 2017), the Illumina library should give 41-fold coverage, and the two PacBio libraries should give 54-fold coverage. The two PacBio libraries were combined, and consensus error-correction was done with CANU (using the options -correct and -genomSize=350 m; Koren et al., 2017; Figure 1, step 1). In most cases, the assemblies were better with error-corrected PacBio reads rather than raw reads.

To make an Illumina library of A. certus, we extracted genomic DNA from a pool of 24 males that were sons of a single, unmated female, using a DNaseasy Blood and Tissue Kit. We used pooled sons from a single, unmated female to reduce polymorphism, which should ease assembly, although there would be some polymorphism even among males from a single female. This being said, the A. certus assembly made with MaSuRCA using this library was actually more contiguous that the A. atriplicis assembly made with MaSuRCA using the library made from a single male. The library was prepared and sequenced at the Delaware Biotechnology Institute, Newark, Delaware, using an Illumina HiSeq 2500. After quality trimming with Trimmomatic (Bolger et al. 2014), the library yielded 203 million reads with a length of 150 nt and a total of 30 billion nt of sequence. Given the 362-Mb size of the A. certus genome estimated from flow cytometry (Gokhman et al. 2017), this library should give 84-fold coverage.

**Initial genome assemblies**

For A. atriplicis, several long-read assemblies were made with the assemblies CSA_assemble [Kuhl et al. 2020; pipeline that runs wtdbg2, Ruan and Li (2020)], Flye (Kolmgorov et al. 2019), MECAT2 (Xiao et al. 2017), Miniasm (Li 2016), Platans-allee (Kajitani et al. 2019), and Raven (Vaser and Šikić 2020). Hybrid assemblies with Illumina and PacBio reads were made with MaSuRCA (Zimin et al. 2013), Wengan (Di Genova et al. 2021), and DBG2OLC (Ye et al. 2016), the latter uses Platans (Kajitani et al. 2014) for the short-read assembly (Figure 1, step 2). Where feasible, each assembler was run with both raw and corrected PacBio reads. Mecat has its own long-read error correction, and MaSuRCA uses short reads to correct long reads; thus, assemblies
made with them were categorized as using error-corrected reads even though run with raw reads. The resulting assemblies were error-corrected with long-read error-correction using Flye (-polish option; Kolmogorov et al. 2019; Figure 1, step 3), to improve performance in downstream steps. (Details of the programs and versions used are in Supplementary Table S1, and parameters, commands, and scripts are in Supplementary Document S1.) These analyses and all the others described below, with the exception of the gene ontology analysis, were done on the BIOMIX computational cluster at the University of Delaware Bioinformatics Core Facility.

Each assembly was run through the HERA pipeline (Du and Liang 2019; Figure 1, step 4). HERA performs gap closing/scaffolding of assemblies using long reads by re-assembly of gaps and resolution of closely related repeats that were unresolved by the original assembler. HERA generally improved contiguity statistics threefold. The performance HERA was evaluated using different combinations of parameters to determine the optimal settings. First HERA scripts were modified to run on the Slurm workload manager on BIOMIX. Second, BWA (Li and Durbin 2009) was replaced with Minimap2 (Li 2018) for read mapping in the pipeline because the latter has similar performance for alignment of long reads but is much faster. Next the parameters evaluated were with a range of mapping settings for Minimap2 ($\sim x$ asm10 vs $\sim x$ asm20, i.e., align up to 10% vs 20% sequence divergence), minPathNum (minimum number of paths for contig pairs), minOverlap_Merge (minimum overlap needed to merge super-contigs and non-scaffolded contigs), and min_ctg_size (minimum contig size used in all vs all and all vs reads mapping). We ended up with parameter settings of $\sim x$ asm20, minPathNum = 6, minOverlap_Merge = 5000, and min_ctg_size of 10,000.

Following preliminary assessments, some low-quality assemblies were dropped. These included Platanus-allee, which made a large number of very short contigs that performed poorly with the HERA and ntJoin in downstream steps. Miniasm with raw reads, which produced a highly inflated assembly size, likely due to sequence duplication or deletion, which can be a problem with some other assembly-reconciliation tools.

The ntJoin algorithm does well at ordering and placing contigs together, but it fills the gaps between adjacent contigs with Ns. To replace these Ns with sequence data wherever possible, we ran the meta-assembly through TGS-GapCloser (Xu et al. 2020; Figure 1, step 7), which uses long reads to fill gaps between contigs on the same scaffold. This final set of scaffolds was polished using short reads with Pilon (Walker et al. 2014; Figure 1, step 8), which fixed local indels and improved the number of genetic markers that could later be uniquely mapped to the assembly.

We tested various parameter values to optimize the performance of ntJoin. For our analysis, the best settings were $w = 1000$ ($w$ is the window size for finding minimizers), no_cut=False (no_cut specifies whether the target assembly should be broken at areas of discrepancy), with target_weight (the weight given to the target assembly when determining the highest scoring path) and reference_weights (weights given to the references when determining the highest scoring path) set equal to each other. This allowed discrepancies in the target assembly to be broken and did not give any preference to any of the reference assemblies.

**Evaluation of assemblies**

We compared the assemblies by evaluating sizes relative to estimates from flow cytometry, contiguity statistics from Quast (Mikheenko et al. 2018), core-gene completeness from BUSCO for Hymenoptera (Hymenoptera_odb10; Seppey et al. 2019), and number of misassemblies (Figure 1, step 5). The number of misassemblies was estimated by counting the number of contigs or scaffolds with genetic markers from different chromosomes. Although this estimate may fail to detect misassemblies where genetic markers are sparse, the A. atripilicus genome was fairly uniformly covered by genetic markers so our estimate should scale well with the true number of misassemblies. To rank the de novo assemblies, we calculated the ratios of the observed values to the best values for seven assembly statistics, including divergence from flow cytometry size, number of contigs, $N_{50}$ (the length of shortest contig at which over half the genome is covered), $L_{50}$ (the number of contigs covering half the genome), the length of the longest contig, the number of misassemblies, and BUSCO scores.

