The First Draft Genome of the Plasterer Bee *Colletes gigas* (Hymenoptera: Colletidae: *Colletes*)

Qing-Song Zhou[1], Arong Luo[1], Feng Zhang [2], Ze-Qing Niu[1], Qing-Tao Wu[1], Mei Xiong[1,3], Michael C. Orr[1], and Chao-Dong Zhu[1,3,*]

[1]Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[2]Department of Entomology, College of Plant Protection, Nanjing Agricultural University, China

[3]University of Chinese Academy of Sciences, Beijing, China

*Corresponding author: E-mail: zhucd@ioz.ac.cn.

Abstract

Despite intense interest in bees, no genomes are available for the bee family Colletidae. *Colletes gigas*, one of the largest species of the genus *Colletes* in the world, is an ideal candidate to fill this gap. Endemic to China, *C. gigas* has been the focus of studies on its nesting biology and pollination of the economically important oil tree *Camellia oleifera*, which is chemically defended. To enable deeper study of its biology, we sequenced the whole genome of *C. gigas* using single-molecule real-time sequencing on the Pacific Bioscience Sequel platform. In total, 40.58 G (150×) of long reads were generated and the final assembly of 326 scaffolds was 273.06 Mb with a N50 length of 8.11 Mb, which captured 94.4% complete Benchmarking Universal Single-Copy Orthologs. We predicted 11,016 protein-coding genes, of which 98.50% and 84.75% were supported by protein- and transcriptome-based evidence, respectively. In addition, we identified 26.27% of repeats and 870 noncoding RNAs. The bee phylogeny with this newly sequenced colletid genome is consistent with available results, supporting Colletidae as sister to Halictidae when Stenotritidae is not included. Gene family evolution analyses identified 9,069 gene families, of which 70 experienced significant expansions (33 families) or contractions (37 families), and it appears that olfactory receptors and carboxylesterase may be involved in specializing on and detoxifying *Ca. oleifera* pollen. Our high-quality draft genome for *C. gigas* lays the foundation for insights on the biology and behavior of this species, including its evolutionary history, nesting biology, and interactions with the plant *Ca. oleifera*.

Key words: Apoidea, PacBio sequencing, genome assembly, genome annotation, gene family evolution.

Introduction

Bees are arguably the most important group of angiosperm-pollinating insects (Klein et al. 2007; Danforth et al. 2013), pollinating nearly 90% of all flowering plants that require pollination (Ollerton et al. 2011). With more than 20,000 described species (Ascher and Pickering 2018), wild bees substantially contribute to crop yields (Garibaldi et al. 2011, 2013; Leonhardt et al. 2013), making them both ecologically and economically invaluable. Among them, the family Colletidae is a diverse group of >2,700 species, ranging from the small, wasp-like *Hylaeus* that carry pollen internally to the more robust, hairy *Colletes* that share their family name (Michener 2007; Ascher and Pickering 2018). This family was traditionally believed to be the most "primitive" taxon within the superfamily Apoidea (according to mouthpart structure, the similarity of their bilobed glossa to closely related apoid wasps), but molecular studies place Melittidae sister to all other bees (Danforth et al. 2006, 2013; Hedtke et al. 2013; Branstetter et al. 2017; Peters et al. 2017; Sann et al. 2018). Instead, it has been suggested that the bilobed colletid glossa actually evolved for adding their characteristic cellophane-like cell lining to nests (Michener 2007; Almeida 2008). This cell lining, unique to Colletidae, has drawn a great deal of prior study, yet the molecular underpinnings of this behavior

remain unknown. *Colletes* specifically nest in the ground and they are the second-largest genus in the family. Though they have been well studied for their systematics and taxonomy (Michener 2007; Kuhlmann and Proshchalykin 2011, 2013; Niu, Kuhlmann, et al. 2013; Niu, Zhu, et al. 2013; Niu et al. 2014a, 2014b), few studies have examined their molecular phylogenetics and evolution (Kuhlmann et al. 2009; Almeida et al. 2012; Ferrari et al. 2020).

