




Genome of the bee *Holcopasites calliopsidis*—a species showing the common apid trait of brood parasitism

Trevor J.L. Sless ^{1,*} Jeremy B. Searle ¹ Bryan N. Danforth ²

¹Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853, USA

²Department of Entomology, Cornell University, Ithaca, NY 14853, USA

*Corresponding author: Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853, USA. Email: tjs328@cornell.edu

Abstract

Brood parasites represent a substantial but often poorly studied fraction of the wider diversity of bees. Brood parasitic bees complete their life cycles by infiltrating the nests of solitary host bees thereby enabling their offspring to exploit the food provisions intended for the host's offspring. Here, we present the draft assembly of the bee *Holcopasites calliopsidis*, the first brood parasitic species to be the subject of detailed genomic analysis. Consistent with previous findings on the genomic signatures of parasitism more broadly, we find that *H. calliopsidis* has the smallest genome currently known among bees (179 Mb). This small genome does not appear to be the result of purging of repetitive DNA, with some indications of novel repetitive elements which may show signs of recent expansion. Nor does *H. calliopsidis* demonstrate any apparent net loss of genic content in comparison with nonparasitic species, though many individual gene families do show significant contractions. Although the basis of the small genome size of this species remains unclear, the identification of over 12,000 putative genes—with functional annotation for nearly 10,000 of these—is an important step in investigating the genomic basis of brood parasitism and provides a valuable dataset to be compared against new genomes that remain to be sequenced.

Keywords: *Holcopasites calliopsidis*; brood parasite; bees; Hymenoptera; genome size; gene family evolution

Introduction

Though bees are particularly well-known for eusociality, this way of life is in fact only seen in about 10% of bee species (Danforth et al. 2019). Nonsocial bees demonstrate a wide range of behavioral strategies including specialized plant associations, diverse nesting strategies, and the parasitic exploitation of other bees. With over 2,700 species, bees include a higher proportion of brood parasites than any animal taxon of comparable size (Sless et al., in review). Yet despite this prevalence, brood parasitic species have received relatively little attention, including in the field of genomics. While some 70 bee species now have publicly available genomes through GenBank, nearly three-quarters of these represent social species (Sayers et al. 2020). Currently, 1 brood parasite (*Nomada fabriciana*) has been sequenced (Wellcome Sanger Institute 2021), however, this genome is unannotated and has not been the subject of further analysis.

Broad shifts in genomic organization and architecture have been discovered that can be associated with the evolution of sociality (Kapheim et al. 2015; Shell et al. 2021) and social parasitism (Schrader et al. 2021) in other groups of hymenopterans, and the evolution of brood parasitism similarly involves phenotypic and behavioral shifts which must have a genomic basis. Some general patterns in genomic evolution have been identified among parasitic animals more generally (reviewed by Poulin and Randhawa 2015). One of the clearest of these is a trend toward reduced genome size in parasitic species, which has also been noted for

other parasitic hymenopterans (Ardila-Garcia et al. 2010; Sundberg and Pulkkinen 2015). The typical explanation for this pattern involves relaxed selection on parts of the genome necessary for survival in a free-living organism, but which may be effectively “off-loaded” by a parasite to its host. While this may involve the loss of metabolic or developmental genes in endoparasites that spend their entire lives within a host, parasitic Hymenoptera (including brood parasites) retain a free-living adult stage without losses in basic functionality. Despite this differing dynamic, brood parasites may be thought of as off-loading behavioral responsibilities rather than metabolic attributes (in the form of parental care through nest building and food provisioning) to the host organisms on which they rely for survival. The question then follows: do brood parasitic bees show similar patterns of broad-scale genomic evolution to other parasites?

The genus *Holcopasites* is a member of the oldest and most diverse group of brood parasitic bees, the subfamily Nomadinae (Apidae) and the only Nearctic representative of the tribe Ammobatoideini. The predominant species in northeastern North America, *Holcopasites calliopsidis*, is a specialist parasite of the solitary mining bee *Calliopsis andreniformis*. Though the study of this single species presents a limited opportunity for inferences about the evolutionary signatures of brood parasitism on insect genomes more generally, it represents an important starting point that can serve as the basis for further comparative study as more genomes become available. Therefore, we here provide broad-scale characterization of the genome of *H. calliopsidis*.

Received: April 09, 2022. Accepted: May 31, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Methods

Specimen collection and DNA extraction

Individuals of *H. calliopsidis* were obtained in June of 2020 from the Lime Hollow Nature Center in Cortland, New York, USA (42.57°N, 76.25°W). Specimens were collected by hand-netting, flash-frozen in liquid nitrogen, and sexed with the assistance of a dissecting microscope before storing at -80°C . Only male specimens were used for subsequent steps, due to the benefit of their haploid genomes. High-molecular weight DNA extractions were conducted using a Qiagen 20/G genomic tip kit (catalog #10223) and associated buffers (catalog #19060). The protocol used followed that recommended by Qiagen's Genomic DNA Handbook (accessed December 4th, 2020 from <https://www.qiagen.com/us/resources/resourcedetail?id=d2b85b26-16dd-4259-a3a7-a08cbd2a08a3&lang=en>). Briefly, whole specimens were mechanically homogenized using a sterile pestle before being treated with RNase A (NEB catalog #T3018L) and proteinase K for 2 h. The resulting lysate was added to the equilibrated genomic tip columns, washed 3 times, and eluted. The eluted genomic DNA was then washed one more time with 70% ethanol before final resuspension in TE buffer. Samples were then assayed for quality and DNA concentration using a Qubit 4 fluorimeter.

