https://doi.org/10.1093/nar/gkac510

NAR Breakthrough Article

De novo sequencing, diploid assembly, and annotation of the black carpenter ant, *Camponotus pennsylvanicus*, and its symbionts by one person for $1000, using nanopore sequencing

Christopher Faulk*

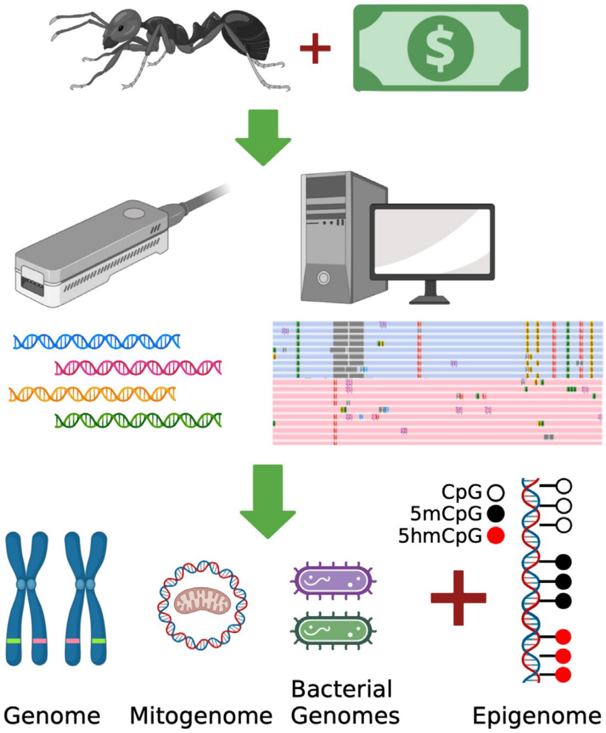
Department of Animal Science, University of Minnesota, College of Food, Agricultural and Natural Resource Sciences, 1334 Eckles Ave, 350 ABLMS, St. Paul, MN 55108, USA

Received April 01, 2022; Revised May 19, 2022; Editorial Decision May 20, 2022; Accepted May 31, 2022

ABSTRACT

The black carpenter ant (*Camponotus pennsylvanicus*) is a pest species found widely throughout North America. From a single individual I used long-read nanopore sequencing to assemble a phased diploid genome of 306 Mb and 60X coverage, with quality assessed by a 97.0% BUSCO score, improving upon other ant assemblies. The mitochondrial genome reveals minor rearrangements from other ants. The reads also allowed assembly of parasitic and symbiont genomes. I include a complete Wolbachia bacterial assembly with a size of 1.2 Mb, as well as a commensal symbiont *Blochmannia pennsylvanicus*, at 791 kb. DNA methylation and hydroxymethylation were measured at base-pair resolution level from the same reads and confirmed extremely low levels seen in the Formicidae family. There was moderate heterozygosity, with 0.16% of bases being biallelic from the parental haplotypes. Protein prediction yielded 14 415 amino acid sequences with 95.8% BUSCO score and 86% matching to previously known proteins. All assemblies were derived from a single MinION flow cell generating 20 Gb of sequence for a cost of $1047 including consumable reagents. Adding fixed costs for equipment brings the total for an ant-sized genome to less than $5000. All analyses were performed in 1 week on a single desktop computer.

GRAPHICAL ABSTRACT



INTRODUCTION

Complete sequencing and annotation of novel eukaryotic genomes has traditionally only been possible with the support of large institutions and budgets. Previous advancements towards a '$1000 genome' were focused on human genomes that required a pre-existing high-quality assembly

*To whom correspondence should be addressed. Tel: +1 612 624 7216; Email: cfaulk@umn.edu and were based on short-read technology. While inexpensive in terms of sample cost, the instrumentation costs remain out of reach of individuals, and requires highly trained specialist staff. With the advent of Oxford Nanopore Technologies MinION instrument, DNA and RNA sequencing has become available to individual researchers for accessible cost. MinION flow cells range below $1000 each and have the capacity to sequence up to 30 Gb and the instrument itself retails for $1000 and includes one flow cell. With minimal additional equipment and reagents, the total cost to build a whole genome sequencing capable lab can be less than $5000. While growing in capability and dropping in cost, the capacity of a MinION flow cell to generate 30× genome coverage, necessary for *de novo* assembly, is currently only feasible for small sized genomes, <500 Mb. Fortunately, one of the most numerous animals on the planet, ants, have both a small genome and are in need of greater genomic resources (1).

The black carpenter ant, *Camponotus pennsylvanicus*, is a prolific pest ranging throughout North America east of the Rocky Mountains and comes into frequent proximity with humans causing significant property damage. Beyond their pest status, the black carpenter ant can reveal biological insights into social behavior, caste determination, and host-endosymbiont interaction, provided we have a high-quality reference genome. For example, within the family Formicidae, a close relative, the Florida Carpenter Ant *C. floridanus*, has been sequenced to high depth using long-read technology (2) and has been used to study plasticity in the biological clock (3), invertebrate calcium regulation (4) and epigenetic control of caste (5). Within the Camponotus tribe, a unique instance of endosymbiosis has occurred where a species of Blochmannia bacteria has become an obligate mutualist and has begun evolving in parallel to the host (6). Additionally, this ant, among many Hymenoptera, hosts endosymbiotic Wolbachia bacteria and can improve the study of this prolific zoonotic (7). Caste determination is still under investigation in Camponotus, which warrants examination of the epigenome. Here we provide base-level measurements of DNA methylation and DNA hydroxymethylation in *C. pennsylvanicus*.

A reference genome for the black carpenter ant will enable study of its ecology, life history traits and evolution. The ease and low expense of assembly creation and analyses opens avenues to other researchers for this ant and other small-genome insects.

MATERIALS AND METHODS

Tissue collection

The author captured a single individual carpenter ant worker (*Camponotus pennsylvanicus*) in the attic structure of a residence in St. Paul, MN on 18 February 2022. Species identification was confirmed by the UMN Extension service. Holotype ant photo is available as Supplementary Figure S1.

DNA extraction

DNA was extracted using a Zymo DNA Plus miniprep kit. The entire worker ant was homogenized using a plastic pestle inside a 1.5 ml Eppendorf tube with 300 μl of Zymo DNA shield. The standard miniprep protocol yielded ∼3 μg of total DNA. Quality was assessed using a nanospectrophotometer (Implen N60, Munich Germany) and run out on a MiniPCR bluegel student electrophoresis rig purchased from Amplyus (Cambridge, MA). Size was >10 000 bp with visible fragmentation smearing (Supplementary Figure S2).

Library preparation

Library prep was performed with an Oxford Nanopore SQK-LSK-110 kit according to manufacturer's instructions with the following changes. The total amount of 3 μg of DNA was used in a single library prep reaction and eluted with 45 μl of elution buffer EB, then split into three aliquots of 15 μl libraries to facilitate reloading of the flow cell.

DNA sequencing

Sequencing was performed using Oxford Nanopore MINKNOW software (v21.11.9) and Guppy (v5.1.15) set to fast basecalling. Post-hoc basecalling was performed using Guppy with the 'super accuracy' model (dna_r9.4.1_450bps_sup.cfg). The read length histogram is indicated in Supplementary Figure S3.

Computational methods

Detailed instructions for the pipeline used here are available as Supplementary File S1. A computer running Linux with 64 Gb of memory and an Nvidia 3060 GPU are minimum requirements. Below are abbreviated methods.

Genome assembly

Genomes were assembled with Shasta (8), Flye (9), NextDenovo (https://github.com/Nextomics/NextDenovo, and Raven (https://github.com/lbcb-sci/raven). The assembly with the highest BUSCO score was chosen for further polishing by multiple rounds of Racon and Medaka polishing. Genome assembly with Shasta v0.8.0 was performed with the following parameters,

`sudo ∼/shasta-Linux-0.8.0 –input ant-total.fastq.gz –config Nanopore-Oct2021 –memoryBacking disk –memoryMode filesystem –Reads.minReadLength 1000 –assemblyDirectory ShastaRun`.

Genome assembly with Flye v2.9 was performed with the following parameters, `flye –nano-hq ant-total.fastq –out-dir flye-results –genome-size 281m –threads 23`. The size estimate of 281 Mb came from estimates for Tsutui *et al.* for members of family Formicidae (10).