Despite general interest in bees, no whole genome has been reported from the family Colletidae until now (Branstetter et al. 2018). Here, we present the whole-genome sequence of *Colletes gigas*, de novo assembled using single-molecule real-time Pacific Bioscience (PacBio) long reads. We annotated essential genomic elements, repeats, protein-coding genes, and noncoding RNAs (ncRNAs), and further compared gene family evolution across major bee lineages. Further, we carried out phylogenomic analyses of bee families using single-copy (Benchmarking Universal Single-Copy Ortholog [BUSCO]) markers for the first time. We also discuss our findings in relation to *C. gigas* specializing on *Camellia oleifera* (Huang et al. 2015), an economically important tea tree with toxic pollen, documented to deplete honey bee colonies when foraged on (Su et al. 2011). Therefore, this and future studies will prove vital for understanding the evolution of floral specialization, especially for chemically defended resources.

Materials and Methods

Sample Collection, Sequencing, and Quality Control

We collected specimens of *C. gigas* in *Ca. oleifera* plantations in East-Central China (Tangchi Village, Shucheng County, Lu'an City, Anhui Province, China). A total of 17 *C. gigas* specimens were collected, including 2 males and 15 females. The species was identified by author Ze-Qing Niu using traditional morphological approaches based on body size, head breadth, facial fovea, clypeus, mesonotum, and wing venations (Wu 1965; Niu, Kuhlmann, et al. 2013), as well as the biology and phenology (Huang et al. 2015). Species identification was also verified by DNA barcoding analyses using COI gene of genus *Colletes* available in National Center for Biotechnology Information (NCBI; supplementary file S1, Supplementary Material online). Upon collection, specimens were brought back to the laboratory alive, flash-frozen in liquid nitrogen, and stored in −80 °C for long-term preservation. As Hymenoptera use a haplodiploid sex-determination system, we used a single male adult *C. gigas* (NCBI taxonomy ID: 935657) (Voucher Code: AHSC1104, supplementary fig. S1, Supplementary Material online) with its gut contents removed for PacBio sequencing, whereas two female specimens with their gut contents removed were used for Illumina whole-genome (Voucher Code: AHSC1105) and Illumina transcriptome (Voucher Code: AHSC1107)

sequencing. The remaining specimens (Female: AHSC1101-03 and AHSC1108-17; Male: AHSC1106) were deposited at the Institute of Zoology, Chinese Academy of Sciences.

Genomic DNA/RNA extraction, library preparation, and sequencing were conducted by the company Nextomics (Wuhan, China). For long-read sequencing, a library was constructed with an insert size of 10 kb and sequenced using P6-C4 chemistry on the PacBio Sequel platform. For short-read sequencing, paired-end libraries were constructed with an insert size of 400 bp and sequenced (2× 150 bp) on the Illumina HiSeq X Ten platform.

Raw Illumina short reads were compressed into clumps, and duplicates were removed with clumpify.sh (BBTools suite v37.93, Bushnell). Quality control was performed with bbduk.sh (BBTools): Both sides were trimmed to Q20 based on Phred scores, reads shorter than 15 bp or with more than 5 Ns were discarded, poly-A or poly-T tails of at least 10 bp were trimmed, and overlapping paired reads were corrected.

Genome Size Estimation

We employed the strategy of short-read k-mer distributions to estimate the genome size. The histogram of k-mer frequencies was first computed with 17-mers and 21-mers using khist.sh (BBTools). Genome size was then estimated with a maximum k-mer coverage of 1,000 using GenomeScope v1.0.0 (Vurture et al. 2017).

Genome, Mitochondrion, and Transcriptome Assembly

We performed de novo genome assembly with long reads using Flye (v2.4.2) (Kolmogorov et al. 2019) and Falcon (pb-assembly v0.0.4) (Chin et al. 2016) (length_cutoff_pr = 7,000, max-diff = 100, max-cov = 100, and min-cov = 2). Both assemblies were first polished by Flye (–polish_target) on raw PacBio sequences. To improve genome contiguity, the two assemblies generated from Flye and Falcon pipelines were merged into one assembly after two rounds of quickmerge v0.3 (Chakraborty et al. 2016) with USAGE 2 (https://github.com/mahulchak/quickmerge/wiki, last accessed November 12, 2016), which was further polished with Illumina short reads using two rounds of Pilon v1.22 (Walker et al. 2014). Subsequently, we filtered possible contaminants by HS-BlastN (Chen et al. 2015) employing BLAST+ v2.7.1 (Camacho et al. 2009) against the NCBI nucleotide database and checked vector contamination using VecScreen against the UniVec database.

