Comparative Genomics Reveals Recent Adaptive Evolution in Himalayan Giant Honeybee *Apis laboriosa*

Dan Lin,[1,†] Lan Lan,[2,†] Tingting Zheng,[1] Peng Shi,[2] Jinshan Xu,[2,*] and Jun Li[1,3,*]

[1]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China

[2]College of Life Sciences, Chongqing Normal University, Chongqing, China

[3]School of Data Science, City University of Hong Kong, Hong Kong, China

[†]These authors contributed equally to this work.

*Corresponding authors: E-mails: xujinshan2008@cqnu.edu.cn; jun.li@cityu.edu.hk.

Accepted: 26 September 2021

Abstract

The Himalayan giant honeybee, *Apis laboriosa*, is the largest individual honeybee with major ecological and economic importance in high-latitude environments. However, our understanding of its environmental adaptations is circumscribed by the paucity of genomic data for *this species*. Here, we provide a draft genome of wild *A. laboriosa*, along with a comparison to its closely related species, *Apis dorsata*. The draft genome of *A. laboriosa* based on the de novo assembly is 226.1 Mbp in length with a scaffold N50 size of 3.34 Mbp, a GC content of 32.2%, a repeat content of 6.86%, and a gene family number of 8,404. Comparative genomics analysis revealed that the genes in *A. laboriosa* genome have undergone stronger positive selection (2.5 times more genes) and more recent duplication/loss events (6.1 times more events) than those in the *A. dorsata* genome. Our study implies the potential molecular mechanisms underlying the high-altitude adaptation of *A. laboriosa* and will catalyze future comparative studies to understand the environmental adaptation of modern honeybees.

Key words: comparative genomics, honeybee, *Apis laboriosa*, whole genome sequencing, evolution.

Significance

The accumulation of knowledge of adaptively evolutionary mechanisms is of great significance for developing strategies for protecting Himalayan giant honeybee, *Apis laboriosa*. We presented high-quality draft genome sequences of *A. laboriosa* for the first time, along with a comparative genomics analysis with its closely related species, *Apis dorsata*. Our study may advance understanding of the molecular basis of environmental adaptation in *A. laboriosa* and provide valuable resources for future comparative/population studies of honeybee evolution.

Introduction

The Himalayan giant honeybee *Apis laboriosa*, the largest individual bee in the genus *Apis*, is a key pollinator and important honey producer in Himalayan regions (Joshi et al. 2004). *A. laboriosa* lives in mountainous areas from Nepal to southwest China and has evolved adaptive behaviors to cope with harsh environments: nesting on inaccessible cliffs above 1,200 m, migrating seasonally and foraging at low temperatures (Batra 1996).

To comprehend its evolutionary dynamics and adaptations to the local environment, *A. laboriosa* is usually compared with the closely related species *Apis dorsata*, a lowland giant honeybee that is widespread throughout tropical and subtropical Asia. These two species were grouped into a "giant honeybees" clade in previous phylogenetic studies (Willis et al. 1992; Engel and Schultz 1997; Arias and Sheppard 2005; Chhakchhuak et al. 2016). Although *A. dorsata* genome has been sequenced recently (Oppenheim et al. 2020), whole-genomic data for *A. laboriosa* are still lacking. An in-depth understanding of the evolution process of *A. laboriosa* necessitates a high-quality reference genome.

In this study, we presented high-quality draft genome sequences of *A. laboriosa*. We compared the draft genome of *A. laboriosa* and *A. dorsata* to describe the recent evolutionary trend because their divergence. The genome sequence and analyses will pave the way for future comparative studies and a mechanistic understanding of honeybee evolution.

