csd alleles in the red dwarf honey bee (Apis florea, Hymenoptera: Apidae) show exceptionally high nucleotide diversity

Zhi-Yong Liu¹,², Zi-Long Wang¹, Xiao-Bo Wu¹, Wei-Yu Yan¹ and Zhi-Jiang Zeng¹

¹Honeybee Research Institute, Jiangxi Agricultural University, ²Experimental Animal Center, Institute of Occupational Disease Prevention, Nanchang, China

Abstract The single locus complementary sex determination (sl-csd) gene is the primary gene determining the gender of honey bees (Apis spp.). While the csd gene has been well studied in the Western honey bee (Apis mellifera), and comparable data exist in both the Eastern honey bee (Apis cerana) and the giant honey bee (Apis dorsata), no studies have been conducted in the red dwarf honey bee, Apis florea. In this study we cloned the genomic region 3 of the A. florea csd gene from 60 workers, and identified 12 csd alleles. Analysis showed that similar to A. mellifera, region 3 of the csd gene contains a RS domain at the N terminal, a proline-rich domain at the C terminal, and a hypervariable region in the middle. However, the A. florea csd gene possessed a much higher level of nucleotide diversity, compared to A. mellifera, A. cerana and A. dorsata. We also show that similar to the other three Apis species, in A. florea, nonsynonymous mutations in the csd gene are selectively favored in young alleles.

Key words  balancing selection, complementary sex determination, csd allele, nucleotide diversity, polymorphism, positive selection

Introduction

In honey bees (Apis spp.), sex is controlled by single locus complementary sex determination (sl-csd) (Cook, 1993; Heimpel & De Boer, 2008), which is an ancestral mode of sex determination found in four superfamilies of the Hymenoptera (Heimpel & De Boer, 2008). Haploids containing a single csd allele will develop into normal males (drones), while diploids, heterozygous for csd become females. When csd alleles are homozygous, a bee will develop into a diploid male, which is consumed at larval stage by workers (Woyke, 1963), because the diploid males usually represent agenetic loads for a population due to sterility, low viability, or production of sterile offspring. Therefore, nucleotide mutations leading to heterozygots at the csd gene are selectively favored. The complementary sex determination mechanism of honey bees remained an hypothesis for a long time (Whiting, 1943). In 2003, Beye et al. (2003) cloned the csd gene from Apis mellifera by positional cloning. They found that no transcription differences exist between the two sexes but suppression of csd in females with double-stranded RNA resulted in male phenotypes. csd was found to encode an arginine serine-rich (SR) type protein, which contains an RS domain in the middle and a proline-rich region at its C terminus; between these two domains is a hypervariable region that differs highly among alleles and contains a variable number of asparagine/tyrosine repeats. The honey bee csd protein is homologous to Drosophila Tra protein, which is involved in Drosophila sex determination (Beye et al., 2003).

Previous research demonstrated that the csd genes of three honey bee species (A. mellifera, A. cerana and...
A. dorsata) have evolved under balancing selection, and several parts of the coding region are possible targets of selection (Hassellmann & Beye, 2004, 2006; Charlesworth, 2004; Cho et al., 2006; Hassellmann et al., 2008b). Moreover, the polymorphic level is approximately seven times higher in the csd region (both coding and non-coding) than in neutral regions (Cho et al., 2006).

Although much evolutionary research on the csd gene has been conducted in A. mellifera, A. cerana and A. dorsata, no research on csd has been conducted in other Apis species. In this study, we analyzed the polymorphism of the csd gene in the red dwarf honey bee, A. florea. A. florea is evolutionarily the furthest removed from other Apis species (Arias & Sheppard, 2005). It has the smallest body size among the nine Apis species. This species is widespread, extending about 7 000 km from east (Vietnam and southeastern China) to west (Iran), covering mainland Asia along and below the southern flanks of the Himalayas (Hepburn et al., 2005). In A. mellifera, the csd gene contains nine exons, which form three clusters separated by two large introns (Beye et al., 2003). The genomic region of the third cluster (from exon 6 to 9) has the highest polymorphism compared to the other two regions. Therefore, we chose region 3 to study its polymorphism.

Materials and methods

Sample collection

Apis florea samples were collected from Wuming County, Guangxi Province, China. Two colonies were sampled, with 30 workers from each colony. The two colonies were about 15 km apart. The samples were first collected into 95% ethanol and stored frozen at −70°C until further use.

DNA extraction

Total genomic DNA was extracted from the cephalothorax of each sampled bee according to the protocol of the Animal Genomic DNA Extraction Kit (BEST ALL-HEAL LLC, New York, NY, USA).

PCR and sequencing

The primers used for amplifying region 3 of the A. florea csd gene in this study were the same as reported by Cho et al. (2006). Expand High Fidelity PCR Systems (Roche, Basel, Switzerland) were used for all polymerase chain reaction (PCR) reactions. PCR conditions were denaturation at 94°C for 3 min, followed by 30 cycles at 94°C for 30 s, annealing at 48°C for 30 s and extension at 72°C for 2 min, with a final extension step at 72°C for 7 min. PCR products were purified using DNA GEL EXTRACTION kits (BEST ALL-HEAL LLC) and cloned into the pEASY-T3 vector (Transgene, Beijing, China). To obtain as many csd alleles as possible, the genomic region 3 of the csd gene was cloned from the cephalothorax of each sampled worker bee, and 1–3 clones of each cloned fragment were subjected to double-strand sequencing. Single-sequencing reads were assembled using the Seqman program in the DNAstar software (Burland, 2000).

