EUROPEAN JOURNAL OF ENTOMOLOGY
ISSN (online): 1802-8829
http://www.eje.cz
Eur. J. Entomol. **115**: 650–657, 2018
doi: 10.14411/eje.2018.064
ORIGINAL ARTICLE

The draft genome sequence of the Japanese honey bee, *Apis cerana japonica* (Hymenoptera: Apidae)

Kakeru YOKOI[1,*], Hironobu UCHIYAMA[2], Takeshi WAKAMIYA[3,], Mikio YOSHIYAMA[4], Jun-ichi TAKAHASHI[3], Tetsuro NOMURA[3], Tsutomu FURUKAWA[5,***], Shunsuke YAJIMA[2,5] and Kiyoshi KIMURA[4,****]**

[1] Insect Genome Research and Engineering Unit, Division of Applied Genetics, Institute of Agrobiological Sciences (NIAS), National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba, Ibaraki 305-8634, Japan; e-mail: yokoi123@affrc.go.jp

[2] NODAI Genome Research Center, Tokyo University of Agriculture, 1-1-1 Sakuragaoka, Setagaya, Tokyo, 156-8502, Japan; e-mails: hu202456@nodai.ac.jp, yshun@nodai.ac.jp

[3] Department of Life Sciences, Kyoto Sangyo University, Kyoto, Motoyama, Kamigamo, Kita-ku, Kyoto, 603-8555 Japan; e-mail: jit@cc.kyoto-su.ac.jp, nomurat@cc.kyoto-su.ac.jp

[4] Honeybee Research Group, Division of Animal Breeding and Reproduction, National Institute of Livestock and Grassland Science (NILGS), National Agriculture and Food Research Organization (NARO), Tsukuba, 2 Ikenodai, Tsukuba, Ibaraki 305-0901, Japan; e-mail: yoshiyam@affrc.go.jp

[5] Department of Bioscience, Tokyo University of Agriculture, 1-1-1 Sakuragaoka, Setagaya, Tokyo 156-8502, Japan

Key words. Hymenoptera, Apidae, *Apis cerana japonica*, genome sequence, transposable elements, innate immune genes

Abstract. Honey bees are not only important for honey production but also as pollinators of wild and cultivated plants. The Eastern honeybee (*Apis cerana*) is more resistant to several pathogens than the Western honeybee (*Apis mellifera)*, and the genomes of two strains of the nominotypical subspecies, *A. cerana cerana*, northern (Korea) and southern (China) strains, have been sequenced. *Apis cerana japonica,* another subspecies of *A. cerana*, shows many specific features (e.g. mildness, low honey production and frequently absconds) and it is important to study the molecular biological and genetic aspects of these features. To accelerate the genetic research on *A. cerana japonica*, we sequenced the genome of this subspecies. The draft genome sequence of *A. cerana japonica* presented here is of high quality in terms of basic genome status (e.g. N50 is 180 kbp, total length is 211 Mbp, and largest contig length is 1.31 Mbp) and BUSCO results. The gene set of *A. cerana japonica* was predicted using AUGUSTUS software and the set of genes was annotated using Blastp and InterProScan, and GO terms were added to each gene. The number of genes is higher than in *A. mellifera* and in the two strains of *A. cerana cerana* sequenced previously. A small number of transposable elements and repetitive regions were found in *A. cerana japonica*, which are also in the genomes of *A. mellifera* and the northern and southern strains of *A. cerana cerana*. *Apis cerana* is resistant to several pathogens that seriously damage *A. mellifera*. We searched for 41 orthologs related to the IMD and Toll pathways, which have key roles in the immune reaction to invading pathogens. Some orthologs were not identified in the genome of the northern strain of *A. cerana cerana*. This indicates that the Toll and IMD pathways function in the same way as in *A. mellifera* and *Drosophila melanogaster.* Use of the draft genome sequence of *A. cerana japonica* provided herein and those of the other *Apis* (sub)species may help to accelerate comparative research on the genome of honey bees.

INTRODUCTION

Honey bees are not only important for honey production but also as pollinators of wild and cultivated plants. In addition, honey bees are used as a model insect for social insect biology. Because of its importance, the western honey bee, *Apis mellifera* (Linnaeus, 1758) (Hymenoptera: Apidae), was first chosen for the sequencing of the whole genome after the fruit fly *Drosophila melanogaster* (Diptera: Drosophilidae) (Meigen, 1830), mosquito *Anopheles gambiae* (Diptera: Anophelinae) (Giles, 1902) and silkworm

* K. Yokoi and H. Uchiyama contributed equally to this work.

** Present address: Graduate School of Life Sciences, Tohoku University, 6-3 Aoba, Aramaki-aza, Aoba-ku, Sendai, Miyagi 980-8578, Japan; e-mail: takeshi.wakamiya.q2@dc.tohoku.ac.jp

*** Present address: Yamazaki University of Animal Health Technology, 4-7-2 Minami-osawa, Hachioji, Tokyo 192-0364, Japan; e-mail: t_furukawa@yamazakicollege.onmicrosoft.com

**** Corresponding author; e-mail: kimura@affrc.go.jp

Final formatted article © Institute of Entomology, Biology Centre, Czech Academy of Sciences, České Budějovice.
An Open Access article distributed under the Creative Commons (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).
OPEN ACCESS

Bombyx mori (Lepidoptera: Bombycidae) (Linnaeus, 1758) (Honeybee Genome Sequencing Consortium 2006, and references therein). Starting with the genome sequence of the western honeybee, genome sequences of other *Apis* species (Hymenoptera: Apidae) including *Apis cerana* subspecies [e.g. *Apis florea* (Fabricius, 1793), *Apis dorsata* (Fabricius, 1793), *Apis cerana cerana* northern (Korea) and southern (China) strains (Fabricius, 1793), *Apis mellifera inermissa* (Buttel-Reepen, 1906) and *Apis mellifera syriaca* (Skorikov, 1829)] were also generated (Wallberg et al., 2014; Park et al., 2015; Haddad et al., 2015, 2016, 2018; Diao et al., 2018).