**Meta-assembly**

The seven de novo assemblers used here employ five algorithms, including overlap-layout-consensus, fuzzy De Bruijn graph, repeat graph, string graph, and De Bruijn graph with overlap-layout-consensus. Five assemblers used long-read PacBio data and two used both short-read Illumina and long-read PacBio data. Lastly, we ran the assemblers with raw and error-corrected PacBio data. This diversity of algorithms, data types, and read processing was designed to capture of the maximum amount of information in our sequence data.

The CSA assembly was chosen as a target for meta-assembly because it ranked highest among the initial assemblies. The CSA assembly was error-corrected with HyPo (long- and short-read error correction for single-nucleotide polymorphisms (SNPs) and small indels; Kundu et al. 2019), which increased the accuracy of gene calls (based on BUSCO scores), as well as the number of genetic markers that mapped to the A. atripilicus genome for genetic map scaffolding. The CSA assembly was then used as the target assembly in ntJoin (Coombe et al. 2020; Figure 1, step 6) with the remaining 10 assemblies serving as references. ntJoin performs alignment-free scaffolding using structural synteny and can be used for several purposes. In the present study, ntJoin identified all of the syntenic contigs among the input assemblies, allowing adjacent contigs to be ordered, oriented, and joined in the target assembly. Errors in the target assembly and discrepancies among the reference assemblies were resolved by consensus (all inputs were given equal weight), allowing misassemblies to be broken and repaired. Only sequences from the target assembly were scaffolded and output, thus there was no chance of sequence duplication or deletion, which can be a problem with some other assembly-reconciliation tools.

The ntJoin algorithm does well at ordering and placing contigs together, but it fills the gaps between adjacent contigs with Ns. To replace these Ns with sequence data wherever possible, we ran the meta-assembly through TGS-GapCloser (Xu et al. 2020; Figure 1, step 7), which uses long reads to fill gaps between contigs on the same scaffold. This final set of scaffolds was polished using short reads with Pilon (Walker et al. 2014; Figure 1, step 8), which fixed local indels and improved the number of genetic markers that could later be uniquely mapped to the assembly.

We tested various parameter values to optimize the performance of ntJoin. For our analysis, the best settings were $w = 1000$ ($w$ is the window size for finding minimizers), no_cut=False (no_cut specifies whether the target assembly should be broken at areas of discrepancy), with target_weight (the weight given to the target assembly when determining the highest scoring path) and reference_weights (weights given to the references when determining the highest scoring path) set equal to each other. This allowed discrepancies in the target assembly to be broken and did not give any preference to any of the reference assemblies.

**Genetic map**

The genetic map used for scaffolding was generated as part of a QTL mapping project using the following procedures. We crossed an A. atripilicus male with an A. certus female and backcrossed 23 $F_1$ females with 23 A. certus males to produce 320 backcross females in 23 families. To generate genetic markers, we made and sequenced reduced-representation libraries (RRL) of each of parent, grandparent, and backcross female. Such libraries can provide large numbers of sequence polymorphisms across many individuals at low cost (Baxter et al. 2011). To make these libraries, we modified a protocol from Baird et al. (2008). Genomic DNA was extracted from individual wasps using Qiagen DNeasy Blood and Tissue Kits and then WGA with REPLI-g Kits, following kit protocols. An aliquot (1 μg) of the resulting DNA was digested with restriction endonucleases using one rare cutter (e.g., NgoMIV
with a 6 bp recognition site) and one frequent cutter (e.g., CviQiL with a 4-bp recognition site), which together determine the number of locations of fragments across the genome and the lengths of these fragments. Custom adaptors, with barcodes for each sample that also serve to register clusters on the Illumina HiSeq during sequencing, were ligated onto the fragments using T4 DNA ligase. The ligation products were pooled and purified using the Qiagen QIAquick PCR purification kit. The pooled ligate was size selected (300–500 bp) using the BluePippin system (Sage Science, Beverly, MA, USA). The size-selected ligate was PCR amplified to both increase copy number at each locus and add more adaptor sequence. The adaptors were designed so only fragments with the rare-common combination of cut sites would amplify. After PCR, the product was purified using AgenCourt AMPure XP beads (Beckman Coulter, Indianapolis, IN, USA), quantitated with qPCR, and sequenced on an Illumina HiSeq 2500 (Illumina) at the Delaware Biotechnology Institute, Newark, Delaware. The read sequences were analyzed using a custom pipeline written in Python that included quality trimming with FastQC (Andrews 2010), barcode deconvolution, mapping of reads to a draft genome assembly of *A. atriplicis* (see details below) with BWA (Li and Durbin 2010), and variant calling using the GATK UnifiedGenotyper (DePristo et al. 2011). The library preparation protocol and bioinformatic pipeline were described in Manching et al. (2017), and the code can be found at https://github.com/UD-CBCB/RedRep. We filtered the variant calls for SNP that (1) were fixed differences between the *A. atriplicis* and *A. certus* individuals used in the crosses, (2) had at least 10× coverage in at least 80% of the backcross females, and (3) had minor allele frequency of at least 30%. We ended up with marker data from 291 backcross females because some females were not scored for enough markers. Using R/qtl (Broman et al. 2003) and the methods described in Broman (2010), we made a genetic linkage map with SNP markers from this RRL (Figure 1, step 9).

**Genetic map scaffolding**

Using the ~100 nt in each RLL locus where SNP markers were found, we mapped these loci to the meta-assemble scaffolds with BWA (Li and Durbin 2010). Eight RLL loci were dropped because they had poor mapping quality, three did not map at all, and 44 were dropped because they only partially aligned, leaving 779 RLL loci to be used for scaffolding. Using Chromomonser (Catchen et al. 2020), we scaffolded the meta-assembly of the *A. atriplicis* genome to the genetic map described above (Figure 1, step 10). This produced a chromosomal-level assembly for much of the *A. atriplicis* genome, which we will refer to as the GLS assembly. The 779 loci considered for scaffolding mapped to 48 contigs from the meta-assembly. Chromomonser used 224 loci to place and order the 48 contigs on 4 chromosomes. The other 555 loci were dropped by the program because their ordering was inconsistent within or between contigs. To get an idea of the completeness of each chromosome-level assembly, we estimated the amount of sequence in each chromosome from the 353-Mb genome size of *A. atriplicis* estimated from flow cytometry and the relative sizes of each chromosome in the karyotype of *A. atriplicis* (Gokhman et al. 2017). This gave size estimates of 141, 92, 64, and 55 Mb for each of the four chromosomes. We compared the amount of sequence assembled in each chromosome to these values to estimate the amount of missing sequence for each chromosome.