Sequencing and assembly

Four samples were sent to the University of Maryland's Institute for Genome Sciences for sequencing. These were processed using Pacific Biosciences' low-input library preparation and analyzed for DNA quality and size distribution. The highest-quality sample was then sequenced with a 30-h run on a Sequel II 8M SMRT Cell in CCS/HiFi mode. Raw data, subreads, and circular consensus sequences were subsequently received from the sequencing facility. The *H. calliopsidis* genome was assembled from the CCS reads described above using SPAdes v3.14.0 (Bankevich et al. 2012) in "assembler only" mode. The resulting assembly was analyzed for completeness and quality using QUAST v5.0.2 (Gurevich et al. 2013) and BUSCO v5.0.0 (Seppy et al. 2019).

Repetitive element identification

The genome was searched for repetitive content using 2 passes of the program RepeatMasker v4.1.0 (Smit et al. 2013–2015) in soft-masking mode. An initial run was conducted using a canonical repeat library from RepBase, DFam 5.0 (Storer et al. 2021). The result was then passed into a second run of RepeatMasker with a custom, species-specific repeat library generated using RepeatModeler v2.0.1 (Flynn et al. 2020).

Gene annotation

The repeat-masked genome was annotated using BRAKER v2.1.5 (Brůna et al. 2021) in "EP" mode, along with a database of orthologous arthropod protein sequences obtained from ensembl.org (Howe et al. 2021). Annotation files from different lines of evidence within the BRAKER2 pipeline (Augustus and GeneMark) were combined using EVIDENCEModeler v1.1.1 (Haas et al. 2008) to generate a final annotation file. Transcripts and protein sequences were then extracted from the genome with the software GffRead (Pertea and Pertea 2020) using this annotation file as a guide.

Ortholog detection

Proteomes from 13 other hymenopteran species (*Habropoda laboriosa*, *Apis mellifera*, *Bombus impatiens*, *Ceratina calcarata*, *Osmia bicornis*, *Megachile rotundata*, *Megalopta genalis*, *Nomia melanderi*,

Camponotus floridanus, *Polistes dominula*, *Chelonus insularis*, *Nasonia vitripennis*, and *Athalia rosae*), with *Drosophila melanogaster* as an outgroup, were obtained from GenBank for comparison with newly annotated protein sequences from the *H. calliopsidis* genome. The software OrthoFinder v2.5.4 (Emms and Kelly 2019) was used to identify orthogroups among these protein sequences, running with default settings and using DIAMOND (Buchfink et al. 2015) for sequence comparisons.

Gene family evolution

Expansion and contraction of gene families was analyzed with CAFE v4.2.1 (Han et al. 2013) using the orthogroups detected by OrthoFinder as the input dataset. The initial species tree produced by OrthoFinder did not produce the correct topology in comparison with established phylogenies and cannot produce an ultrametric tree when provided with a preset topology. As a result, the time-calibrated tree required by CAFE was generated manually using divergence time estimates from a previous phylogenomic study (Peters et al. 2017).

Functional analysis

Initial gene ontology annotations were conducted with both InterProScan v5.51-85.0 (Jones et al. 2014) and DIAMOND v2.0.9 (Buchfink et al. 2015) using the Uniref90 reference dataset. The resulting .xml files from both programs were passed on to Blast2GO v1.5.1 (Götz et al. 2008). Functional enrichment analyses were conducted in Cytoscape (Shannon et al. 2003) using the BiNGO application v3.0.4 (Maere et al. 2005) with a custom annotation file modified from the Blast2GO output.

Results

Genome assembly

The final genome assembly after running SPAdes on the circular consensus sequencing data included 5,627 contigs, with 491 > 1 kb in length (Fig. 1). This resulted in a total assembly size of just over 179 Mb, corresponding to an expected C-value of 0.183 pg. Though the estimated coverage of 47 \times was lower than some other recent assemblies, the N50 value of nearly 4.8 Mb compares favorably (Table 1). The L50 value of 14 indicates that the largest several contigs likely represent entire chromosomes or major sections thereof, however the exact chromosome count and structure cannot be accurately determined from the assembly. The considerable number of extremely small scaffolds can be effectively ignored, as the smallest 5,188 of these (<1 kb) contain just 0.56% of the overall assembly despite making up 92.2% of all scaffolds. This genome assembly appears to be largely complete, with 97% of the 5,991 highly conserved single-copy orthologs in BUSCO's Hymenoptera dataset identified (the remainder comprising 0.3% duplicated orthologs, 0.6% fragmented, and 2.1% not detected). It is also noteworthy that the genome of *H. calliopsidis* has one of the highest percentages of GC content of any known bee at approximately 42.5%.