We assembled the mitochondrial genome of *C. gigas* based on Illumina short reads using Mitobim v1.9.1 (Hahn et al. 2013) and with reference to the published mitochondrial genome of *C. gigas* (KM978210, Huang et al. 2016), which was then annotated with MITOS webserver (Bernt et al. 2013). We performed transcriptome assembly under a genome-guided method via HISAT2 v2.1.0 (Kim et al. 2015), mapping RNA sequencing (RNA-seq) reads to our

assembled genome, and then assembled with StringTie v1.3.4 (Pertea et al. 2015). Redundant isoforms were removed with Redundans v0.13c (Pryszcz and Gabaldón 2016) under default parameters. We finally evaluated the completeness of all assemblies using the BUSCO (Waterhouse et al. 2018) analyses against the insect data set ($n = 1,658$).

Genome Annotation

We generated a custom library by combining a de novo species-specific repeat library constructed by RepeatModeler version open-1.0.11 (Smit and Hubley 2008–2015 www.repeatmasker.org, last accessed April 8, 2020) with the Dfam 3.0 (Hubley et al. 2016) and RepBase-20181026 databases (Bao et al. 2015). Repeats were identified and masked using RepeatMasker v4.0.9 (Smit AFA, Hubley R, Green P. 2013-2015 www.repeatmasker.org/faq.html, last accessed April 9, 2019) together with the custom library.

Gene prediction was conducted with the MAKER v2.31.10 pipeline (Holt and Yandell 2011) by integrating ab initio, transcriptome-based, and protein homology-based evidence. Ab initio gene predictions were performed with Augustus v3.3 (Stanke et al. 2004) and GeneMark-ET v4.33 (Lomsadze et al. 2005). We trained two predictors using BRAKER v2.1.0 (Hoff et al. 2016) with RNA-seq data and used previously assembled, genome-guided transcripts as transcriptome-based evidence.

Homology-based gene functions were assigned using Diamond v0.9.18 (Buchfink et al. 2015) and the UniProtKB (SwissProt + TrEMBL) (–sensitive -e 1e–5). Protein domains, gene ontology, and pathway annotations were searched with InterProScan 5.34-73.0 (Finn et al. 2017) against the Pfam (Finn et al. 2014), PANTHER (Mi et al. 2017), Gene3D (Lewis et al. 2018), Superfamily (Wilson et al. 2009), and CDD (Marchler-Bauer et al. 2017) databases. Protein sequences of *Tribolium castaneum*, *Acyrthosiphon pisum*, *Apis mellifera*, *Drosophila melanogaster*, *Bombus impatiens*, and *Bombyx mori* were downloaded as protein homology-based evidence from Ensembl (Flicek et al. 2014).

ncRNAs were identified with Infernal v1.1.2 (Nawrocki and Eddy 2013) against the Rfam v14.0 (Kalvari et al. 2018) database. Transfer RNAs were further refined with tRNAscan-SE v2.0 (Lowe and Eddy 1997).

Phylogenomic Analyses

We generated a phylogeny of Apoidea using two data types. The first part is public genomic data from 17 species (see supplementary table S1, Supplementary Material online) with 2 species from Vespidae and Bethylidae selected as outgroups. The second component is RNA-seq data from six other species from GenBank (see table 1). We assembled the transcripts using Trinity v2.8.6 (Grabherr et al. 2011), combined highly similar transcripts, and extracted the longest transcripts using CD-HIT-EST (Li and Godzik 2006). Complete, single-copy genes were selected using BUSCO assessments against the Hymenoptera data set ($n = 4,415$). For phylogenetic analyses, single-copy genes matrices were then generated following Zhang et al. (2019) using MAFFT v7.394 (Katoh and Standley 2013), trimAl v1.4.1 (Capella-Gutiérrez et al. 2009), and FASconCAT-G v1.04 (Kück and Longo 2014).

Phylogenomic tree reconstructions were made using maximum likelihood (ML) and coalescent-based species tree (ASTRAL) methods. ML reconstructions were performed using IQ-TREE v1.6.3 (Nguyen et al. 2015) with 1,000 ultrafast bootstraps (UFBoot, Hoang et al. 2018) and 1,000 SH-aLRT replicates (Guindon et al. 2010). Partitioning schemes and substitution models were estimated with ModelFinder (built into IQ-TREE; Kalyaanamoorthy et al. 2017). Species trees were estimated using ASTRAL-III v5.6.1 (Zhang et al. 2018) based on gene trees generated with IQ-TREE on individual gene alignments. Local branch supports were estimated from quartet frequencies (Sayyari and Mirarab 2016).