Results

Apis laboriosa and *A. dorsata* have Similar Genome Compositions

A total of 74.16 gigabases (Gbp) of paired-end DNA reads were generated from whole genome shotgun sequencing of *A. laboriosa* worker bees. The short reads were de novo assembled into 4,376 scaffolds (total size: 226.1 Mbp, N50: 3,339,770 bp; fig. 1*A*). The GC content of the *A. laboriosa* assembly was 32.2% (fig. 1*A*), similar to that of *A. dorsata* (31.9%). A genome completeness assessment (fig. 1*A*) showed that the assembly of *A. laboriosa* was 99.2% complete, similar to the *A. dorsata* assembly (98.9%). The predicted gene number in *A. laboriosa* genome (11,466 genes) is higher than that in the *A. dorsata* genome (9,910 genes) (fig. 1*A*). Additionally, an analysis of repeat elements (fig. 1*B*) showed that the genome of *A. laboriosa* contained a higher proportion of transposable elements than that of *A. dorsata* (*A. laboriosa*: 15,518,024 bp, 6.86% of the genome; *A. dorsata*: 14,367,696 bp, 6.24% of the genome).

Subsequently, we examined the shared gene families between these two species. A total of shared 8,404 gene families were identified based on the gene sets from *A. laboriosa* (9,857 of 11,466 genes), *A. dorsata* (9,001 of 9,910 genes), and seven other *Hymenoptera* species (fig. 1*C*). All these gene

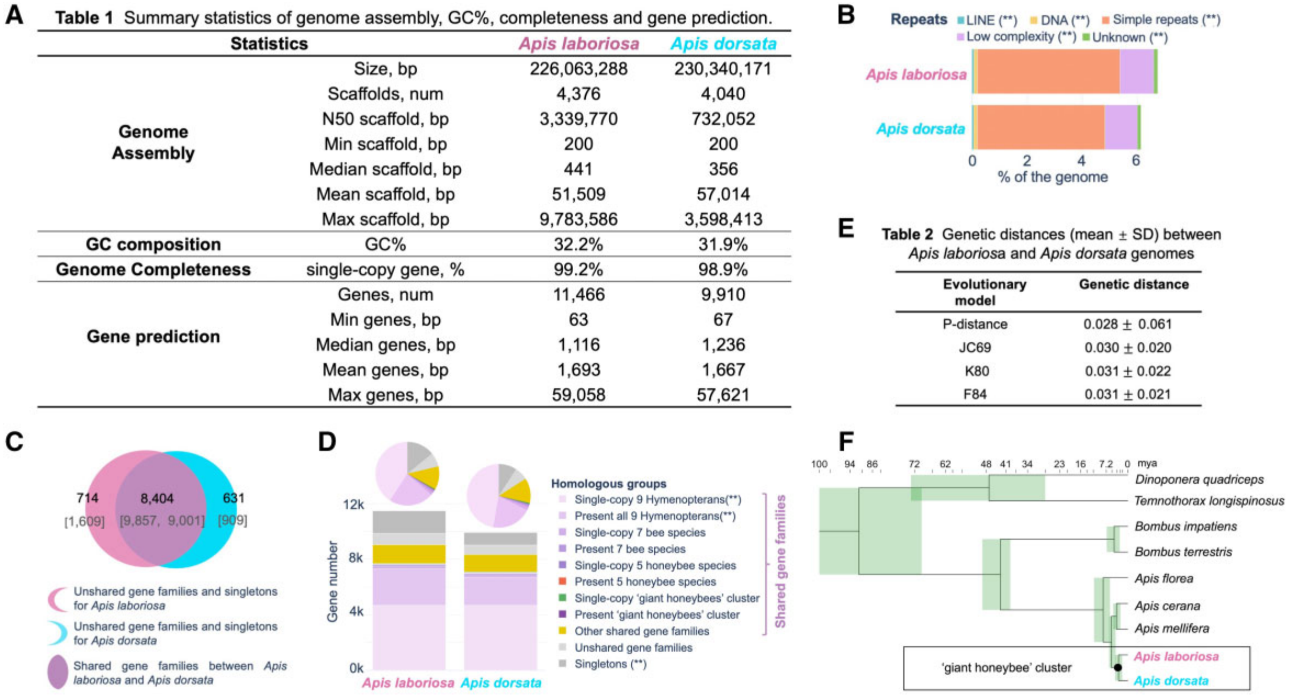

Table 1 Summary statistics of genome assembly, GC%, completeness and gene prediction.