Genome size

Based on a comparison of other assemblies published in GenBank, the 179 Mb *H. calliopsidis* genome assembly is the smallest of any bee species (Fig. 2), and among other Hymenopterans is exceeded in this respect only by some parasitoid wasps of the family Braconidae. The high proportion (97%) of orthologs detected by the BUSCO analysis above indicates that this result is not simply an artifact due to missing DNA, coverage problems, or a poor assembly. The true genome size should therefore be very

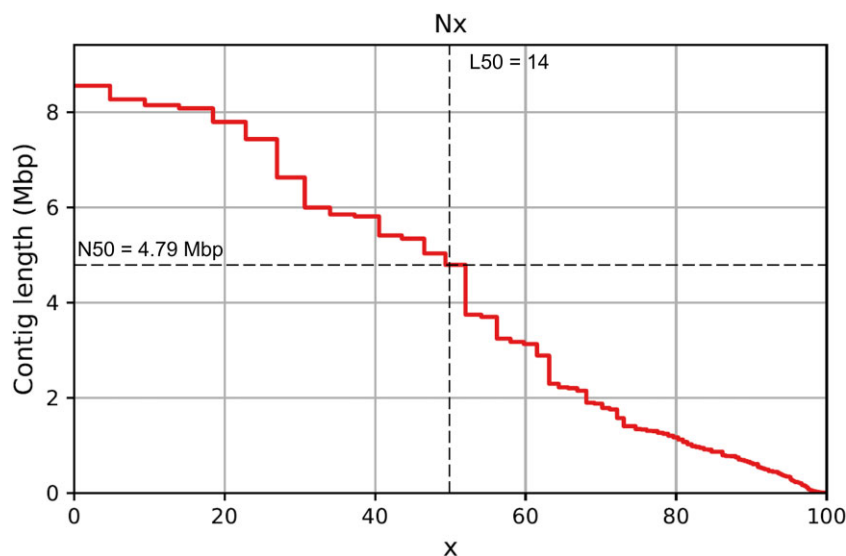


Fig. 1. “Nx plot” generated by QUAST (Gurevich et al. 2013) showing contigs ranked by size vs cumulative assembly completeness. N50 and L50 values (minimum contig size and minimum number of contigs containing 50% of the assembly, respectively) for the *H. calliopsidis* genome are indicated by horizontal and vertical dashed lines.

Table 1. Comparison of assembly statistics for *H. calliopsidis* and recent assemblies from a selection of nonparasitic species spanning the diversity of bees (in approximate order of increasing phylogenetic distance from *H. calliopsidis*).

Species	Family: subfamily	Assembly size (Mb)	No. of scaffolds	N50 (bp)	Coverage (×)	Source
<i>Holcopasites calliopsidis</i>	Apidae: Nomadinae	179	5,627 (491 > 1 kb)	4,790,652	~47	This study
<i>Habropoda laboriosa</i>	Apidae: Anthophorinae	297	27,566	1,784,116	113	Kapheim et al. (2015)
<i>Apis mellifera</i>	Apidae: Apinae	225	177	13,619,445	192	Wallberg et al. (2019)
<i>Bombus impatiens</i>	Apidae: Apinae	248	5,559	1,399,493	108	Sadd et al. (2015)
<i>Megachile rotundata</i>	Megachilidae: Megachilinae	273	6,266	1,699,680	272	Kapheim et al. (2015)
<i>Nomia melanderi</i>	Halictidae: Nomiinae	299.6	3,194 > 1 kb	2,054,768	75	Kapheim et al. (2019)
<i>Colletes gigas</i>	Colletidae: Colletinae	273	326	8,109,000	147.5	Zhou et al. (2020)

close to the size of this assembly, and possibly even smaller if some trailing contigs represent microbial sequences rather than *H. calliopsidis* DNA.

Repetitive content

The initial run of RepeatMasker using the Dfam canonical repeat library identified a relatively small proportion of repetitive sequences in the *H. calliopsidis* genome. Less than 1% of the assembly is composed of known noninterspersed elements including satellite sequences and short simple repeats (SSRs), in contrast to ~4% in the genome of the Western honey bee, *A. mellifera*. Similarly, *H. calliopsidis* has proportionately about half as many canonical retroelements than *A. mellifera* (2.12% vs 4.18%), though slightly more DNA transposons in direct comparison (1.79% vs 0.57%). However, the second run using a species-specific library produced by RepeatModeler identified a sizeable proportion of the *H. calliopsidis* genome as “unclassified” repeats not yet represented in any database of repetitive sequences, comprising a total of nearly 14% of the assembly. These unclassified repetitive elements specifically included a total of 698 distinct families with an average size of 200.9 bp and average of 154.6 copies masked throughout the genome.

Gene annotation

The BRAKER2 pipeline identified 13,028 putative protein sequences derived from 12,364 predicted genes throughout the *H. calliopsidis* genome. Overall, this is similar to the number of genes reported from most other bee species. Though exact identity of these genes

is often not possible to ascertain, the overall number and density indicates no net loss of genes in the *H. calliopsidis* genome in contrast to nonparasitic bees. Additionally, the average number and length of introns is similar between *H. calliopsidis* and other species. For example, *H. calliopsidis* averages 5.34 introns of 524.3 bp each per transcript, while the comparable figures for the genome of *Colletes gigas* average 4.99 introns of 665 bp each (Zhou et al. 2020), resulting in comparable proportions of intronic content (62% and 67% of overall gene content, respectively).