Gene Family Identification and Evolution

We identified gene families using 14 public genome protein sequences of insect species, including 13 Hymenoptera species, five Apidae species (*Apis mellifera*, *Bombus impatiens*, *Ceratina calcarata*, *Habropoda laboriosa*, and *Melipona quadrifasciata*), two Megachilidae species (*Megachile rotundata* and *Osmia bicornis*), one Halictidae species (*Dufourea novaeangliae*), and one species each from Formicidae (*Harpegnathos saltator*), Vespidae (*Polistes dominula*), Braconidae (*Diachasma alloeum*), Pteromalidae (*Nasonia vitripennis*), and Diprionidae (*Neodiprion lecontei*). *Drosophila melanogaster* was selected as the outgroup. OrthoFinder v2.2.7 (Emms and Kelly 2015) was used to infer orthogroups with Diamond (Buchfink et al. 2015) as the sequence aligner. Gene family evolution (gain and loss) was analyzed using CAFE v4.2 (Han et al. 2013) with the lambda parameter used to calculate birth and death rates. The ultrametric tree generated from OrthoFinder was transformed using r8s (Sanderson 2003) and time calibrated by the divergence time (99 Ma) of *Habropoda laboriosa* and *Dufourea novaeangliae* from the TimeTree database (Kumar et al. 2017).

Results and Discussion

Genome Sequencing and Assembly

We generated 6,251,585 subreads on the PacBio Sequel platform totaling 40.58 Gb (150×). The mean and N50 length of long subreads were 6.49 and 11.44 kb, respectively. A total of 39.1 Gb (147.5×) and 7.77 Gb clean data were produced on the Illumina HiSeq X Ten platform for whole-genome and transcriptome sequencing, respectively.

The estimated genome size varied from 299.45 to 322.07 Mb at a maximum k-mer coverage of 1,000

Table 1

Summary of Each Assembly at Each Step for *Colletes gigas*

Assembly	Total Length (Mb)	No. Scaffolds	N50 Length (kb)	Longest Scaffold (Mb)	GC (%)	BUSCO (*n* = 1,658) (%)			
						C	D	F	M
Flye	317.355	4,252	5,882	12.25	39.38	99.3	0.7	0.0	0.7
Falcon	274.246	377	4,809	10.8	39.72	88.6	0.2	6.6	4.8
Merged	274.984	343	7,254	13.274	39.68	98.9	1.4	0.2	0.9
Pilon	275.056	343	7,253	13.274	39.66	99.1	1.4	0.1	0.8
Final genome assembly	**273.056**	**326**	**8,109**	**13.274**	**39.69**	**94.4**	**1.2**	**1.0**	**4.6**
Transcript assembly	**50.080**	**18,407**	**5.41**	**0.05781**	**40.56**	**92.1**	**4.6**	**3.7**	**4.2**

NOTE.—Values of final assemblies are bold. C, complete BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; M, missing BUSCOs.

(supplementary table S2, Supplementary Material online). The overall rate of heterozygosity (0.176–0.298%) and a distinct first peak occurred at a mean k-mer coverage of 29.33–37.34 in the k-mer plots (supplementary fig. S2, Supplementary Material online). Unique (nonrepetitive) length estimates were generally consistent among analyses, ranging from 194.77 to 245.50 Mb. Ranging from 65.85 to 126.89 Mb, our genome repetitive length estimates varied with the maximum k-mer coverage cutoff, indicating that the assembled genome may include a high number of repeated regions.

The size of the Flye assembly was 317.36 Mb including 4,260 contigs, whereas that of the Falcon assembly was 274.25 Mb with an N50 of 4.81 Mb (table 1). The Flye and Falcon assemblies were merged into 343 contigs with N50 = 7.25 Mb after two rounds of quickmerging. Following polishing with Illumina reads and checking for possible contaminants, our final draft assembly of *C. gigas* had 326 scaffolds, a total length of 273.06 Mb, an N50 length of 8.11 Mb, a maximum scaffold length of 13.274 Mb, and 39.69% GC content. With the genome-guided strategy, there were a total of 18,405 transcripts assembled with a mean and N50 length of 2.72 and 5.41 kb, respectively.