Assembly with NextDenovo v2.5.0 was run with the following command `nextDenovo run.cfg` where configuration parameters are in File S5.

Assembly with Raven v1.8.1 was performed with the following command, `raven -t 23 ../cattotal.fastq > raven-assembly.fasta`.

BUSCO v5.2.2 scores were calculated using the following command, `busco -i assembly.fasta -o busco -m genome –lineage hymenoptera-c 23`.

Polishing

Racon was run with the following parameters. `racon –cudapoa-batches 40 -t 23 ant-total.fastq.gz ant-total.sorted.sam assembly.fasta > assembly-racon1.fasta`. This step was repeated 4 times. This version of Racon was compiled with CUDA support which sped up processing to under 1 hour per cycle. After each Racon polish step, the raw reads must be re-aligned to the resulting assembly with minimap2 prior to the next polishing step (11). Medaka similarly has a GPU mode which greatly decreases polishing time to <4 h. It was run with the following parameters. Environment variable was set with `export TF_FORCE_GPU_ALLOW_GROWTH = true`, and polishing performed with `medaka_consensus -b 100 -i ant-total.fastq -d consensus-racon.fasta -o medaka-results/ -t 23 -m r941_min_sup_g507`.

Contamination removal

NCBI's megablast was used to determine contig identity. Blobtools2 was used to visualize a density plot of GC content versus genomic coverage with interactive selection of contigs by phylum and species (12). Blobtools2 local installation was used to generate the blobplot, snail plot, cumulative count and BUSCO plot as well as filtered table output. The blobtoolkit requires the consensus, coverage statistics generated with samtools, and the blast output with species and taxonomy. Details on configuration are provided in supplementary methods. Alien contigs flagged for removal were deleted manually with the nano text editor.

Commensal and organelle assembly

Total reads were individually aligned to reference genomes for *Blochmannia pennsylvanicus* (GCF_000011745.1), *Wolbachia pipientis* (GCF_014107475.1), and the mitogenome of the red fire ant (*Solenopsis invicta*, NC_014672.1) using minimap2. Flye required the '–meta' tag for the mitochondrial assembly and the '–asm-coverage 50' flag for the *B. pennsylvanicus* assembly.

Coverage assessment

Depth statistics were generated with mosdepth v0.3.3 (13).

Repeat identification

Since annotated repeats in insect genomes are sparse, repeat identification requires two stages. First *de novo* repeat identification was performed with RepeatModeler2 v2.0.2a (14). Second, the libraries generated with RepeatModeler2 were used as input to RepeatMasker v4.1.0 to create a complete genome annotation of repeats and classified using existing names for classes, families, and subfamilies from the Dfam v3.5 open source repeat library where known and novel IDs for previously unknown families (14,15). The family consensus sequences are found in Supplementary File S2, 'C_pennsylvanicus-repeat-families.fa'.

Table 1. *C. Pennsylvanicus* read summary

Number of reads	6 448 773
Number of bases	20 751 962 095
N50 read length	5113
Longest read	901 114
Shortest read	22
Mean read length	3217
Median read length	2038
Mean read quality	14.38
Median read quality	14.32

Gene annotation

Augustus v3.4.0 was used for *ab initio* protein prediction using the honeybee as nearest species (16). The masked consensus was used for prediction to eliminate false positives from open reading frames present in transposons. To identify the proteins, I used DIAMOND (17) to match predicted proteins against the NCBI non-redundant protein database, 'nr'. The complete models and annotation matches are in Supplementary File S3.

GeMoMa v1.8 was used for homology-based protein prediction using the reference transcriptome of *C. floridanus*. As with Augustus, the masked genome was used, and protein identification was internally generated by homology to IDs from the query transcriptome. Complete annotation matches and .gff file are in Supplementary File S4. All protein models were scored using BUSCO in protein mode.

Variants and diploid construction

Sequence variants were called against the haploid consensus using the PEPPER-Margin-DeepVariant v0.7 pipeline with GPU acceleration and phased haplotype output (18). Whatshap v1.3 was used to calculate statistics from the resulting 'vcf' file (19). The original consensus assembly is considered the primary haplotype. The secondary haplotype consensus was built by swapping biallelic variants from the vcf file into the primary consensus using 'bcftools consensus PEPPER_MARGIN_DEEPVARIANT_FINAL_OUTPUT.phased.vcf.gz ⟩ haplotype2.consensus.fasta'.

Epigenetic marks

DNA methylation (5mC) and hydroxymethylation (5hmC) were determined using megalodon v2.4.2. I used a base calling model capable of detecting both marks on cytosines in any context natively, res_dna_r941_min_modbases_5mC_5hmC_v001, calling the same reads as used to generate the assembly. Post processing yielded files containing 5mC or 5mC marks at CpG sites and CH sites (i.e. non-CpG sites) within the consensus assembly. Complete parameters are in Supplementary File S1.

RESULTS

Sequencing

I extracted DNA from a single diploid individual worker ant. Extraction yielded ∼3 μg of total DNA with high

Table 2. Assembly statistics

Assembly	Size (Mb)	Contigs	N50	Avg_len	BUSCO	Cov	Technique	Assembler
C. floridanus_7.5	284 009 182	657	1 278 439	432 282	96.7	53×	PacBio	Canu
C. floridanus_1.0	232 685 334	10 791	19 487	21 563	96.0	100×	Illumina	SOAPdenovo v. 1.0
***C. pennsylvanicus* assembled using all reads**								
NextDenovo	278 594 827	1225	560 840	227 424	90.1		Nanopore	NextDenovo 2.5.0
Shasta	281 462 227	3277	430 265	85 890	94.1		Nanopore	Shasta 0.8.0
Raven	295 081 168	1774	333 350	166 337	95.7		Nanopore	Raven 1.8.1
Flye	272 267 557	5516	520 016	49 360	96.8		Nanopore	Flye 2.9-b1768
***C. pennsylvanicus* assembled using reads >1000 bp**								
Flye	309 753 621	1655	565 278	187 162	96.9		Nanopore	Flye
Flye-Racon-1X	309 865 371	1648	565 400	187 681	96.9		Nanopore	Flye + racon_1X
Flye-Racon-2X	309 298 248	1641	565 935	188 189	96.8		Nanopore	Flye + racon_2X
Flye-Racon-3X	308 443 459	1633	565 664	188 882	96.9		Nanopore	Flye + racon_3X
Flye-Racon-4X	308 109 057	1625	565 657	189 606	97.0		Nanopore	Flye + racon_4X
Flye-R4X-Medaka	308 881 295	1625	566 325	190 081	97.0		Nanopore	Flye + R4X + Medaka1X
Final assembly	306 426 343	1609	565 603	190 445	97.0	60×	Nanopore	Manually curated

molecular weight, >10 kb, but with high fragmentation, as visualized on a gel (Supplementary Figure S2). DNA was sequenced using a single MinION flow cell. Base calling was performed using a GPU optimized version of Oxford Nanopore's Guppy base caller. Base calling yielded 6.4 million reads and 20.7 Gb of sequence with an N50 of 5113 bp (Table 1).

Assembly and filtering

Since no single standard exists for long-read assembly, I performed a comparative analysis on assembly and polishing pipelines and assessed efficacy by BUSCO scoring (Table 2) (20). The genome sequence of a closely related species, the Florida carpenter ant (*Camponotus floridanus*) was used as a benchmark. There are two available assemblies for *C. floridanus*, one created with Illumina short reads (C_flo 1.0) and one with PacBio long-read sequencing (C_flo 7.5). Both assemblies have >50× coverage and >96% BUSCO scores.

In order to create an unbiased assembly for *C. pennsylvanicus*, I chose a *de novo* approach, without aligning to any existing data set. I compared several assemblers capable of running with low memory overhead including Shasta, Raven, NextDenovo, and Flye (Table 2.) These programs created assemblies ranging from 272 to 295 Mb, with between 1200 and 5500 contigs. All programs assembled genomes in <6 h with BUSCO scores above 90%. Flye produced the highest quality genome with a 96.8% BUSCO score and this method was explored for further improvements to contiguity and polishing.