Orthology detection

OrthoFinder identified a total of 21,872 orthogroups across the proteomes of the included species (14 hymenopterans + *D. melanogaster*), which together accounted for 97.5% of all genes. The mean number of genes per orthogroup was 15.4 (median = 7), and a total of 4,298 orthogroups contained genes from every included species (of which 101 were single-copy orthogroups). A total of 10,913 putative genes from the *H. calliopsidis* genome were placed into exactly 9,700 orthogroups which were conserved at a range of phylogenetic scales (Fig. 3). Although this is similar to the average number of orthogroups for all species (mean = 9,910.47), *H. calliopsidis* had a noticeably higher fraction of genes which could not be assigned to any orthogroup (16.2%) than other species.

A total of 108 orthogroups containing 393 genes were identified that appear to be unique to the *H. calliopsidis* genome—henceforth referred to as “*H. calliopsidis*-specific” genes. An important caveat of this terminology is that, due to the taxonomic bias of proteomes available for comparison, it is not possible to

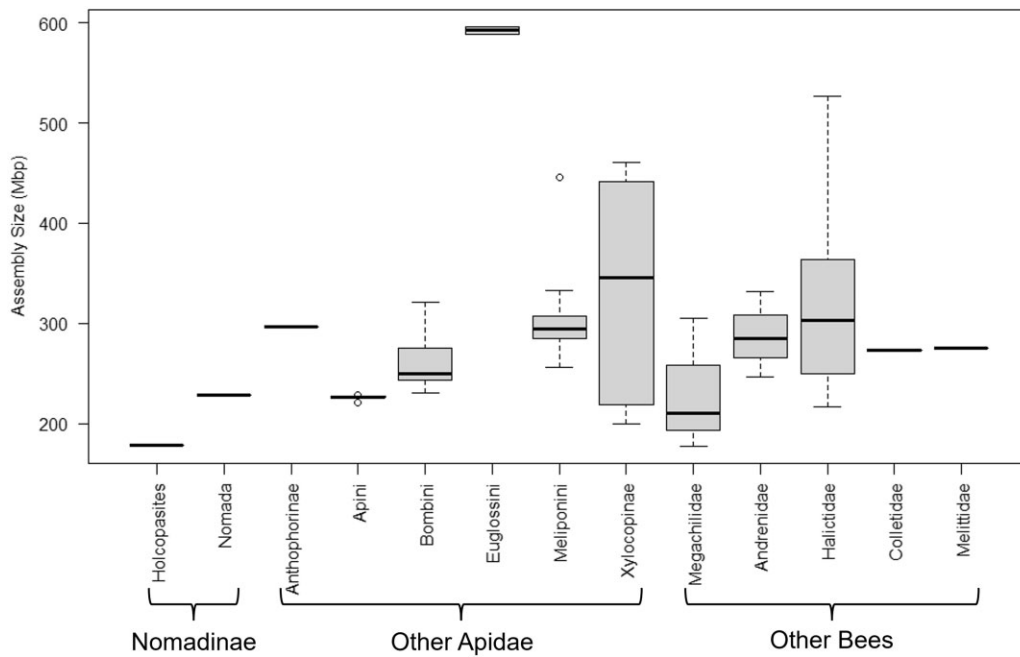


Fig. 2. Size comparison in megabases of all bee species with genome assemblies available on GenBank ($n = 70$), arranged by approximate phylogenetic distance from *H. calliopsidis*.