**Interspecies scaffolding**

To determine whether interspecies scaffolding would be useful, we made a *de novo* assembly of the genome of *A. certus* using MaSuRCa with Illumina paired-end reads. We scaffolded this assembly with njoing using the GLS assembly of *A. atriplicis*. We evaluated the assemblies of *A. certus* using the methods described above for the assemblies of *A. atriplicis*.

**Bacterial contamination**

To find any bacterial contamination in the final assemblies, we ran blastn with the assembly contigs against the NCBI nt database restricted to taxa in Insecta or Bacteria (using parameters -evalue 1e-25, -max_hsps 1, and -max_target_seqs 10; Altschul et al. 1997; Coordinators 2016). We then ran Blobtools (Laetsch and Blaxter, 2017) to assign the most likely taxa for the hits, and to visualize the contig GC contents and mapping depths compared to the assigned taxa. We removed the contigs with bacterial sequences from the final assemblies for downstream analyses.

**Repeats and transposable elements**

We searched for repetitive DNA with GenomeScope (Ranallo-Benavidez et al. 2020), which uses input from Jellyfish (Marcais and Kingsford 2011). We also searched for transposons with the Extensive de novo TE Annotator (EDTA) pipeline (Ou et al. 2019). This pipeline runs LTR_FINDER (Xu and Wang, 2007, Ou and Jiang, 2019), LTRharvest (Ellinghaus et al. 2008), and LTR_retriever (Ou and Jiang, 2018) to find all of the LTRs in the genome. It also runs Generic Repeat Finder (Shi and Liang, 2019), HelitronScanner (Xiong et al. 2014), and TIR-Learner (Su et al. 2019) to find non-LTR transposons. It then runs a series of filters and RepeatModeler (Flynn et al. 2020) to remove duplicates, identify full-length transposable elements (TEs), and create a final TE library. The EDTA pipeline identifies repeats *de novo* based on their sequences and the attributes of known TE families.

**Mitochondrial genomes**

To find the mitochondrial genome in the *A. atriplicis* assembly, we used blastn to compare the *A. atriplicis* assembly with the published assembly of the mitochondrial genome of *Encarsia formosa* (Zhu et al. 2018), which is in another genus of Aphelinidae which is the same family as Aphelinus. We identified one contig (ctg280) in the assembly of *A. atriplicis* that appeared to harbor the complete mitochondrial genome. We submitted this contig to Mitos2 (Donath et al. 2019) to identify mitochondrial features and determine their lengths and order.

To find the mitochondrial genome in the *A. certus* assembly, we first used blastn to compare contig 280 from the *A. atriplicis* final assembly with the *A. atriplicis*-scaffolded assembly of *A. certus*. We found 12 contigs in the *A. certus* assembly that together appeared to harbor the complete mitochondrial genome. We concatenated these contigs in the order in which they occurred in contig 280, and we then used Mitos2 to identify mitochondrial features and determine their lengths and order.

**Genome annotation**

Using Augustus with the Nasonia gene model (Stanke et al. 2008), we identified protein coding regions in the CSA, meta, and GLS assemblies of *A. atriplicis* and in the MaSuRCa and GLS-scaffolded assemblies of *A. certus* (Figure 1, step 11). To determine the likely functions of these genes, we compared their amino acid sequences to proteins in the nr database (ncbi.nlm.nih.gov) using blastp (with the BLOSUM62 scoring matrix, E-value = 0.001, and the default values for other parameters; Altschul et al. 1997; Coordinators 2016). We analyzed gene ontology annotations with blast2GO (Götz et al. 2008) and functional domains using
InterProScan (Jones et al. 2014). We then compared the numbers of genes, functional annotations, and intron/exon structure among the assemblies.

**Results**

**Initial genome assemblies**

The *de novo* assemblies of *A. atriplicis* differed in values for seven representative metrics of genome completeness (Figure 2 and Table 1). Other metrics are given in Supplementary Table S2. The assemblies had better values for some metrics but not others. For example, the DBG2OLC assembly made with error-corrected data was closest in size to the estimate from flow cytometry and had the fewest contigs compared to other assemblies made using error-corrected data, but the CSA assembly made with error-corrected data had the highest values for five of the seven metrics and had the highest average ratio for these metrics for both raw and error-corrected data. The numbers of complete, single-copy BUSCO genes of Hymenoptera were high and differed little among assemblies (Table 1) and other BUSCO metrics also differed little among assemblies (Supplementary Table S2).

**Meta-assembly**

Meta-assembly of the *A. atriplicis* genome using results from four assemblers with raw data and seven assemblers with error-corrected data improved contiguity over the best single assembly, assemblers with raw data and seven assemblers with error-corrected data had the highest values for five of the seven BUSCO metrics of single-copy hymenopteran BUSCO genes that were found in RLL loci, left 1328 markers that formed four linkage groups with >200 markers per group, while the remaining 18 markers were either unlinked or formed linkage groups with <5 markers. Ordering the markers on linkage groups and filtering for genotyping errors, left 834 markers in a genetic map 1776 cM long with a mean of 2 cM between markers and relative sizes of the linkage groups of 39%, 27%, 18% and 15% of the genome. These sizes matched well the relative sizes of the four chromosomes (40%, 26%, 18%, 16% of the genome) in the *A. atriplicis* karyotype (Gokhman et al. 2017). Given the genetic map size and a flow-cytometry genome size of 353 Mb (Gokhman et al. 2017), the mean recombination rate was 5 cM/Mb.

**Genetic linkage scaffolding**

Of the 834 markers in our genetic map of *A. atriplicis*, we used 224 markers to place 48 meta-assembly scaffolds on chromosomes (Figure 4 and Supplementary Figure S3). Together these scaffolds comprised 290 Mb, which is 85% of both the meta-assembly and GLS assembly sizes and 82% of the 353-Mb genome size estimated from flow cytometry. Of the sequence data in the meta-assembly, 53 Mb, i.e., 15%, could not be placed on chromosomes because 536 meta-assembly scaffolds lacked genetic markers. This left an 11-Mb difference between the sizes of the meta- and GLS assemblies and the 353 Mb estimated from flow cytometry. The percent of sequence captured on each chromosome in the GLS assembly differed among the chromosomes from flow-cytometry and karyotype: 75%, 81%, 95%, 82%, respectively, for chromosomes 1, 2, 3, and 4.