Assembly improves dramatically if smaller, lower quality reads are filtered out. For example, Flye with the full unfiltered reads generated a 5516 contig assembly with an N50 of 520 kb. However, when using half as many reads, by filtering for reads >5 kb and dropping the 10% of reads with the worst quality, Flye generated an assembly with only 1655 contigs, and an N50 of 565 kb, despite having nearly identical BUSCO scores. Therefore, I chose to use the filtered read set assembly with Flye for further stages. Subsequent polishing steps used the complete read set.

Four rounds of Racon polishing yielded an assembly with the highest N50 and BUSCO score and was polished a final time with Medaka. After Medaka, the BUSCO score improved to 97.0%. The Flye-Racon4X-Medaka-assembly was 309 Mb in length spread over 1625 contigs. This nanopore-only assembly had the second highest BUSCO score compared to all other currently available ant assemblies, even those produced with hybrid short and long-reads (Figure 1).

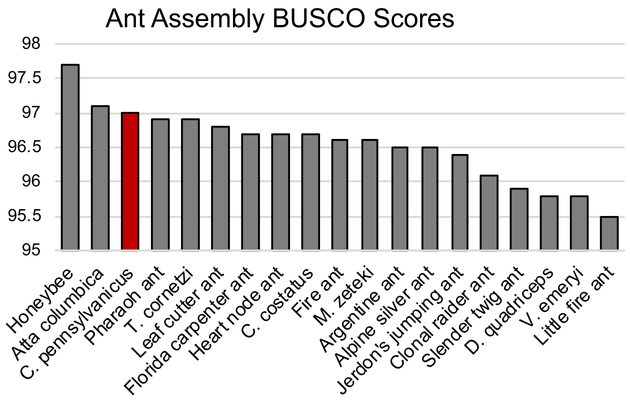


Figure 1. Ant genomes BUSCO scores. *C. Pennsylvanicus* has the second highest ant BUSCO completeness score. BUSCO scores were calculated using the reference genome for at least one member of every genus of ant available. Honeybee is included as the most complete hymenopteran available. Two ant assemblies with less than 95% completeness are not shown.

Contaminants

Careful curation of the assembly is vital to ensure that all contigs represent the true host nuclear genome. I used several measures including GC content, contig coverage, and sequence similarity to known species to filter for any spurious contigs. To identify any known sequence similarity, I first used NCBI's megablast tool to query the consensus contigs against the NCBI non-redundant nucleotide database ('nt' database). Next, blobtools2 was used to determine the extent of any extra-species inclusion into the assembly contig set (Figure 2). Of the 1625 contigs in the Flye-Racon4X-Medaka assembly, 1580 had a match to an *Arthropoda* species and 31 had no hit to any species and were kept for assembly (Supplementary Table S1). The remaining hits were considered alien contigs (i.e. contigs within an assembly that do not match to the target species).

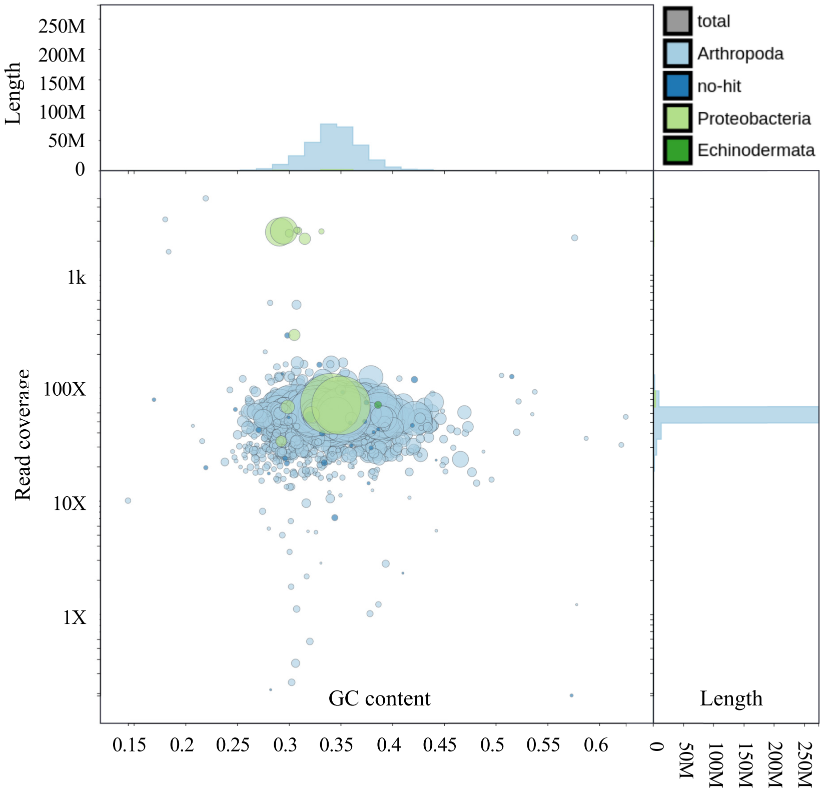


Figure 2. Blobplot of unfiltered consensus contigs. 1625 contigs are shown. Light and dark blue contigs matched to Arthropoda or 'no hit' respectively. Higher y-axis blobs have higher coverage. The x-axis is GC content. A single uniform blob indicates single genome origin, rather than multiple species origins. The green blobs with high coverage above the primary blob are of bacterial parasitic origin. The large green blob in the middle is Wolbachia.

There were 13 alien contigs matching to bacterial species and 1 with a small partial hit to an Echinodermata starfish.

To determine origin and filtering criteria of the bacterial and alien contigs, I applied thresholds for identification and removal. Anomalously high coverage proved a reliable marker of non-host contigs. Seven contigs had over 2000× coverage and were of bacterial origin, including nearly the complete genome of *Blochmannia pennsylvanicus*, an endosymbiont of the black carpenter ant. These contigs were removed from the assembly. There were six contigs between 30× and 300× coverage that were annotated as bacteria, all *Blochmannia pennsylvanicus*, however close examination of Blast ID showed only fragmentary matches to <5% of the query length. The 1 starfish contig was also a <5% fragmentary ID match. These were left in the assembly in the assumption that they were misannotated and were true *C. pennsylvanicus* contigs with the following exception. There was a single suspicious contig with typical ant level nuclear DNA coverage. It was a 1.5 Mb contig matching the full length of the *Wolbachia pipientis* genome with no flanking ant sequences. This is consistent with infection rather than an endogenous horizontal insertion, despite this contig having only 69× coverage (Supplementary Figure S4). The list complete list of alien contigs, identification, and removal decision is available in Supplementary Table S2.

There were also 11 contigs with over 1000× coverage that were identified as either 'Arthropoda' or 'no hit' that proved to be worthy of examination. A group of 3 contigs matched mitochondria genomes from other ant species and were removed. Another was a large rRNA subunit, known to be highly repetitive and thus attract numerous matches collapsed into a single contig. It was kept. There were five contigs with anomalously low coverage of <1× and they were removed. The manually curated ant assembly contained 1609 contigs.

The generation of so few alien contigs, despite using a whole ant is unsurprising since non-host DNA is unlikely to generate enough consistent coverage to build any consensus fragments. Exceptions were all intracellular high-copy number commensals such as mitochondria, symbiotes and parasites present. However, when assessing contamination, it is important to also examine the complete unassembled read set and not just the assembly. For this purpose, I used the Kraken2 microbiome database. In the ∼20 Gb of total reads, a small fraction environmentally derived reads were discovered, with the highest non-ant DNA being human origin at 0.002% frequency.

