Letter

Evolution of the complementary sex-determination gene of honey bees: Balancing selection and trans-species polymorphisms

Soochin Cho,[1] Zachary Y. Huang,[2] Daniel R. Green,[1] Deborah R. Smith,[3] and Jianzhi Zhang[1,4]

[1]*Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA;* [2]*Department of Entomology, Michigan State University, East Lansing, Michigan 48824, USA;* [3]*Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas 66045, USA*

The mechanism of sex determination varies substantively among evolutionary lineages. One important mode of genetic sex determination is haplodiploidy, which is used by ~20% of all animal species, including >200,000 species of the entire insect order Hymenoptera. In the honey bee *Apis mellifera*, a hymenopteran model organism, females are heterozygous at the *csd* (complementary sex determination) locus, whereas males are hemizygous (from unfertilized eggs). Fertilized homozygotes develop into sterile males that are eaten before maturity. Because homozygotes have zero fitness and because common alleles are more likely than rare ones to form homozygotes, *csd* should be subject to strong overdominant selection and negative frequency-dependent selection. Under these selective forces, together known as balancing selection, *csd* is expected to exhibit a high degree of intraspecific polymorphism, with long-lived alleles that may be even older than the species. Here we sequence the *csd* genes as well as randomly selected neutral genomic regions from individuals of three closely related species, *A. mellifera*, *Apis cerana*, and *Apis dorsata*. The polymorphic level is approximately seven times higher in *csd* than in the neutral regions. Gene genealogies reveal trans-species polymorphisms at *csd* but not at any neutral regions. Consistent with the prediction of rare-allele advantage, nonsynonymous mutations are found to be positively selected in *csd* only in early stages after their appearances. Surprisingly, three different hypervariable repetitive regions in *csd* are present in the three species, suggesting variable mechanisms underlying allelic specificities. Our results provide a definitive demonstration of balancing selection acting at the honey bee *csd* gene, offer insights into the molecular determinants of *csd* allelic specificities, and help avoid homozygosity in bee breeding.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. DQ324907–DQ325278.]

Sexual differentiation is a fundamental process of life, yet the mechanism of sex determination varies substantially among species (Bull 1983). Detailed molecular genetic studies have been conducted only in a few model organisms such as the fruit fly, nematode, and mouse, leaving most sex-determination mechanisms uncharacterized (Marin and Baker 1998). One important mode of genetic sex determination is haplodiploidy, which is used by ~20% of all animal species, including >200,000 species of the entire insect order Hymenoptera (e.g., ants, bees, wasps, and sawflies) (Cook 1993). In honey bees, females (either queen or worker) develop from fertilized diploid embryos, whereas males (drones) develop from unfertilized haploid embryos. This discovery, first made in the mid-nineteenth century (Dzierzon 1845), led to the belief that sex is determined by ploidy or fertilization process in honey bees (Nachtsheim 1916).

Later studies using controlled mating of queens and drones of the western honey bee *Apis mellifera* revealed that it is actually the allelic composition of a single locus that determines the sex (Mackensen 1951; Woyke 1963; Crozier 1971). This peculiar mode of sex determination, called complementary sex determination (*csd*), from which the locus was named, was first discovered in the braconid wasps (Whiting 1943), and is the primary mode of sex determination in hymenopteran insects (Cook 1993; Beye 2004). More precisely, diploid individuals heterozygous for *csd* become females, whereas haploids develop as drones. If a queen of the genotype A_1A_2 mates with a drone of genotype A_1 at the *csd* locus, half of the diploid offspring will be A_1A_1. However, in honey bees, these homozygous larvae are eaten by workers a few hours after they hatch (Woyke 1963). If taken from the hive before they are eaten and reared in artificial conditions, the homozygotes develop into sterile males (Woyke 1986).

Because homozygotes have zero fitness, *csd* is expected to be subject to strong overdominant selection, or heterozygote advantage. Furthermore, because common alleles are more likely than rare alleles to be in homozygotes, *csd* should also be subject to negative frequency-dependent selection. Population genetic theories predict that these selective forces, together known as balancing selection, will generate a high intraspecific polymorphism at *csd*, with long-lived alleles that may be even older than the species (trans-species polymorphism).

Recent molecular identification of the *csd* gene in *A. mellifera* (Beye et al. 2003) opens the door for testing these hypotheses at the molecular genetic level. The honey bee csd is a distant

[4]**Corresponding author.**
E-mail jianzhi@umich.edu; fax (734) 763-0544.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4695306. Freely available online through the *Genome Research* Open Access option.