Statistics		*Apis laboriosa*	*Apis dorsata*
Genome Assembly	Size, bp	226,063,288	230,340,171
	Scaffolds, num	4,376	4,040
	N50 scaffold, bp	3,339,770	732,052
	Min scaffold, bp	200	200
	Median scaffold, bp	441	356
	Mean scaffold, bp	51,509	57,014
	Max scaffold, bp	9,783,586	3,598,413
GC composition	GC%	32.2%	31.9%
Genome Completeness	single-copy gene, %	99.2%	98.9%
Gene prediction	Genes, num	11,466	9,910
	Min genes, bp	63	67
	Median genes, bp	1,116	1,236
	Mean genes, bp	1,693	1,667
	Max genes, bp	59,058	57,621

Table 2 Genetic distances (mean ± SD) between *Apis laboriosa* and *Apis dorsata* genomes

Evolutionary model	Genetic distance
P-distance	0.028 ± 0.061
JC69	0.030 ± 0.020
K80	0.031 ± 0.022
F84	0.031 ± 0.021

Fig. 1.—The summary for genome assembly, GC, genes, gene families, repeats, homologous groups, nucleotide divergency and phylogeny relationship of *Apis laboriosa* and *Apis dorsata* genome (*A*) Summary statistics of the genome assembly, GC%, completeness and gene prediction of two species genomes; (*B*) Bar charts describing the repeat elements of two genomes. ** indicates a statistical significance (Chi-squared test, $P < 0.05$) between two genomes; (*C*) Venn diagram describing the number of gene families and genes in two genomes; (*D*) Bar charts and pie charts describing homologous groups of two genomes in different categories. ** indicates a statistical significance (Chi-squared test, $P < 0.05$) between two genomes; (*E*) Summary statistics (mean ± standard deviation/SD) of genetic distances between single-copy genes from *Apis laborisa* genome and *Apis dorsata* genome. The distances were calculated using R package 'ape'; (*F*) Phylogenetic tree of the nine Hymenopterans species. The tree was reconstructed using maximum-likelihood method under the GTR model based on the concatenated alignments of the single-copy genes and the dating time with 95% Confidence Interval was assessed using Markov chain Monte Carlo analysis under the GTR model based on the concatenated alignments of fourfold degenerate sites from single-copy genes. The bootstrap values of all nodes were 100%. *Apis laboriosa* was indicated by pink label, and *Apis dorsata* was denoted by blue label. Branch length represented time of divergence in million years.