We generated a circular mitochondrial genome of 15,888 bp (GenBank No. MN841004), which is slightly longer and higher A + T content (86.47%) than the previously published one (KM978210, 15,885 bp in length with 86.29% A + T content).

Assembly completeness was assessed by querying the genome for the insect BUSCO marker set (*n* = 1,658). We identified 88.6–99.3% complete, 0.0–6.6% fragmented, and 0.7–4.8% missing BUSCOs across all versions of the assembly (table 1).Therefore, the BUSCO analysis indicates that our assembly is near-complete. Genome-guided transcriptome assemblies also show similar completeness. In addition, 92.78% of PacBio long reads, 94.63% of Illumina short reads, as well as all (18,405) assembled transcripts could be mapped to the final genome assembly using the bwa-mem command (Li 2013). All statistics suggest that our assembly is highly complete and reliable.

Genome Annotation

RepeatMasker identified 378,335 repeats, masking 26.27% of the genome assembly. The top-five, most-abundant repeat types were unclassified repeats (11.05%), Helitron transposable elements (2.95%), DNA/TcMar-Tc1 transposons (2.20%), Gypsy long terminal repeat retrotransposons (1.21%), and DNA/PiggyBac (PB) transposons (0.94%) (supplementary table S3, Supplementary Material online).

A total of 11,016 protein-coding genes were identified by the MAKER pipeline with means of 5.99 exons and 4.99 introns per gene. The mean length of exons and introns was 271.91 and 665.33 bp, respectively, whereas the gene density of the *C. gigas* genome was 40.19 genes/Mb. Among predicted genes, 10,851 (98.50%) were supported by protein-based evidence and 9,336 (84.75%) were supported by transcriptome-based evidence. BUSCO analysis identified 1,518 (91.6%) complete, 21 (1.3%) duplicated, 32 (1.9%) fragmented, and 108 (6.5%) missing BUSCOs. InterProScan identified protein domains for 9,495 (86.19%) genes, among which there were 6,577 assigned with gene ontology terms, and 597,486 and 2,729 ones matching the Kyoto Encyclopedia of Genes and Genomes, MetaCyc, and Reactome pathway databases, respectively.

For ncRNAs, we identified 122 rRNAs, 258 tRNAs, 52 micro-RNAs, 52 small nuclear RNAs, 11 ribozymes, 366 cis-regulatory elements, 1 antisense, 2 lncRNAs, 3 sRNAs, and 3 other ncRNAs. A total of 21 tRNA isotypes were identified, excepting the SelCys-isotype. (See details in supplementary table S4, Supplementary Material online.)