16:1366–1375 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

homolog of the *Drosophila* Tra protein (Beye et al. 2003), which is involved in *Drosophila* sex determination. The *csd* gene has nine exons, which form three clusters separated by two large introns (Fig. 1). We name these three clusters regions 1, 2, and 3. Region 3 has an R (arginine)- and S (serine)-rich domain and a P (proline)-rich domain. These domains are known to mediate protein–protein interactions, suggesting that csd functions in coordination with other proteins. Between these domains is a hypervariable region (HVR), which harbors variable numbers of short repetitive sequences (Fig. 1).

Recently, (Hasselmann and Beye 2004) obtained cDNA sequences from 34 *A. mellifera csd* alleles, grouped into the more variable type 1 and more conservative type 2 alleles. These investigators found elevated levels of nonsynonymous differences between *csd* alleles in some parts of the gene and long branches in the gene genealogy, suggesting the action of balancing selection. However, because the polymorphic level of neutral genomic regions is unknown in honey bees, it is unclear whether *csd* is significantly more polymorphic than neutral regions. It is also unclear whether *csd* exhibits trans-species polymorphisms, a hallmark of balancing selection as previously shown in the major histocompatibility complex genes of jawed vertebrates (Klein 1987; Hughes and Nei 1988; Takahata and Nei 1990; Takahata et al. 1992) and the self-incompatibility genes of flowering plants (Ioerger et al. 1990; Dwyer et al. 1991; Richman et al. 1996; Charlesworth and Awadalla 1998) and fungi (Wu et al. 1998; May et al. 1999). Whether the molecular mechanisms determining *csd* allelic specificity are conserved across species is another unanswered question.

Here we address these questions by an evolutionary analysis of *csd* alleles obtained from individuals of *A. mellifera*, its probable sister species, *Apis cerana*, and another closely related species, *Apis dorsata*. We also compared *csd* with six randomly chosen non-coding regions in the genome. Our results provide a definitive demonstration of balancing selection and offer insights into the molecular determinants of *csd* allelic specificities.

Results

Trans-species polymorphisms at *csd*

To test the hypothesis of trans-species polymorphism at honey bee *csd*, we determined the genomic sequences of the gene (Fig. 1) in 10 *A. cerana* and 11 *A. dorsata* workers. *A. cerana* is known as the eastern hive bee and is distributed across Asia from eastern Iran to Japan, and south to India, Malaysia, Indonesia, and the Philippines. *A. dorsata* is known as the giant honey bee; its range extends from Pakistan, India, and Sri Lanka in the west, across southern China to the Philippines in the east, and south through Indochina, Indonesia, and Malaysia. The sampling locations of the bees used here are listed in Supplemental Table S1. *A. mellifera* and *A. cerana* are known to be phylogenetically closer to each other than either is to *A. dorsata* (Alexander 1991; Arias and Sheppard 2005). Previous molecular dating suggested that *A. mellifera* and *A. cerana* were separated during the Miocene (6–8 million years ago [Mya]) (Sheppard and Berlocher 1989; Garnery et al. 1991). We first targeted region 1 of the gene (Fig. 1) because insertions and deletions, which may lower the reliability of phylogenetic analyses, are less frequent in this region than in other regions of the gene, especially region 3. We amplified the genomic region using polymerase chain reaction (PCR) and cloned the PCR product into a vector. Only one clone per individual was sequenced to minimize a potential sampling bias between two alleles in a diploid worker (see Methods). Using these sequences along with 34 *A. mellifera csd* alleles that were sequenced by Hasselmann and Beye (2004) and are available at GenBank, we reconstructed a gene tree using the nucleotide sequences (Fig. 2). Six well-supported clusters are found in the tree. Alleles of *A. mellifera* can be grouped into two distinct clusters (type 1 and type 2), as previously reported (Hasselmann and Beye 2004). Similarly, alleles of *A. cerana* and *A. dorsata* also form two distinct clusters per species. We tentatively call these two clusters "type 1" and "type 2" as in *A. mellifera*. All alleles of *A. cerana* form one cluster. But alleles of *A. mellifera* type 2 and *A. dorsata* type 2 form a cluster with a high bootstrap support (99%), in exclusion of type 1 alleles of either species. This result indicates that the divergence between type 1 and type 2 alleles in *A. mellifera* and *A. dorsata* predated the divergence of the two species, thus demonstrating trans-species polymorphisms at *csd*. The divergence between *A. mellifera* and *A. cerana* is known to post-date that between *A. mellifera* and *A. dorsata* (Arias and Sheppard 2005). Thus, it is expected that *A. cerana* should also have alleles similar to the type 2 alleles found in *A. mellifera* and *A. dorsata*. However, it is possible that these alleles have been lost in *A. cerana*, as there are other highly divergent alleles present in the species. Alternatively, these alleles exist in *A. cerana*, but are not detected because of the limitation of our sample size. At any rate, *A. mellifera* harbor *csd* allelic lineages that originated before the origin of the species.