Repetitive DNA results

Ant genomes, as with most animals, consist of a large proportion of repetitive sequence made of interspersed transposable elements (TEs) and more simple repeats. Since repeat databases are skewed towards mammals, I chose to use a *de novo* repeat identification pipeline with RepeatModeler to find repeats, and RepeatMasker to classify and anno-

Table 3. Repetitive element content of *C. pennsylvanicus*

	Element counts	Total length (bp)	% of Genome covered
Total bases masked		104678206	34.16
Retroelements	30521	21771960	7.11
SINEs:	2025	500013	0.16
Penelope	3155	1490951	0.49
LINEs:	10719	6928742	2.26
CRE/SLACS	152	18220	0.01
L2/CR1/Rex	1114	719754	0.23
R1/LOA/Jockey	3564	2525973	0.82
R2/R4/NeSL	120	35432	0.01
RTE/Bov-B	2232	1692132	0.55
LTR elements	17777	14343205	4.68
BEL/Pao	2379	1846404	0.6
Ty1/Copia	1053	376038	0.12
Gypsy/DIRS1	14030	11817732	3.86
Retroviral	46	51129	0.02
DNA transposons	101469	36141816	11.79
hobo-Activator	7468	2780257	0.91
Tc1-IS630-Pogo	44001	15963966	5.21
PiggyBac	2021	228011	0.07
Tourist/Harbinger	2099	514354	0.17
Mirage	782	389406	0.13
Rolling-circles	5504	3278131	1.07
Unclassified	102298	33140531	10.82
Total interspersed elements	91054307	29.71	
Small RNA	2098	572887	0.19
Simple repeats	170891	8762741	2.86
Low complexity	27386	1510153	0.49

tate them. In total 34.16% of the genome was occupied by repetitive elements, including simple repeats and low complexity regions, at 2.86% and 0.49% of the genome, respectively (Table 3). Overall, there were 1604 distinct families detected, and not surprisingly, most of the repeats are of DNA transposon origin. Unlike mammals, with retroelement predominant genomes, insect genomes tend to have a greater proportion of DNA transposons (21). All major groups of TEs in this study are distributed in a similar fashion to the TEs described in the Harvester ant (*Pogonomyrmex californicus*) genome (22). To benchmark the *de novo* detected families, I masked the *C. floridanus* genome using the same library and found a slightly lower level, 27.33% of the genome, which may reflect the larger assembly size of *C. pennsylvanicus* (307 Mb) versus *C. floridanus* (284 Mb), or a more sensitive detection of repeats in my pipeline. The complete list of families is available in Supplementary File S2.

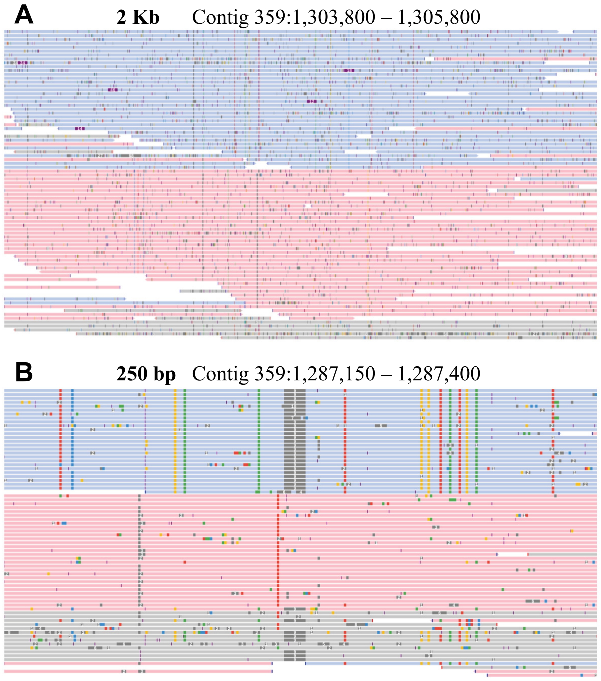


Figure 3. Phased genome view. Depiction of phased reads indicating parental haplotypes at two scales. (**A**) This 2 kb region illustrates large scale phasing and random sequencing errors. Haplotypes 1 is in blue, haplotype 2 is in red. (**B**) This 250 bp region illustrates haplotype-specific SNPs and a 9 bp indel in dark gray in haplotype 1. Phased reads show linkage between variants. Reads in gray were not assigned to a specific haplotype.

Diploid genome variation

The genome was ∼0.16% heterozygous based on biallelic SNPs divided by total callable bases (i.e. the full consensus genome), and 0.21% when based on all heterozygous alleles combined (e.g. indels and SNPs). There were 397 295 transitions (Ts) and 157 601 transversions (Tv) across all the SNPs, leading to a 2.5 Ts/Tv ratio. I generated haploid phased output to better visualize the two haplotypes and their resulting haplotype specific variants such as SNPs and indels. Of 759,295 variants detected by PEPPER-Margin-DeepVariant, most (526 154) were phased, i.e. belonging to a specific parentally derived haplotype. Most variants were also heterozygous. Of 641 913 heterozygous variants, most (482 430) were biallelic SNPs. Examples of phased SNPs and indels can be seen in Figure 3. These figures are consistent with a diploid animal genome where most variation between parental haplotypes will be small, e.g. SNVs, followed by larger indels, and fewer inversions, however the variant caller used here does not detect inversions. The use of a single ant allowed parental haplotype resolution, instead of population-based SNP prevalence which would have been produced by pooled samples. A phased variant call file is included with the assembly to distinguish the haplotypes (Supplementary File S5).

Epigenetics

DNA methylation (5mC) in Hymenoptera is low, generally <2% and is not reliably associated with either eusociality or caste determination (23). The nanopore instrument can directly detect DNA methylation and hydroxymethylation at cytosines based on divergent signal intensity compared to unmodified cytosines. Megalodon was used to detect cytosine modifications in all contexts. Here in *C. pennsylvanicus* at CpG cytosines, DNA methylation was 0.359%, and at non-CpG cytosines 5mC was 0.003%. The 5mC level detected by nanopore correlates well with Bewick *et al.*'s study using whole genome bisulfite sequencing.(23) There are no previous reports describing genome-wide, base-level hydroxymethylation in insects. Here, at CpG cytosines, DNA hydroxymethylation was 0.999% and at non-CpG cytosines 5hmC was 0.101%. With all measures of cytosine methylation and hydroxymethylation below 1%, or near 0%, the detected modified bases are not likely to be biologically relevant.

Annotation

Ab initio protein prediction was performed using Augustus against both the masked and non-masked consensus assembly to compare prediction of native proteins and transposon-derived protein contributions to the transcriptome (16). There were 20 209 amino acid sequences, representing proteins, detected in the masked genome, made up of 99 891 exons. To determine the identity of these proteins, I compared their sequences to the NCBI 'nr' non-redundant database which combines protein sequences from GenPept, Swissprot, PIR, PDF, PDB and NCBI RefSeq. Of the predicted transcripts from the masked genome, 17 419 identity matches were found, representing 86% of the putative proteins identified by Augustus, suggesting high accuracy and a low false positive rate. The protein model scored at 90.5% for BUSCO completeness in protein mode with for the Hymenoptera lineage.

Without masking, there were 40 077 transcripts detected, made up of 150 404 exons. Of the transcripts, 33 286 were identified when compared to the 'nr' database, an ID match rate of 83% and a BUSCO score of 90.3%. This suggests that the majority of transposon derived transcripts were also positively identified. All the transcripts and IDs are found in Supplementary file S3.

Homology-directed protein prediction was also performed since Hymenoptera transcriptomes are readily available from NCBI. For this method, I used Gene Model Mapper (GeMoMa) which uses related species' amino acids, transcriptomes, and intron positioning to predict protein-coding transcripts (24). The unmasked assembly yielded 14 415 amino acid predictions with a 95.8% BUSCO score, a substantial improvement over *ab initio* prediction and higher than all 103 Hymenoptera transcriptomes with BUSCO scores reported in Waterhouse et al.(25) All the transcripts and IDs are found in Supplementary file S4.

Complete *C. pennsylvanicus* genome deposited

The final assembly of the *C. pennsylvanicus* had a length of 306 426 343 bp, spread over 1609 contigs, with an N50 of 565 603 bp, and an average coverage of 60×. A snail plot describing contig length and coverage is shown in Figure 4. The GC content was relatively low for an animal at 34.45%, but in line with other Hymenoptera (26). The CG to GC ratio is a nearly even 1.2, reflecting the lack of DNA methylation and concomitant lack of accelerated CpG deamination loss seen in mammals. The diploid assembly was deposited as BioProject PRJNA820489 for the primary haplotype and PRJNA821232 for the secondary haplotype assembly.