Genetic scaffolding increased $N_{50}$ eight-fold from 9 to 74 Mb, and decreased $L_{50}$ six-fold from 12 to 2 (Table 2). $L_{75}$ was 4, number of chromosomes in *A. atriplicis* and other species in the varipes complex of *Aphelinus* (Gokhman et al. 2017). Although the average recombination rate across the genome was 5 cM/Mb, recombination rate varied within Chromonomer scaffolds from 1 to 60 cM/Mb with a median of 9 cM/Mb.

Genetic scaffolding did not increase the amount of the genome captured compared to the size estimated from flow cytometry, but it did slightly increase the number of complete, single-copy hymenopteran BUSCO genes that were found (Table 2; Supplementary Table S2).

**Interspecific scaffolding**

Scaffolding the *de novo* assembly of *A. certus* (made with Illumina data and MaSuRCa) using the GLS assembly of the *A. atriplicis* genome enormously improved the contiguity of the *A. certus* assembly (Table 3). Scaffolding reduced the number of contigs by half, increased the $N_{50}$ 2316-fold, decreased the $L_{50}$ 1303-fold, and $L_{75}$
Table 1. Seven completeness metrics for *Aphelinus atriplicis* genome assembled with seven assemblers. BUSCO values are for complete, single-copy orthologs in the hymenopteran set.

| Assembler | Flye | CSA | DBG2OLC | Meiac | Mecat | Miniasm | Raven | Raw PacBio | Corrected PacBio |
|-----------|------|-----|---------|-------|-------|---------|-------|------------|------------------|
| Size (Mb) | 346  | 346 | 346     | 346   | 346   | 346     | 346   | 346        | 346              |
| Deviation from flow cytometry size (Mb) | 7    | 7   | 7       | 7     | 7     | 7       | 7     | 7          | 7                |
| Contigs | 887  | 887 | 887     | 887   | 887   | 887     | 887   | 887        | 887              |
| nCONTIGs | 2.1  | 2.1 | 2.1     | 2.1   | 2.1   | 2.1     | 2.1   | 2.1        | 2.1              |
| N50 (Mb) | 47   | 47  | 47      | 47    | 47    | 47      | 47    | 47         | 47               |
| Longest contig (Mb) | 9.1 | 9.1 | 9.1     | 9.1   | 9.1   | 9.1     | 9.1   | 9.1        | 9.1              |
| BUSCO Number | 5350 | 5350 | 5350 | 5350 | 5350 | 5350 | 5350 | 5350 | 5350 |
| BUSCO Percent | 88.9 | 88.9 | 88.9 | 88.9 | 88.9 | 88.9 | 88.9 | 88.9 | 88.9 |

The assembly of the *A. atriplicis* genome had 332 Mb of arthropod sequence, 9 Mb of unidentified sequence, and 1.9 Mb of viral sequence, the latter being 0.6% of the assembly size. The GC-by-coverage plot for *A. atriplicis* revealed two clusters of contigs with viral contamination (Supplementary Figure S4a), one which is Wolbachia (order Rickettsiales), a known endosymbiont of these *Aphelinus* species, and the other is Buchnera (order Enterobacterales), an endosymbiont of aphids. There were 66 contigs with Wolbachia sequences, together comprising 1.3 Mb, which is in the size range of the complete Wolbachia genome (Scholz et al. 2020). There were six contigs with Buchnera sequences, together comprising 0.6 Mb, which is the size of the Buchnera genome (Moran and Mira 2001). The Buchnera contamination probably resulted from residual material in the parasitoid guts of the large numbers of parasitoids used for PacBio sequencing.

The assembly of the *A. certus* genome had 292 Mb of arthropod sequence, 42 Mb of unidentified sequence, and 1.9 Mb of viral sequence, the latter being 0.6% of the assembly size of 334 Mb. The GC-by-coverage plot for *A. certus* revealed several clusters of contigs with bacterial contamination (Supplementary Figure S4b), one which is Wolbachia. There were 338 contigs with Wolbachia sequences comprising 1.87 Mb, which is 0.2 Mb larger than the maximum size reported for the complete Wolbachia genome (Scholz et al. 2020). There were 24 contigs with other bacterial contaminants in the *A. certus* final assembly that comprised 0.07 Mb of sequence but did not include Buchnera. The difference in the presence of Buchnera between the *A. atriplicis* and *A. certus* assemblies likely results from the different numbers and sexes of individuals involved in library preparation: for the *A. certus* libraries, a small number of males were used, while for the *A. atriplicis* PacBio libraries, hundreds of females and males were used. Because *Aphelinus* females host-feed to get nutrients for egg production, adult females may have remains of aphids and aphid symbionts in their guts. Contigs annotated as harboring bacteria were removed from the final assemblies of the *A. atriplicis* and *A. certus* genomes and submissions of these assemblies to NCBI.

Repeats and TEs

Analyses of repetitive content with Jellyfish and GenomeScope identified 70 Mb of the *A. atriplicis* assembly and 62 Mb of the *A. certus* assembly as repetitive. Analysis with EDTA of the assembly of the *A. atriplicis* genome revealed 95 Mb or 28% of the assembly in TEs, including 53 Mb in DNA TEs, 41 Mb in LTR retrotransposons, and 1 Mb in MITEs. The major transponson families included the DNA TE mutator family (7%), helitrons (4%), and the LTR Gypsy family (7%). Almost all of the TEs were non-intact, with only 1291 intact TEs out of 228,000 found, i.e., less than 1% were intact. This suggests most of the TEs are decaying and may be subsequently removed from the genome. In the assembly of the *A. certus* genome 67 Mb or 20% in TEs, including 52 Mb in
DNA TEs, 13 Mb in LTR retrotransposons, and 2 Mb in MITEs. The lower amount of sequence in LTR retrotransposons in *A. certus* compared to *A. atriplicis* may be because LTRs are long and not fully assembled in *A. certus*.