Although trans-species polymorphisms are usually taken as strong evidence for balancing selection, they may occur by chance when the species is very young, such as the Lake Victoria cichlid fishes (Nagl et al. 1998). To exclude this possibility, we sequenced six randomly selected noncoding genomic regions (each of ~1 kb) from 18 *A. mellifera* and 11 *A. cerana* workers. These individuals were randomly chosen and represented the entire spectrum of *csd* variation, including both types in each species. We then constructed gene trees using the allelic sequences at the six noncoding regions (Fig. 3A). It is clear that for these presumably neutral regions, all alleles from each species cluster in an unequivocal species-specific branch (100% bootstrap support) with a long internal branch linking the two species. This lack of trans-species polymorphisms at neutral

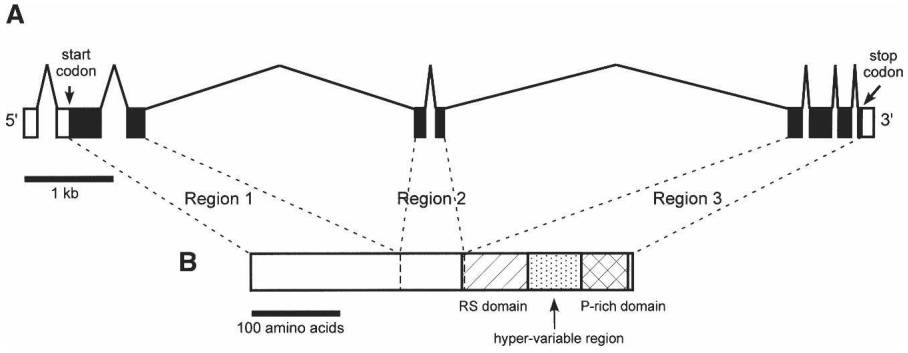

Figure 1. Diagram of the (*A*) gene and (*B*) protein structures of *csd* drawn to scale. The figure is drawn based on Beye et al. (2003).

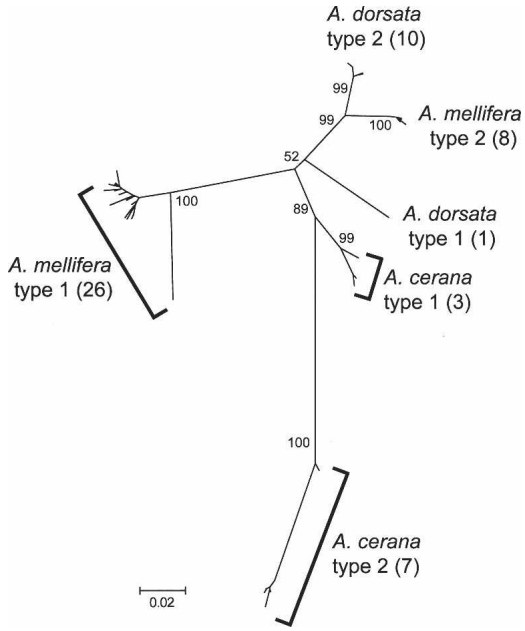

Figure 2. The gene genealogy of *csd* alleles from three honey bee species. A total of 426 nucleotide sites in the coding regions of region 1 are used in tree-making. The neighbor-joining method with Kimura's two parameter distances is used. The number of alleles in each cluster is given in parentheses. Bootstrap percentages (from 2000 replications) for major clusters are shown on internal branches. Scale bars show the number of nucleotide changes per site.

genomic regions strongly suggests that the phenomenon of trans-species polymorphisms at *csd* is not due to an exceptionally young age of the species, but rather is due to balancing selection. We did not sequence the randomly selected neutral regions from *A. dorsata* because the species is so divergent from *A. mellifera* that our primers designed based on the *A. mellifera* genome sequence did not work in *A. dorsata*, as the primers were not targeted for conserved regions as in the case of *csd*. Obviously, no trans-species polymorphisms are expected for any of these neutral regions between *A. dorsata* and *A. mellifera*.