Mitochondrial genome

The total read fastq file was also mapped against the mitochondrial genome of the red fire ant (*Solenopsis invicta*, NC_014672.1) to identify reads that were likely mitochondrial in origin. There were 89 Mb of sequence mapping to the reference mitogenome. Resulting hits were extracted, assembled and polished with Flye, Racon, and Medaka. The mitochondrial genome of the black carpenter ant is 16 536 bp, and was covered to 5385× (Figure 5). It was substantially larger than the 15 549 bp fire ant reference but close in size to the pharaoh ant (*Monomorium pharaonis*) and little fire ant (*Wasmannia auropunctata*) mitogenomes. During annotation 1 frameshift SNP was discovered in a homopolymeric 't' stretch and was manually deleted to create a contiguous coding region. The gene order was broadly similar to mitogenomes in other ants, however several tRNA genes were in different order than even the most similar species available on NCBI's assembly database. The most similar species were the fire ant and the pharaoh ant, and given their 97 MYA divergence, a discordance in mitochondrial gene order of the observed magnitude is not unexpected. The *C. pennsylvanicus* mitogenome created here was deposited as part of BioProject PRJNA820489.

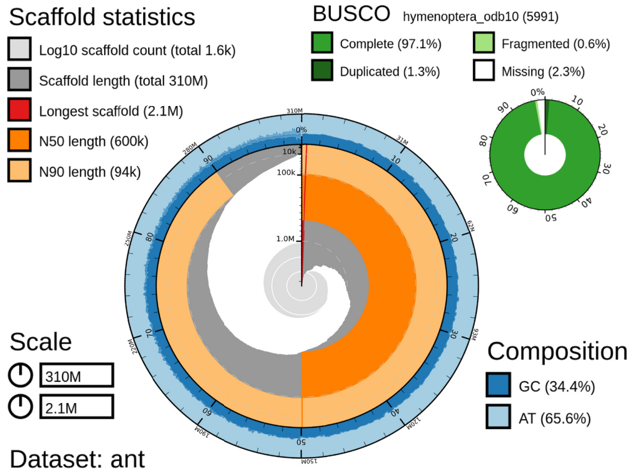


Figure 4. Snail plot of the unfiltered *C. pennsylvanicus* assembly. GC content is shown in the outer ring. The N90 and N50 scores and sizes are shown in light and dark orange respectively.

Commensal bacteria *B. pennsylvanicus*

The tribe of Camponotus ant species have a unique and unusual relationship with the *Blochmannia* genus of endosymbiotic bacteria, without which they do not develop properly. Unsurprisingly then, I found high coverage of reads specifically matching *Blochmannia pennsylvanicus* which has a publicly available sequence. To create a complete assembly of *B. pennsylvanicus* from this ant, I mapped the total reads to the published *B. pennsylvanicus* reference. A total of 192 Mb of sequence mapped to the reference. As with the nuclear genome, Flye was used to assemble a consensus, followed by two rounds of Racon polishing and one round of Medaka polishing. The consensus sequence was 791 499 bp long, with 2434× coverage. My assembly was within 200 bp of the reference length. It matched the reference genome (NC_007292.1) with 99.84% identity over 100% of the length. It was deposited as BioProject PRJNA824867.

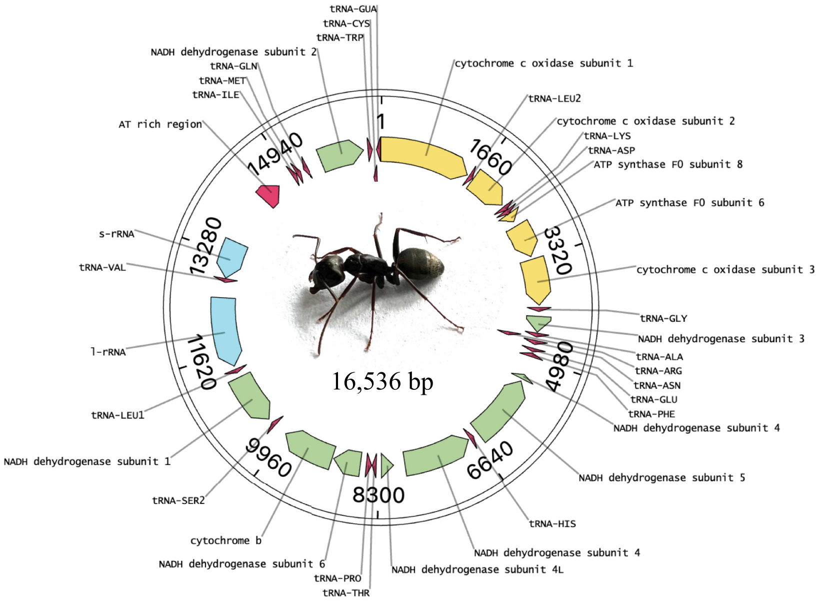

tRNA-GUA
tRNA-CYS
tRNA-TRP
NADH dehydrogenase subunit 2
tRNA-GLN
tRNA-MET
tRNA-ILE
AT rich region
cytochrome c oxidase subunit 1
tRNA-LEU2
cytochrome c oxidase subunit 2
tRNA-LYS
tRNA-ASP
ATP synthase F0 subunit 8
ATP synthase F0 subunit 6
cytochrome c oxidase subunit 3
tRNA-GLY
NADH dehydrogenase subunit 3
tRNA-ALA
tRNA-ARG
tRNA-ASN
tRNA-GLU
tRNA-PHE
NADH dehydrogenase subunit 4
NADH dehydrogenase subunit 5
tRNA-HIS
NADH dehydrogenase subunit 4
NADH dehydrogenase subunit 4L
s-rRNA
tRNA-VAL
l-rRNA
tRNA-LEU1
NADH dehydrogenase subunit 1
tRNA-SER2
cytochrome b
NADH dehydrogenase subunit 6
tRNA-PRO
tRNA-THR
1
1660
3320
4980
6640
8300
9960
11620
13280
14940
16,536 bp

Figure 5. Mitochondrial genome of *C. pennsylvanicus*. The circular length of 16 536 bp was assembled with 5385× coverage.

Wolbachia

Ants, along with the majority of arthropods, are parasitized by Wolbachia bacteria. Within the nuclear assembly, I found and removed a full length Wolbachia genome, and targeted it for separate assembly. A total of 159 Mb of sequence mapped to the *Wolbachia pipientis* contig extracted from the assembly. Flye, Racon, and Medaka were used to assemble and polish as before. The strain I assembled here had a 1 582 319 bp genome and had only 76× coverage, indicating a 1:1 ratio of Wolbachia to ant cells. My assembly was ∼400 kb longer than the *W. pipientis* reference. It was deposited as BioProject PRJNA824870.

Costs and processing time

As well as expense for the sequencing, time is an important consideration in both sequencing and in data processing. Given the constraints of one person, $1000 in consumables, and one individual ant sequenced, it is appropriate to limit computational hardware to maintain the accessibility of the pipeline described here. Therefore, I restricted all the analyses to using only commodity hardware housed locally in a single desktop computer. The computer had 64 Gb of memory, 4 Tb of SSD storage space, with a 12 core AMD Ryzen 3900x processor and an NVIDIA RTX 2080 Ti GPU to accelerate base-calling and other processing stages.

Costs for sequencing were reduced by using classroom grade equipment that is also suitable for fieldwork. Consumables totaled $1748 and amortized to $1047 per sample when accounting for multiple use reagents and kits (Table 4). The largest expense was the flow cell, followed by the library preparation kit. Fixed costs included all equipment necessary for DNA extraction to library preparation, sequencing, and data analysis. The MinION sequencer instrument and the GPU processor were equally expensive at $1000 each. The NVIDIA GPU is required to reduce the base calling stage by several orders of magnitude in time to process. It also accelerates Medaka polishing and variant calling with PEPPER-Margin-DeepVariant. The Guppy base caller requires a minimum of 8 Gb of GPU memory with current neural network models which governs the choice of GPU. A computer running Linux with a minimum of 1 Tb SSD is also required and is assumed to be pre-existing equipment. Optionally the addition of a nanospectrophotometer aids in determining DNA concentration and purity. These instruments can be sourced for less than $5000.

The sequencing took 72 h on an Oxford Nanopore MinION instrument using an r9.4.1 flow cell with a nuclease flush and library reload every 24 h. Base calling took 6 hours using Guppy software. Assembly with Flye took 2 h, followed by four rounds of Racon and one round of Medaka, totaling ∼6 h. Repeat identification was the longest stage, taking 28 h for RepeatModeler2 and 1 h for RepeatMasker. Modified base calling for DNA methylation took 10 h with Megalodon. Annotation with Augustus took 6 h and GeMoMa took 2 h. Variant calling and read phasing took 4 h with PEPPER-Margin-DeepVariant. The data processing pipeline was performed by the author in less than a week.