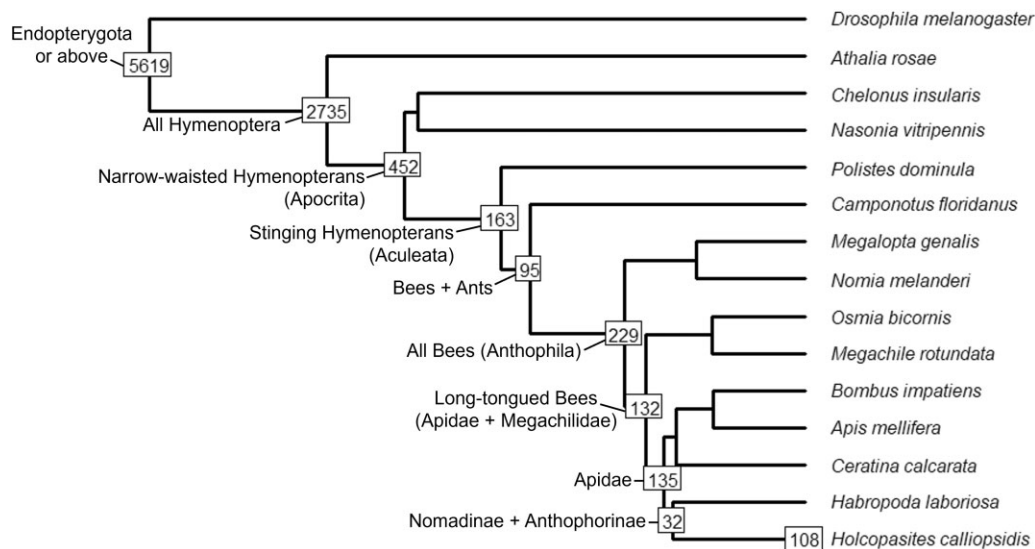


Fig. 3. Phylogeny of *H. calliopsidis* and 14 additional species used in orthology analysis. Boxes at nodes indicate the number of orthogroups containing at least 1 *H. calliopsidis* gene (9,700 in total) which were phylogenetically conserved at the level of the indicated clade. The box at the tip of the *H. calliopsidis* branch indicates orthogroups not found in any other species. Branch lengths are to scale, with approximate divergence times taken from [Peters et al. \(2017\)](#).

determine whether these genes are truly unique to *H. calliopsidis*, or more broadly conserved across the entire genus *Holcopasites*, tribe Ammobatoidini, subfamily Nomadinae, or an intermediate clade between these. However, these numbers are somewhat smaller than the overall mean of 390.6 species-specific orthogroups (and 1,570 species-specific genes) among all 14 hymenopteran genomes, despite the longer branch length of *Holcopasites* in contrast to several other included bees. Outside of the Nomadinae, 32 orthogroups were shared with their closest sister group (digger bees in subfamily Anthophorinae), 135 conserved across the family Apidae, 229 conserved across all bees, and 2,735 shared across all Hymenoptera. However, the majority

of orthogroups containing *H. calliopsidis* genes (5,619) were identified as conserved at the level of holometabolous insects (represented by *D. melanogaster*) or higher.

Gene family evolution

The orthogroups described above were further analyzed with CAFE to identify nodes in the phylogeny of the 15 included species with significant gene family expansions or contractions ([Fig. 4](#)). A single birth-death rate parameter ($\lambda = 0.00259067$) was estimated for the entire tree. Significantly elevated rates of gene family evolution (family-wide P -value < 0.01) were detected in 831 (3.80%) of orthogroups, with 212 rapidly evolving gene families identified in *H.*

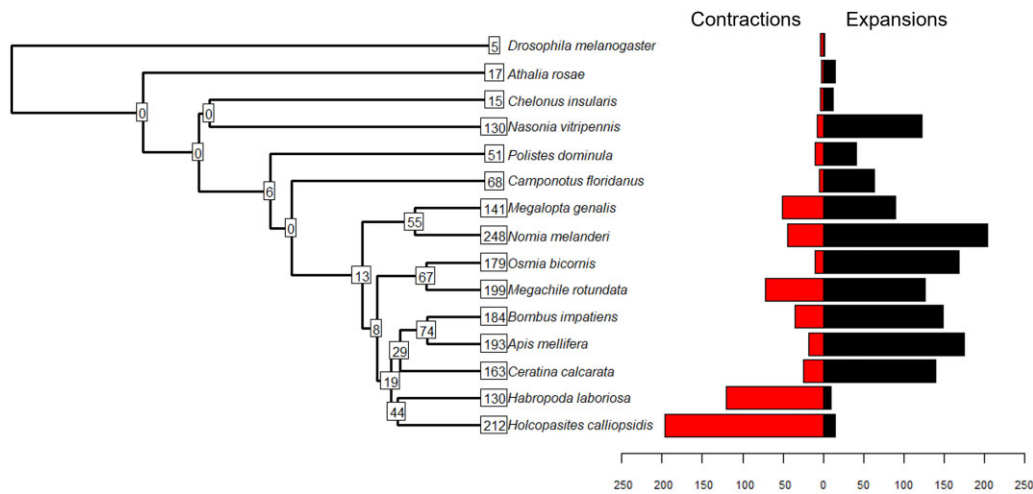


Fig. 4. Phylogeny of *H. calliopsidis* and 14 additional species used in analysis of gene family evolution. Numbers at nodes and tips indicate how many rapidly evolving gene families were detected by CAFE for each species/lineage, respectively. Bars further separate rapidly evolving gene families for each tip into contractions or expansions in gene family size.

calliopsidis specifically (15 expansions, 197 contractions). In comparison with the other included species, this represents the second-highest number of rapidly evolving gene families (after *Nomia melanderi* with 248), but by far the most gene family contractions. Interestingly, when ranked by statistical significance, 9 of the top 10 gene families which have undergone the most rapid evolution in *H. calliopsidis* are expansions rather than contractions. Of these, several families include genes with blast2GO annotations suggesting involvement with transposable or retroviral elements, which may indicate a recent or ongoing spread of such features in the genome.

Functional analysis

In total, InterProScan and blast2GO assigned over 60,000 GO terms to 9,943 genes annotated from the *H. calliopsidis* genome, leaving about one-quarter of predicted genes without functional annotation. Analysis of the distribution of level 3 GO terms indicates few surprising features; various metabolic processes account for the most commonly assigned biological processes, and the most common molecular functions included enzymatic activity and binding of proteins and other compounds.