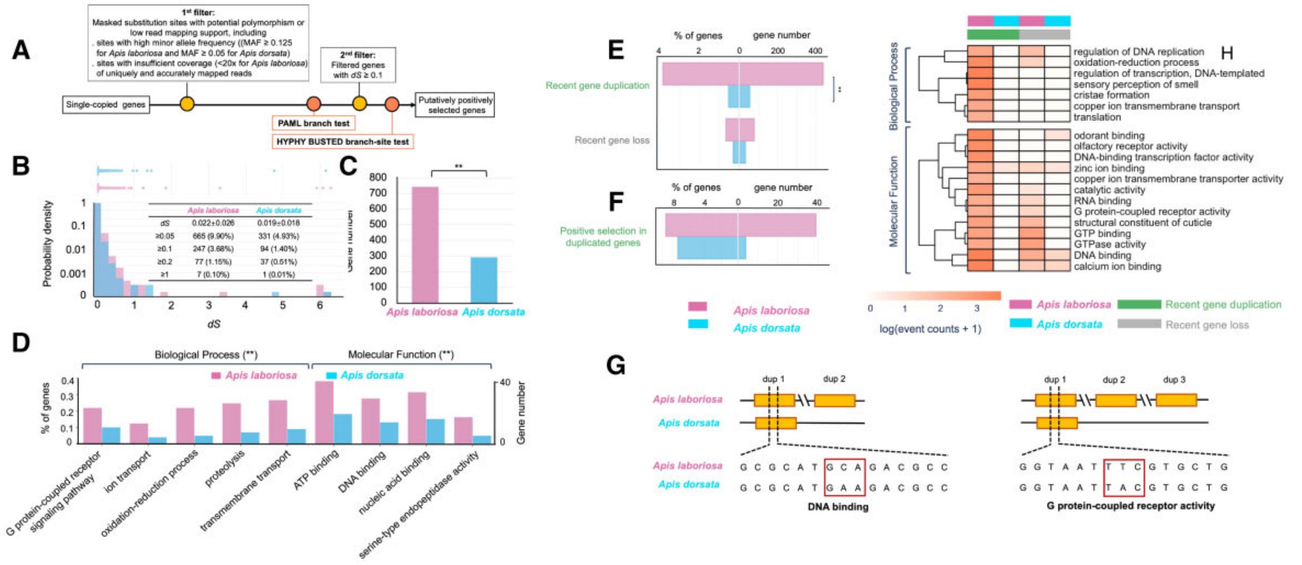

FIG. 2.—Gene family evolution based on nucleotide substitutions model and recent duplication/loss model (*A*) The quality control procedure to achieve more accurate estimation of positive selected genes; (*B*) The frequency distribution of dS in 6,718 single-copy gene families, derived from the results of branch model in PAML. The sub-table showed the mean and standard deviation of dS, the number and percent of single-copied genes with dS equal or larger than specific thresholds (0.05, 0.1, 0.2 and 1); (*C*) Total number of positively selected genes in 6,718 single-copy gene families, derived from the results of branch-site model in BUSTED (Chi-squared test, $P < 0.05$); (*D*) Enriched functional categories and the frequency of positively selected genes derived from the results of branch-site model in BUSTED (Chi-squared test, adjusted $P < 0.1$); (*E*) The overall summary (number and percent) of the DL events in two species. ** indicates that there is statistical significance (Chi-squared test, $P < 0.05$) between two genomes; (*F*) Frequency of the positive selection on the duplicated genes in two species; (*G*) Representative examples of positive selection on the duplicated genes in *Apis laboriosa*. Yellow boxes represented putatively functional genes. Double slashes indicated that genes were retrieved from different scaffolds. Red boxes represented nonsynonymous mutations on the duplicated genes in *Apis laboriosa*. The homolog sequences from the other seven Hymenopterans species had the same nucleotide composition as *Apis dorsata*. Note that each example just represents one gene family; (*H*) Heatmap of the total number of DL events in each functional category in two species. Each functional category is statistically significantly different (Chi-squared test, adjusted $P < 0.1$) between two species in at least one of the DL events. All the terms were enriched in recently duplicated genes, and only 2 terms were enriched in recent losses in *Apis laboriosa* (GTP binding: 0 vs 13, adjusted $P = 0.013$; GTPase activity: 0 vs 13, adjusted $P = 0.013$).

families were further classified into different categories based on the phylogenetic composition and the copy number of the genes in each species (see Materials and Methods). We found that *A. laboriosa* and *A. dorsata* presented similar frequency distribution of gene family categories (χ^2 test, $P > 0.05$; fig. 1*D*), indicating that the genomes of these two species shared a high amount of core genes families.

The nucleotide divergency in single-copy genes between *A. laboriosa* and *A. dorsata* was around 0.03 with relatively large standard deviations (P-distance: 0.028 ± 0.061; JC69: 0.030 ± 0.020; K80: 0.031 ± 0.022; F84: 0.031 ± 0.021) (fig. 1*E*), demonstrating a close evolutionary relationship of these two species. The estimated divergence time between *A. laboriosa* and *A. dorsata* was 2.37 million years (Myr) (95% CI 1.9–4.3 Myr) (fig. 1*F*).

