G3, 2021, 11(1), jkaa019
DOI: 10.1093/g3journal/jkaa019
Advance Access Publication Date: 4 December 2020
Genome Report

Genome assembly and annotation of the California harvester ant *Pogonomyrmex californicus*

Jonas Bohn [1,†] Reza Halabian [1] Lukas Schrader [2] Victoria Shabardina [1,‡] Raphael Steffen,[2] Yutaka Suzuki [3] Ulrich R. Ernst [2] Jürgen Gadau [2,*] and Wojciech Makałowski [1,*]

[1]Faculty of Medicine, Institute of Bioinformatics, University of Münster, 48149 Münster, Germany
[2]Faculty of Biology, Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany
[3]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8562, Japan
[†]Present address: Division of Medical Informatics for Translational Oncology, Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
[‡]Present address: Institute of Evolutionary Biology, CSIC-University Pompeu Fabra, Marítim de la Barceloneta 39, 08003 Barcelona, Spain.

*Corresponding authors: gadauj@uni-muenster.de (J.G.); wojmak@uni-muenster.de (W.M.)

Abstract

The harvester ant genus Pogonomyrmex is endemic to arid and semiarid habitats and deserts of North and South America. The California harvester ant *Pogonomyrmex californicus* is the most widely distributed Pogonomyrmex species in North America. *Pogonomyrmex californicus* colonies are usually monogynous, *i.e.* a colony has one queen. However, in a few populations in California, primary polygyny evolved, *i.e.* several queens cooperate in colony founding after their mating flights and continue to coexist in mature colonies. Here, we present a genome assembly and annotation of *P. californicus*. The size of the assembly is 241 Mb, which is in agreement with the previously estimated genome size. We were able to annotate 17,889 genes in total, including 15,688 protein-coding ones with BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness at a 95% level. The presented *P. californicus* genome assembly will pave the way for investigations of the genomic underpinnings of social polymorphism in the number of queens, regulation of aggression, and the evolution of adaptations to dry habitats.

Keywords: genome assembly; genome annotation; Nanopore sequencing; polygyny; social insect; Hymenoptera

Introduction

Ants (Hymenoptera: Formicidae) are important components of almost all terrestrial ecosystems and more than 16,000 species have been described so far (AntWeb, version 8.41, California Academy of Science, online at https://www.antweb.org; accessed on August 19, 2020). The majority of them, over 6900, belong to the highly diverse subfamily Myrmicinae ants (AntWeb, version 8.41, California Academy of Science, online at https://www.antweb.org; accessed on August 19, 2020). Currently, 40 assembled ant genomes are available at NCBI (Entrez "Genome" accessed on August 19, 2020).

The harvester ant genus *Pogonomyrmex* is endemic to arid and semiarid habitats and deserts of North and South America (Buckley 1867; Cole 1968; Snelling *et al.* 2009). This genus thrives in extremely dry habitats, *e.g.* Death Valley or Anza Borega, and evolved seed harvesting behavior independently from the Old World harvester ant genus *Messor*. Members of the genus *Pogonomyrmex* are a very conspicuous element of the deserts in the Southwest of the USA and have been studied extensively (De Vita 1979; Rissing *et al.* 2000; Lighton and Turner 2004; Clark and Fewell 2014; Helmkampf *et al.* 2016; Overson *et al.* 2016). Within this genus, several interesting traits have evolved, such as social parasitism, genetic caste determination, and social polymorphism in terms of the queen number (Cole 1968; Rissing *et al.* 2000; Julian *et al.* 2002). Arguably, the most widely distributed *Pogonomyrmex* species in North America is *P. californicus* (Johnson 2002). *Pogonomyrmex californicus* colonies are usually monogynous, *i.e.* a colony has one queen. However, in a few populations in California, primary polygyny has evolved, *i.e.* several queens cooperate in colony founding after their mating flights and continue to coexist in mature colonies (Rissing *et al.* 2000; Johnson 2004; Shaffer *et al.* 2016). The Red Imported Fire Ant, *Solenopsis invicta*, and several other *Formica* ant species have a similar social polymorphism, which has been shown to be due to a supergene (Wang *et al.* 2013; Yan *et al.* 2020). This discovery was only possible by next-generation sequencing and the availability of genomic information for these species. Of approximately 70 described *Pogonomyrmex* species, only *Pogonomyrmex barbatus* (AntWeb, version 8.41, California Academy of Science, online at https://www.antweb.org; accessed on August 19, 2020) has its genome sequenced, assembled, and annotated (Smith *et al.* 2011). Five other species of this genus (*Pogonomyrmex anergismus*, *Pogonomyrmex colei*, *Pogonomyrmex imberbiculus*, *Pogonomyrmex occidentalis*, and *Pogonomyrmex rugosus*) have nuclear genomes partially sequenced but none have been so far processed and only raw reads are available in NCBI's Sequence Reads Archive (SRA). Sequences of *P. rugosus*, *P. anergismus*, and *P. colei* have been aligned to the

Received: September 02, 2020. **Accepted:** November 18, 2020

P. barbatus genome for a study of gene gains/losses in socially parasitic ants (Smith *et al.* 2015).

The genome sequencing and annotation of *P. californicus* will result in a better understanding of genomic sequence and structural variations and evolution in Formicidae in general and in the myrmicine genus *Pogonomyrmex* in particular. It will also pave the way for investigations of the genomic underpinnings of social polymorphism in the queen number, regulation of aggression, and the evolution of adaptations to dry habitats in *P. californicus*.

Materials and methods

Samples and transcriptome data source

The nuclear DNA was extracted from 13 haploid males from a single polygynous colony collected in 2016 from Pine Valley, CA, USA (32.819761, −116.521512; N32 49 11 – W116 31 17). Previously published transcriptome data based in parts on queens from the same population/area (Helmkampf *et al.* 2016) were downloaded from NCBI's Sequence Read Archive (BioProject accession number PRJDB4319). In addition, Oxford Nanopore sequencing (MinION) was performed on RNA extracted from workers of five polygynous colonies also collected in 2016 from Pine Valley, CA, USA. These reads are accessible at NCBI with BioProject accession number PRJNA622899.