Different hypervariable regions of csd in the three species

In *A. mellifera* csd, region 3 contains an RS domain and a P-rich domain that are likely involved in protein–protein interactions (Beye et al. 2003; Beye 2004). Between the two domains is a hypervariable region that has a potential of conferring allelic specificities (Beye et al. 2003; Beye 2004). Consistent with this view, both synonymous and nonsynonymous nucleotide variations are higher in region 3 than in other regions of the protein among *A. melllifera* type 1 alleles (Hasselmann and Beye 2004). Thus, region 3 may be directly subject to balancing selection and have deeper coalescence than other regions. To test these hypotheses, we sequenced region 3 in 42 *A. mellifera*, 45 *A. cerana*, and 13 *A. dorsata* workers sampled from various geographic locations (Supplemental Table S1). Although two distinct alleles should be present in each worker, this was found only in 32 *A. mellifera*, 38 *A. cerana*, and 10 *A. dorsata* individuals; only one allele per individual was found for the other workers, possibly because of the failure to amplify the other allele using our PCR primers (see Methods). A phylogenetic tree was made with the genomic sequences of region 3 obtained from the three species (Fig. 4). Alleles from *A. mellifera* are divided into two types, with the majority belonging to type 1. Alleles from *A. cerana* are also divided into two distinct types. However, unlike in *A. mellifera*, similar numbers of alleles exist in each *A. cerana* type. Interestingly, alleles in one *A. cerana* type are more variable than those in the other type, as is observed in *A. mellifera* (Hasselmann and Beye 2004). Therefore, we named the more variable group "type 1" and the other "type 2." Unlike region 1, region 3 sequences of *A. dorsata* appear to form a single cluster. However, the possibility of two distinct types present in natural populations cannot be ruled out, because our *A. dorsata* sample is not large (13 individuals) and all individuals were collected from one geographic location (four colonies). More extensive sampling would help clarify this issue. For region 3, all alleles from a species form a species-specific cluster in the gene tree, without apparent trans-species polymorphisms. However, it is noteworthy that the grouping of all *A. mellifera* alleles has a poor bootstrap support (64%) and the

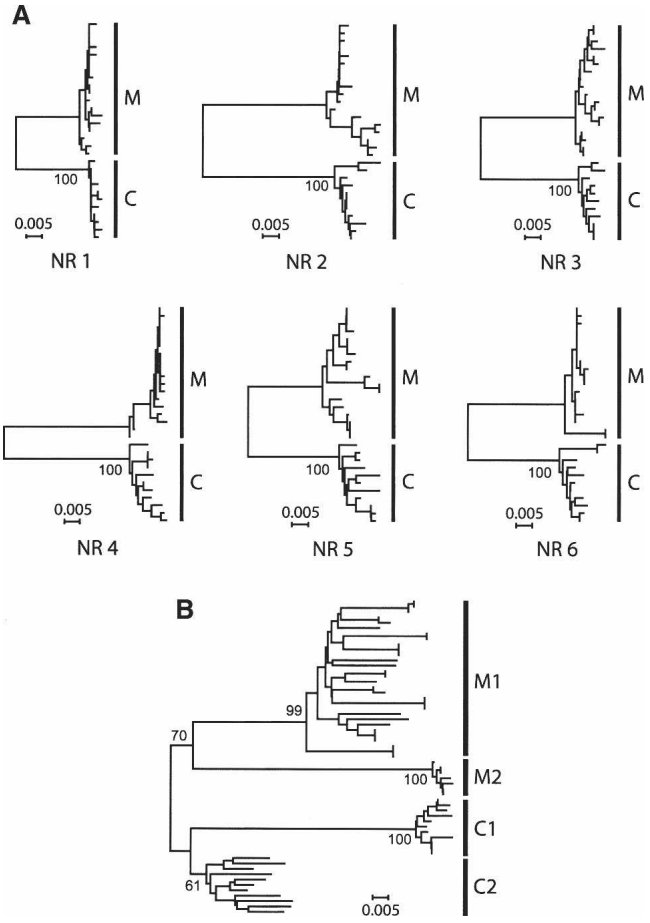

Figure 3. Gene genealogies of allelic sequences at (*A*) the six neutral genomic regions and (*B*) the *csd* region 3 from *A. mellifera* and *A. cerana*. The neighbor-joining method with Kimura's two parameter distances is used. Bootstrap percentages (from 2000 replications) for major groups are shown on the internal branches. Major groups are labeled as follows: (M) *A. mellifera*, (C) *A. cerana*, (M1) *A. mellifera* type 1, (M2) *A. mellifera* type 2, (C1) *A. cerana* type 1, and (C2) *A. cerana* type 2. The same individuals are used for all seven gene trees. For easy comparison, the scale (in units of the number of nucleotide changes per site) is the same for all the trees. (NR 1–6) Neutral regions 1–6. For tree-making, 583 (for *csd*), 934 (NR 1), 785 (NR 2), 842 (NR 3), 770 (NR 4), 864 (NR 5), and 835 (NR 6) nucleotide sites are used.