A key feature of *Apis* genomes is a higher AT-rich content compared with other model insect species, and this feature is also reported in other insects. For example, the genomes of *A. mellifera, A. cerana cerana* northern and southern strains, *Apis florea* and *Polistes dominula* (Hymenoptera: Vespidae) (Christ, 1791) (their genome sizes are about 235 Mbp, 238 Mbp, 226 Mbp, 230 Mbp, and 230 Mbp, respectively) with approximately 67%, 62%, 60%, 65%, 67% and 69% AT contents, respectively (Honeybee Genome Sequencing Consortium, 2006; Wallberg et al., 2014; Park et al., 2015; Standage et al., 2016; Diao et al., 2018), whereas *D. melanogaster* and *A. gambiae* genomes have 58% and 56% AT contents, respectively (Adams et al., 2000; Holt et al., 2002).

Another feature of honey bee genomes is the low level of transposable elements (TEs) and repetitive sequences. The values for the percentages of TEs and repetitive sequences in genomes vary between insect species, from 16% in *A. gambiae* (Holt et al., 2002), 47% in *Aedes aegypti* (Linnaeus, 1762) (Diptera: Culicidae) (Nene et al., 2007), 33% in *Tribolium castaneum* (Herbst, 1797) (Coleoptera: Tenebrionidae) (Tribolium Genome Sequencing Consortium, 2008) and over 10% in *P. dominula* (Standage et al., 2016) to only 3%, 6.48% and 4.2% in *A. mellifera*, *A. cerana* northern and southern strains, respectively, which indicates *Apis* species have one of the lowest contents of TEs in the animal kingdom (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015; Diao et al., 2018).

The western honey bee genome contains fewer genes involved in innate immunity, detoxification enzymes and gustatory (taste) receptors, although not surprisingly, it contains more genes for olfactory receptors and novel genes for nectar and pollen utilization (Evans et al., 2006; Honeybee Genome Sequencing Consortium, 2006). The honey bee genome is more similar to that of vertebrates than insects for genes involved in circadian rhythm, as well as biological processes involved in turning genes on or off. These features accord with the ecology and social structure of honey bees. Such reductions appear to be especially pervasive in the immune system. Although *A. mellifera* has 1 : 1 orthologs related to the signal transduction phase and NF-κB transcription factors in the Toll and IMD pathways, *A. mellifera* has fewer immune effector gene, such as antimicrobial peptide genes, compared with *D. melanogaster, T. castaneum* and *A. gambiae* (Evans et al., 2006). The immune flexibility in bees may be associated with either the strength of social barriers to disease, or a tendency for bees to be attacked by a limited set of highly coevolved pathogens (Evans & Spivak, 2010).

The genomes of the two subspecies of the Asian honey bee *A. cerana cerana* northern and southern strains have been sequenced (Park et al., 2015; Diao et al., 2018). The general features of the genome of the western honey bee are also conserved in these species. Compared with *A. mellifera*, *A. cerana* is known to be more resistant to varroa mites (Peng et al., 1987) and American foulbrood (Chen et al., 2000). On the other hand, it is known that the eastern honey bee is vulnerable to tracheal mites, *Acarapis woodi* (Rennie, 1921) (Tarsonemoidea: Tarsonemidae) (Sakamoto et al., 2017), and susceptible to Chinese sacbrood (Shan et al., 2017). The Japanese honey bee *A. cerana japonica* (Radoszkowski, 1877) (Hymenoptera: Apidae), a subspecies of *A. cerana*, differs in characteristics such as mildness, low honey production, hot defensive bee ball and frequent absconding, from Western honey bees (Matsuura, 1988; Ono et al., 1995; Yoshida, 2000). Even though *A. cerana japonica* is a relatively well investigated species compared with the other subspecies of *A. cerana cerana*, there is little genetic or molecular level evidence for these behaviours and features (Ugajin et al., 2012). Genetic variation among subspecies of the Eastern honey bee has been studied (Smith et al., 2000). *Apis cerana cerana* and *A. cerana japonica* are considered to be one species because the average genetic distance of 13 mitochondrial protein-coding genes in the two subspecies is 0.006 (Takahashi et al., 2016). To accelerate genetic research on *A. cerana japonica*, we determined the whole genome sequence of *A. cerana japonica*.

There are two reasons for generating the whole genome sequence of *A. cerana japonica*. One is to investigate the immune genes in more detail, as the genome sequences could contribute to further understanding of the characteristic behaviour and features of this system. Furthermore, we assume that the features of *A. cerana japonica* related to bacterial resistance could be reflected in differences in the immune-related genes. As described above, the IMD and Toll pathways have key roles. However, several important orthologs consisting of the IMD and Toll pathways have not been identified in the genome of *A. cerana cerana* from Korea, the northern strain (*Pelle*, *BG4*, *Kenny* and *Dredd*) (Park et al., 2015). To confirm whether orthologs related to the IMD and Toll pathways exist in the *A. cerana* genome, we searched for these orthologs in *A. cerana japonica* using a local Blast method.

Another reason for a whole genome sequencing of *A. cerana japonica* was to survey for TEs. The lower number of TE or depletion of TEs in *Apis* species is one of the mysteries of *Apis* genomes (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015; Diao et al., 2018). We know that methylation cannot occur in TEs in *Apis* species and it is the very important to solve this mystery (Lyko et al., 2010; Foret et al., 2012). For this we need to increase our understanding. From the characteristics of the frequency on an evolutionary time scale, comparison

of TEs in closely related species is critical, and a whole genome analysis of *A. cerana japonica* is most suitable for this purpose.

In this report, we summarize the features of the Japanese honey bee genome by comparing it mainly with the genomes of the two strains of *A. cerana cerana* and *A. mellifera*, with particular focus on innate immune genes and TEs.

MATERIAL AND METHODS

Bee sample

A drone pupa was collected from a hive in an apiary at Kyoto Sangyo University, Kyoto, Japan in 2012 (35°04′13″N, 135°45′28″E).