Phylogenomic Analyses

Nucleotide and protein matrices of 147 shared, single-copy genes had 212,277 and 70,268 sites that were divided by ModelFinder into 49 and 50 partitions, respectively. ML trees from proteins generated the same topologies as species trees generated by ASTRAL-III using both nucleotide and protein matrices, which were similar to the ML ones generated using nucleotide matrices, except for the position of *Andrena fulva*. All interior nodes were supported with high values (fig. 1). The
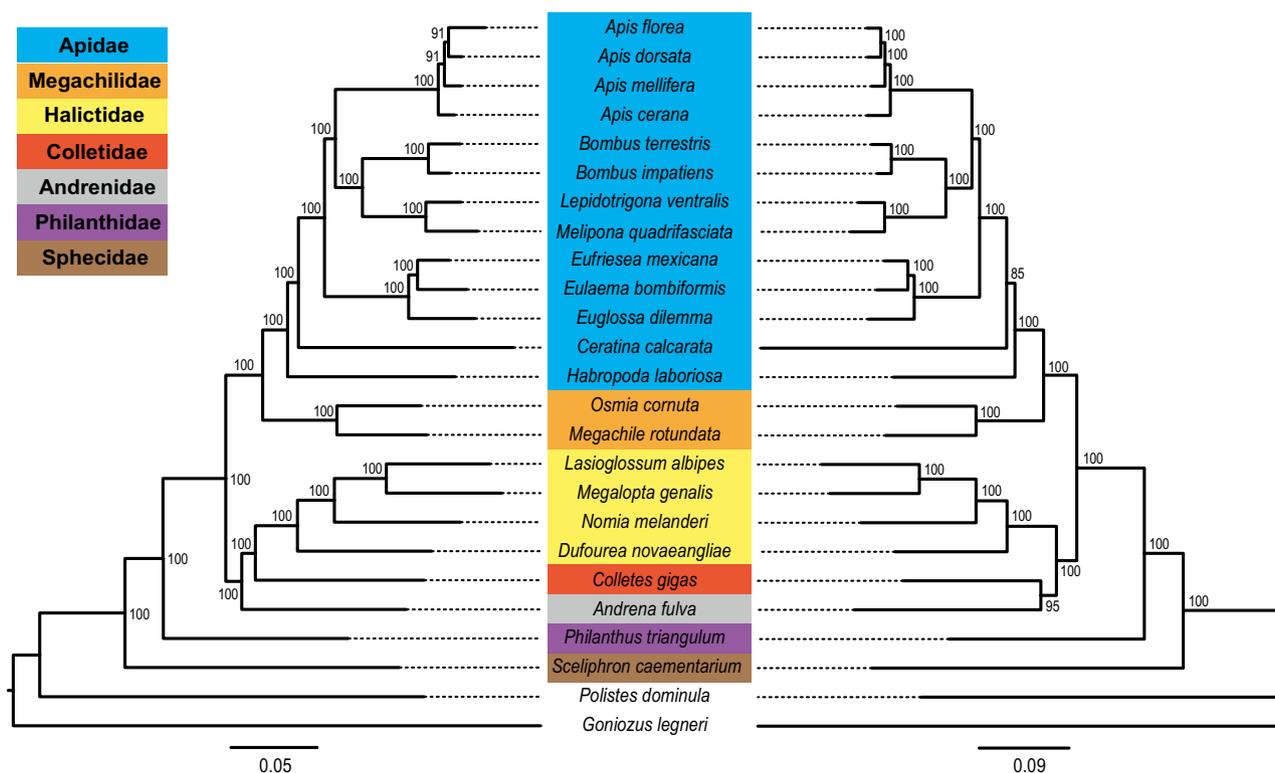
Fig. 1—Phylogenomic trees of Apoidea constructed based on protein (left) and nucleotide (right) matrices of single-copy genes from 19 published whole genomes and six RNA-seq data sets. Support values are given at nodes. Species in blue belong to family Apidae, orange belong to Megachilidae, yellow belong to Halictidae, red belong to Colletidae, gray belong to Andrenidae, purple belong to Philanthidae, and brown belong to Sphecidae. *Gonniozus legneri* and *Polistes dominula* were used as outgroups.

phylogeny of Apoidea derived from protein data shows the sister relationship between species of (Apidae + Megachilidae) sister to ((Colletidae + Halictidae) + Andrenidae), supporting the results of numerous recent phylogenies (Hedtke et al. 2013; Branstetter et al. 2017; Peters et al. 2017; Sann et al. 2018).

Gene Family Evolution

Gene families were identified using OrthoFinder based on 14 hymenopterans and *D. melanogaster*. Overall, 91.4% (184,939) of genes were assigned into 10,994 gene families with a mean orthogroup size of 16.8. Among 5,254 families shared by all species, 1,473 were single-copy orthogroups. For *C. gigas*, 10,926 (94.20%) genes were clustered into 10,269 gene families, and only one family and seven genes were specific to *C. gigas* (supplementary table S5, Supplementary Material online).

We analyzed gene family evolution (gain and loss) using CAFE and estimated gene birth rate (lambda) at 0.00120 when accounting for duplications/gene/Ma. We found that 968 gene families experienced significant expansion or contraction events (family-wide P value < 0.05, supplementary