Genome sequencing and assembling

DNA from 13 male ants was isolated using Qiagen MagAttract HMW DNA Kit with the protocol for tissue DNA extraction according to the protocol from October 2017. DNA was extracted from the whole body and all individuals were pooled together. This resulted in 4575 ng of DNA of which 1.2 ng was used for a 10x Genomics Chromium sequencing approach. The sequencing library was prepared according to the Chromium™ Genome Library Kit standard protocol (Manual Part Number CG00022) and the Illumina HiSeq 3000 system was used to sequence the library. The quality of the produced reads was checked using FastQC software, version 0.11.5 (Andrews 2010). We performed neither filtering nor trimming on these linked-reads to avoid losing any information. We used the *de novo* assembler Supernova, version 2.1 (Weisenfeld *et al.* 2017), with the following parameters: –maxreads = 156111200 and –accept-extreme-coverage. The maximum number of reads was set to a 75× effective coverage of the genome, which was chosen based on a set of Supernova runs with different coverages to obtain optimal parameters as a trade-off between genome size, BUSCO assessment (see below), and N50 coverage (see Figure 1) and assuming *P. californicus* has a genome size of 244.5 Mb (http://www.genomesize.com). Subsequently, the resulting genome assembly was polished by three rounds of Pilon, version 1.23, processing (Walker *et al.* 2014). For this step, 678,988,626 one-hundred bp reads from an independent Illumina sequencing run from the same initial DNA extraction were added to the 269,953,173 linked-reads used for the genome assembly, bringing the number of reads used in the polishing step to almost 1 billion. For this additional sequencing standard, an Illumina protocol was used to prepare a sequencing library, which was sequenced using the Illumina HiSeq 3000 system.

Transcript sequencing and assembling

For transcriptome analysis, MinION long-read RNA sequencing of the entire bodies of worker ants from a laboratory colony (pleometrotic colony from Pine Valley, NJ, USA) was performed. We extracted RNA using Monarch® Total RNA Miniprep Kit (New England BioLabs GmbH, Frankfurt, D, E2010). Material was grounded in Mixer Mill 200 (Retsch GmbH, Haan, D) in a protection reagent. A quality check was performed with Bioanalyser, Nanophotometer, and Qubit. The library was prepared from 5 µg of the total RNA using cDNA-PCR sequencing kit SQK_PCS_9035_v108_revD_26.6.17 (Oxford Nanopore Technologies, Oxford, UK). The library was sequenced using MinION and the flow cell FLO-MIN107 R9 (Oxford Nanopore Technologies, Oxford, UK). ONT's albacor software, version 2.3.1 with standard parameters, was used for base calling and only sequences that passed a standard quality check (placed in "pass" folder by basecaller) were used for further analyses.

RNA Illumina reads from Helmkampf *et al.* (2016) were aligned employing Hisat, version 2.1.0 (Kim *et al.* 2015), for a genome-guided assembly. A genome-independent transcript assembly was done using Trinity, version 2.8.4 (Grabherr *et al.* 2011), on the next generation sequencing (NGS) RNA-Seq data, using the Trinity assembly provided by Helmkampf *et al.* (2016). In addition, minimap2, version 2.17 within FLAIR pipeline version 1.4, was used for aligning nanopore long reads and the Trinity assemblies to the genome. Finally, StringTie2, version 2.0.1 (Kovaka *et al.* 2019), was employed in order to link the different transcript assemblies filtered by a minimum FPKM of 0.14, as also performed by Helmkampf *et al.* (2016).

Repeat annotation

We used two independent pipelines for *de novo* repeats discovery, namely RepeatModeler, version 1.0.11 (http://repeatmasker.org/RepeatModeler/), and REPET, version 2.5 (Flutre *et al.* 2011). The obtained libraries were merged with Hymenoptera-specific repeats from RepBase, version 22.07 (Bao *et al.* 2015). TEclass software, version 2.1.3 (Abrusán *et al.* 2009), was used for classification of consensus sequences lacking TE-family assignment. Finally, we removed sequences sharing more than 90% identity by employing CD-HIT, version 4.7 (Fu *et al.* 2012). This resulted in the library consisting of 2595 consensus sequences, which were used to annotate repeats in the *P. californicus* genome using RepeatMasker, version 4.0.7 (Smit *et al.* 2013).

Protein-coding gene annotation

The identification of protein-coding genes (PCGs) in the newly assembled genome of *P. californicus* was carried out by GeneModelMapper (GeMoMa), version 1.6.1 (Keilwagen *et al.* 2018), followed by MAKER2, version 2.31.10 (Holt and Yandell 2011) (see Figure 1). We used annotation of four insect species (*P. barbatus*, *S. invicta*, *Camponotus floridanus*, and *A. mellifera*) to run GeMoMa. These annotations were downloaded from NCBI (see Supplementary Table S1). GeMoMa was run for each reference species separately and the results were merged using the GeMoMa annotation filter (GAF). Next, four runs of MAKER2 were used to refine genome annotation. MAKER2 was used with the following data: GeMoMa predictions, transcript assembly, transcript and protein annotations from relative species, and RepeatMasker annotation (see above). AUGUSTUS (Stanke and Morgenstern 2005), which is a part of the MAKER2 pipeline, was trained on the AUGUSTUS reference model from *Nasonia* for the first run and trained on the created *P. californicus* reference model by applying BUSCO, version 3.0.2 (Waterhouse *et al.* 2018), for the third run. In addition, SNAP (Korf 2004) was performed for the last three MAKER2 runs and trained on Hidden Markov Model (HMM) reference models from gene predictions of the previous run with a minimum length of 50 amino acids and a maximum annotation edit distance of 0.25. Redundant identical transcripts and proteins within the final predictions of MAKER2 were filtered with CD-Hit, version 4.7 (Fu *et al.* 2012).