DISCUSSION

Rationale

Along with the black carpenter ant, several other insect genomes have benefitted greatly from long-read sequencing (27–30). Single insect sequencing has also been previously performed. A single Drosophila fly individual has been sequenced to full genome status with great success by using a combination of long and short read sequencing with 3D conformation to bring the consensus to chromosome level (31). Similarly, sequencing 101 Drosophila species using

Table 4. Reagent and equipment costs

Unit Cost	Consumable	Catalog number	Total cost	Company	Workflow
$3	Zymo Quick-DNA Plus Kit	D4068	$140 (50 reactions)	Zymo Research	Tissue to DNA
$4	Gel materials (Agarose, TBE, gelRed, loading dye, ladder)		$85 (20 reactions)	MiniPCR	Quality check DNA
$100	Ligation Sequencing Kit	SQK-LSK110	$600 cost includes 6 library prep rxns	Oxford Nanopore	Prepare DNA library
$2	Axygen AxyPrep Mag PCR Clean Up Kit- 5 ML	MAG-PCR-CL-5	$175 for smallest volume (5ml or 96 rxns)	Corning	Prepare DNA library
$9	NEBNext® FFPE DNA Repair Mix	M6630	$170 (20 reactions)	New England Biolabs	Prepare DNA library
$13	NEBNext® Ultra™ II End Repair/dA-Tailing Module	E7546	$250 (20 reactions)	New England Biolabs	Prepare DNA library
$16	NEBNext® Quick Ligation Module	E6056	$328 (20 reactions)	New England Biolabs	Prepare DNA library
$900	Flow cell (R9.4.1)			Oxford Nanopore	Sequence DNA
$1,047	**Total Reagent Cost Per Sample**		**$1748**	**Total Batch Cost of Reagents**	
Fixed costs	**Equipment**				
$1,000	Minion sequencer*		Oxford Nanopore Technologies	*Sequencer comes with library kit (6 reactions) & 1 flow cell.	
$300	BlueGel rig & transilluminator QP-1500-01		MiniPCR		
$165	Micropipettes (10, 200, 1000)		MiniPCR		
$1,000	External Nvidia GPU, minimum of RTX 3060 with 12 Gb RAM		Various		
Assumed	Computer with Ubuntu linux		1–4 Tb SSD available		
$2,465	**Total Equipment Cost**				

only nanopore reads has also been achieved, revealing many facets of evolution in insects (32). However, these achievements still required large teams and funding. With advancements in nanopore base calling accuracy, cost reductions, and assembly and analysis tools, a $1000 assembly from a single insect and sole researcher has now become possible.

Assembly

Assembly methods are improving rapidly, as indicated by numerous benchmarking studies (33–35). For small scale laboratories, a major bottleneck is the computational power needed to assemble a genome from reads. Assemblers working on mammalian size genomes at 3 Gb tend to require high memory, on the order of 2 Tb of RAM. While such machines exist, they are currently out of price range for most laboratories and available only at high performance computing facilities. Time is also an issue. If using a high accuracy assembler like Canu, even small genomes take two orders of magnitude longer than a nearly equivalent consensus generated with Flye. For instance, a ∼10 Mb genome at 70× coverage takes over 22 h with Canu and <1 h with Flye (36). Mammalian scale assemblies can take weeks to months with Canu. The black carpenter ant genome, at 300 Mb, was able to be assembled in reasonable time frame on consumer level hardware, taking just 2 h using Flye. Progress towards faster, equally accurate assemblers such as Shasta (8), combined with falling prices of high memory computers will make human scale assemblies available even to small scale laboratories in the near future.

It is critically important to determine whether the assembly is of high quality before undertaking additional analyses as it will become the publicly available reference genome of this species. Long-reads generated by nanopore sequencing are generally random in errors but do contain some systematic errors that can be overcome through polishing. Vaser et. al recommend four rounds of Racon polishing followed by Medaka polishing (37), which is what yielded the best results here. Evaluating assemblies beyond read lengths and quality scores has been greatly enhanced through the benchmarking of universal single copy orthologs (BUSCO) scoring program (20). The BUSCO score is a more biologically relevant determination of whether an assembly is accurate since it detects genes that should be present, rather than relying simply on N50 size or contig number. My comparison reveals that a nanopore-only assembly provided equal or better BUSCO scores than hybrid assemblies of other ant species.

Contamination

Removal of contamination is vital to generate an accurate public reference genome. Sequencing of whole animals will result in inevitable contamination through a variety of sources. In this case, a wild ant was captured from the environment and may host microbial contamination. Consensus coverage was the most useful filter of non-host DNA. There was a gradation of contig coverage from the mean, 60×, to a high of 566×. Above that range, only 11 contigs exist, and all were above 1600× coverage. Therefore, a cutoff above 1000× was reasonable to identify alien contigs that are not likely to be nuclear in origin. Due to the use of a long-read assembly, contigs should be made of reads derived from single molecules that have substantial overlap. The five contigs below 1× coverage imply chimeric joining of unrelated reads into a false contig, so were also removed. GC content has also been used to identify contaminants which often diverge from nuclear genome GC percentage, however in this case, no contigs were flagged as divergent. The use of the Blobtoolkit allowed easy visualization of contig coverage vs. GC content (12).

Finally, during the process of DNA extraction and library preparation, the sample may become contaminated with human or lab-sources of DNA. Cross-checking raw reads using kraken2 efficiently detected human contaminants and detects microbiome reads simultaneously (38).

Repetitive DNA results

There is a paucity of information on repetitive DNA content in ants. Transposons, retroelements, and simple repeats, are major components of all animal genomes, yet they are not well described in repeat databases. For this reason, it is important to use *de novo* methods of repeat identification before masking the genome to obtain the most accurate percentage of repeat content. Less accurate identification may explain the lower percentage found in the Florida Carpenter ant (27%) vs. its close relative, the black carpenter ant studied here (34%). However, repeat element content can be low in ants, with two lineages of the heart node ant, *Cardiocondyla obscurior,* both having 21% repeat content (39).

Diploid genome variation

An important advancement will be to develop diploid-aware assemblers to better incorporate heterozygosity seen in diploid species and pooled samples (40). The use of a single ant here allowed detection of phased haplotypes of parental origin based on the biallelic nature of SNPs and indels. The variant call format (vcf) file associated with the primary consensus provides alternative haplotype information, encompassing the full diploid variation of this genome. Phasing reveals linkage of variants from parental haplotigs.

Heterozygosity has traditionally been assessed with only a small number of highly variable markers due to technical constraints, so little data is available to compare full genome wide SNP heterozygosity as I was able to do here. Increasingly measures of whole genome heterozygosity are being adopted. When compared to 101 different drosophila species also sequenced by nanopore and calculated the same way, the SNP heterozygosity of the black carpenter ant (0.16%) fell in the lower middle of Drosophilid range of 0.00035% to 1.1% (32). The SNP transition transversion ratio was found to be 2.5, higher than the 2.1 seen in humans and most mammals (41). This is a natural outcome of mammalian genome methylation at CpG sites which has lead to rapid evolutionary loss of CpG sites, resulting in fewer opportunities for transitions to occur in mammals today. Thus, our ant, without DNA methylation, retains a nearly equal CG to GC ratio of 1.2, unlike humans with 0.25 CG:GC.

Methylation

Here I replicated the previous reports of extremely low DNA methylation at a base pair specific resolution, and I am the first to report the nearly complete lack of DNA hydroxymethylation in this animal. The lack of DNA methylation was unsurprising, given the generally low levels seen in other studies of Formicidae (42). DNA methylation in the honeybee is known to partially govern caste development, though its level and activity are widely variable in other Hymenoptera, and do not generally correlate with caste development or eusociality (23). For example, in the Florida carpenter ant, histone acetylation appears to be the dominant driver of caste development as seen by Simola *et al.* (5). DNA methyltransferase proteins (DNMTs) do appear to be conserved across Hymenoptera species, however, and DNMTs can have essential functions despite the lack of CpG methylation in other insects (43). An important limitation here is that these results were limited to a single worker, and final determination of the role of DNA methylation in caste development should include queens and other castes in this species.