table S6, Supplementary Material online), with details for the 15 species shown in supplementary figure S3, Supplementary Material online. Among them, *C. gigas* has 70 (33 expansions and 37 contractions) rapidly evolving families (supplementary table S7, Supplementary Material online). The top-five of the largest expanded families were reverse transcriptase (RNA-dependent DNA polymerase) (112), transposase (90), zinc finger (32), carboxylesterase family (17), and olfactory receptor (15). Among them, olfactory receptors are a large family of membrane-associated G-protein-coupled receptors that play crucial roles in insect survival and reproductive success, mediating responses to food, mates, and oviposition sites (Hallem et al. 2006), and carboxylesterase is a multifunctional superfamily widely distributed in nature, many as enzymes participating in catalyzing chemical reactions involving compounds such as toxins or drugs, meaning they play important roles in xenobiotic detoxification (Yu et al. 2009; Aranda et al. 2014), which could be directly beneficial for foraging on the toxic nectar and pollen of *Ca. oleifera*. Similarly, more olfactory receptors should make *C. gigas* better at selecting the specific floral resources it is best adapted to, or perhaps even enable measurement of toxins between specific flowers or at stages of

bloom such that this species could minimize its exposure, but more study is necessary to determine the major genomic elements related to the specialization of this species on the chemically defended *Ca. oleifera*.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Author Contributions

Q.-S.Z. and C.-D.Z. designed the study. Q.-S.Z., Z.-Q.N., and Q.-T.W. collected the samples. Q.-S.Z., F.Z., and A.L. performed the analyses and wrote the paper. Z.-Q.N. and M.C.O. provided subject matter expertise throughout. All authors edited and approved the final manuscript.

Literature Cited

Almeida EAB. 2008. Colletidae nesting biology (Hymenoptera: Apoidea). Apidologie 39(1):16–29.

Almeida EAB, Pie MR, Brady SG, Danforth BN. 2012. Biogeography and diversification of colletid bees (Hymenoptera: Colletidae): emerging patterns from the southern end of the world. J Biogeogr. 39(3):526–544.

Aranda J, et al. 2014. The catalytic mechanism of carboxylesterases: a computational study. Biochemistry 53(36):5820–5829.

Ascher JS, Pickering J. 2018. Discover life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila). Available from: http://www.discoverlife.org/mp/20q? guide=Apoidea_species.

Bao WD, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1):11.

Bernt M, et al. 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 69(2):313–319.

Branstetter MG, et al. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr Biol. 27(7):1019–1025.

Branstetter MG, et al. 2018. Genomes of the Hymenoptera. Curr Opin Insect Sci. 25:65–75.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Capella-Gutiérrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44(19):e147.

Chen Y, Ye WC, Zhang YD, Xu YS. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 43(16):7762–7768.

Chin CS, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 13(12):1050–1054.

Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D. 2013. The impact of molecular data on our understanding of bee phylogeny and evolution. Annu Rev Entomol. 58(1):57–78.

Danforth BN, Sipes S, Fang J, Brady SG. 2006. The history of early bee diversification based on five genes plus morphology. Proc Natl Acad Sci U S A. 103(41):15118–15123.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157.

Ferrari RR, Onuferko TM, Monckton SK, Packer L. 2020. The evolutionary history of the cellophane bee genus *Colletes* Latreille (Hymenoptera: Colletidae): molecular phylogeny, biogeography and implications for a global infrageneric classification. Mol Phylogenet Evol. 146:106750.

Finn RD, Miller BL, Clements J, Bateman A. 2014. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic Acids Res. 42(D1):D364–D373.

Finn RD, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45(D1):D190–D199.

Flicek P, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42(D1):D749–D755.

Garibaldi LA, et al. 2011. Stability of pollination services decreases with isolation from natural areas despite honey bee visits. Ecol Lett. 14(10):1062–1072.

Garibaldi LA, et al. 2013. Wild pollinators enhance fruit set of crops regardless of honey bee abundance. Science 339(6127):1608–1611.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 29(7):644–652.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res. 41(13):e129.

Hallem EA, Dahanukar A, Carlson JR. 2006. Insect odor and taste receptors. Annu Rev Entomol. 51(1):113–135.

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30(8):1987–1997.

Hedtke SM, Patiny S, Danforth BN. 2013. The bee tree of life: a super-matrix approach to apoid phylogeny and biogeography. BMC Evol Biol. 13(1):138.

Hoang DT, Chernomor O, Von HA, Minh BQ, Le SV. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 35(2):518–522.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32(5):767–769.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12(1):491.

Huang DY, et al. 2015. The study on bionomics character of *Colletes gigas* (Hymenoptera, Colletidae). J Environ Entomol. 37(1):133–138.