Extraction of genomic DNA

Genomic DNA (gDNA) was extracted from thoracic muscles using the standard phenol/chloroform method, which is based on Sambrook et al. (1989). To avoid contamination, the sample was washed with 99.5% ethanol and the muscle removed using dissecting scissors and sterilized tweezers. Lysis buffer (100 mM Tris-HCl pH 8.0, 10 mM EDTA, 0.5% SDS), Proteinase K (Qiagen, Hilden, North Rhine-Westphalia, Germany) and RNase A (Qiagen) were used to dissolve the tissue, which was then incubated overnight at 55°C. The solution was then washed with Tris-EDTA-saturated phenol, phenol/chloroform/isoamyl alcohol (25 : 24 : 1) and chloroform/isoamyl alcohol (24 : 1). Finally, gDNA was purified by precipitating in ethanol and re-eluting in Tris-EDTA buffer.

DNA preparation and sequencing

Using the extracted gDNA, we prepared three libraries for the construction of the *Apis cerana japonica* genome. A flow diagram of the procedure is shown in Fig. 1. The three libraries were: a short-read paired-end library using a NEB Ultra DNA Library Preparation Kit (New England BioLabs, Beverly, MA, USA), a synthetic long-read DNA library (Illumina, San Diego, CA, USA) using a Truseq Long-Read Synthetic DNA library preparation Kit, and a long-read library using a rapid sequencing kit (RAD002) (Oxford Nanopore Technologies, Oxford, UK). The short-read paired-end library and the synthetic long-read DNA library were sequenced using Illumina HiSeq 2500 (Illumina), and the long-read library was sequenced using MinION with R9.4 flow cells and a MinKnow v1.4 (Oxford Nanopore Technologies). The raw sequence data from the short-read paired-end library, the synthetic long-read DNA library and the long-read library were deposited in the DDBJ sequence read archives (accession numbers DRR095708, DRR095707, and DRR095709, respectively in DRA005890).

Bioinformatics methods

Genome assembly, gene prediction and gene annotation

Approximately two-hundred-and-fifty-thousand synthetic long-reads were constructed using BaseSpace (Illumina). Primary contigs were constructed using Spades ver. 3.10.1 software (Bankevich et al., 2012; Bankevich & Pevzner, 2016), with default settings, from the synthetic long-reads and short-read paired-end reads. Finally, the draft genome sequences were constructed from the primary contigs and the long-read sequences using the Finishing module in CLC genomic workbench ver. 1.7 (Qiagen) with default settings. The work flow for constructing the *A. cerana japonica* draft genome is shown in Fig. 1. The draft genome sequences were deposited in DDBJ and the accession numbers of each contig are BDUG01000001 to BDUG01003315. The gene

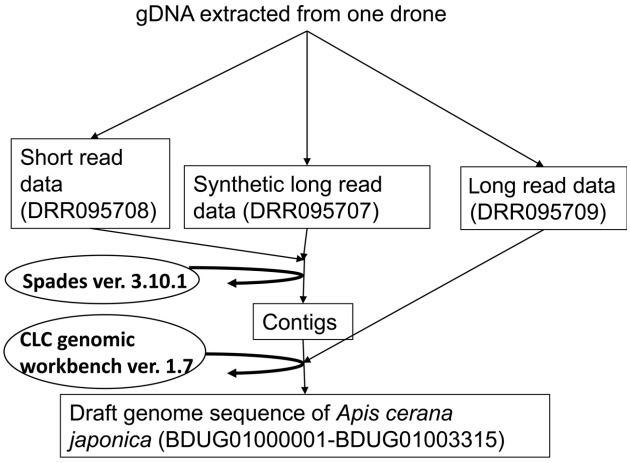

Fig. 1. Flow diagram of the stages in the construction of the *A. cerana japonica* draft genome. The *A. cerana japonica* draft genome was constructed from multiple types of sequence data (square boxes) using several types of software (circular boxes). This figure shows the types of data and software used. Numbers in brackets indicate the accession numbers of each sequence data set in DDBJ.

set of *A. cerana japonica* was predicted using AUGUSTUS ver. 3.2.3 assisted by the *A. mellifera* library and with default settings (Stanke & Waack, 2003). For validation of the predicted gene set, the predicted amino acid sequence data for *A. mellifera* (version Amel_4.5), *A. cerana cerana* northern strain (version ACSNU-2.0), *A. cerana cerana* southern strain (BioProject Accession number: PRJNA239323) and *A. cerana japonica* were evaluated using BUSCO.v3 with Insecta_odp9 and default settings (Simão et al., 2015: http://busco.ezlab.org/).

The amino acid sequences of the gene set were analyzed using the Blastp method against the NCBI-nr dataset, with default settings except e-value < 1e−3 and the three top descriptions limited. Additionally, InterProScan was used with the gene set using Blast2GO version 5.1.12 with default settings (Conesa et al., 2005). Using the Blastp and InterProScan results, the Gene Ontology (GO) terms were added using Blast2GO version 5.1.12 with default settings.

Searching for transposable elements and repetitive sequences

The genome sequence of *A. cerana japonica* was searched for TEs and repetitive sequences using Repeat masker with default settings (version open 4.0.5 with *Apis mellifera* sequence library in Repbase 20160829 update version with RMBlastn version 2.2.27+: Smit A.F.A., Hubley R. & Green P. RepeatMasker Open 4.0. 2013–2015, http://www.repeatmasker.org).

Identification of orthologs related to the Toll and IMD pathways

A Blastp database of amino acid sequences of the predicted gene sets in *A. cerana japonica* and *A. cerana cerana* southern strain was compiled (sequence data are available at https://www.ncbi.nlm.nih.gov/genome/proteins/12051?genome_assembly_id=328750). Query amino acid sequences related to the Toll and IMD pathways were extracted from the *A. mellifera* genome database, BeeBase, and *D. melanogaster*, FlyBase or NCBI GenBank (each query sequence and accession ID is listed in Supplementary data 6). Almost all *A. mellifera* sequence IDs in the query sequences, which begin with "GB1….", were extracted from the amel_OGSv1.0 data file (http://hymenopteragenome.org/drupal/sites/hymenopteragenome.org.beebase/files/data/amel_OGSv1.0_pep.fa.gz). If the Blastp result of each query

showed sequences with e-values < 1e−10, we regarded the top hit sequence as an ortholog of the corresponding query gene.

Supplementary data

All six supplementary data sets and the description of each data set can be accessed through DOI: 10.6084/m9.figshare.6964550 or URL: https://figshare.com/articles/Supplemental_data/6964550.