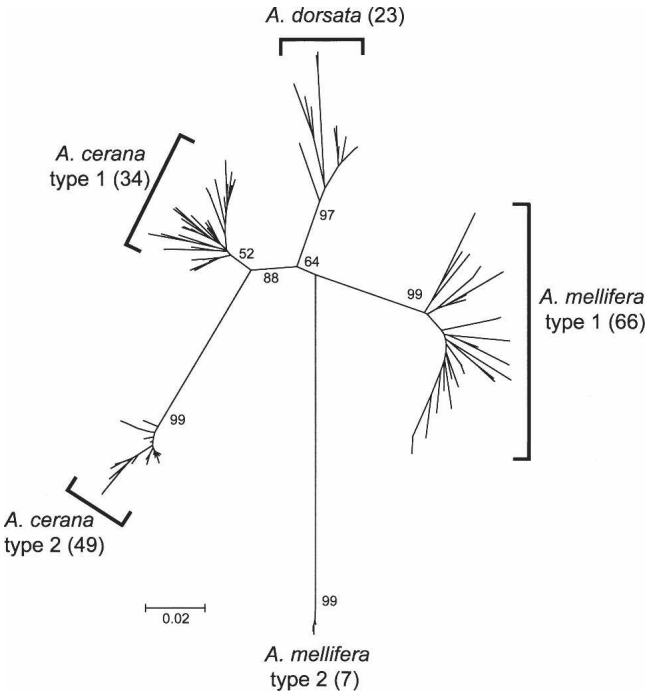

Figure 4. Gene genealogy of *csd* region 3 sequences from the three honey bee species. A total of 508 nucleotide sites are used. The neighbor-joining method with Kimura's two parameter distances is used to make the tree. Bootstrap percentages (from 2000 replications) are shown for major clusters on the internal branches. Scale bars show the number of nucleotide changes per site.

interior branch supported by this bootstrap is short. These observations suggest that the possibility of trans-species polymorphism cannot be ruled out for region 3 (see Discussion).

The RS domain and P-rich domain found in *A. mellifera* are conserved in *A. cerana* and *A. dorsata* (Supplemental Fig. S1). In a previous study, various numbers of "$(N)_{1-4}Y$" repeats located between the RS and P-rich domains were found in *A. mellifera* type 1 alleles (Beye et al. 2003). This hypervariable region (HVR) was suggested to contribute to the determination of allelic specificities in complementary sex determination (Beye 2004; Hasselmann and Beye 2004). The HVR is absent in *A. mellifera* type 2 alleles. Interestingly, we found that different types of repetitive units are present in *csd* alleles from different species (Fig. 5). For instance, in *A. cerana*, type 1 alleles have the $[(N)_{1-4}Y]_n$ repeats, and some of them also have additional $[KHYN]_n$ repeats. As in *A. mellifera* type 2, *A. cerana* type 2 alleles do not have any repeats in this region, but instead they have $[RRERSRN]_n$, in another region of the protein. All of the *A. dorsata* alleles seem to have $[(N)_{1-4}Y]_n$ repeats, and most of them also have additional $[KHYN]_n$ and/or $[KHEHYN]_n$ repeats.

Higher-than-neutral levels of polymorphism in *csd* region 3

Population genetic theories predict higher levels of polymorphism at genetic loci under balancing selection than those under neutral evolution. To test this prediction for *csd*, we calculated the nucleotide diversity (π) and nucleotide polymorphism (Watterson's θ) at *csd* region 3 and compared them with those in the six aforementioned genomic regions that are presumably neutral. For both *A. mellifera* and *A. cerana*, we found that the level of polymorphism is, indeed, much higher in both the coding region and introns of *csd* than in any of the neutral regions examined (Fig. 6). However, one may argue that this could also result from a higher mutation rate in *csd* than in the neutral regions. To exclude this possibility, we performed an HKA neutrality test comparing the intra- and interspecific sequence variations between loci, because it is known that mutation-rate variation among loci would not result in significant HKA test results (Hudson et al. 1987). The test yielded very significant results when either the coding region ($P = 0.0013$) or introns ($P = 2.33 \times 10^{-7}$) of *csd* were compared with all six neutral regions combined in *A. mellifera* (Table 1). Similar results were obtained in *A. cerana* (Table 1). We also made a gene tree using region 3 sequences of *csd* (Fig. 3B) and compared it with the gene trees of the neutral regions (Fig. 3A). The coalescence of the *csd* alleles is at least 10 times deeper than that of the neutral regions in both *A. mellifera* and *A. cerana*. If only type 1 alleles of *A. mellifera* are considered, the coalescence of the *csd* alleles is approximately five times deeper than that of the neutral regions. Overall, these results corroborate the notion that balancing selection has maintained *csd* alleles for a long time in both *A. mellifera* and *A. cerana*, resulting in higher-than-neutral levels of polymorphism at this locus. Similar results were found for region 1 (Supplemental Table S2; Supplemental Figs. S2 and S3).