Sequence analysis

The exons, introns and coding regions of our sequences were determined by consulting the sequences of the genomic region 3 of A. mellifera csd gene reported by Cho et al. (2006) and complementary DNA (cDNA) sequences of A. mellifera csd gene reported by Hassellmann et al. (2004). Coding sequences of A. mellifera, A. cerana and A. dorsata csd genomic region 3 (n = 228 sequences) published by Cho et al. (2006) and Hassellmann et al. (2008b) were downloaded from GenBank under accession numbers DQ324946–DQ325026, DQ325038–DQ325133 and EU100885–EU100935. Sequences belonging to type II alleles (fem gene) and repetitive sequences were removed and the remaining sequences (n = 129) were used for polymorphism analysis together with our sequences. Nucleotide sequence alignment was performed using Clustal X version 1.8 (Thompson et al., 1997) and alignment results were adjusted manually for obvious alignment errors. Phylogenetic trees were constructed using MEGA version 4.0 (Tamura et al., 2007). The minimum evolution (ME) method and Kimura’s 2-parameter distances were adopted to obtain an unrooted tree with 2 000 bootstrap replications. The number of nonsynonymous changes per synonymous site (dN) and the synonymous changes per synonymous site (dS) were calculated using the complete deletion-of-gaps option. All population genetic analyses were performed using DNASP version 5.0 (Librado & Rozas, 2009).

Results

Polymorphism of A. florea csd alleles

From the 60 A. florea workers, we obtained a total of 54 sequences belonging to 37 sequence variants; no sequences were obtained in several individuals because of failure in PCR amplification or sequencing. A genealogy tree was constructed based on the 37 sequence variants;
High nucleotide diversity in *Apis florea* csd gene

all sequences fell into three clades, I, II and III (Fig. 1). Figure 2 shows the nucleotide diversity ($\pi$) in the exon, intron and coding region of the three clades of sequences. Both clade II and clade III sequences showed extremely low diversity, while the nucleotide diversity of the exon, intron and coding region of clade I was more than 10 times higher than those of clade II and clade III. We speculate that clade II and clade III sequences might belong to other genes. Previous studies once identified two major types (type I and type II) of csd alleles (Hasselmann & Beye, 2004; Cho et al., 2006), but more recently type II alleles were thought to be a different gene, *fem*, which is a paralog of csd with sex determination function downstream of csd (Hasselmann et al., 2008a, b). Moreover, several different fragments (including *fem* gene) were highly similar to the

---

**Fig. 1** The gene genealogy of csd region 3 alleles (genomic region including both introns and exons) from *Apis florea*. The minimum evolution method and Kimura’s 2-parameter distances were used to construct the tree. Bootstrap percentages are shown on internal branches. The scale bar represents the number of nucleotide changes per site.

**Fig. 2** Nucleotide diversity ($\pi$) (mean ± SD) of the exons, introns and coding regions of clades I, II and III sequences in *Apis florea*. $N = 12, 15$ and 10 sequences for clades I, II and III, respectively.


csd gene in the *A. mellifera* genome when we aligned the cDNA sequence of the *A. mellifera* csd gene against the *A. mellifera* genome sequence. Other genomic fragments similar to csd must have been amplified simultaneously when we used the PCR primers of the csd gene. Therefore, we used only clade I sequences/alleles (*n* = 12) for further analysis because they are most likely the actual csd gene.

We compared the nucleotide diversity (\(\pi\)) of the csd coding region of different *Apis* species. Data from *A. florea* was from this study, while data from species other than *A. florea* were derived from Cho *et al.* (2006) and Hasselmann *et al.* (2008b). The polymorphism of the *A. florea* csd gene was significantly higher than the polymorphism of the csd gene of the other three species (two-tailed \(Z\)-test, \(P < 0.01\)) (Table 1).

### Hypervariable region of *A. florea* CSD protein

By consulting the sequences of the *A. mellifera* csd gene reported by Cho *et al.* (2006) and Hasselmann *et al.* (2004), we determined the exons, introns and coding sequences of the obtained clade I allele fragments. The coding sequence is only part of the full-length coding region because the fragments we obtained only contained exons 6–9 of the csd gene. Similar to *A. mellifera*, *A. cerana* and *A. dorsata*, sequence analysis showed the coding region of this part also contains an RS domain at the N terminal and a P-rich domain at the C terminal, with a hypervariable region between these two domains (Fig. 3). The hypervariable region is rich in asparagine (N) and tyrosine (Y); they form a basic (N)\(^{1–4}\)Y repeat in each allele, and about half of the alleles have other (KHYN)\(^{1–4}\) repeats following the (N)\(^{1–4}\)Y repeats. The (N)\(^{1–4}\)Y and (KHYN)\(^{1–4}\) repeats are two important motifs found in the hypervariable region, which are often terminated with KK (lysine), KQ (glutamine), KP (proline) or KH (histidine).