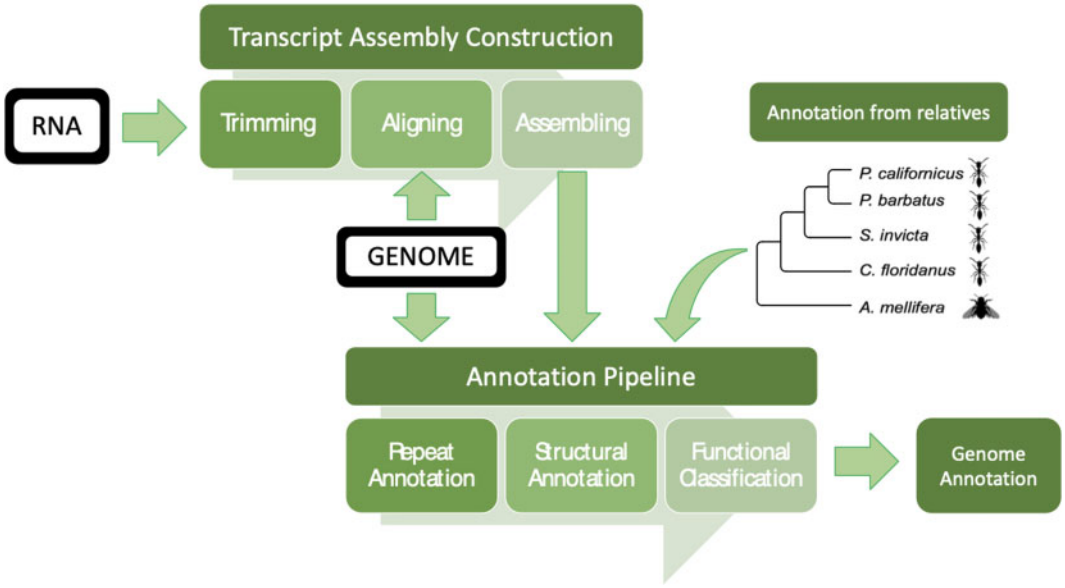

Figure 1 Overview of the annotation workflow. The workflow includes a construction of the transcript assembly (upper part) and a pipeline for the genome annotation (lower part). The transcript assembly and annotations from related species are providing evidence for the annotation of PCGs.

Functional classification of PCGs

The functional classification of the unique PCGs was based on sequence similarity. NCBI's non-redundant (nr) protein database was searched using BLASTP, version 2.2.31 (Altschul *et al.* 1990), with default settings except *e*-value set to 1e−6 and coverage threshold as described below. We considered three possibilities for query and reference sequences overlap. First, the exact matches of the BLAST alignment cover more than 70% to the reference protein and the query protein. In this case, the query protein is similar to the reference protein. Second, if the query protein is just a part of the reference protein, the BLAST matches will cover more than 70% of the query sequence but less than 70% of the reference sequence. Lastly, the reference sequence might be included in the query protein. In this case, the BLAST matches are covering more than 70% of the reference protein but less than 70% of the query protein. This allowed addition of the functional description from the reference protein (annotated protein in the nr database) to the protein query (*P. californicus* protein predicted by MAKER2 annotation). Further downstream analysis was done with Interproscan, version 5.30 (Jones *et al.* 2014), for deletion of protein domain residues in classified and non-classified proteins. This analysis includes several pipelines including PANTHER, Pfam, Gene3D, SUPERFAMILY, MobiDBLite, ProSiteProfiles, SMART, CDD, Coils, PRINTS, TIGRFAM, PIRSF, Hamap, ProDom, and SFLD.

Odorant receptor annotation

We annotated odorant receptors (ORs) for the genomes of *P. californicus* and *P. barbatus* using manually curated OR gene models from three other ant species: *Acromyrmex echinatior*, *Atta cephalotes*, and *S. invicta* (McKenzie *et al.* 2016). Initial OR gene models were annotated with exonerate, version 2.4.0, and GeMoMa, version 1.4, and combined with Evidence Modeler, version 1.1.1 (Haas *et al.* 2008). All models were screened for the 7tm_6 protein domain typical for insect OR proteins with PfamScan, version 1.5. All genes were further assigned to different OR protein subfamilies by aligning the protein sequence against a set of OR subfamily reference sequences (S. McKenzie, personal communication).

Protein alignment was calculated with MAFFT (Katoh *et al.* 2002) using the following parameters: –globalpair = T, –keeporder = T, –maxiterate = 16. The resulting alignment was trimmed employing trimal with the parameters: –keepheader = T –strictall = T (Capella-Gutiérrez *et al.* 2009). The phylogenetic tree of all predicted OR gene models in both ant species was inferred with FastTreeMP (Price *et al.* 2010) with the following settings: –pseudo –lg –gamma.

Annotation of non-PCGs

In addition to the PCGs, non-coding genes were annotated as well. Genes for tRNAs have been annotated with tRNAscan-SE, version 2.0.3 (Chan *et al.* 2019). Other types of ncRNAs, including rRNAs, snRNAs, snoRNAs, miRNAs, and lncRNAs, were predicted by Infernal, version 1.1.2 (Nawrocki and Eddy 2013). To this end, we downloaded the Rfam library, release 14.1, of the covariance models along with the Rfam clan file (https://rfam.xfam.org). Afterwards, cmscan, a built-in Infernal program, was used to annotate the RNAs represented in the Rfam library in the genome under study. Eventually, the lower-scoring overlaps were removed and the final results were used to generate the gff file containing the annotation of non-coding RNA genes. In addition, we searched for homologs of lncRNA genes from *P. barbatus* (based on the annotation of assembly from Supplementary Table S1) in *P. californicus* using Splign, version 2.1.0 (Kapustin *et al.* 2008). Genes where the exons detected by Splign cover more than 90% of *P. barbatus* lncRNA genes were classified as lncRNA genes in the *P. californicus* genome assembly.