Nonsynonymous mutations are selectively favored in young alleles

The balancing selection operating on *csd* not only maintains divergent allelic lineages in a population for a long time, but should also favor functionally distinct rare alleles over common alleles, because rare alleles are less likely than common ones to be in homozygotes. Thus, nonsynonymous (amino-acid-altering) nucleotide changes generating functionally distinct alleles may be positively selected, particularly when the alleles are still young and rare. To test this prediction, we plotted the number of nonsynonymous changes per nonsynonymous site (d_N) against the number of synonymous changes per synonymous site (d_S) in region 3 between all possible pairs of type 1 *csd* alleles from *A. mellifera* (Fig. 7A). For about half of the pairs, d_N is greater than d_S, and the average d_N/d_S ratio of all pairs is 1.01. This result alone does not provide convincing evidence for positive selection. Because it is expected that positive selection is strongest for new alleles, we plotted the d_N/d_S ratios of all these pairs against their d_S values, which is a proxy of time of divergence between alleles (Fig. 7B). Indeed, we found that d_N/d_S declines as d_S increases. For example, all nine gene pairs with $d_S < 0.025$ show $d_N/d_S > 1$, with an average d_N/d_S of 2.22 ± 0.30. In contrast, the 181 gene pairs with $d_S > 0.025$ have an average d_N/d_S ratio of 0.95 ± 0.02. The $d_S - d_N/d_S$ plot is fitted best to the logarithmic curve with a strong R^2 value (0.57). Hence, positive selection for nonsynonymous mutations is detected between closely related *csd* alleles, but not between divergent alleles. An alternative interpretation of the above observation is that the rate of nonsynonymous substitution is constant with time, but nonsynonymous substitutions become saturated between relatively divergent alleles owing to the structural constraint of csd. However, because the majority of d_S and d_N values are lower than 0.1 among the *csd* alleles (Fig. 7A), saturation is unlikely. In fact, of all the *csd* sequences we examined, ~80% of amino acid positions are variable in region 3. Analyses using the *A. cerana* type 1 (Fig. 7C,D) and *A. dorsata* (Fig. 7E,F) alleles yielded similar results, suggesting that similar selective forces govern the evolution of *csd* in these species. Intrigu-

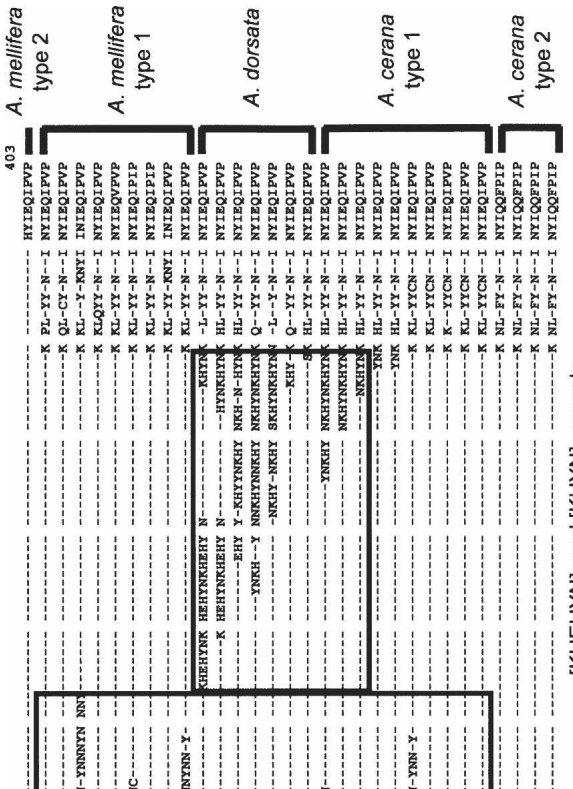

Figure 5. Three different kinds of repetitive sequences exist in csd of three honey bee species. The name of each repeat type is shown *under* the boxed area. The names of the bee samples, from which the sequences are obtained, are shown to the *left* of the sequences. Numbers in the first row indicate the positions of the first and last amino acid residues in the sequence of MW2B.