**Mitochondrial genomes**

Using Mitos2 with contig 280 which harbored the mitochondrial genome of *A. atriplicis*, we found that the feature lists were almost identical between *A. atriplicis* and *E. formosa*, the three exceptions being tRNAs (Supplementary Figure S5a). The mitochondrial feature list for *A. certus* was almost identical to that of *E. formosa*, the only exceptions being one tRNA present in one assembly and not in the other and another tRNA with the opposite pattern. The tRNAs involved were different from those in the *A. atriplicis* assembly. The feature lengths were identical or very similar between *A. certus* and *E. formosa*, the one exception being the sequences for 16S rRNA which was shorter in the *A. atriplicis* assembly, and the feature orders were almost identical between *A. atriplicis* and *E. formosa*, the three exceptions being tRNAs (Supplementary Figure S5a). The mitochondrial feature list for *A. certus* was almost identical to that of *E. formosa*, the only exceptions being one tRNA present in one assembly and not in the other and another tRNA with the opposite pattern. The tRNAs involved were different from those in the *A. atriplicis* assembly. The feature lengths were identical or very similar between *A. certus* and *E. formosa*, the two exceptions being the sequences for 16S rRNA and NADH5 which were shorter in the *A. certus* assembly, and the orders of protein-coding genes and rRNAs were almost identical between *A. certus* and *E. formosa*.
We found about 29,000 genes in the CSA, meta, and GLS assemblies of *A. atriplicis* (Table 4). For each assembly, the sequences in these genes comprised 25–26% of the genome assemblies. Although meta-assembly and genetic scaffolding increased contiguity metrics, these improvements had little effect on the number of genes called, changing this by <2% (Table 4). However, 237 more genes were called in the meta-assembly than in the CSA assembly, and some of these genes could be important for evolution. Fewer genes were called in the GLS assembly than

![Genetic and physical maps of the Aphelinus atriplicis genome. Genetic map (shown in red) gives physical positions of 224 genetic markers that were used to scaffold 48 contigs (shown in alternating gray and black) in our meta-assembly that comprise 82% of the genome of *A. atriplicis*, as estimated from flow cytometry.](image)

**Table 3** Change in assembly metrics with interspecific scaffolding

| Metric                                | de novo (MaSuRCA) | A. atriplicis scaffolded (ntJoin) |
|---------------------------------------|-------------------|----------------------------------|
| Total length (Mb)                     | 331               | 334                              |
| Deviation from flow cytometry size (Mb)| 31                | 28                               |
| n Contigs                             | 27,241            | 13,750                           |
| N50 (kb)                              | 23                | 54,006                           |
| Largest contig (Mb)                   | 0.27              | 96                               |
| L50                                   | 3908              | 3                                |
| L75                                   | 8993              | 4                                |
| BUSCO Number                          | 4974              | 5383                             |
| Percent                               | 83                | 90                               |

De novo assembly of *Aphelinus certus* using Illumina data compared to that assembly scaffolded with GLS assembly of *Aphelinus atriplicis*. BUSCO values are for complete single-copy orthologs in the hymenopteran set.

**Genome annotation**

We found about 29,000 genes in the CSA, meta, and GLS assemblies of *A. atriplicis* (Table 4). For each assembly, the sequences in these genes comprised 25–26% of the genome assemblies. Although meta-assembly and genetic scaffolding increased contiguity metrics, these improvements had little effect on the number of genes called, changing this by <2% (Table 4). However, 237 more genes were called in the meta-assembly than in the CSA assembly, and some of these genes could be important for evolution. Fewer genes were called in the GLS assembly than
in meta-assembly presumably because the GLS assembly had the Wolbachia genome removed. Increased contiguity had almost no effect on the proportion of genes with homologs on GenBank, gene ontology mappings and annotations, or InterProScan annotations (Table 4).

Scaffolding the de novo assembly of *A. certus* with the GLS assembly of *A. atriplicis* greatly increased the contiguity of the *A. certus* assembly, and this affecting gene annotation. Although the number of genes decreased by 2003 or 7% from the meta-assembly to the atriplicis-scaffolded assembly of *A. certus*, the total length in genes increased by 7 Mb or 9% because the gene lengths increase by 17%. However, there was little or no change in the proportion of these genes with homologs on GenBank, GO mappings and annotations, or InterProScan annotations (Table 4).

**Discussion**

Meta-assembly, also known as assembly reconciliation, is an increasingly popular approach to providing higher quality genome assemblies (Wenczes and Schatz 2015; Babb et al. 2017; Torresen et al. 2017; Patel et al. 2018; Song et al. 2019; Tang et al. 2020). Assembly reconciliation takes advantage of differences in the sequence information captured by assemblers that differ in how well they deal with coverage level, repetitive or duplicated sequence information captured by assemblers that differ in how assembly reconciliation programs did not consistently improve quality as cover gaps between contigs produced by other assemblers. Tang et al. (2020) reviewed some of the same assembly-reconciliation programs, proposed a new one, and found that theirs gave modest increases in contiguity and decreases in misassembly compared to the best de novo assembly and other assembly-reconciliation programs. Our results with ntJoin gave a meta-assembly with a threefold increase in contiguity and no misassemblies, and these are greater improvements than described by Alhakami et al. (2017) and Tang et al. (2020). However, ntJoin (Coombe et al. 2020) was published after Alhakami et al. (2017) and Tang et al. (2020) published their comparisons and were not included in their analyses. ntJoin uses an alignment-free method of minimizer graphs and resolves discrepancies using the path with support in the highest number of input assemblies, whereas other meta-assemblers are alignment based. We used seven assemblers, involving long-read PacBio data and short-read Illumina plus long-read PacBio data, and raw vs error-corrected PacBio reads to capture diverse information. Meta-assembly contiguity statistics improved as the number input assemblies was increased (Supplementary Figure S6). When only using two or three input assemblies, a meta-assembly can give results worse than any of the input assemblies because discrepancies between the assemblies cannot be resolved under equal weighting.