Genes in *A. laboriosa* had Undergone More Frequently Positive Selection

We performed a series of quality control procedures at gene/sites levels to achieve more accurately estimated selective pressure (fig. 2*A*, see Materials and Methods for details). The result showed that the synonymous substitution rate, *dS* values, in 6,718 single-copy gene families (core-gene ortholog families) were similar between these two species using PAML branch model (free-ratios) (*A. laboriosa*: 0.022 ± 0.026; *A. dorsata*: 0.019 ± 0.018; fig. 2*B*), indicating a similar neutral nucleotide substitution rate with a relatively large random fluctuations in these two lineages.

To test whether the evolutionary process of the coding regions in both species was deviated from the neutral evolution because the divergence from the latest common ancestor, we analyzed the selective pressure of single copy genes. The BUSTED branch-site models and likelihood ratio tests showed that compared with *A. dorsata*, a statistically significantly higher amount of genes in *A. laboriosa* have undergone positive selection at both gene families level (742 vs 291, χ^2 test: $P < 0.001$) (fig. 2*C*) and functional categories level (a total of nine enriched functional categories in the *A. laboriosa* genome, χ^2 test: adjusted $P < 0.1$; fig. 2*D*). Around 88.9% and 77.8% enriched gene ontology (GO) categories with positively selected genes are overlaid (697 and 507 genes overlaid) between PAML branch-site model and BUSTED models, and between BUSTED and HYPHY MNMs models, respectively (supplementary fig. S2, Supplementary Material online).

Gene Families in *A. laboriosa* had Undergone More Frequent Duplications

We detected a total of 515 and 85 duplication or loss (DL) events in the *A. laboriosa* and *A. dorsata* genomes, respectively, by reconciling 8,404 gene trees with the reference species tree (fig. 1*F*) using RANGERDTL v2.0. We found that there was a significantly higher frequency of recent duplication events in *A. laboriosa* than in *A. dorsata* (438 vs 54 genes, χ^2 test: $P < 0.001$; fig. 2*E*).

Furthermore, 39 positive selection events in these recently duplicated genes were identified in the *A. laboriosa* genome, whereas only four such events were identified in the *A. dorsata* genome (fig. 2*F*). For instance, key positively selected sites can be detected in the duplicated genes related to DNA binding or G protein-coupled receptor activity in *A. laboriosa* (fig. 2*G*).

We finally investigated the enriched/depleted functional categories of duplication/loss events in *A. laboriosa* compared with those in *A. dorsata* (fig. 2*H*). A total of 20 statistically significantly (χ^2 test, adjusted $P < 0.1$) GO terms with enriched duplications were detected. The recent evolution of these gene families in *A. laboriosa* potentially contributes to the adaptation of harsh living environment with strengthened pollination, enhanced learning, foraging, etc. (Honeybee Genome Sequencing Consortium 2006; Scheiner et al. 2006). The pattern of significantly higher proportion of recent duplication in *A. laboriosa* was confirmed by another duplication-loss reconciliation software, TREERECS (90% of enriched functional categories identified by RANGERDTL v2.0 were also identified by TREERECS; supplementary fig. S4, Supplementary Material online).

Several caveats are worth discussing. First, the small divergence between the genomes of *A. laboriosa* and *A. dorsata* (fig. 1*E*) could potentially introduce a high amount of random errors into the results. The genetic drift in a small population (e.g., giant bee clade) further complicated the interpretation of potential adaptive selection, with a high rate of false positives and false negatives in the positive selection and gene duplication/loss analyses. Also, the sequencing errors, alignment errors (Anisimova 2001), gene tree discordance (Mendes and Hahn 2016), polymorphism, and multinucleotide mutations (Venkat 2018) could introduce bias into positive selection analysis. Thus, we adopted several filtering procedures the minimize the bias and achieve more accurately estimated positive selection (fig. 2*A*). Besides, we found that the enrichment pattern at functional categories level were consistent with both PAML branch-site model and a branch-site model that incorporates multinucleotide mutations (HYPHY MNMs) (supplementary figs. S2 and S3*A*, *B*, Supplementary Material online), which further indicated the robustness of our positive selection analysis.

In conclusion, our comparative study has identified some genomic evolutionary processes that are possibly related to the local adaptation of wild *A. laboriosa*. In the future, more in-depth comparative and population-level studies will help to describe the landscape of local adaptation in this species, thereby facilitating the development of feasible strategies for protecting this important pollinator.