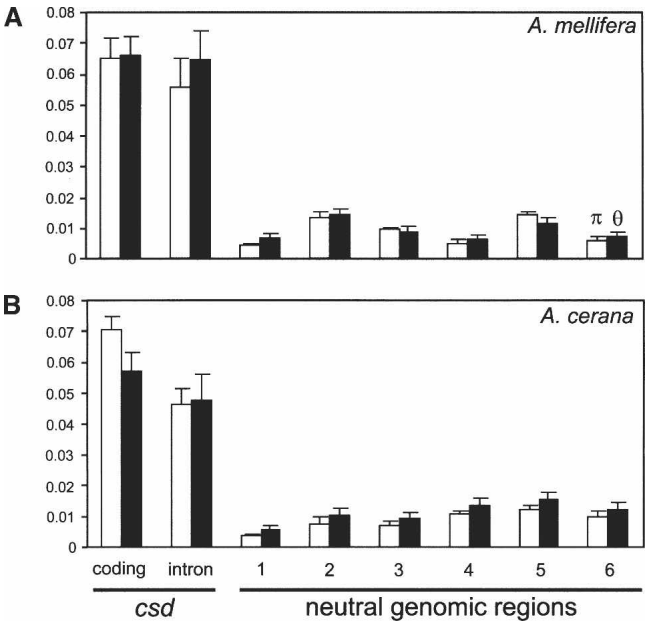

Figure 6. The level of nucleotide diversity (π, open bars) and polymorphism (Watterson's θ, solid bars) in region 3 of *csd* and randomly selected neutral genomic regions of (*A*) *A. mellifera* and (*B*) *A. cerana*. Error bars indicate one standard deviation of the measurement.

ingly, d_N is greater than d_S in most of the comparisons between *A. dorsata* alleles. It should be noted that all *A. dorsata* alleles used in this study were sampled from a single location and d_S values among them are smaller than those in the other species (Fig. 7). Type 2 alleles are too homogeneous in both *A. mellifera* and *A. cerana* for this type of analysis, and therefore are not included in the analyses. Similar patterns of d_S and d_N were observed in region 1 sequences (Supplemental Fig. S4).

Discussion

In this work, we sequenced the *csd* gene from *A. cerana* and *A. dorsata*, extending the understanding of complementary sex determination from the model organism of *A. mellifera* to other honey bees. Similar patterns of high polymorphism and balancing selection at *csd* among the three species strongly suggest that the single-locus complementary sex-determination system involving multiple *csd* alleles is common to all three species. We presented three lines of evidence for the action of balancing selection at *csd* of honey bees. First, we detected trans-species polymorphisms at *csd*. Second, we showed that the level of polymorphism at *csd* is five to 10 times that at the neutral regions. This difference cannot be explained by an elevation of the mutation rate at *csd*, but rather is due to selection. Finally, we found $d_N/d_S > 1$ between closely related *csd* alleles, reflecting positive selection for functionally distinct new alleles.

The occurrence of trans-species polymorphisms serves as strong evidence for the antiquity of alleles, which, in turn, suggests the action of balancing selection (Klein et al. 1998). However, the presence of trans-species polymorphisms by itself is insufficient for establishing balancing selection, because trans-species neutral alleles are occasionally found if the species concerned is young, as previously reported in cichlid fishes of East African great lakes (Nagl et al. 1998), which are believed to have arisen since 12,000 yr ago. Nevertheless, the lack of trans-species polymorphisms at six neutral genomic regions surveyed (Fig. 3A) and the detection of positive selection in *csd* (Fig. 7) indicate that the trans-species polymorphisms at *csd* are results of balancing selection. Introgressive hybridization, another possible explanation for trans-species polymorphisms, is unlikely, because the three honey bee species cannot crossbreed (Koeniger and Koeniger 2000), and *A. cerana* and *A. mellifera*, the two sister species, have been geographically separated since they speciated (Ruttner 1987). Indeed, no geographic clustering of alleles from different species was found in our data. One notable observation

Table 1. Intra-and interspecific DNA sequence variations in *csd* and neutral genomic regions of *A. mellifera* and *A. cerana*