Comparative genomic analysis

The LAST aligner, version 909, was used for whole-genome alignments (Kiełbasa *et al.* 2011). The *P. californicus* and *P. barbatus* genome assemblies were aligned in order to find cognate genes and to search for conserved synteny. We used BEDTools intersect, version 2.27.1 (Quinlan and Hall 2010), to compare the genome annotations and estimate the proportion of shared genes.

Assembly and annotation quality assessment

The nucleotide-level quality of final assembly was evaluated using Merqury software (Rhie *et al.* 2020). We assessed the completeness of our assembly and annotation using BUSCO, version 3.0.2 (Waterhouse *et al.* 2018), and DOGMA web server (Dohmen *et al.* 2016; Kemena *et al.* 2019). For BUSCO analyses, we used the Hymenoptera-specific single-copy orthologous genes from OrthoDB, version 9 (Zdobnov *et al.* 2017). For DOGMA, we employed the insect domain core set from Pfam, version 32.

Data availability

All analyses, including the assembly and the annotation pipeline, are available at http://www.bioinformatics.uni-muenster.de/publication_data/P.californicus_annotation/index.hbi. The raw sequencing data are available at the NCBI Sequence Read Archive under accession number PRJNA622899 (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA622899).
Supplementary Material is available at figshare https://doi.org/10.25387/g3.13259183.

Results and discussion

Sequencing results

We performed two rounds of NGS genomic sequencing and transcriptome sequencing using nanopore long reads technology. We obtained 339,494,313 of 100 bp (about 34 Gb) pair-end reads after standard Illumina sequencing and 269,953,173 of 150 bp (about 40.5 Gb) linked-reads using a 10x Genomics technology. Only the latter reads were used for the genome assembly. Additionally, MinION transcriptome sequencing resulted in 394,085 reads ranging between 49 and 6182 bp. N50 of the set was 660 bp and the total size of 241.6 Mb with a median read Phred had a quality score of 7.8, which translates to about 85% accuracy.

Genome assembly and evaluation

We assembled a draft *P. californicus* genome using a linked-read 10x Genomics approach and the Supernova assembler. Assuming 244.5 Mb as the genome size of *P. californicus* (http://www.genomesize.com), our 10x Genomics data coverage was 162×. Supernova was originally designed for *de novo* assemblies of human genomes (Weisenfeld *et al.* 2017). Nevertheless, recently it has been successfully used for non-human genome assemblies (Ozerov *et al.* 2018; Wang *et al.* 2019; Lu *et al.* 2020). For human genomes, a 56× coverage is recommended. However, since there is not much information on the optimal coverage for non-model genomes, we performed a series of assemblies. We resampled our sequencing data to obtain coverages ranging from 47× to 162× (see Figure 2). In order to minimize the number of artificially duplicated and missing BUSCO genes, we decided that a coverage of 75× is optimal for the assembly of the *P. californicus* genome (see Figure 2). Based on this assembly (75× coverage), the *P. californicus* draft genome consisted of 6793 contigs totaling in 240,287,203 bp with about 13 undetermined nucleotides (Ns) per 1 kb. The Supernova assembly was followed by three rounds of polishing by Pilon. This resulted in further improvement of the assembly, with a final assembly of 241,081,918 bp and a reduced number of N characters (see Supplementary Table S2). The nucleotide-level quality value of the final assembly evaluated by Merqury (Rhie *et al.* 2020) was 45.56, which corresponds to 99.99% accuracy (error rate = 2.78e−05).

By comparing the genome assemblies of relative ants used in the annotation pipeline, our genome assembly seems to have a very small N character coverage. This means that we have shorter regions between contigs within scaffolds and less portions of input sequencing reads contain N characters (see Table 1). This impact is very much noticeable by comparing assemblies of the congeners (*P. californicus* and *P. barbatus*) in our set of insects. The N50 of the scaffolds is five times higher because of the about six times higher N character coverage in the *P. barbatus* assembly. Additionally, by considering the assembly size difference of 5 Mb between these two ants, we believe that we present a more complete assembly and consequently a better annotation of the *P. californicus* genome in comparison to the *P. barbatus* one. Moreover, because of a more fragmented genome assembly, the latter may include more erroneous transcript models.

Annotation of repetitive sequences

Annotation of repetitive sequences was performed in two stages. First, we built a library of repetitive elements, which was later used to annotate individual repeats and mask the genome for annotation of different gene types. We used two different *de novo* pipelines to compile consensus sequences of *P. californicus* repetitive sequences, namely RepeatModeler (http://repeatmasker.org/RepeatModeler/) and REPET (Flutre *et al.* 2011). After adding 1240 Hymenoptera-specific repeats from RepBase, our library consisted of 3156 sequences, which were subjected to redundancy filtering using CD-HIT with the cutoff level set at 90%. The final library contained 2595 sequences ranging from 42 to 28,331 bp (median equal to 988 bp). Three hundred forty-five sequences in this dataset were unclassified and TEclass was employed to classify these sequences. We were able to classify most of them and only 71 sequences in our TE library remained unclassified. This library was used as a TE-reference set for a RepeatMasker run. In total, 20.25% of the genome was occupied by repetitive elements, including simple repeats and low complexity regions, 3.98% and 0.53% of the genome, respectively. Not surprisingly, most of the repeats are of TE-origin and all major groups of TEs are represented in the *P. californicus* genome. DNA elements are most common, followed by LTR retroposons and LINEs (see Table 2). Interestingly, SINEs are very rare in the genome. However, it is possible that most of unclassified interspersed repeats are actually SINEs.

Annotation of PCGs

A homology-based GeMoMa annotation followed by four runs of MAKER2 resulted in 15,688 PCGs, which included 170 exact duplicates of potential transcripts. All following downstream analyses were based on a non-redundant set of 15,518 transcripts and translated proteins of this set were referred to as being unique. Additionally, 2288 unique isoforms were annotated by our pipeline based on RNA-seq data. Detailed information on the number of predicted genes at different stages is provided in Supplementary Figure S1. The missing gene numbers presented in this figure come from a BUSCO assessment on unique *P. californicus* transcripts. Isoforms from the MAKER annotation are referred as alternative transcripts with different intron/exon decomposition (Campbell *et al.* 2014). For an annotation with MAKER, it is recommended to run it at least three times. There is a drastic increase of annotation in the second MAKER run, based on the training used SNAP from filtered annotations of the first MAKER run. The high reduction of annotations in the third MAKER run is based on the training of Augustus on the *P. californicus* genome using BUSCO and forcing detection of start and stop codons in order to predict complete genes.