Materials and Methods

Sample Collection, DNA Extraction, Library Preparation, and Sequencing

Four *A. laboriosa* workers were collected from the same wild colony in Shangri-La, Yunnan Province, China in February 2019 (supplementary fig. S1*A*, Supplementary Material online). DNA was extracted from its thorax by CTAB immediately for library preparation. A total of 0.2 μg DNA was used for each sample as the input material for DNA library preparation. Sequencing library was generated using NEB Next Ultra DNA Library Prep Kit for Illumina (NEB) following the manufacturer's recommendations (see supplementary data, Supplementary Material online for more details).

Whole genome shotgun sequencing was applied to generate short paired-end reads (150 bp) libraries with a series of short-insert length (11.05 and 35.79 G for 250 and 500 bp, respectively) and long-insert length (14.01 and 13.30 G for 2 and 5 kb, respectively). The sequencing was performed on a Hiseq platform (Illumina, San Diego, CA).

Quality Control and Genome Assembly

The quality control of raw reads (removed the adapters, low-quality bases/reads and PCR duplicates) was performed as previously described (Heshiki et al. 2017; Zheng et al. 2019). Afterwards, SOAPdenovo v2.04 (Li et al. 2010) was used to perform the de novo genome assembly by constructing and simplifying de Bruijn graph with default parameters (max_rd_len [150]; avg_ins [350]; reverse_seq [0]; asm_flags [3]; pair_num_cutoff [3]; map_len [32]). BUSCO v4 (lineage_dataset [metazoa_odb9]; mode [genome]) was used to assess the genome assembly completeness (Simão et al. 2015).

Predictions of Gene and Repeats

Genes were predicted via a strategy combining ab initio gene prediction and homology-based gene prediction (Li et al. 2019). The coding sequences (CDS) of *A. laboriosa* were identified using EVidenceModeler v1.1.1 (weight of ab initio gene predictions: 8; weight of homology-based gene predictions: 10) (Haas et al. 2008) by incorporating gene predictions results (see supplementary data, Supplementary Material online for more details).

The transposons of *A. laboriosa* and *A. dorsata* genome were identified using RepeatMasker v4.1 (Tarailo-Graovac and Chen 2009) with the library "honeybee" and the default cutoff "225." The identified transposons were grouped according to the annotated repeat class information.

Construction and Annotation of Gene Families

The gene families among *A. laboriosa* and eight *Hymenopterans* species were constructed using OrthoMCL v2.0.9 (Li et al. 2003) with default parameters (dbVendor [mysql]; percentMatchCutoff [50]; evalueExponentCutoff [−5]). Gene families were further classified into 11 categories based on the copy number and presence/absence information of genes in various bee lineages: 1) single-copy gene families among all nine *Hymenopterans*, among all seven bee species, among all five honeybee species, or among "giant honeybees" cluster; 2) nonsingle-copy gene families containing all nine *Hymenopterans*, all seven bee species, all five honeybee species, or "giant honeybees" cluster; 3) other shared gene families; 4) unshared gene families (*Apis laborisa* and *A. dorsata* genes are not present simultaneously); and 5) singletons (fig. 1*D*).

GO annotations of *A. laboriosa* genes were performed using InterProscan v5.29 (Jones et al. 2014) with default parameters (appl [Pfam]; goterms). The GO annotation of the *A. laboriosa* genes within a gene family was used to represent the function of that gene family.

Construction of Species Tree and Estimation of Divergence Time

Phylogenetic trees of *A. laboriosa* and eight *Hymenopterans* species were reconstructed based on the single-copy genes using the maximum likelihood method via PhyML v3.0 package (Guindon et al. 2010), with the parameters of 1,000 bootstrap replicates, GTR model with gamma-distributed rate heterogeneity. The topology of the resulting species tree (fig. 1*F*) was consistent with the phylogenies in previous studies (Arias and Sheppard 2005; Kapheim et al. 2015; Chhakchhuak et al. 2016; Branstetter et al. 2017, 2018; Diao et al. 2018).