RESULTS

Assembled genome sequence of *A. cerana japonica* and its predicted gene sets

We used a drone pupa from a single hive as a sample for the whole genome sequencing, because honey bees exhibit haplodiploid reproduction, in which males (drones) are haploid whereas females are diploid. We constructed the draft genome of *A. cerana japonica* from short-read data, synthetic long-read data and single molecule long-read data (Fig. 1). Table 1 shows the basic status of the *A. cerana japonica* draft genome sequence we constructed. The genome sequence consisted of 3,315 contigs. Of these contigs, 3,295 contigs were over 1 kbp in length (the largest was about 1.31 Mbp). The coverage number was about 291.8, N50 (the shortest contig length required for covering 50% of the genomes, which indicates the quality of the whole genome sequence, was 180,259 bp, and the number of Ns (undetermined or ambiguous nucleotides in genome sequences) per 100 kbp was 3.24. These numbers indicate that the constructed contigs were sufficient for use as a draft genome sequence of *A. cerana japonica.* The genome sequences contained 33.1% GC, which was similar to *A. mellifera* (32.5%), *A. cerana cerana* northern strain (30.0%) and southern strain (38.0%) (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015). Using the draft genome sequence, we obtained a predicted gene set. The predicted total number of genes was 13,222. The detailed data on the genes was derived from AUGUSTUS as a gff3 file and is available in Supplementary data 1. Data on the DNA and amino acid sequences of the predicted genes in fasta formats based on the gff3 files are in Supplementary data 2 and 3, respectively. The assessment of the predicted gene sets of *A. cerana japonica* and the three *Apis* species (*A. mellifera* and the northern and southern strains of *A. cerana cerana*) using BUSCO software and comparison of the numbers in the three genomes (Table 2) provided additional support for the draft genome. Almost all genes of *Apis* species extracted using BUSCO belong to the Complete category, in which genes are of high-identity, full-length homologs and conserved in insects. The results indicate that the sequenced genome encoded highly conserved genes and the data on the genome is of good quality and suitable for further analyses.

Table 1. Basic status of the draft genome sequence of *A. cerana japonica*.

Number of contigs	3,315
(> 1000 bp)	3,293
Largest contig length	1,316,057 bp
Total length	211,196,132 bp
N50	180,259 bp
GC content	33.04%
Ns per 100 kbp	3.24
Number of genes (predicted by Augustus)	13,222

Table 2. Results of the validation of the predicted gene sets of *A. mellifera*, *A. cerana* (Korea) and *A. cerana japonica* using BUSCO.

Category[a]	*A. m.*[c]	*A. c. c.* (N)[d]	*A. c. c.* (S)[e]	*A. c. j.*
Complete	96.5%	97.8%	97.0%	97.2%
Single-copy[b]	64.4%	66.0%	94.9%	95.5%
Duplicated[b]	34.4%	31.8%	2.1%	1.7%
Fragmented	0.4%	1.3%	1.1%	1.8%
Missing	0.8%	0.9%	1.9%	1.0%

Percentages of the genes in all the predicted gene sets assigned to each category against 1,658 candidate proteins in Insecta_odb9 data. *A. m., A. c. c.* (N), *A. c. c.* (S) and *A. c. j.* are the abbreviations for *Apis mellifera*, *Apis cerana cerana* northern strain, *Apis cerana cerana* southern strain, and *Apis cerana japonica*, respectively. [a, b] Detailed descriptions of each category including Single-copy and Duplicated, subcategory of Complete, are provided in Simão et al. (2015) and Waterhouse et al. (2018). [c] Protein sequence data of the Representative genome of *A. mellifera* (version Amel_4.5) in NCBI genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/195/GCF_000002195.4_Amel_4.5/GCF_000002195.4_Amel_4.5_protein.faa.gz). [d] Protein sequence data of the Representative genome of the northern strain of *A. cerana cerana* (version ACSNU-2.0) in NCBI genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/442/555/GCF_001442555.1_ACSNU-2.0/GCF_001442555.1_ACSNU-2.0_protein.faa.gz). [e] 9,936 Protein Sequence data in NCBI Bioproject Accession Number PRJNA239323 (https://www.ncbi.nlm.nih.gov/bioproject/239323).

To annotate the functions of the predicted genes, Blastp and InterProScan were used. Among the 13,222 predicted genes, 11,352 and 13,200 genes were annotated using Blastp and InterProScan, respectively (Fig. 2A). Consequently, using Blastp and InterProScan, GO terms were added to 5,060 genes (raw output results from Blast2GO are available in Supplementary data 4). Level 1 GO terms consisting of Biological process (Bp), Cellular component (Cc) and Molecular function (Mf) accounted for 3,275, 3,046, and 3,706 genes respectively (Fig. 2B). The frequencies of Level 2 GO terms for Bb, Cc and Mf were investigated. Among 29 level 2 Bp terms, "cellular process" and "metabolic process" were added to half of the 3,275 Bp-annotated genes (Fig. 2C). "Biological regulation" and "localization" were annotated to over 10% of the 3,275 genes. The other terms were annotated to less than10% of the genes or no genes. Among 21 level 2 Cc terms, 50% of 3,046 Cc-annotated genes were annotated with "membrane part" and "cell part" (Fig. 2D). "Organelle" "protein-containing complex", "organelle part", and "membrane" were annotated to about 20%, 20%, 15% and 10% of the 3,046 genes, respectively. The other 15 terms were annotated to less than 5% of the genes or no genes. Among the 15 level 2 Mf terms, "binding" and "catalytic activity" were annotated to over 50% of the 3,706 Mf-annotated genes (Fig. 4E). The other level 2 Mf terms were annotated to less than 10% of the 3,706 genes. Seven Mf term were not added.