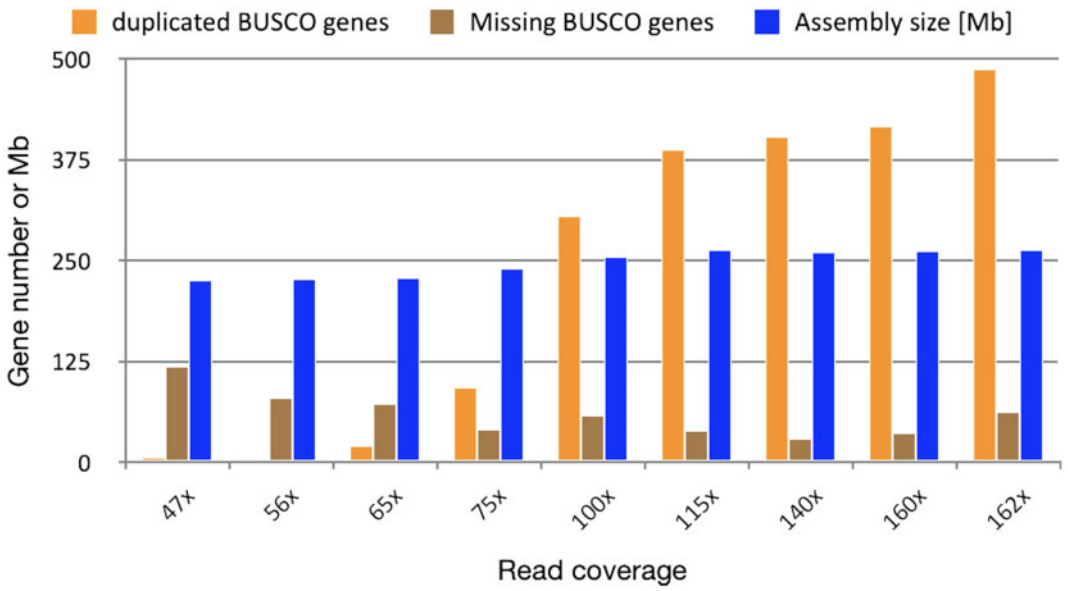

Figure 2 Raw read coverage effect on assembly size and quality. Please note that assembly size is provided in mega base pairs.

Table 1 Comparison of genome assemblies of related insect species

Parameter	*P. californicus*	*P. barbatus*	*S. invicta*	*C. floridanus*	*A. mellifera*
Assembly size	241 Mb	236 Mb	399 Mb	284 Mb	250 Mb
Scaffold N50	208,871 bp	819,605 bp	621,039 bp	1,585,631 bp	997,192 bp
Scaffold N90	16,229 bp	117,988 bp	1,950 bp	211,219 bp	147,519 bp
Number of scaffolds	6,793	4,645	66,904	657	5,644
Percent of N characters in the assembly	1.15	6.60	8.31	0.62	8.45
GC content	36.7	36.5	36.2	34.3	32.7
RefSeq assembly ID	n/a	GCF_000187915.1	GCF_000188075.2	GCF_003227725.1	GCF_000002195.4

With the exception of *P. californicus*, the data were taken from NCBI.

Table 2 Transposable elements present in the *P. californius* genome

TE-class	Number of elements	Fraction of the genome
LTR	15,391	4.38%
LINE	9,525	1.42%
SINE	596	0.03%
DNA	72,737	8.69%
Unclassified	13,610	1.37%

The functional classification of predicted genes was done employing BLASTp against NCBI's nr protein database. We distinguish three categories of functional annotation: (1) 8807 predicted genes were similar to a protein present in nr database with the alignment coverage on query and target of at least 70% of the protein length; (2) 3129 predicted genes were similar to an nr-protein with an alignment coverage of query or target with less than 70% of the protein length, and (3) 2047 predicted proteins where neither the query nor the target fulfill the 70% alignment coverage threshold. These predictions show some similarity to proteins but may be novel proteins as they are not clearly classified. About 1535 predicted proteins did not have any cognate protein in nr database. Therefore, in total, we classified about 90% of all predicted proteins. These include also 54 proteins that consist of multiple domains potentially representing individual proteins. This may be the result of protein fusion or erroneous gene prediction.

Interestingly, from these 1535 potential orphan genes from *P. californicus*, 544 are apparently present in the *P. barbatus* genome; however, they are missing from the current *P. barbatus* annotation. The number of orphan genes or TSG/LSG (taxon-specific/lineage-specific genes) in *P. californicus* is what would be expected for two relatively closely related ant species but is much lower than what has been shown in leaf cutter ants (Wissler *et al.* 2013). In comparison to other relative insect genomes, we have annotated more PCGs (see Table 3). This may suggest that our pipeline resulted in some false-positive predictions. Interestingly but not surprisingly, non-classified proteins are on average significantly shorter than classified proteins: non-classified proteins are on average 108 amino acid long versus a 536 amino acid average length for classified proteins (see Supplementary Figure S2).