Genetic distance between *A. laboriosa* and *A. dorsata* was calculated based on the single-copy genes using R "ape" package. Divergence time among species was estimated based on the concatenated alignments of 4-fold degenerate sites from single-copy genes using MCMCtree in PAML v4.9 (Yang 2007) (see supplementary data, Supplementary Material online for more details).

Analysis of Positive Selection

Positive selection drives the adaptive evolution of the species through promoting the spread of beneficial alleles in the population (Sabeti et al. 2006). In this study, we assessed substitution rates using branch model (seqtype [1]; clock [0]; model [1]; NSsites [0]; fix_omega [0]; fix_blength [0]) in PAML v4.9 package (Yang 2007) and gene-wide positive selection signals using BUSTED (Branch-site Unrestricted Statistical Test for Episodic Diversification) (Murrell et al. 2015) in HyPhy (Hypothesis testing using Phylogenies) (Pond et al. 2005) in both *A. laboriosa* and *A. dorsata* branches after the divergence from their latest common ancestor.

To identify a more robust pattern of positive selection, two preprocessing and one postprocessing procedures were adopted for the calculation of selective pressure using the PAML branch models and HYPHY BUSTED branch-site models (fig. 2*A*): 1) masked the sites (without consideration of substitutions between two giant honeybees) with minor allele frequencies ≥ 0.125 (1/8) for *A. laboriosa* and ≥ 0.05 (1/19) for *A. dorsata* to avoid the false substitution induced by polymorphisms; 2) masked the substitution sites with coverage $<20\times$ of uniquely and accurately mapped paired-end reads in *A. laboriosa* genome to avoid substitutions resulted from low quality or misassembled sites/regions.

To determine the influence of discordance between species tree and gene tree on the positive selection, we compared the results of positive selection with a phylogeny either consisting of all nine *Hymenopterans* species, or only seven *Hymenopterans* species (two closely related species, *Apis mellifera* and *Apis cerana* were removed to avoid unresolved phylogeny). Only nine (1.4%) positively selected genes in the *A. laborios*a genome were uniquely detected by using the phylogenetic trees consisting of all nine *Hymenopterans* species, suggesting a very low false-positive rate produced due to the discordance between species and gene trees.

Detection of Gene Duplications and Losses

Inferring the DL (D: duplication; L: loss) events within a given gene family is important to understand the macro evolution of a gene family (Bansal et al. 2018). In this study, RANGERDTL v2.0 (Bansal et al. 2018) was employed to deduce all putative DL events related to *A. laboriosa* and *A. dorsata* by reconciling the reconstructed gene trees with the species tree (fig. 1*F*) using the cost parameters which only considers gene duplication/loss (duplication cost [2]; transfer cost [1,000]; loss cost [1]). TREERECS (Comte et al. 2020) was also adopted to confirm the robustness of enrichment/depletion patterns for the duplication/loss events (duplication cost [2]; loss cost [1]) (supplementary fig. S4, Supplementary Material online). For each gene family, after the multiple alignment of CDS using MUSCLE (Edgar 2004), a gene tree was constructed using PhyML v3.0 (Guindon et al. 2010) with bootstrap number of 100 and substitution model of "GTR" (see supplementary data, Supplementary Material online for more details).

Statistical Analysis

Chi-squared tests were used to detect the significant differences regarding repeats, homologous groups, and evolutionary events between *A. laboriosa* and *A. dorsata*. Categories with a false discovery rate (Benjamini–Hochberg procedures) below 0.1 were deemed to be significant.

Supplementary Material

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

Data Availability

The drafted genome of the wild *A. laboriosa* is available at NCBI BioProject PRJNA647849. All command lines and configuration files for programs used in this study are available at https://github.com/lindan1128/Apis-Laboriosa-Project.

Acknowledgments

This project was supported by the APRC grant (9676008) from the City University of Hong Kong, and the earmarked fund for China Agriculture Research System (No. CARS-44-KXJ21). We thank Prof. Haiou Kuang from Yunnan Agricultural University for giving the help to collect samples.