Annotation

Homology-directed protein prediction by GeMoMa was the most accurate method (95.8% BUSCO) but relies upon the availability and quality of transcriptome data from related species. *Ab initio* prediction by Augustus still outperformed 103 previously published other Hymenoptera transcriptomes using Augustus without hints from other species (25). Both GeMoMa and Augustus identified nearly the expected number of 20 000 genes. Some 86% of putative transcripts in the masked genome matched to previously known genes as well as a majority of the transposon ORFs in the unmasked genome. The improvement over other insect transcriptomes is likely a combination of good assembly quality and rapidly improving prediction algorithms.

Mitochondrial genome

The full mitogenome was extracted and sequenced marking the first available reference mitogenome for this species. The layout was similar but not exactly convergent with other ant species. Previous nanopore sequencing has recovered mitogenomes from even highly degraded samples such as primate feces, and generated accurate assemblies and so was expected here (44). Copy number of the *C. pennsylvanicus* mitochondria was found to be relatively low compared to the nuclear genome with a ratio of 77:1 mtDNA genomes per haploid nuclear genome. This is significantly lower than in human muscle with varies from 3000–6000 copies per cell or the hundreds to thousands of copies per cell seen in drosophila (45,46). The low number is likely due to the fact that mtDNA is circular and must be linearized before it can pass through a nanopore for sequencing. Therefore, only fragmented or sheared mitogenomes would be present in the reads, with the majority of circular mitogenomes absent from the data set.

Commensal bacteria

Ants also contain parasites and symbiotes as well as non-nuclear DNA from mitochondria whose genomes can be assembled from nanopore reads (47). Both Wolbachia and *Blochmannia pennsylvanicus* are known symbiotes of the black carpenter ant and were extracted for full assembly, along with the mitogenome. The Wolbachia assembly was surprisingly much longer than the reference *W. pipientis*, though strains are known to vary widely in co-evolution with their insect host. Copy number was approximately 1:1 based on read coverage. This is lower than is seen in Drosophila, that have a mean of 5.3 copies of Wolbachia per cell (48). In other Hymenoptera, the Wolbachia load can vary from less than 1 to over 10 per cell depending on host factors (49).

The black carpenter ant, and all members of the Camponotus tribe, host specialized endosymbionts of the Blochmannia genus that co-evolve with each species. They are known to have especially slow rates of evolution within species, which aids identification. Here we found the highly abundant coverage of a full-length *B. pennsylvanicus* genome in its own contig. It differed by only 200 bp in length from the reference genome with a nearly perfect 99.84% sequence identity, confirming both the species of bacteria as well as the host ant.

This individual ant did not yield any contigs deriving from any other bacteria. This could be due to several reasons. This ant had recently emerged from a long winter dormancy in Minnesota and therefore may not have had much gut microbiota remaining. Or there were potentially too few cells of any particular species to build a contig.

Project cost and processing time

To complete this assembly for ∼$1000, some DNA quality control steps were simplified. Sizing and fragment validation were performed on an agarose gel rather than the recommended Agilent TapeStation or Bioanalyzer (50). The gel rig and pipettes were robust classroom grade equipment available from Amplyus that are less expensive than laboratory grade hardware. Quantitation was performed on a nanophotometer that requires no reagents instead of a more sensitive qubit fluorometer (ThermoFisher Inc.) and still yielded reliable results. Oxford Nanopore Technologies makes the MinION nanopore sequencer and associated flow cells and reagents available to purchase affordably by individual labs. I chose to use the Ligation Sequencing kits SQK-110 for higher throughput, but Rapid Sequencing kits and Field Sequencing Kits can eliminate additional reagents or allow cold-chain free sequencing respectively. Limiting the project to the use of commodity computer hardware makes this pipeline more accessible to small scale labs. The use of classroom grade gel equipment makes sequencing more accessible in the field. Combined, these optimizations put full genome assembly, at least for small size genomes, in reach of a larger share of the scientific community.

DATA AVAILABILITY

DNA reads have been deposited in fastq format in the Sequence Read Archive (SRA) repository (https://www.ncbi.nlm.nih.gov/sra) under SRA submission SRX14660682 and methylation modified base calls under SRX14660818. Assembly is available as BioProject PRJNA820489 for the primary haplotype and PRJNA821232 for the secondary haplotype. The mitogenome was deposited as part of BioProject PRJNA820489. *B. pennsylvanicus* assembly was deposited as BioProject PRJNA824867. The *Wolbachia pipientis* strain associated with *C. pennsylvanicus* was deposited as BioProject PRJNA824870.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Thanks to A.K. Barks for critically reading the manuscript.
Author contributions: C.F. captured the ant. C.F. performed sequencing and all analyses. C.F. wrote the manuscript.

FUNDING

USDA-NIFA HATCH, AES [MIN-16-12 to C.F.]; UMN College of Natural Resource Sciences. Funding for open access charge: Indirect Cost Recovery funds from my college.
Conflict of interest statement. None declared.

REFERENCES

1. Hotaling,S., Sproul,J.S., Heckenhauer,J., Powell,A., Larracuente,A.M., Pauls,S.U., Kelley,J.L. and Frandsen,P.B. (2021) Long reads are revolutionizing 20 years of insect genome sequencing. *Genome Biol. Evol.*, **13**, evab138.
2. Shields,E.J., Sheng,L., Weiner,A.K., Garcia,B.A. and Bonasio,R. (2018) High-Quality genome assemblies reveal long Non-coding RNAs expressed in ant brains. *Cell Rep.*, **23**, 3078–3090.
3. Das,B. and de Bekker,C. (2022) Time-course RNASeq of camponotus floridanus forager and nurse ant brains indicate links between plasticity in the biological clock and behavioral division of labor. *BMC Genomics*, **23**, 57.
4. Shi,Y., Bethea,J.P., Hetzel-Ebben,H.L., Landim-Vieira,M., Mayper,R.J., Williams,R.L., Kessler,L.E., Ruiz,A.M., Gargiulo,K., Rose,J.S.M. *et al.* (2021) Mandibular muscle troponin of the florida carpenter ant camponotus floridanus: extending our insights into invertebrate Ca^{2+} regulation. *J. Muscle Res. Cell Motil.*, **42**, 399–417.
5. Simola,D.F., Graham,R.J., Brady,C.M., Enzmann,B.L., Desplan,C., Ray,A., Zwiebel,L.J., Bonasio,R., Reinberg,D., Liebig,J. *et al.* (2016) Epigenetic (re)programming of caste-specific behavior in the ant camponotus floridanus. *Science*, **351**, aac6633.
6. Degnan,P.H., Lazarus,A.B. and Wernegreen,J.J. (2005) Genome sequence of blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res.*, **15**, 1023–1033.
7. Treanor,D. and Hughes,W.O.H. (2019) Limited female dispersal predicts the incidence of wolbachia across ants (Hymenoptera: formicidae). *J. Evol. Biol.*, **32**, 1163–1170.
8. Shafin,K., Pesout,T., Lorig-Roach,R., Haukness,M., Olsen,H.E., Bosworth,C., Armstrong,J., Tigyi,K., Maurer,N., Koren,S. *et al.* (2020) Nanopore sequencing and the shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.*, **38**, 1044–1053.
9. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.