In addition to the sequence similarity classification, we also performed further protein domain analysis with Interproscan. Ninety-one percent of classified proteins include predictions from Interproscan (see Supplementary Figure S3), while only 24% of non-classified proteins show some Interproscan predictions (see Supplementary Figure S4). The Interproscan results from classified predictions include mostly predictions from PANTHER (Protein ANalysis THrough Evolutionary Relationships), which is a protein classification system (Thomas *et al.* 2003) and Pfam, which is a large collection of protein domains (El-Gebali *et al.* 2019). These predictions promote the evidence of the classified proteins. Most predictions of the non-classified proteins are coming from MobiDBLite, which is included in the Interpro database and is used for detection of long intrinsically disordered regions (Necci *et al.* 2017). Based on Intrinsic disorder (ID) and missing

Table 3 Comparison of *P. californicus genome* annotation with selected Hymenopteran genomes

Species	Assembly size	Protein coding	tRNA	lncRNA	Other RNA	Total	Assembly version	Annotation version
Acromyrmex echinatior	296 Mb	11,219	159	1,210	449	13,037	Aech_3.9	100
Camponotus floridanus	233 Mb	12,512	208	1,243	696	14,659	Cflo_v7.5	102
Dinoponera quadriceps	260 Mb	11,048	212	570	493	12,323	ASM131382v1	100
Harpegnathos saltator	335 Mb	12,654	230	1,385	928	15,197	Hsal_v8.5	102
Linepithema humile	220 Mb	11,610	178	1,411	655	13,854	Lhum_UMD_V04	100
Monomorium pharaonis	326 Mb	14,019	186	3,126	1,318	18,649	ASM1337386v2	102
Ooceraea biroi	224 Mb	11,907	202	1571	970	14,650	Obir_v5.4	100
Pogonomyrmex barbatus	236 Mb	11,348	201	1,138	406	13,093	Pbar_UMD_V03	101
Pogonomyrmex californicus	241 Mb	15,688	1,180	931	79	17,878	n/a	n/a
Pseudomyrmex gracilis	283 Mb	11,572	193	935	558	13,258	ASM200609v1	100
Solenopsis invicta	399 Mb	14,820	227	1,376	691	17,114	Si_gnH	103
Apis mellifera	225 Mb	9,935	218	3,146	1,295	14,594	Amel_HAv3.1	104

All the data are taken from NCBI's genome database except *P. californicus*.

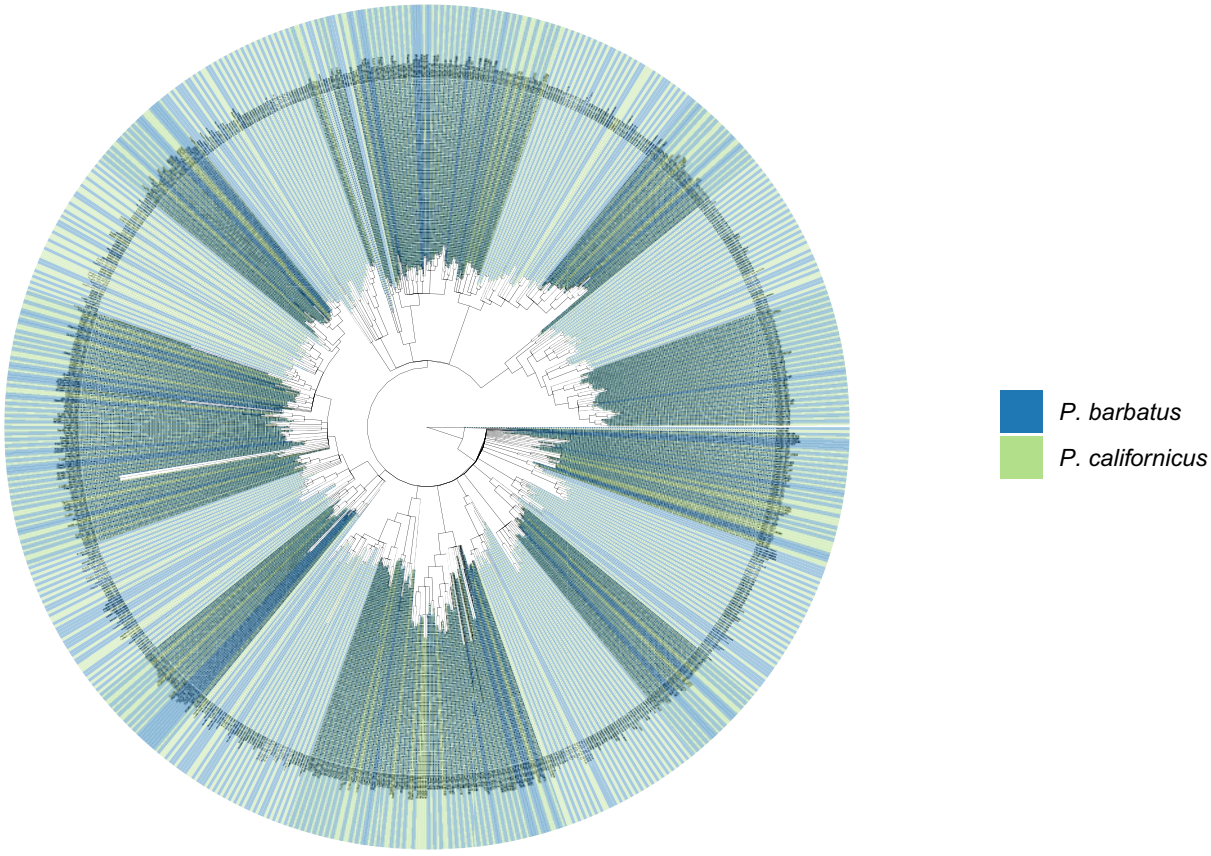

Figure 3 OR gene repertoires are similar between *P. californicus* ($N = 417$ genes) and *P. barbatus* ($N = 453$ genes). Most gene models have their closest relative in the other species. The gene tree shows no large clusters containing genes exclusively of one of the two species. This is evidence for a close relatedness between the species and an equally high quality of the two genome assemblies.

Domains in the Pfam database, at least 10% of the human proteome are missing protein domain detections (Mistry *et al.* 2013). This suggests that these proteins are non-classified based on ID and/or incomplete databases. Based on the length of the non-classified proteins (see Supplementary Figure S2), they seem to include several small proteins which are very much of biological importance but not annotated by most annotation pipelines (Su *et al.* 2013).