### Balancing selection of *A. florea* csd alleles

Previous studies have shown that balancing selection favors the accumulation of nonsynonymous changes in young alleles in *A. mellifera*, *A. cerana* and *A. dorsata* (Cho *et al.*, 2006; Hasselmann *et al.*, 2008b). To determine whether this is also the case in *A. florea*, we analyzed the relationship between the number of nonsynonymous changes per synonymous site (dN) and the synonymous changes per synonymous site (dS). In general, dS represents the time of divergence among alleles because it accumulates over evolutionary time; low dS thus indicates a time when alleles have newly diverged from each other, whereas high dS indicates a time when alleles have diverged a long time ago. Therefore, we plotted dN against dS for all allele pairs (Fig. 4). For most of the pairs, dN is higher than dS. The average dN/dS ratio of all allele pairs is 1.66. Moreover, for newly diverged alleles, the regression line is above the dN/dS = 1 ratio (dashed line in Fig. 4), whereas for anciently diverged alleles, the regression line drops below dN/dS = 1. These results indicate that nonsynonymous mutations are selectively favored in young alleles.

### Discussion

Previous studies showed that the csd genes in *A. mellifera*, *A. cerana* and *A. dorsata* have a very high level of polymorphism (Cho *et al.*, 2006; Hasselmann *et al.*, 2008b). In this study we found that the *A. florea* csd gene had a higher polymorphism than the csd gene of *A. mellifera*, *A. cerana* and *A. dorsata*. This result further confirmed

<table>
<thead>
<tr>
<th>Species</th>
<th>(n)</th>
<th>L(^\dagger)</th>
<th>Mean ± SD% of (\pi)(^\dagger)</th>
<th>Mean ± SD% of (\theta)(^\dagger)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. florea</em></td>
<td>12</td>
<td>468</td>
<td>10.107 ± 1.159(^A)</td>
<td>9.764 ± 0.831(^{AB})</td>
</tr>
<tr>
<td><em>A. mellifera</em></td>
<td>49</td>
<td>443</td>
<td>4.685 ± 0.172(^B)</td>
<td>7.189 ± 0.603(^{BA})</td>
</tr>
<tr>
<td><em>A. cerana</em></td>
<td>49</td>
<td>446</td>
<td>3.546 ± 0.170(^C)</td>
<td>4.978 ± 0.500(^{MB})</td>
</tr>
<tr>
<td><em>A. dorsata</em></td>
<td>31</td>
<td>407</td>
<td>6.475 ± 0.338(^D)</td>
<td>7.626 ± 0.685(^{NA})</td>
</tr>
</tbody>
</table>

Data from *A. florea* was from this study, while data from species other than *A. florea* were derived from Cho *et al.* (2006) and Hasselmann *et al.* (2008b).

\(^A\) \(n\), the sequence number used for analysis.

\(^B\) L, the sequence length excluding alignment gaps.

\(^C\) In this column, means followed by different uppercase letters differ significantly at \(P < 0.01\) by two-tailed \(Z\)-test.

\(^D\) In this column, means followed by different lowercase letters differ significantly at \(P < 0.05\) by two-tailed \(Z\)-test.
High nucleotide diversity in Apis florea csd gene

Fig. 3 Amino acid sequence alignments of Apis florea csd alleles. Asterisks indicate conserved residues, and “–” indicates alignment gaps. Lysine (R) and serine (S) residues in the RS domains are labeled blue, and proline (P) residues in the P-rich domain are labeled red. Hypervariable regions are boxed.

that balance selection of csd gene may be common for all honey bee species (Cook, 1993). One reason for the higher polymorphism of the csd gene in A. florea may be that the csd gene, as the primary gene of sex determination in bees, has to maintain a high diversity to compensate for the lower effective mating frequency of A. florea, because the mating number of A. florea is the lowest among these four Apis species (Tarpy et al., 2004). The low mating frequencies will tend to produce more diploid males, increasing the genetic load diploid males pose on populations, because there may not be enough fertile haploid males left to sustain the population. However, with an increasing diversity at the sex locus, fewer diploid males are produced, which may represent an evolutionary advantage on the population level.

Previous studies identified two types (type I and type II) of csd alleles in A. mellifera and A. cerana, but the type II alleles were considered to be the fem gene
In conclusion, we found a high polymorphism level of the csd gene in *A. florea* by molecular analysis, and further verified that the sexes of bees are determined by the csd gene. The present study expanded our understanding of the sex determination mechanism in bees.

Acknowledgments

We thank Jun-Jun Hu from Guangxi for sampling the honey bees and Zachary Huang for suggesting this study and help with revising the manuscript. This work was supported by the National Modern Honeybee Industrial Technique System of China (No. ncytx-43-kxj15).

Authors' contributions

ZJZ conceived and designed the experiments. ZYL, XBW, WYY performed the experiments. ZLW analyzed the data. ZLW, ZYL and ZJZ wrote the paper. All authors read and approved the final manuscript.

References


Accepted April 1, 2011