A BiNGO analysis of the 393 *Holcopasites*-specific genes identified through orthology detection compared to the genome as a whole identified several GO terms that appear enriched for these loci (Fig. 5). Several of these (e.g. transposition, DNA integration, and various classes of endonuclease activity) may in fact be signatures of the large fraction of the genome containing presumed *Holcopasites*-specific repetitive elements as identified above.

A second BiNGO analysis compared the entire proteomes of our *H. calliopsidis* assembly with those of *Ha. laboriosa* (the phylogenetically closest well-annotated genome to *H. calliopsidis*) as well as *A. mellifera* to identify the GO terms which appear enriched in *H. calliopsidis* in contrast to its nonparasitic relatives or vice versa. Overall, this analysis was inconclusive. Many GO terms were identified as being significantly over- or underrepresented in *H. calliopsidis* in direct comparison to the other 2 species, yet in some cases these findings may be artifactual. Differences in the depth of gene ontology annotations achieved for the 3 genomes can result in artificial overrepresentation of GO terms which may not appear in all species, but which are simply child nodes of other well-represented terms. For example, GO terms

for synthesis of all essential amino acids were identified as over-represented in *H. calliopsidis* over *Ha. laboriosa*/*A. mellifera*, though it is difficult to explain this biologically. Some terms (e.g. neurogenesis/neuron development, oogenesis) were underrepresented in *H. calliopsidis* in ways which appear superficially consistent with a priori expectations about adaptations to the brood parasitic life history strategy. However, this must be considered with the caveat that enrichment of certain ontology terms, which may represent expansion of genes involved in a given biological process, is still separated from the actual control of such processes by several layers of regulation (e.g. transcriptional regulation, splicing, post-translational modifications).

Discussion

The most striking feature of our broad-scale characterization of the *H. calliopsidis* genome is its small genome size, and consequently we focus the discussion on this attribute. However, as the first brood parasitic bee for which a well-characterized genome has been produced, there is much which remains to be learned. Brood parasitism has evolved at least 20 times independently in bees, with *H. calliopsidis* representing the oldest and most diverse such clade in the subfamily Nomadinae (Sless et al. in review). In this sense, our study has opened a new avenue of genomic research in relation to a well-known and fascinating behavioral strategy. This work is also complementary to nascent genomic studies on brood parasitic birds (such as cuckoos and cowbirds), which share a common life history despite their great phylogenetic distance from bees, and similarly include recently sequenced genomes available through GenBank but awaiting in-depth analysis (Wolf et al. unpublished data; Wuitchik et al. unpublished data; Zhang unpublished data).

Genome size reduction

While most bee genomes fall into a fairly consistent range of ~250–350 Mb, with a mean of 287.05 Mb for all 70 species with genomes available on GenBank, our assembly for *H. calliopsidis* is noticeably smaller at just 179 Mb. Methodological explanations for this finding seem unlikely, but corroboration of this estimate with cytological techniques could serve as an additional line of evidence. Assuming

especially when they occur over brief time periods, have often been attributed to the expansion or purging of repetitive elements (e.g. Hancock et al. 2021). Indeed, some bee lineages such as the orchid bees (Euglossini) appear to have undergone massive genomic expansions as a result of increased repetitive content (Brand et al. 2017). However, the reverse situation in which repeat sequences are purged from the genome does not appear to be the case in *H. calliopsidis*, which shows comparable or even higher levels of repetitive content to other species as a fraction of the genome. The large fraction of “unclassified” repetitive DNA which could not be matched to existing databases is likely due to the fact that this de novo genome assembly has no close relatives which have been previously analyzed for repetitive content. A similar phenomenon has been reported for at least one other phylogenetically distinct genome, that of *Colletes gigas*, which also consisted of >10% “unclassified” repetitive elements (Zhou et al. 2020).

Similarly, the reduced genome size in *H. calliopsidis* does not appear to be a result of net loss in genic content. Though many genes found in other species did not have orthologues annotated in *H. calliopsidis*, the overall number of putative genes identified was similar to other known bees at approximately 12,000. The average number and length of introns is also similar between *H. calliopsidis* and *A. mellifera*, which would seem to rule out any substantial purging of noncoding sequences within genes. However, this certainly does not mean that genic content is static or identical to that of other species. Many gene families are identified as having experienced rapid evolution in *H. calliopsidis*, though with the caveat that differing methodologies for gene annotation used across the compared species may artificially influence this analysis. It may be the case that *H. calliopsidis* does in fact exhibit the loss of a large number of genes associated with nonparasitic life histories, but that these losses are compensated to some extent by a number of rapidly evolving gene families. This finding may in fact be related to the large amount of “unclassified” repetitive elements discussed above, since transposable elements containing functional genes may be included among the most rapidly expanding gene families. It remains unclear exactly which types of material account for the reduction in *H. calliopsidis*’ genome size compared to the ancestral bee genome.

Other features

Our genome assembly for *H. calliopsidis* is noticeably more GC rich (42.5%) than most other bee genomes. Of those available on GenBank, the mean GC content is 37.6%, and only 2 species (*Andrena dorsata* and *haemorrhhoa*) surpass *H. calliopsidis* in this regard (Sayers et al. 2020). It is unclear what relation if any this may have to the evolution of brood parasitism, however. Proportion of GC bases is not generally correlated with genome size in animals (Li and Du 2014), though a positive correlation between GC content and recombination rates has been noted in honey bees (Beye et al. 2006).