Transposable elements and repeat contents

We searched for TEs and repetitive regions in the assembled genome of *A. cerana japonica* using RepeatMasker. A summary of the RepeatMasker results is shown in Table 3, and detailed data on each identified element is available in

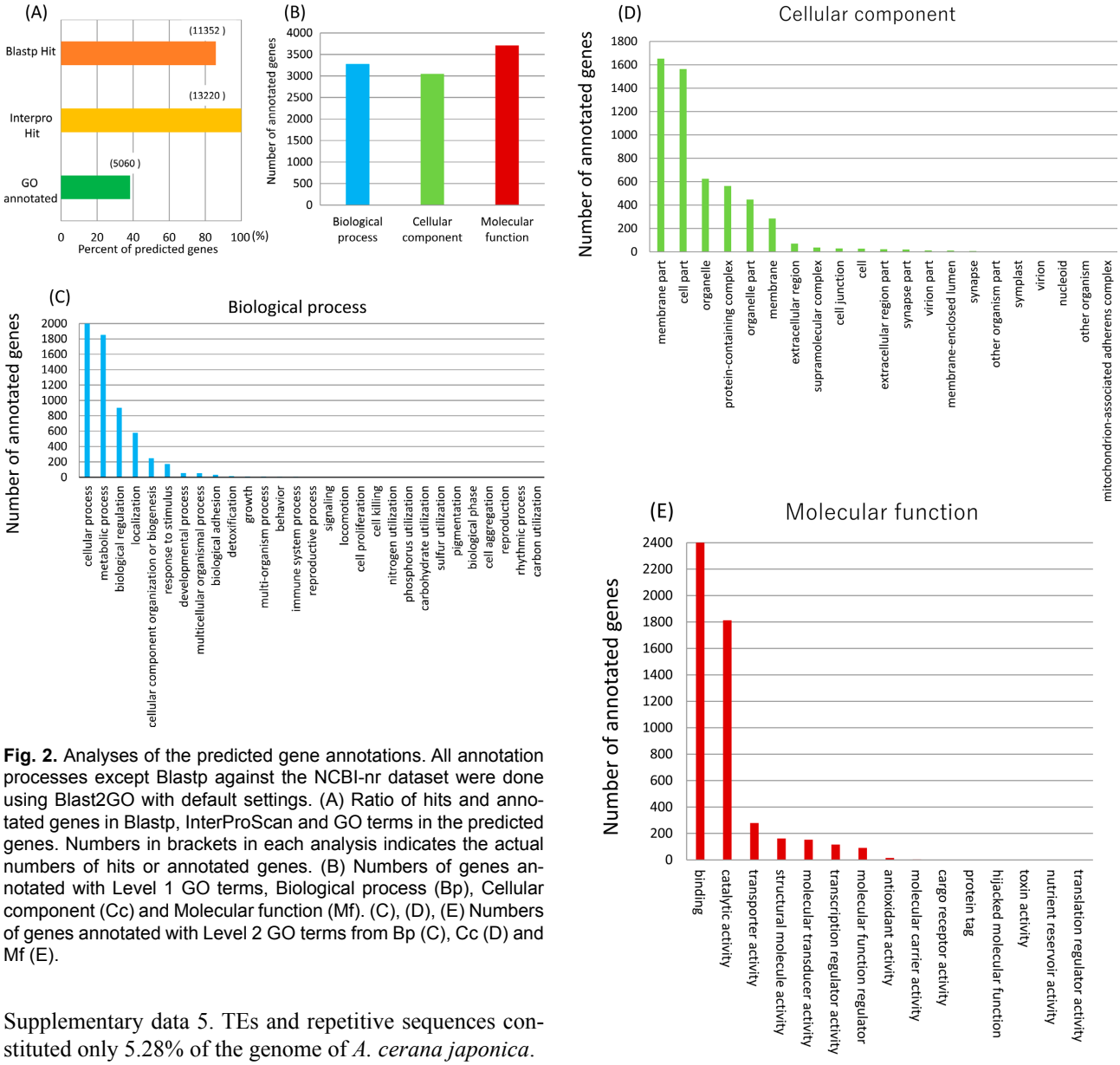

Fig. 2. Analyses of the predicted gene annotations. All annotation processes except Blastp against the NCBI-nr dataset were done using Blast2GO with default settings. (A) Ratio of hits and annotated genes in Blastp, InterProScan and GO terms in the predicted genes. Numbers in brackets in each analysis indicates the actual numbers of hits or annotated genes. (B) Numbers of genes annotated with Level 1 GO terms, Biological process (Bp), Cellular component (Cc) and Molecular function (Mf). (C), (D), (E) Numbers of genes annotated with Level 2 GO terms from Bp (C), Cc (D) and Mf (E).

Supplementary data 5. TEs and repetitive sequences constituted only 5.28% of the genome of *A. cerana japonica*.

TEs were classified into two main classes, class I retrotransposable elements and class II DNA transposons (Piégu et al., 2015). Class I retrotransposable elements use mRNA as an intermediate and transpose in a "copy and paste" manner. Class I TEs are further classified into two subclasses according to whether or not long terminal repeats (LTR) are present in the TE: non-LTR elements and LTR elements. Non-LTR elements consist of long repeats interspersed with nuclear elements (LINEs) and short repeats interspersed with nuclear elements (SINEs). As shown in Table 3, only 44 class I retrotransposable elements were found in the genome of *A. cerana japonica* (total length 75,753 bp; 0.04% of the genome). The 44 class I elements were categorized into the LINEs R2/R4/NeS. Class II DNA transposons move in a "cut and paste" manner, and only 149 class II TEs were identified (total length 57,679 bp). The 144 class II TEs were classified in the family Tc1/Mariner-IS630-Pogo. Tc1/Mariner-IS630-Pogo elements are widely found in multiple species, including humans and *Drosophila*, and the sequence of the element varies. Because of this variability, the element is often referred to as a Mariner-like element (MLE) (Oosumi et al., 1995). RepeatMasker also identified 105 "Small RNAs" (total length 311,143 bp), 194,204 "Simple repeats" (total length 8,635,243 bp) and 44,509 "Low complexity" elements (total length 2,369,637 bp), which make up 0.01%, 4.08%, and 1.12%, respectively, of the genome.

Identification of orthologs involved in the Toll and IMD pathways in *A. cerana japonica*

We conducted searches for Toll and IMD pathway orthologs in *A. cerana japonica*. First, we selected 41 genes comprising Toll and IMD pathways genes from other species, most of which were from *A. mellifera* or *D. melanogaster*. Detailed results are shown in Supplementary data 5, and see the Query sequence description column in these results. These gene sequences were used as queries against the *A. cerana japonica* amino acid Blastp database we constructed. The Blastp database consisted of the amino acid sequence data of the predicted gene data set described

Table 3. Transposable elements and repetitive regions in the genome of *A. cerana japonica*.