Author Contributions

J.X. and J.L. designed this study. P.S. performed the experiments. D.L. and L.L. analyzed the data. D.L. drafted the manuscript. All authors commented on the manuscript and agreed to the submission.

Literature Cited

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 18(8):1585–1592.

Arias MC, Sheppard WS. 2005. Phylogenetic relationships of honey bees (Hymenoptera: Apinae: Apini) inferred from nuclear and mitochondrial DNA sequence data. Mol Phylogenet Evol. 37(1):25–35.

Bansal MS, Kellis M, Kordi M, Kundu S. 2018. RANGER-DTL 2.0: rigorous reconstruction of gene family evolution by duplication, transfer and loss. Bioinformatics 34(18):3214–3216.

Batra SWT. 1996. Biology of *Apis laboriosa* smith, a pollinator of apples at high altitude in the great Himalaya range of Garhwali, India (Hymenoptera: Apidae). J Kans Entomol Soc. 69:177–181.

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

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.

Chhakchhuak L, et al. 2016. Complete mitochondrial genome of the Himalayan honey bee, *Apis laboriosa*. Mitochondrial DNA A DNA Mapp Seq Anal. 27(5):3755–3756.

Comte N, et al. 2020. Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. Bioinformatics 36(18):4822–4824.

Diao Q, et al. 2018. Genomic and transcriptomic analysis of the Asian honeybee *Apis cerana* provides novel insights into honeybee biology. Sci Rep. 8(1):1–14.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Engel MS, Schultz TR. 1997. Phylogeny and behavior in honeybees (Hymenoptera: Apidae). Ann Entomol Soc Am. 90(1):43–53.

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.

Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using evidence-modeler and the program to assemble spliced alignments. Genome Biol. 9(1):R7.

Heshiki Y, et al. 2017. Toward a metagenomic understanding on the bacterial composition and resistome in Hong Kong banknotes. Front Microbiol. 8:632.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443(7114):931.

Jones P, et al. 2014. Interproscan 5: genome scale protein function classification. Bioinformatics 30(9):1236–1240.

Joshi SR, Ahmad F, Gurung MB. 2004. Status of *Apis laboriosa* populations in Kaski district, Western Nepal. J Apic Res. 43(4):176–180.

Kapheim KM, et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. Science 348(6239):1139–1143.

Li F, et al. 2019. Insect genomes: progress and challenges. Insect Mol Biol. 28(6):739–758.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.

Li R, et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20(2):265–272.

Mendes FK, Hahn MW. 2016. Gene tree discordance causes apparent substitution rate variation. Syst Biol. 65(4):711–721.

Murrell B, et al. 2015. Gene-wide identification of episodic selection. Mol Biol Evol. 32(5):1365–1371.

Oppenheim S, Cao X, Rueppel O, et al. 2020. Whole genome sequencing and assembly of the Asian honey bee *Apis dorsata*. Genome Biol Evol. 12(1):3677–3683.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676–679.

Sabeti PC, et al. 2006. Positive natural selection in the human lineage. Science 312(5780):1614–1620.

Scheiner R, Baumann A, Blenau W. 2006. Aminergic control and modulation of honeybee behaviour. Curr Neuropharmacol. 4(4):259–276.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E, Zdobnov E. 2015. BUSCO: assessing genome assembly and annotation completeness with single copy orthologs. Bioinformatics 31(19):3210–3212.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 25(1):4–10.

Venkat A, Hahn MW, Thornton JW. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat Ecol Evol. 2(8):1280–1288.

Willis LG, Winston ML, Honda BM. 1992. Phylogenetic relationships in the honeybee (genus *Apis*) as determined by the sequence of the cytochrome oxidase ii region of mitochondrial DNA. Mol Phylogenet Evo. 1(3):169–178.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Zheng T, et al. 2019. Mining, analyzing, and integrating viral signals from metagenomic data. Microbiome 7(1):1–15.

Associate editor: Dennis Lavrov