10. Tsutsui,N.D., Suarez,A.V., Spagna,J.C. and Johnston,J.S. (2008) The evolution of genome size in ants. *BMC Evol. Biol.*, **8**, 64.
11. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.*, **34**, 3094–3100.
12. Challis,R., Richards,E., Rajan,J., Cochrane,G. and Blaxter,M. (2020) BlobToolKit - interactive quality assessment of genome assemblies. *G3 Bethesda Md*, **10**, 1361–1374.
13. Pedersen,B.S. and Quinlan,A.R. (2018) Mosdepth: quick coverage calculation for genomes and exomes. *Bioinforma. Oxf. Engl.*, **34**, 867–868.
14. Flynn,J.M., Hubley,R., Goubert,C., Rosen,J., Clark,A.G., Feschotte,C. and Smit,A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 9451–9457.
15. Storer,J., Hubley,R., Rosen,J., Wheeler,T.J. and Smit,A.F. (2021) The dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA*, **12**, 2.
16. Nachtweide,S. and Stanke,M. (2019) Multi-Genome annotation with AUGUSTUS. *Methods Mol. Biol.*, **1962**, 139–160.
17. Buchfink,B., Reuter,K. and Drost,H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.
18. Shafin,K., Pesout,T., Chang,P.-C., Nattestad,M., Kolesnikov,A., Goel,S., Baid,G., Kolmogorov,M., Eizenga,J.M., Miga,K.H. *et al.* (2021) Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods*, **18**, 1322–1332.
19. Garg,S., Rautiainen,M., Novak,A.M., Garrison,E., Durbin,R. and Marschall,T. (2018) A graph-based approach to diploid genome assembly. *Bioinforma. Oxf. Engl.*, **34**, i105–i114.
20. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.
21. Petersen,M., Armisén,D., Gibbs,R.A., Hering,L., Khila,A., Mayer,G., Richards,S., Niehuis,O. and Misof,B. (2019) Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol. Biol.*, **19**, 11.
22. Bohn,J., Halabian,R., Schrader,L., Shabardina,V., Steffen,R., Suzuki,Y., Ernst,U.R., Gadau,J.R. and Makałowski,W. (2020) High-Quality Genome Assembly and Annotation of the California Harvester Ant *Pogonomyrmex californicus* (Buckley, 1867) Genomics.
23. Bewick,A.J., Vogel,K.J., Moore,A.J. and Schmitz,R.J. (2017) Evolution of DNA methylation across insects. *Mol. Biol. Evol.*, **34**, 654–665.
24. Keilwagen,J., Hartung,F. and Grau,J. (2019) GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. In: Kollmar,M. (ed). *Gene Prediction, Methods in Molecular Biology*. Springer New York, NY, Vol. **1962**, pp. 161–177.
25. Waterhouse,R.M., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2019) Using BUSCO to assess insect genomic resources. In: Brown,S.J. and Pfrender,M.E. (eds). *Insect Genomics, Methods in Molecular Biology*. Springer New York, NY, Vol. **1858**, pp. 59–74.
26. Branstetter,M.G., Childers,A.K., Cox-Foster,D., Hopper,K.R., Kapheim,K.M., Toth,A.L. and Worley,K.C. (2018) Genomes of the hymenoptera. *Curr. Opin. Insect Sci.*, **25**, 65–75.
27. Satoh,A., Takasu,M., Yano,K. and Terai,Y. (2021) De novo assembly and annotation of the mangrove cricket genome. *BMC Res. Notes*, **14**, 387.
28. Urban,J.M., Foulk,M.S., Bliss,J.E., Coleman,C.M., Lu,N., Mazloom,R., Brown,S.J., Spradling,A.C. and Gerbi,S.A. (2021) High contiguity de novo genome assembly and DNA modification analyses for the fungus fly, sciara coprophila, using single-molecule sequencing. *BMC Genomics*, **22**, 643.
29. Abeynayake,S.W., Fiorito,S., Dinsdale,A., Whattam,M., Crowe,B., Sparks,K., Campbell,P.R. and Gambley,C. (2021) A rapid and cost-effective identification of invertebrate pests at the borders using MinION sequencing of DNA barcodes. *Genes*, **12**, 1138.
30. Baldwin-Brown,J.G., Villa,S.M., Vickrey,A.I., Johnson,K.P., Bush,S.E., Clayton,D.H. and Shapiro,M.D. (2021) The assembled and annotated genome of the pigeon louse columbicola columbae, a model ectoparasite. *G3*, **11**, jkab009.
31. Adams,M., McBroome,J., Maurer,N., Pepper-Tunick,E., Saremi,N.F., Green,R.E., Vollmers,C. and Corbett-Detig,R.B. (2020) One fly-one genome: chromosome-scale genome assembly of a single outbred drosophila melanogaster. *Nucleic Acids Res.*, **48**, e75.
32. Kim,B.Y., Wang,J.R., Miller,D.E., Barmina,O., Delaney,E., Thompson,A., Comeault,A.A., Peede,D., D'Agostino,E.R., Pelaez,J. *et al.* (2021) Highly contiguous assemblies of 101 drosophilid genomes. Elife, **10**, e66405.
33. Murigneux,V., Rai,S.K., Furtado,A., Bruxner,T.J.C., Tian,W., Harliwong,I., Wei,H., Yang,B., Ye,Q., Anderson,E. *et al.* (2020) Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience*, **9**, giaa146.
34. Wick,R.R. and Holt,K.E. (2019) Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research*, **8**, 2138.
35. Chen,Z., Erickson,D.L. and Meng,J. (2020) Benchmarking long-read assemblers for genomic analyses of bacterial pathogens using oxford nanopore sequencing. *Int. J. Mol. Sci.*, **21**, E9161.
36. Wang,J., Chen,K., Ren,Q., Zhang,Y., Liu,J., Wang,G., Liu,A., Li,Y., Liu,G., Luo,J. *et al.* (2021) Systematic comparison of the performances of de novo genome assemblers for oxford nanopore technology reads from piroplasm. *Front. Cell. Infect. Microbiol.*, **11**, 696669.
37. Vaser,R., Sović,I., Nagarajan,N. and Šikić,M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, **27**, 737–746.
38. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with kraken 2. *Genome Biol.*, **20**, 257.
39. Errbii,M., Keilwagen,J., Hoff,K.J., Steffen,R., Altmüller,J., Oettler,J. and Schrader,L. (2021) Transposable elements and introgression introduce genetic variation in the invasive ant cardiocondyla obscurior. *Mol. Ecol.*, **30**, 6211–6228.
40. Guiglielmoni,N., Houtain,A., Derzelle,A., Van Doninck,K. and Flot,J.-F. (2021) Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinf.*, **22**, 303.
41. Wang,J., Raskin,L., Samuels,D.C., Shyr,Y. and Guo,Y. (2015) Genome measures used for quality control are dependent on gene function and ancestry. *Bioinforma. Oxf. Engl.*, **31**, 318–323.
42. Bonasio,R., Zhang,G., Ye,C., Mutti,N.S., Fang,X., Qin,N., Donahue,G., Yang,P., Li,Q., Li,C. *et al.* (2010) Genomic comparison of the ants camponotus floridanus and harpegnathos saltator. *Science*, **329**, 1068–1071.
43. Schulz,N.K.E., Wagner,C.I., Ebeling,J., Raddatz,G., Diddens-de Buhr,M.F., Lyko,F. and Kurtz,J. (2018) Dnmt1 has an essential function despite the absence of CpG DNA methylation in the red flour beetle tribolium castaneum. *Sci. Rep.*, **8**, 16462.
44. Wanner,N., Larsen,P.A., McLain,A. and Faulk,C. (2021) The mitochondrial genome and epigenome of the golden lion tamarin from fecal DNA using nanopore adaptive sequencing. *BMC Genomics*, **22**, 726.
45. Miller,F.J., Rosenfeldt,F.L., Zhang,C., Linnane,A.W. and Nagley,P. (2003) Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Res.*, **31**, e61.
46. Salminen,T.S., Oliveira,M.T., Cannino,G., Lillsunde,P., Jacobs,H.T. and Kaguni,L.S. (2017) Mitochondrial genotype modulates mtDNA copy number and organismal phenotype in drosophila. *Mitochondrion*, **34**, 75–83.
47. Petrone,J.R., Muñoz-Beristain,A., Glusberger,P.R., Russell,J.T. and Triplett,E.W. (2022) Unamplified, long-read metagenomic sequencing approach to close endosymbiont genomes of low-biomass insect populations. *Microorganisms*, **10**, 513.
48. Signor,S. (2017) Population genomics of wolbachia and mtDNA in drosophila simulans from california. *Sci. Rep.*, **7**, 13369.
49. Funkhouser-Jones,L.J., van Opstal,E.J., Sharma,A. and Bordenstein,S.R. (2018) The maternal effect gene wds controls wolbachia titer in nasonia. *Curr. Biol.*, **28**, 1692–1702.
50. Zascavage,R.R., Hall,C.L., Thorson,K., Mahmoud,M., Sedlazeck,F.J. and Planz,J.V. (2019) Approaches to whole mitochondrial genome sequencing on the oxford nanopore MinION. *Curr. Protoc. Hum. Genet.*, **104**, e94.