Total length of contigs		211,196,132 bp		
Total number of masked bases		11,157,260 bp (5.28%)		
Element classifications	Repeat class / family	No. of elements*	Length occupied	% of sequence
Class I retrotransposable elements				
Non-LTR elements:				
SINEs:		0	0 bp	0%
Penelope		0	0 bp	0%
LINEs:		44	75,753 bp	0.04%
CRE/SLACS		0	0 bp	0%
L2/CR1/Rex		0	0 bp	0%
R1/LOA/Jockey		0	0 bp	0%
R2/R4/NeSL		44	75,753 bp	0.04%
RTE/Bov-B		0	0 bp	0%
LTR elements:		0	0 bp	0%
L1/CIN4		0	0 bp	0%
BEL/Pao		0	0 bp	0%
Ty1/Copia		0	0 bp	0%
Gypsy/DIRS1		0	0 bp	0%
Retroviral		0	0 bp	0%
Class II DNA transposons				
Hobo-Activator		0	0 bp	0%
Tc1/Mariner-IS630-Pogo		149	59,679 bp	0.03%
En-Spm		0	0 bp	0%
MuDR-IS905		0	0 bp	0%
PiggyBac		0	0 bp	0%
Tourist/Harbinger		0	0 bp	0%
Other		0	0 bp	0%
Unclassified		2	102	0.00%
Small RNA		105	31,113 bp	0.01%
Simple repeats		194,204	8,635,243 bp	4.09%
Low complexity		44,509	2,369,637 bp	1.12%

*Most of the repeats fragmented by insertions or deletions were counted as one element.

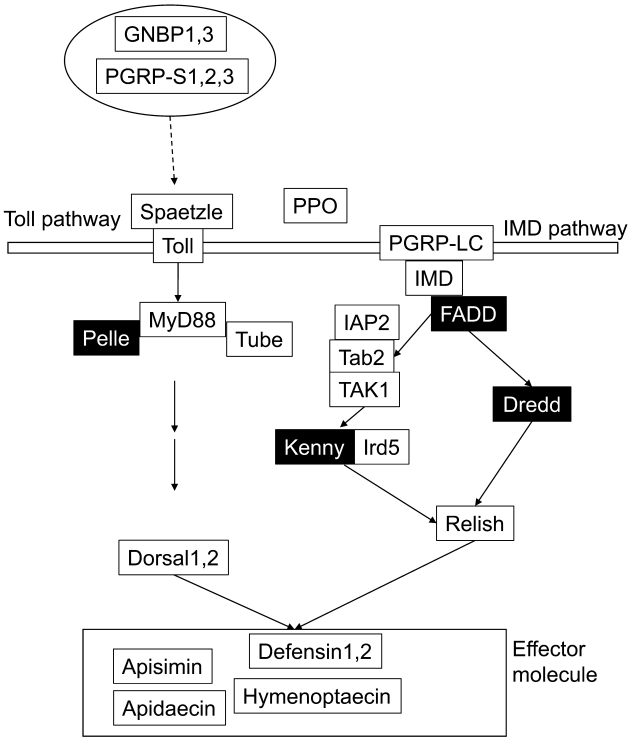

Fig. 3. IMD and Toll pathway-related genes in the genomes of *A mellifera*, northern and southern strains of *A. cerana cerana*, and *A. cerana japonica*. The IMD and Toll pathways and the genes in these pathways are known to be important in the immune reactions of several species of insects. We searched for 41 orthologs of the genes that make up the IMD and Toll pathways in the southern strain of *A. cerana cerana* and *A. cerana japonica*. The genes in the figure are the orthologs identified in the gene sets of the southern strain of *A. cerana cerana* and *A. cerana japonica*; information on these orthologs in *A. mellifera* and the northern strain of *A. cerana cerana* is reported in Evans et al. (2006) and Park et al. (2015) respectively. Genes whose names are written in black were identified in all four (sub)species of *Apis*, whereas those in white in black boxes were identified only in the genomes of *A. mellifera*, southern strain of *A. cerana cerana* and *A. cerana japonica*.

above (Supplementary data 3). In *A. cerana japonica*, all orthologs except *abaecin* were found, and e-values of the gene orthologs were < 2e–29 (highest number of the e-values: *D. melanogaster* DREDD) (Supplementary data 6). In the case of *abaecin*, we have previously identified and determined the sequence of *A. cerana japonica abaecin* (Yoshiyama & Kimura, 2010).

We also searched for the 41 orthologs in the genome of the southern strain of *A. cerana cerana*, and all 41 orthologs were found (data not shown). Taken together, these 41 immune-related genes are conserved in *A. cerana japonica* and the southern strain of *A. cerana cerana*. Several identified orthologs for the Toll and IMD pathway genes in *A. cerana japonica* and the southern strain of *A. cerana cerana* are summarized in Fig. 3.

DISCUSSION

In this study, we constructed a draft genome for the Japanese honeybee, *A. cerana japonica*. Compared with the basic genome sequence status of *A. mellifera* and the southern and northern strains of *A. cerana cerana*, e.g. contig N50 numbers, total length and BUSCO results (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015; Diao et al., 2018), the *A. cerana japonica* draft genome sequence is of higher or similar quality, respectively. For construction of the draft genome sequence, we used three different low-reads, which consisted of Illumina short-reads, Illumina short-reads from a Long-Read Synthetic DNA library, and long-reads from MinION, and utilized different software, as shown in Fig. 1. The N50 value was approximately 180 kbp, which was sufficient for using the sequences as whole a genome draft sequence, compared with other *Apis* genome sequences (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015; Diao et al., 2018). This indicates that the present method is useful for the construction of high-quality draft genomes.

A total of 13,222 genes were predicted in *A. cerana japonica* using Blastp and InterProScan. Based on the results, 5,060 genes were annotated with GO terms. Although both the ratios of Blastp and InterProScan hits in the predicted gene set were over 80%, the ratio of GO annotated genes was less than 40%. This may be because GO terms are mainly based on human, mouse or the other model species, but not non-model organisms (e.g. insects), implying that the genes that are not conserved in model species are not annotated with GO terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2017). On the other hand, among Level 2 GO term distributions in the three basic GO terms (Bp, Cc, and Mf), the GO terms with many sub terms (e.g.

cellular process and metabolic process for Bp, membrane part and cell part for Cc, and binding and catalytic activity for Mf) were annotated to many genes. These results indicate that over half of the genes in *A. cerana japonica* are not conserved in mammals or other model organisms, whereas other genes functions were conserved between *A. cerana japonica*, mammals and the model organisms.