Odorant receptors

Chemical communication and perception of olfactory cues via ORs is essential for the performance of many tasks in ant colonies (Trible *et al.* 2017; Yan *et al.* 2017). Given the biological significance of this gene family in ants, we generated in-depth annotations of OR genes in the two closely related Pogonomyrmex species, *P. californicus* and *P. barbatus*, for which assembled genomes are available. Our custom pipeline predicted 417 OR gene models in the *P. californicus* genome and 454 OR gene models in the *P. barbatus* genome. Of these, 303 gene models were complete in *P. californicus* and 342 were complete in *P. barbatus* (see Figure 3, Supplementary Table S3). This nearly doubles the number of originally predicted OR gene models (274) published for *P. barbatus* (Smith *et al.* 2011). Classifying our gene models by known OR gene families showed that most of them fall into the

Table 4 Comparison of completeness and quality of Hymenopteran insects used for the annotation of *P. californicus*

Species	BUSCO genome completeness (%)	BUSCO genome duplication (%)	BUSCO transcript completeness (%)	DOGMA transcript completeness (%)
P. californicus	95.80	2.20	91.60	94.80
P. barbatus	94.20	0.10	95.80	97.60
S. invicta	85.70	0.30	94.10	96.50
C. floridanus	85.90	0.30	99.20	98.10
A. mellifera	97.10	0.20	98.60	98.30

BUSCO and DOGMA analyses are based on unique sets of transcripts without duplicated sequences and soft-masked genomes were used in BUSCO assessments.

9-exon (9E) family, with the next biggest families being L, V, E, and U in both *P. barbatus* and *P. californicus* (see Supplementary Table S4). This is in line with previous studies about OR genes in ants (Engsontia *et al.* 2015; McKenzie and Kronauer 2018). A phylogenetic analysis of Pogonomyrmex ORs showed that most ORs can be considered as single-copy orthologs, as expected when comparing two closely related species. Clusters in the gene phylogeny of multiple genes from the same species would indicate either very recent gene duplications or losses (*i.e.* after the species split) or could hint at assembly errors in either genome. The lack of extensive same-species clusters (largest cluster: seven genes, no other cluster exceeding four genes) thus suggests that the assemblies are of equally high quality, with few signs of gene duplication through assembly errors.

Annotation of non-PCGs

There are several categories of functional RNAs, including tRNAs, lncRNAs, rRNAs, snoRNAs, snRNAs, and rRNAs. Annotation of some of these is pretty straightforward thanks to the conserved secondary structure, *e.g.* tRNA or snRNA genes, and some are more difficult to annotate, *e.g.* lncRNA genes. Nevertheless, we were able to annotate more than 2000 such genes in the *P. californicus* genome (see Table 3). In general, numbers of non-PCGs detected in the *P. californicus* genome are similar to those in other insect genomes with the exception of tRNA genes exceeding more than five times the usual number of tRNA genes in insect genomes. Upon close inspection, it appeared that the excess of tRNA genes is due to unusual number of $tRNA^{Thr}$ genes and in particular its GGT isotype. Moreover, these genes are identical to each other including 200-bp flanking regions, thus suggesting that they might be an artifact of faulty assembly and not a real biological phenomenon.

Assembly and annotation quality assessment

BUSCO and DOGMA programs were used for quality assessment. These programs work on different signatures in order to estimate the completeness of genome assemblies and the resulting annotation of transcripts and proteins. Duplicated transcripts and proteins within annotations of relative genomes were detected using cd-hit as it was done for the *P. californicus* annotation (see Table 4). In general, results from the two programs are in good agreement. The small differences are consequences of different methodology employed by the software; while BUSCO is searching for single-copy orthologous hymenopteran genes, DOGMA searches for Conserved Domain Arrangements (CDA) from an insect reference set. Our annotation of *P. californicus* is comparable to or exceeds annotation of published ant genomes. The only parameter that seems to be significantly different in our assembly is the level of genome duplication reported by BUSCO—over 2% comparing to less than 0.4% in other genomes. This is also reflected in the number of duplicated transcript but interestingly not in the number of duplicated proteins (see Table 4). However, at this point, it is difficult to evaluate if this phenomenon reflects the intrinsic biological feature of the *P. californicus* genome or results from a less-than-perfect assembly of the genome.

Conclusions

With the availability of a genome assembly and annotation for *P. californicus*, we can now start to analyze the genetic architecture of the intraspecific social polymorphism, differences in aggressive behavior of founding queens, and adaptations to desert life in this widely distributed harvester ant. This will also allow us to test whether a supergene, similar to other cases of intraspecific social polymorphism, is responsible for this trait variation. We should also be able to demonstrate that the evolution of OR genes in both *Pogonomyrmex* species proceeded at approximately the same rate without any obvious major gene losses or gains.

Funding

This research was partly funded by the German Research Foundation (DFG) as part of the SFB TRR 212 (NC^3)—project numbers 316099922 and internal fund of the Institute of Bioinformatics.

Conflicts of interest: None declared.

Literature cited

Abrusán G, Grundmann N, DeMester L, Makalowski W. 2009. TEclass–a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics. 25:1329–1330.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Andrews. 2010. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (Accessed: 2017 November 20).