Huang DY, et al. 2016. The complete mitochondrial genome of the *Colletes gigas* (Hymenoptera: Colletidae: Colletinae). Mitochondrial DNA Part A 27(6):3878–3879.

Hubley R, et al. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44(D1):D81–D89.

Kalvari I, et al. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46(D1):D335–D342.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler AV, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Kim D, Landmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 12(4):357–360.

Klein A-M, et al. 2007. Importance of pollinators in changing landscapes for world crops. Proc R Soc B 274(1608):303–313.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5):540–546.

Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 11(1):8.

Kuhlmann M, Almeida EAB, Laurenne N, Quicke DLJ. 2009. Molecular phylogeny and historical biogeography of the bee genus *Colletes* Latreille, 1802 (Hymenoptera: Apiformes: Colletidae), based on mitochondrial COI and nuclear 28S sequence data. Insect Syst Evol. 40(3):291–318.

Kuhlmann M, Proshchalykin MY. 2011. Bees of the genus *Colletes* Latreille 1802 of the Asian part of Russia, with keys to species (Hymenoptera: Apoidea: Colletidae). Zootaxa 3068(1):1–48.

Kuhlmann M, Proshchalykin MY. 2013. The genus *Colletes* (Hymenoptera: Apoidea: Colletidae) in Central Asia. Zootaxa 3750(5):401–449.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34(7):1812–1819.

Leonhardt SD, Gallai N, Garibaldi LA, Kuhlmann M, Klein A-M. 2013. Economic gain, stability of pollination and bee diversity decrease from southern to northern Europe. Basic Appl Ecol. 14(6):461–471.

Lewis TE, et al. 2018. Gene3D: extensive prediction of globular domains in proteins. Nucleic Acids Res. 46(D1):D1282.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

Li WZ, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33(20):6494–6506.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.

Marchler-Bauer A, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45(D1):D200–D203.

Mi HY, et al. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45(D1):D183–D189.

Michener CD. 2007. The bees of the world. 2nd ed. Baltimore (MD): Johns Hopkins University Press.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Niu ZQ, Kuhlmann M, Zhu CD. 2013. A review of *Colletes succinctus*-group (Hymenoptera: Colletidae: Colletinae: *Colletes*) from China with redescription of the male of *C. gigas*. Zootaxa 3626(1):173–187.

Niu ZQ, Zhu CD, Kuhlmann M. 2013. Bees of the *Colletes clypearis*-group (Hymenoptera: Apoidea: Colletidae) from China with description of seven new species. Zootaxa 3745(2):101–151.

Niu ZQ, Zhu CD, Kuhlmann M. 2014a. Bees of the *Colletes flavicornis*-group from China with description of one new species (Hymenoptera: Apoidea: Colletidae). Zootaxa 3780(5):534–546.

Niu ZQ, Zhu CD, Kuhlmann M. 2014b. The Bees of the Genus *Colletes* (Hymenoptera: Apoidea: Colletidae) from China. Zootaxa 3856(4):451–483.

Ollerton J, Winfree R, Tarrant S. 2011. How many flowering plants are pollinated by animals? Oikos 120(3):321–326.

Pertea M, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33(3):290–295.

Peters RS, et al. 2017. Evolutionary history of the Hymenoptera. Curr Biol. 27(7):1013–1018.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44(12):e113.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19(2):301–302.

Sann M, et al. 2018. Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. BMC Evol Biol. 18(1):71.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7):1654–1668.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32(Web Server):W309–W312.

Su R, Li H, Dong K, He S. 2011. The situation and utilization of oil tea (*Camellia oleifera*) as nectar source in China. Apic China 63:48–50.

Vurture GW, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33(14):2202–2204.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.

Waterhouse RM, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 35(3):543–548.

Wilson D, et al. 2009. SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37(Suppl 1):D380–D386.

Wu YR. 1965. Economic insect fauna of China. Fasc. 9, Hymenoptera: Apoidea. Beijing (China): Science Press, 83 pp.

Yu QY, Lu C, Li WL, Xiang ZH, Zhang Z. 2009. Annotation and expression of carboxylesterases in the silkworm, *Bombyx mori*. BMC Genomics. 10(1):553.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19(S6):153.

Zhang F, et al. 2019. Phylogenomics from low-coverage whole-genome sequencing. Methods Ecol Evol. 10(4):507–517.

Associate editor: Dorothée Huchon