Long-distance scaffolding with the Hi-C and other technologies has become available, is relatively cheap, and provides a much-used alternative to scaffolding with genetic linkage data (Neely et al. 2011; Burton et al. 2013). However, during the last decade, over 200 papers, including over 30 in 2020–2021, have been published on scaffolding assemblies with genetic linkage data in a wide variety of organisms, like fungi (Celton et al. 2010; Jiang et al. 2021), wild and cultivated plants (Song et al. 2016; Bellinger et al. 2020; Bernhardsson et al. 2021; Paritosh et al. 2021), aquatic mollusks and arthropods (Dukic et al. 2016; Yin et al. 2020; Penalozza et al. 2021), fish (de los Rios-Perez et al. 2020; Kai et al. 2011), insects (Davey et al. 2016; Niehuis et al. 2010; Kosova et al. 2019; Wallberg et al. 2019), reptiles (Yurchenko et al. 2020), and wild and domesticated birds and mammals (Dodson et al. 2011; Li et al. 2016; Batra et al. 2020; Hagen et al. 2020). Genetic maps have been published for nonparasitic species of Hymenoptera species (Hunt and Page 1995; Beye et al. 1999; Wilfert et al. 2006, 2007; Mezmar et al. 2010; Sirvio et al. 2011; Stolle et al. 2011), and for hymenopteran parasitoids in particular (Laurent et al. 1998; Holloway et al. 2000; Niehuis et al. 2010; Pannebakker et al. 2011; Desjardins et al. 2013; Matthey-Doret et al. 2019; Ma et al. 2021; Ulrich et al. 2021). Genetic map size varied among these hymenopteran species from ~450 cM for *Nasonia vitripennis* (Hymenoptera: Pteromalidae; Niehuis et al. 2010; Pannebakker et al. 2011; Desjardins et al. 2013) to ~3500 cM for *Apis mellifera* (Hunt and Page 1995; Beye et al. 1999). The genetic map size (1776 to 1806 cM) for *Aphelinus atriplicis* (Table 4).

![Table 4 Gene annotation metrics of assemblies](image-url)
we measured for *A. atriplicis* was intermediate between these extremes, and resembles the 1300 cM found for *Cotesia typhae* (Hymenoptera: Braconidae; *Benoist et al. 2020*) and the 1330 cM found for *Trichogramma brassicae* (Hymenoptera; *Trichogrammatidae; Laurent et al. 1998*). Most of the variation in genetic map sizes arises from differences in recombination rates, which vary an order of magnitude from 1.5 cM/Mb for *N. vitripennis* to 19 cM/Mb for *A. mellifera*, rather from differences in the physical sizes of the genomes, which are similar for these two species. At 5 cM/Mb, the mean recombination rate for our GLS assembly of *A. atriplicis* was intermediate among the published values (for review, see *Wilfert et al. 2007*). In many cases, genetic maps were generated for reasons other than improving genomes assembly and primarily for mapping QTL. However, the linkage data have been successfully repurposed for scaffolding genomic assemblies, and indeed the genetic map we describe here was generated for mapping QTL. Genetic linkage has been used to scaffold an assembly of *N. vitripennis* to its chromosomes (*Desjardins et al. 2013*). Thus, although starting from scratch, Hi-C and other molecular technologies are cheaper than doing crosses or pedigrees to generate linkage maps, if one already has a genetic map the only cost for using it for scaffolding an assembly is the cost of bioinformatic analysis.

Although detection and decontamination of the final assemblies revealed the complete genomes of *Wolbachia* in each *Aphelinus* species, it had very little effect on the statistics of genome assembly or gene annotation. This is because the contaminant DNA represented less than 1% of each final assembly. The greater fragmentation of both the *Wolbachia* and mitochondrial genomes in the *A. certus* assembly compared to the *A. atriplicis* assembly not doubt reflects using only Illumina short-read sequencing for *A. certus* vs Illumina short-read plus PacBio long-read sequencing for *A. atriplicis*. This is particularly true because genetic-scaffolding did not help in assembly of either the *Wolbachia* genome or the mitochondrial genome. Both genomes ended up as unscaffolded contigs in both species, which is not surprising given that the genetic map used for scaffolding was based on meiotic recombination absent in *Wolbachia* or mitochondria.

Increased contiguity of the assemblies of *A. atriplicis* had little effect on assembly completeness, as measured by percentage of complete BUSCO hymenopteran genes. This lack of effect arose either because most, if not all, of the complete BUSCO genes in *A. atriplicis* were found in the *de novo* assembly with CSA or because a small number of such genes were present but not found in the more contiguous assemblies either. The BUSCO gene set for Hymenoptera (odb10) comprises 5991 genes, which is one-fifth of all the genes we identified in *A. atriplicis*. Whether these core genes would be identified more readily than the remaining genes is unclear. However, increased contiguity also led to little or no increases in the numbers of other genes we identified or the numbers with annotations. Again, this could be because all genes were identified in the *de novo* assembly with CSA, leaving few additional genes to call in more contiguous assemblies or because genes present, but not found, in the *de novo* assembly were also not found in the more contiguous assemblies. The CSA assembly did not include 3.2 Mb of sequence (about 1%) that was in the GLS assembly. This included 310 kb of gene sequence with 43% being intronic. This suggests that the more contiguous assembly did not add much coding sequence. Although there were 45 gaps in the meta-assembly and an additional 44 gaps in the GLS assembly, together these comprised about 0.7 Mb of sequence data. Thus, these gaps were neither numerous nor large enough to have much effect on gene identification or annotation.

Comparative genomics can inform assembly of genomes of closely related species (*Coome et al. 2020, Waterhouse et al. 2020*). In the present study, scaffolding a low-contiguity *de novo* assembly of *A. certus* with the high-contiguity GLS assembly of the *A. atriplicis,* greatly improved contiguity of the *A. certus* assembly. In a way, this is not surprising, given that these two species have similar genome sizes (*Gokhman et al. 2017*), are close phylogenetically (*Heraty et al. 2007*), and can hybridize in the laboratory (*Heraty et al. 2007*). However, the amount of improvement in contiguity surprised us: over 2000-fold increase in *N₀* and over 1000-fold decrease in *L₅₀*. Increased contiguity provided modest improvements in assembly completeness. The percentage of complete single-copy BUSCO hymenopteran genes increased by 7% or 410 genes. Furthermore, the total sequence identified as genes increased, and while the number of genes declined, gene length increased, which together suggests greater accuracy of gene models and less gene fragmentation.