The orthology detection analysis with OrthoFinder identified fewer unique orthogroups containing only *H. calliopsidis* genes than for any other included species besides *Ha. laboriosa*. This may initially seem to suggest that *H. calliopsidis* has experienced less evolution of novel genes than most other Hymenoptera. However, over 2,000 putative *H. calliopsidis* genes (a higher fraction than any other species) were not able to be assigned to any orthogroup, and some of these are surely also unique (but single-copy) genes that arose sometime after the split between Nomadinae and its sister group. Unsurprisingly though, the lack of interspecific orthologs for these unassigned genes also meant that the vast majority of them received no functional annotations from blast2GO, and so their identity remains unclear.

Future directions

The annotated draft genome of *H. calliopsidis* presented herein is a major first step toward understanding the genomic basis of brood parasitism in the Nomadinae and in bees more broadly. However, as already addressed throughout this study, the amount of information which can be gained from a single genome is inherently limited. From this dataset alone, it is possible to identify many interesting features of genomic content in *H. calliopsidis*, but not their age or phylogenetic distribution as orthologs. Some such features may in fact be ancient signatures of the initial transition to brood parasitism in the subfamily Nomadinae approximately 100 MYA, while others could be much more recent innovations related to *H. calliopsidis*’ status as a narrow host specialist.

Future investigation into the genomics of brood parasitism should focus on sampling other species representing a wider phylogenetic distribution of parasites, including other members of the Nomadinae as well as bees representing independent origins of parasitism. Comparative genomic investigation has already yielded interesting results in studying the genomic basis of social behavior in bees (e.g. Kapheim et al. 2015; Shell et al. 2021), and it is our hope that the *H. calliopsidis* genome will serve as a starting point for the parallel exploration of the equally fascinating life history innovation of brood parasitism.

Data availability

Our *H. calliopsidis* genome assembly is available through GenBank (accession # JALHAQ000000000). Data files resulting from major analyses (repetitive elements, gene annotations, functional annotations, and orthology/gene family investigations) have been uploaded to the GSA Figshare portal in association with this manuscript. [Supplementary data files](#) are available through the GSA figshare portal: <https://doi.org/10.25387/g3.19522369>. [Supplementary File 1](#) (“Assembly”) includes the main contig-level assembly for *H. calliopsidis* as well as the results of a QUAST quality assessment and BUSCO analysis of genome completeness. [Supplementary File 2](#) (“Repetitive Elements”) includes the masked assemblies resulting from 2 runs of RepeatMasker and RepeatModeler as described in the Methods section. [Supplementary File 3](#) (“Annotation”) contains output gene annotation files from BRAKER2, protein, and transcript sequences extracted with EvidenceModeler and GffRead, and functional annotation output from InterProScan and Blast2GO, respectively. Finally, [Supplementary File 4](#) (“Orthology”) provides the results of an OrthoFinder run to identify orthogroups among *H. calliopsidis* and other species, as well as an analysis of gene family evolution with CAFE (including the dated phylogeny used as input for this software based on Peters et al. 2017).

Acknowledgments

We thank Luke Tallon of the Institute for Genomic Sciences at the University of Maryland as well as Dr. Margarita López-Urbe of Penn State University for providing valuable advice on genomic DNA extraction and sequencing methods. We also thank Dr. Qi Sun of Cornell University’s Bioinformatics Facility for his expertise and assistance with genome assembly and various in silico analyses. Additionally, we wish to acknowledge that Cornell University, where this study was conducted, is located on the traditional homelands of the Gayogo-hó:nö people within the Haudenosaunee Confederacy.

Funding

TJL Sless was supported by a Postgraduate Scholarship - Doctoral through the Natural Sciences and Engineering Research Council of Canada for part of the duration during which this work was completed. Library preparation and sequencing costs were paid by a U.S. National Science Foundation grant (DEB-1555905) awarded to B.N. Danforth, S.G. Brady, J.P. Pitts, and R. Ross.

Conflicts of interest

None declared.