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

Buckley SB. 1867. Descriptions of new species of North American Formicidae. Proc Entomol Soc Philadelphia. 6:335–350.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics. 48:4.11.1–4.11.39.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Chan PP, Lin BY, Mak AJ, Lowe TM. 2019. tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. BioRxiv.doi: 10.1101/614032

Clark RM, Fewell JH. 2014. Social dynamics drive selection in cooperative associations of ant queens. Behav Ecol. 25:117–123.
Cole AC. 1968. Pogonomyrmex Harvester Ants; a Study of the Genus in North America. Knoxville: University of Tennessee Press.
De Vita J. 1979. Mechanisms of interference and foraging among colonies of the harvester ant *Pogonomyrmex californicus* in the Mojave Desert. Ecology. 60:729–737.
Dohmen E, Kremer LP, Bornberg-Bauer E, Kemena C. 2016. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics. 32:2577–2581.
El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, *et al.* 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47: D427–D432.
Engsontia P, Sangket U, Robertson HM, Satasook C. 2015. Diversification of the ant odorant receptor gene family and positive selection on candidate cuticular hydrocarbon receptors. BMC Res Notes. 8:380.
Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. PLoS One. 6:e16526.
Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28: 3150–3152.
Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644–652.
Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, *et al.* 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.
Helmkampf M, Mikheyev AS, Kang Y, Fewell J, Gadau J. 2016. Gene expression and variation in social aggression by queens of the harvester ant *Pogonomyrmex californicus*. Mol Ecol. 25:3716–3730.
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491.
Johnson RA. 2002. Semi-claustral colony founding in the seed-harvester ant *Pogonomyrmex californicus*: a comparative analysis of colony founding strategies. Oecologia. 132:60–67.
Johnson RA. 2004. Colony founding by pleometrosis in the semi-claustral seed-harvester ant Pogonomyrmex californicus (Hymenoptera: Formicidae). Anim Behav. 68:1189–1200.
Jones P, Binns D, Chang HY, Fraser M, Li W, *et al.* 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236–1240.
Julian GE, Fewell JH, Gadau J, Johnson RA, Larrabee D. 2002. Genetic determination of the queen caste in an ant hybrid zone. Proc Natl Acad Sci U S A. 99:8157–8160.
Kapustin Y, Souvorov A, Tatusova T, Lipman D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct. 3:20.
Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics. 19:189.
Kemena C, Dohmen E, Bornberg-Bauer E. 2019. DOGMA: a web server for proteome and transcriptome quality assessment. Nucleic Acids Res. 47:W507–W510.
Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res. 21:487–493.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 12:357–360.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics. 5: 59.
Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, *et al.* 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20:278.
Lighton JR, Turner RJ. 2004. Thermolimit respirometry: an objective assessment of critical thermal maxima in two sympatric desert harvester ants, *Pogonomyrmex rugosus* and *P. californicus*. J Exp Biol. 207:1903–1913.
Lu L, Zhao J, Li C. 2020. High-quality genome assembly and annotation of the big-eye mandarin fish (*Siniperca knerii*). G3 (Bethesda). 10:877–880.
McKenzie SK, Fetter-Pruneda I, Ruta V, Kronauer DJC. 2016. Transcriptomics and neuroanatomy of the clonal raider ant implicate an expanded clade of odorant receptors in chemical communication. Proc Natl Acad Sci U S A. 113:14091–14096.
McKenzie SK, Kronauer DJC. 2018. The genomic architecture and molecular evolution of ant odorant receptors. Genome Res. 28: 1757–1765.
Mistry J, Coggill P, Eberhardt RY, Deiana A, Giansanti A, *et al.* 2013. The challenge of increasing Pfam coverage of the human proteome. Database (Oxford). 2013:bat023.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 29:2933–2935.
Necci M, Piovesan D, Dosztanyi Z, Tosatto SCE. 2017. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics. 33:1402–1404.
Overson R, Fewell J, Gadau J. 2016. Distribution and origin of intraspecific social variation in the California harvester ant *Pogonomyrmex californicus*. Insect Soc. 63:531–541.
Ozerov MY, Ahmad F, Gross R, Pukk L, Kahar S, *et al.* 2018. Highly continuous genome assembly of Eurasian perch (*Perca fluviatilis*) using linked-read sequencing. G3 (Bethesda). 8:3737–3743.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 5:e9490.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.
Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21:245.
Rissing SW, Johnson RA, Martin JW. 2000. Colony founding behavior of some desert ants: geographic variation in metrosis. Psyche. 103:95–101.
Shaffer Z, Sasaki T, Haney B, Janssen M, Pratt SC, *et al.* 2016. The foundress's dilemma: group selection for cooperation among queens of the harvester ant. Sci Rep. 6:29828.
Smit AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. (http://repeatmasker.org).
Smith CR, Cahan SH, Kemena C, Brady SG, Yang W, *et al.* 2015. How do genomes create novel phenotypes? Insights from the loss of the worker caste in ant social parasites. Mol Biol Evol. 32: 2919–2931.
Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, *et al.* 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc Natl Acad Sci U S A. 108:5667–5672.
Snelling RR, Snelling GC, Schmidt JO, Cover SP. 2009. The sexual castes of *Pogonomyrmex anzensis* Cole (Hymenoptera: Formicidae). J Hymen Res. 18:315–321.
Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33:W465–467.

Su M, Ling Y, Yu J, Wu J, Xiao J. 2013. Small proteins: untapped area of potential biological importance. Front Genet. 4:286.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, *et al.* 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 13:2129–2141.

Trible W, Olivos-Cisneros L, McKenzie SK, Saragosti J, Chang NC, *et al.* 2017. Orco mutagenesis causes loss of antennal lobe glomeruli and impaired social behavior in ants. Cell. 170:727–735.e10.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.

Wang J, Wurm Y, Nipitwattanaphon M, Riba-Grognuz O, Huang YC, *et al.* 2013. A Y-like social chromosome causes alternative colony organization in fire ants. Nature. 493:664–668.

Wang W, Yan HJ, Chen SY, Li ZZ, Yi J, *et al.* 2019. The sequence and de novo assembly of hog deer genome. Sci Data. 6:180305.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, *et al.* 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 35:543–548.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. Genome Res. 27: 757–767.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biology and Evolution. 5:439–455.

Yan H, Opachaloemphan C, Mancini G, Yang H, Gallitto M, *et al.* 2017. An engineered orco mutation produces aberrant social behavior and defective neural development in ants. Cell. 170: 736–747.e9.

Yan Z, Martin SH, Gotzek D, Arsenault SV, Duchen P, *et al.* 2020. Evolution of a supergene that regulates a trans-species social polymorphism. Nat Ecol Evol. 4:240–249.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simao FA, *et al.* 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 45:D744–D749.

Communicating editor: E. Betran