Although increased contiguity had no effects on the numbers of genes called or annotated for *A. atriplicis* and only modest effects for *A. certus*, increase contiguity would be very useful for other purposes. As *Waterhouse et al. 2020* have recently pointed out, increased contiguity would aid in identifying genes associated with QTL and determining how evolution has affected chromosome structure, including fusions, splits, and other rearrangements, which could in turn affect reproductive isolation and thus speciation. Indeed, the reasons we have assembled these and other genomes of *Aphelinus* are to find QTL and the underlying genes involved in the evolution of host specificity and climatic adaptation and to study mechanisms of reproductive isolation among the many closely related species in this genus.

**Conclusions**

Our results showed that meta-assembly of multiple assemblies based on short-read and long-read libraries with different assemblers increases assembly contiguity. However, genetic-linkage scaffolding of the meta-assemblies increases contiguity by orders of magnitude. We cannot say whether other methods of scaffolding would give similar or better effects on contiguity.

**Data availability**

DNA sequence data and final assemblies are archived at NCBI (www.ncbi.nlm.nih.gov) under BioProject number PRJNA767831. Supplementary material is available at G3 online.

**Acknowledgments**

We thank Karol Miaskiewicz at University of Delaware Bioinformatics Core Facility for help with the BIOMIX computational cluster and Bruce Kingham, Olga Shevchenko, and Erin Bernberg at the DNA Sequencing & Genotyping Center of the Delaware Biotechnology Institute for assistance with preparing and sequencing DNA libraries. We thank Kathryn Lanier for maintaining the cultures of *Aphelinus certus* and its host and assisting with the crosses.
**Funding**

This research was supported by the United States Department of Agriculture, Agricultural Research Service Project Number 8010-22000-032-00D, as well as USDA-NIFA grant 2017-67013-26526 to K.R.H. Support from the University of Delaware Bioinformatics Core Facility, including use of the BIOMIX computational cluster, was made possible by the Delaware IDEA Network of Biomedical Research Excellence (National Institutes of Health grant P20 GM103446).

**Conflicts of interest**

The authors declare that there are no conflicts of interest.

**Literature cited**

Alhakami H, Mirebrahim H, Lonardi S. 2017. A comparative evaluation of genome assembly reconciliation tools. Genome biology. 18:93. doi: 10.1186/s13059-017-1213-3.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Andrews S. 2010. Fastqc: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, et al. 2017. The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. Nat Genet. 49: 895–903.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced rad markers. PLoS One. 3.e3376.

Batra SS, Levy-Sakin M, Robinson J, Guillory J, Durinck S, et al. 2020. Accurate assembly of the olive baboon (Papio amnis) genome using long-read and hi-c data. Gigascience. 9:giaa134. 10.1093/gigascience/giaa134

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, et al. 2011. Linkage mapping and comparative genomics using next-generation rad sequencing of a non-model organism. PLoS One. 6:e19315.

Bellinger MR, Fautel R, Starrnes S, Kambic L, Kantar MB, et al. 2020. Taro genome assembly and linkage map reveal qtls for resistance to taro mosaic blight. G3 (Bethesda). 10:2763–2775.

Benoist R, Capdevielle-Dulac C, Chantre C, Jeannette R, Calatayud V, et al. 2020. Quantitative trait loci involved in the reproductive success of a parasitoid wasp. Mol Ecol. 29:3476–3493.

Bernhardsson C, Zan YJ, Chen ZQ, Ingvarsson PK, Wu HX. 2021. Development of a highly efficient 50k single nucleotide polymorphism genotyping array for the large and complex genome of Norway spruce (Picea abies I. Karst) by whole genome resequencing. Gigascience. 10:19315.

Beyer M, Hunt GJ, Page RE, Fondrk MK, Grohmann L, et al. 1999. Unusually high recombination rate detected in the sex locus region of the honey bee (Apis mellifera). Genetics. 153:1701–1708.

Bole AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 30:2114–2120.

Bradshaw CJA, Leroy B, Bellard C, Roiz D, Albert C, et al. 2016. Massive yet grossly underestimated global costs of invasive insects. Nat Commun. 7:12986.

Broman KW. 2010. Genetic map construction with r/qtl. Madison, Wisconsin, technical Report 214. University of Wisconsin, Department of Biostatistics & Medical Informatics.

Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: qtl mapping in experimental crosses. Bioinformatics. 19:889–890.

Burton JN, Adey A, Patwardhan RP, Qui RL, Kitzman JO, et al. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 31:1119–1125.

Catchen J, Amores A, Bassham S. 2020. Chromonomer: a tool set for repairing and enhanced assembled genomes through integration of genetic maps and conserved synteny. G3 (Bethesda). 10:4115–4128.

Celton JM, Christoffels A, Sargent DJ, Xu XM, Rees DJG. 2010. Genome-wide SNP identification by high-throughput sequencing and selective mapping allows sequence assembly positioning using a framework genetic linkage map. BMC Biol. 8:155.

Cock MJW, Murphy ST, Kairo MTK, Thompson E, Murphy RJ, et al. 2016. Trends in the classical biological control of insect pests by insects: an update of the biocat database. Biocontrol. 61:349–363.

Coombe L, Nikolić V, Chu J, Birol I, Warren RL. 2020. Ntojoin: fast and lightweight assembly-guided scaffolding using minimizer graphs. Bioinformatics. 36:3885–3887.

Coordinators NR. 2016. Database resources of the national center for biotechnology information. Nucleic Acids Res. 44: D7–D19.

Davey JW, Chouteau M, Barker SL, Maroja L, Baxter SW, et al. 2016. Major improvements to the Heliconius melpomene genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. G3 (Bethesda). 6:695–708.

de los Rios-Perez L, Nguinkal JA, Verlieh M, Rebl A, Brunner RM, et al. 2020. An ultra-high density SNP-based linkage map for enhancing the pike perch (Sander lucioperca) genome assembly to chromosome-scale. Sci Rep. 10.

DePristo MA, Banks E, Poplin R, Garmire KV, Maguire JR, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491–498.

Desjardins CA, Gadou J, Lopez JA, Niehuis O, Avery AR, et al. 2013. Fine-scale mapping of the nasonia genome to chromosomes using a high-density genotyping microarray. G3 (Bethesda). 3:205–215.