Literature cited

- Ardila-Garcia AM, Umphrey GJ, Gregory TR. An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. *Insect Mol Biol.* 2010;19(3):337–346. doi:10.1111/j.1365-2583.2010.00992.x.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455–477. doi:10.1089/cmb.2012.0021.
- Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlupalius D, Mougél F, Emore C, Rueppell O, et al. Exceptionally high levels of recombination across the honey bee genome. *Genome Res.* 2006;16(11):1339–1344. doi:10.1101/gr.5680406.
- Brand P, Saleh N, Pan H, Li C, Kapheim KM, Ramírez S. The nuclear and mitochondrial genomes of the facultatively eusocial orchid bee *Euglossa dilemma*. *G3 (Bethesda).* 2017;7(9):2891–2898. doi:10.1534/g3.117.043687.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform.* 2021;3(1):lqaa108. doi:10.1093/nargab/lqaa108.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12(1):59–60.
- Danforth B, Minckley R, Neff J. *The Solitary Bees: Biology, Evolution, Conservation.* Princeton/Oxford: Princeton University Press; 2019.
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. doi:10.1186/s13059-019-1832-y.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA.* 2020;117(17):9451–9457. doi:10.1073/pnas.1921046117.
- Götz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 2008;36(10):3420–3435. doi:10.1093/nar/gkn176.
- Gurevich A, Saveliev S, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29(8):1072–1075. doi:10.1093/bioinformatics/btt086.
- Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* 2008;9(1):r7. doi:10.1186/gb-2008-9-1-r7.
- Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol.* 2013;30(8):1987–1997. doi:10.1093/molbev/mst100.
- Hancock ZB, Hardin FO, Murthy A, Hillhouse A, Johnston JS. Rapid genomic expansion and purging associated with habitat transitions in a clade of beach crustaceans (Amphipoda: Haustoriidae). *J Crust Biol.* 2021;41(3):ruab042. doi:10.1093/jcabi/ruab042.
- Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. Ensembl 2021. *Nucleic Acids Res.* 2021;49(1):884–891. doi:10.1093/nar/gkaa942.
- Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 2014;30(9):1236–1240. doi:10.1093/bioinformatics/btu031.
- Kapheim KM, Pan H, Li C, Blatti C, Harpur BA, Ioannidis P, Jones BM, Kent CF, Ruzzante L, Sloofman L, et al. Draft Genome Assembly and Population Genetics of an Agricultural Pollinator, the Solitary Alkali Bee (Halictidae: *Nomia melanderi*). *G3 (Bethesda).* 2019;9(3):625–634. doi:10.1534/g3.118.200865.
- Kapheim M, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, et al. Genomic signatures of evolutionary transitions from solitary to group living. *Science.* 2015;348(6239):1139–1144. doi:10.1126/science.aaa4788.
- Li X-Q, Du D. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PLoS One.* 2014;9(2):e88339. doi:10.1371/journal.pone.0088339.
- Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics.* 2005;21(16):3448–3449. doi:10.1093/bioinformatics/bti551.
- Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Res.* 2020;9:304. doi:10.12688/f1000research.23297.2.
- Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, Kozlov A, Podsiadlowski L, Petersen M, Lanfear R, et al. Evolutionary history of the Hymenoptera. *Curr Biol.* 2017;27(7):1013–1018. doi:10.1013/j.cub.2017.01.027.
- Poulin R, Randhawa HS. Evolution of parasitism along convergent lines: from ecology to genomics. *Parasitology.* 2015;142(S1):S6–S15. doi:10.1017/S0031182013001674.
- Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsieck CG, Gadau J, Grimmelikhuijzen CJP, Hasselmann M, Lozier JD, et al. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.* 2015;16:76. doi:10.1186/s13059-015-0623-3.
- Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. *Nucleic Acids Res.* 2020;48(D1):D84–D86. doi:10.1093/nar/gkz956.
- Schrader L, Pan H, Bollazzi M, Schjøtt M, Larabee FJ, Bi X, Deng Y, Zhang G, Boomsma JJ, Rabeling C. Relaxed selection underlies genome erosion in socially parasitic ant species. *Nat Commun.* 2021;12(1):2918. doi:10.1038/s41467-021-23178-w.
- Seppy M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol.* 2019;1962:227–245. doi:10.1007/978-1-4939-9173-0_14.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498–2504. doi:10.1101/gr.1239303.
- Shell WA, Steffen MA, Pare HK, Seetharam AS, Severin AJ, Toth AL, Rehan SM. Sociality sculpts similar patterns of molecular

- evolution in two independently evolved lineages of eusocial bees. *Commun Biol.* 2021;4(1):253. doi:[10.1038/s42003-021-01770-6](https://doi.org/10.1038/s42003-021-01770-6).
- Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. [accessed 2021 Apr 8]. <http://www.repeatmasker.org>.
- Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* 2021;12(1):2. doi:[10.1186/s13100-020-00230-y](https://doi.org/10.1186/s13100-020-00230-y).
- Sundberg L-R, Pulkkinen K. Genome size evolution in macroparasites. *Int J Parasitol.* 2015;45(5):285–288. doi:[10.1016/j.ijpara.2014.12.007](https://doi.org/10.1016/j.ijpara.2014.12.007).
- Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics.* 2019;20(1):275. doi:[10.1186/s12864-019-5642-0](https://doi.org/10.1186/s12864-019-5642-0).
- Wellcome Sanger Institute. Tree of Life QC: *Nomada fabriciana*; 2021. https://tolqc.cog.sanger.ac.uk/darwin/insects/Nomada_fabriciana/.
- Zhou Q-S, Luo A, Zhang F, Niu Z-Q, Wu Q-T, Xiong M, Orr MC, Zhu C-D. The first draft genome of the plasterer bee *Colletes gigas* (Hymenoptera: Colletidae: Colletes). *Genome Biol Evol.* 2020;12(6):860–866. doi:[10.1093/gbe/evaa090](https://doi.org/10.1093/gbe/evaa090).

Communicating editor: M.-A. Félix