Most genes involved in the Toll and IMD pathways, which were found in the genome of *A. mellifera*, were identified in that of *A. cerana japonica*. This indicates that the Toll and IMD pathways function in same way as in *A. mellifera* and *D. melanogaster* (Evans et al., 2006). Several genes from the intracellular Toll and IMD pathways, specifically *Pelle*, *BG4 (FADD)*, *Kenny* and *Dredd*, are not present in the genome of the northern strain of *A. cerana cerana* (Park et al., 2015). We identified these genes in the genome of *A. cerana japonica*. This might be because the quality of the genome sequence of the northern strain of *A. cerana cerana* is lower than that of *A. cerana japonica*. The genome of the northern strain of *A. cerana cerana* was constructed from sequence reads obtained using 454 Roche and Illumina sequencers, which could have resulted in incorrect sequence regions or undetermined sequence regions. *Pelle*, *BG4 (FADD)*, *Kenny* and *Dredd* in the genome of the northern strain of *A. cerana cerana* may be located in such regions and could not be identified.

Analyses showed that repetitive regions in the genome of *A. cerana japonica* made up about 5% of the whole genome sequence, which is consistent with the 3%, 6%, and 4.2% in *A. mellifera*, and the northern and southern strains of *A. cerana cerana*, respectively (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015; Diao et al., 2018). TE depletion can also be shown in *A. cerana japonica*. An interesting point is that there are only 149 MLEs in the genome of *A. cerana japonica* (Table 3), whereas there are 1,130 MLEs, 924 DNA transposons, and no MLEs in *A. mellifera* and the northern and southern strains of *A. cerana cerana*, respectively (Honeybee Genome Sequencing Consortium, 2006; Park et al., 2015). We searched for TEs in the genome of *A. cerana japonica* using RepeatMasker and the *A. mellifera* transposon library. The differences could be explained by the differences in qualities of sequences of the genomes of species of *Apis* and *A. cerana japonica*. The number of MLEs in the genomes of other species of *Apis* may be not determined correctly, because the numbers of contig N50 in the genomes of other species of *Apis* are lower than in *A. cerana japonica*. When the genome sequence was constructed from only short read data, sequences containing repeat regions could not be determined precisely. On the other hand, even if these estimated numbers may contain false positives, the number of MLEs may differ between closely related species, because among species, the number of MLEs differ by several orders of magnitude.

In this study, we determined the draft genome sequence of a subspecies of *A. cerana*: *A. cerana japonica.* Using multiple sequence techniques, a high-quality draft genome sequence was constructed. Moreover, the gene set and repetitive elements, including TEs, were predicted using the draft genome. Consequently, it was possible to analyze this data and obtain basic and comprehensive genomic insights, which indicate that the genome of *A. cerana japonica* is similar to that of the other subspecies of *A. cerana* and *A. mellifera*. However, there are several differences between these species of *Apis* (e.g. in TE numbers). Further comparative analyses will accelerate the development of the comparative research on species of *Apis*.

ACKNOWLEDGEMENTS. We express our thanks to A. Joraku at NARO, NIAS and H. Bono at DataBase Center for Life Science (DBCLS) for bioinformatic and data handling advice and assistance. This work was supported by a Cooperative Research Grant in Genome Research for BioResources from NODAI Genome Research Center, Tokyo University of Agriculture.

REFERENCES

Adams M.D., Celniker S.E., Holt R.A., Evans C.A., Gocayne J.D., Amanatides P.G., Scherer S.E., Li P.W., Hoskins R.A., Galle R.F. et al. 2000: The genome sequence of *Drosophila melanogaster*. — *Science* **287**: 2185–2195.

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T. et al. 2000: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. — *Nat. Genet.* **25**: 25–29.

Bankevich A. & Pevzner P.A. 2016: TruSPAdes: barcode assembly of TruSeq synthetic long reads. — *Nat. Methods* **13**: 248–250.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D. et al. 2012: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. — *J. Comput. Biol.* **19**: 455–477.

Chen Y., Wang C., An H. & Kai-Kuang H. 2000: Susceptibility of the Asian honey bee, *Apis cerana*, to American foulbrood, *Paenibacillus larvae*. — *J. Apic. Res.* **39**: 169–175.

Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M. & Robles M. 2005: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. — *Bioinformatics* **21**: 3674–3476.

Diao Q., Sun L., Zheng H., Zeng Z., Wang S., Xu S., Zheng H., Chen Y., Shi Y., Wang Y. et al. 2018: Genomic and transcriptomic analysis of the Asian honeybee *Apis cerana* provides novel insights into honeybee biology. — *Sci. Rep.* **8**: 822, 14 pp.

Evans J.D. & Spivak M. 2010: Socialized medicine: individual and communal disease barriers in honey bees. — *J. Invertebr. Pathol.* **15**: 645–656.

Evans J.D., Aronstein K., Chen Y.P., Hetru C., Imler J.L., Jiang H., Kanost M., Thompson G.J., Zou Z. & Hultmark D. 2006: Immune pathways and defence mechanisms in honey bees *Apis mellifera*. — *Insect Mol. Biol.* **15**: 645–656.

Foret S., Kucharski R., Pellegrini M., Feng D.S., Jacobsen S.E., Robinson G.E. & Maleszka R. 2012: DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. — *Proc. Natl. Acad. Sci. USA* **109**: 4968–4973.

Haddad N.J., Loucif-Ayad W., Adjlane N., Saini D., Manchiganti R., Krishnamurthy V., Alshagoor B., Batainh A.M. & Mugasimangalam R. 2015: Draft genome sequence of Algerian bee *Apis mellifera intermissa*. — *Genomics Data* **4**: 24–25.