Di Genova A, Buena-Atenza E, Ossowski S, Sagot M-F. 2021. Efficient hybrid de novo assembly of human genomes with Wengan. Nat Biotechnol. 39:422–430.

Dodgson JB, Delany MF, Cheng HH. 2011. Poultry genome sequences: progress and outstanding challenges. Cytogenet Genome Res. 134:19–26.

Donath A, Jähling F, Al-Arab M, Bernhart SH, Reinhardt F, et al. 2019. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Res. 47:10543–10552.

Du H, Liang C. 2019. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nat Commun. 10:1–10.

Dukic M, Berner D, Roesti M, Haag CR, Ebert D. 2016. A high-density genetic map reveals variation in recombination rate across the genome of Daphnia magna. BMC Genet. 17.

Ellinghaus D, Kurtz S, Willhoef U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC bioinformatics. 9:18. doi: 10.1186/1471-2105-9-18.

Flyn JM, Hubley R, Goubert C, Rosen J, Clark AG, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America. 117:9451–9457. doi: 10.1073/pnas.1921046117.
plant systems. In: MR Berenbaum, editor. Annual Review of Entomology, Volume 60. p. 621–645.
Neely RK, Deen J, Hofkens J. 2011. Optical mapping of DNA: single-molecule-based methods for mapping genomes. Biopolymers. 95:298–311.
Niehuis O, Gibson JD, Rosenberg MS, Pannebakker BA, Koevoets T, et al. 2010. Recombination and its impact on the genome of the haplodiploid parasitoid wasp nasonia. PLoS One. 5:e8597.
Noyes JS. 2019. Universal chalcidoidea database. World Wide Web electronic publication. www.nhm.ac.uk/entomology/chalcidoids/index.html.
Ou S, Jiang N. 2019. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mobile DNA. 10:48. Doi: 10.1186/s13100-019-0193-0.
Ou S, Jiang N. 2018. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant physiology. 176:1410–1422. Doi: 10.1104/pp.17.01510.
Ou S, Su W, Liao Y, Chougule K, Agda JRA, et al. Benchmarking transposable element annotation methods for creation of a streamlined, defined pipeline. Genome biology. 20:275. Doi: 10.1186/s13059-019-1905-y.
Pannebakker BA, Watt R, Knott SA, West SA, Shuker DM. 2011. The construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly glyma1.01. BMC Genomics. 17.
Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cdna alignments to improve de novo gene finding. Bioinformatics. 24:637–644.
Stolle E, Wilfert L, Schmid-Hempel R, Schmid-Hempel P, Kube M, et al. 2011. A second generation genetic map of the bumblebee bombus terrestris (linnaeus, 1758) reveals slow genome and chromosome evolution in the apidae. BMC Genomics. 12.
Su W, Xu X, Peterson T. 2019. TIR-Learner, A New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. Molecular plant. 12:447–460. Doi: 10.1016/j.molp.2019.02.008.
Tang L, Li M, Wu FX, Fan Y, Wang JX. 2020. Mac: merging assemblies by using adjacency algebraic model and classification. Front Genet. 10:1396.
Torresen OK, Star B, Jenstoøt S, Reinar WB, Grove H, et al. 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics. 18.
Ulrich GF, Zemp N, Vorburger C, Boulain H. 2021. Quantitative trait locus analysis of parasitoid counteradaptation to symbiont-conf erred resistance. Heredity. 127:219–232. Doi: 10.1038/s41437-021-00444-7.
Vaser R, Sikić M. 2020. Raven: a de novo genome assembler for long reads. bioRxiv. 2–10. 10.1101/2020.08.07.242461.
Walker BJ, Abeel T, Shea T, Priest M, Abouelhail A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.
Wallberg A, Bunikis I, Petersson OV, Mosbeck MB, Childers AK, et al. 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genomics. 20.
Waterhouse RM, Aganezov S, Anselmetti Y, Lee J, Ruzzante L, et al. 2020. Evolutionary superscaffolding and chromosome anchoring to improve anopheles genome assemblies. BMC Biol. 18:1.
Wences AH, Schatz MC. 2015. Metassembler: merging and optimizing de novo genome assemblies. Genome Biol. 16.
Wilfert L, Gadau J, Schmid-Hempel P. 2006. A core linkage map of the bumblebee bombus terrestris. Genome. 49:1215–1226.
Wilfert L, Gadau J, Schmid-Hempel P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. Heredity (Edinb). 98:189–197.
Wu Z, Hopper KR, O’Neil RJ, Voegtlin DJ, Prokrym DR, et al. 2004. Reproductive compatibility and genetic variation between two strains of Aphelinus alipidus (hymenoptera: aphelinidae), a parasitoid of the soybean aphid, Aphis glycines (homoptera: aphi- dae). Biol Control. 31:311–319.
Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, et al. 2017. Mecat: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods. 14:1072–1074.
Xiong W, He L, Lai J, Dooner HK, Du C. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences of the United States of America. 111:10263–10268. Doi: 10.1073/pnas.140068111.
Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research. 35:W265–W268. Doi: 10.1093/nar/gkm286.
Xu M, Guo L, Gu S, Wang O, Zhang R, et al. 2020. Tgs-gapcloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience. 9:1–11.
Ye C, Hill CM, Wu S, Ruan J, Ma Z. 2016. Dbg2olc: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep. 6:31900–31909.
Yin XS, Arias-Perez A, Kitapci TH, Hedgecock D. 2020. High-density linkage maps based on genotyping-by-sequencing (gbs) confirm a chromosome-level genome assembly and reveal variation in recombination rate for the pacific oyster Crassostrea gigas. G3 (Bethesda). 10:4691–4705.

Yurchenko AA, Recknagel H, Elmer KR. 2020. Chromosome-level assembly of the common lizard (Zootoca vivipara) genome. Genome Biol Evol. 12:1953–1960.

Zhu JC, Tang P, Zheng BY, Wu Q, Wei SJ, Chen XX. 2018. The first two mitochondrial genomes of the family aphelinidae with novel gene orders and phylogenetic implications. Int J Biol Macromol. 118:386–396.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, et al. 2013. The masurca genome assembler. Bioinformatics. 29:2669–2677.

Communicating editor: J. Hesselberth