Haddad N.J., Batainh A.M., Migdadi O.S., Saini D., Krishnamurthy V., Parameswaran S. & Alhamuri Z. 2016: Next gen-

eration sequencing of *Apis mellifera syriaca* identifies for *Varroa* resistance and beneficial bee keeping trains. — *Insect Sci.* **23**: 579–590.

Haddad N.J., Adjlane N., Saini D., Menon A., Krishnamurthy V., Jonklaas D., Tomkins J.P., Loucif-Ayad W. & Horth L. 2018: Whole-genome sequencing of north African honey bee *Apis mellifera syriaca* to its beneficial trains. — *Entomol. Res.* **48**: 174–186.

Holt R.A., Subramanian G.M., Halpern A., Sutton G.G., Charlab R., Nusskern D.R., Wincker P., Clark A.G., Ribeiro J.M., Wides R. et al. 2002: The genome sequence of the malaria mosquito *Anopheles gambiae*. — *Science* **298**: 129–149.

Honeybee Genome Sequencing Consortium 2006: Insights into social insects from the genome of the honeybee *Apis mellifera.* — *Nature* **443**: 931–949.

Lyko F., Foret S., Kucharski R., Wolf S., Falckenhayn C. & Maleszka R. 2010: The honey bee epigenomes: differential methylation of brain DNA in queens and workers. — *PLoS Biol*. **8**: e1000506, 12 pp.

Matsuura M. 1988: Ecological study on vespine wasps (Hymenoptera: Vespidae) attacking honeybee colonies. I. Seasonal changes in the frequency of visits to apiaries by vespine wasps and damage inflicted, especially in the absence of artificial protection. — *Appl. Entomol. Zool.* **23**: 428–440.

Nene V., Wortman J.R., Lawson D., Haas B., Kodira C., Tu Z.J., Loftus B., Xi Z., Megy K., Grabherr M. et al. 2007: Genome sequence of *Aedes aegypti*, a major arbovirus vector. — *Science* **316**: 1718–1723.

Ono M., Igarashi T., Ohno E. & Sasaki M. 1995: Unusual thermal defence by a honeybee against mass attack by hornets. — *Nature* **377**: 334–336.

Oosumi T., Belknap W.R. & Garlick B. 1995: Mariner transposons in humans. — *Nature* **378**: 672.

Park D., Jung J.W., Choi B.S., Jayakodi M., Lee J., Lim J., Yu Y., Choi Y.S., Lee M.L., Park Y. et al. 2015: Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. — *BMC Genomics* **16**: 1, 16 pp.

Peng Y.S., Fang Y., Xu S. & Ge L. 1987: The resistance mechanism of the Asian honey bee *Apis cerana* Fabr., to an ectoparasitic mite, *Varroa jacobsoni* Oudemans. — *J. Invertebr. Pathol.* **49**: 54–60.

Piégu B., Bire S., Arensburger P. & Bigot Y. 2015: A survey of transposable element classification systems – a call for a fundamental update to meet the challenge of their diversity and complexity. — *Mol. Phylogenet. Evol*. **86**: 90–109.

Sakamoto Y., Maeda T., Yoshiyama M. & Pettis J. 2017: Differential susceptibility to the tracheal mite *Acarapis woodi* between *Apis cerana* and *Apis mellifera.* — *Apidologie* **48**: 150–158.

Sambrook J., Fritsch E.F. & Maniatis T. 1989: *Molecular Cloning: A Laboratory Manual*. *Vol. 2.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 9.14–9.19.

Standage D.S., Berens A.J., Glanstad K.M., Severin A.J., Brendel V.P. & Toth A.L. 2016: Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. — *Mol. Ecol.* **25**: 1769–1784.

Shan L., Liuhao W., Jun G., Yujie T., Yanping C., Jie W. & Jilian L. 2017: Chinese Sacbrood virus infection in Asian honey bees (*Apis cerana cerana*) and host immune responses to the virus infection. — *J. Invertebr. Pathol.* **150**: 63–69.

Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V. & Zdobnov E.M. 2015: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. — *Bioinformatics* **31**: 3210–3212.

Smith D., Villafuerte L., Otis G. & Palmer M. 2000: Biogeography of *Apis cerana* F. and *A. nigrocincta* Smith: insights from mtDNA studies. — *Apidologie* **31**: 265–279.

Stanke M. & Waack S. 2003: Gene prediction with a hidden Markov model and a new intron submodel. — *Bioimformatics* **19** (Suppl. 2): ii215–ii225.

Takahashi J., Wakamiya T., Kiyoshi T., Uchiyama H., Yajima S., Kimura K. & Nomura T. 2016: The complete mitochondrial genome of the Japanese honeybee, *Apis cerana japonica* (Insecta: Hymenoptera: Apidae) — *Mitchon. DNA (B)* **1**: 156–157.

The Gene Ontology Consortium 2017: Expansion of the Gene Ontology knowledgebase and resources. — *Nucl. Acids Res.* **45**: D331–D338.

Tribolium Genome Sequencing Consortium 2008: The genome of the model beetle and pest *Tribolium castaneum*. — *Nature* **452**: 949–955.

Ugajin A., Kiya T., Kunieda T., Ono M., Yoshida T. & Kubo T. 2012: Detection of neural activity in the brains of Japanese honeybee workers during the formation of a "hot defensive bee ball". — *PLoS ONE* **7**: e32902, 12 pp.

Wallberg A., Han F., Wellhagen G., Dahle B., Kawata M., Haddad N., Simões Z.L., Allsopp M.H., Kandemir I. & De la Rúa P. 2014: A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. — *Nat. Genet.* **46**: 1081–1088.

Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G., Kriventseva E.V. & Zdobnov E.M. 2018: BUSCO applications from quality assessments to gene prediction and phylogenomics. — *Mol. Biol. Evol*. **35**: 543–548.

Yoshida T. 2000: *Methods of Rearing and Ecology of Japanese Honey Bee*. Tamagawa University Press, Machida, Tokyo, 135 pp. [in Japanese].

Yoshiyama M. & Kimura K. 2010: Characterization of antimicrobial peptide genes from Japanese honeybee *Apis cerana japonica* (Hymenoptera: Apidae). — *Appl. Entomol. Zool.* **46**: 1081–1088.

Received June 4, 2018; revised and accepted September 25, 2018
Published online November 14, 2018