							HKA with coding		HKA with introns	
Species	Regions	L^a (nucleotides)	π (%)	θ (%)	d^b (%)	θ/d	χ^2	*P*-value[c]	χ^2	*P*-value[c]
A. mellifera	*csd* coding region	435	6.33	6.88	14.8	0.465				
	csd introns	182	5.55	6.88	8.34	0.825				
	Neutral region 1	954	0.42	0.64	4.81	0.133	2.67	0.102	5.00	**0.025**
	Neutral region 2	908	1.33	1.41	8.53	0.165	2.46	0.117	5.46	**0.020**
	Neutral region 3	889	0.92	0.85	6.79	0.125	4.01	**0.045**	7.37	**0.007**
	Neutral region 4	875	0.49	0.63	9.13	0.069	7.14	**0.008**	12.6	**4.0×10^{-4}**
	Neutral region 5	919	1.39	1.14	6.38	0.178	1.57	0.210	3.75	0.053
	Neutral region 6	986	0.54	0.74	6.80	0.108	3.97	**0.046**	7.29	**0.007**
	All six neutral regions	5531	0.84	0.90	6.99	0.129	21.8	**0.001**	41.5	**2.3×10^{-7}**
A. cerana	*csd* coding region	432	7.19	5.85	14.8	0.395				
	csd introns	190	4.61	4.20	8.34	0.504				
	Neutral region 1	996	0.35	0.55	4.81	0.114	2.79	0.095	4.01	**0.045**
	Neutral region 2	822	0.76	1.04	8.53	0.122	2.31	0.129	3.77	0.052
	Neutral region 3	931	0.70	0.95	6.79	0.140	1.64	0.201	2.67	0.102
	Neutral region 4	791	1.06	1.34	9.13	0.146	7.14	**0.008**	3.26	**0.071**
	Neutral region 5	925	1.23	1.55	6.38	0.243	0.43	0.513	0.91	0.341
	Neutral region 6	874	0.99	1.21	6.80	0.178	0.95	0.329	1.69	0.193
	All six neutral regions	5339	0.84	1.09	6.99	0.157	15.3	**0.018**	16.3	**0.012**

[a] *L* is the sequence length excluding alignment gaps. Note that only region 3 of *csd* is considered in this table.
[b] *d* is the average number of nucleotide substitutions per site between the two species.
[c] Statistically significant *P*-values are in bold.

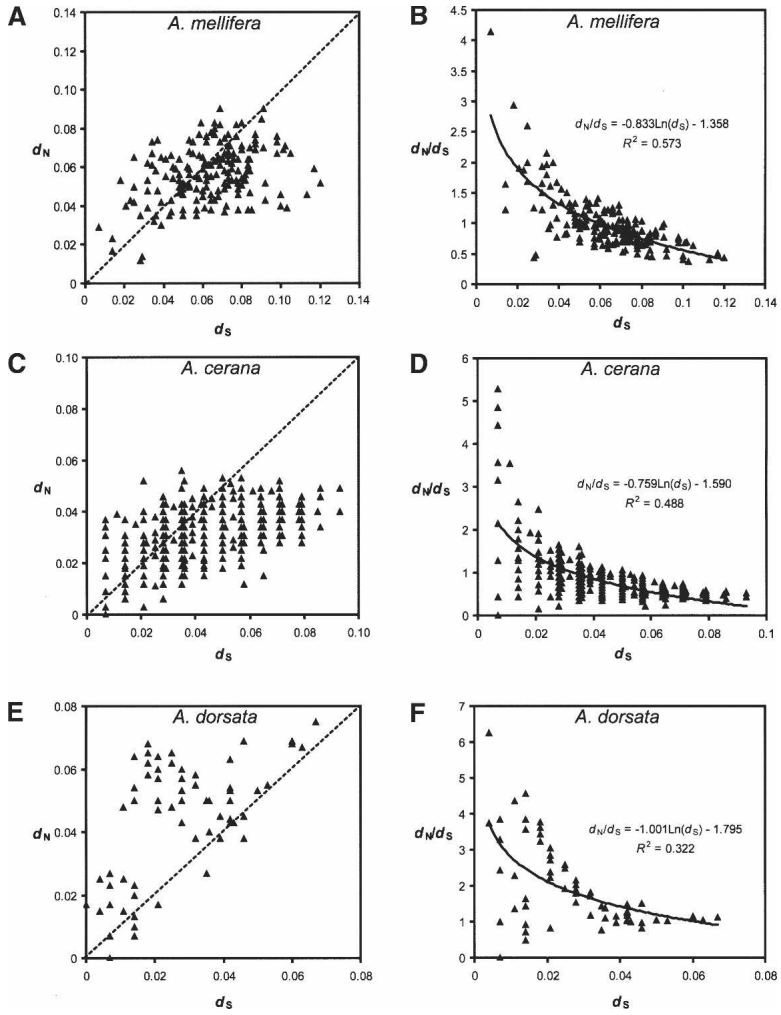

Figure 7. Patterns of nucleotide changes in region 3 of *csd*. Synonymous (d_S) and nonsynonymous (d_N) nucleotide distances are shown for all pairs of *csd* alleles in (*A*) *A. mellifera*, (*C*) *A. cerana*, and (*E*) *A. dorsata*. Decline of d_N/d_S with d_S is shown for (*B*) *A. mellifera*, (*D*) *A. cerana*, and (*F*) *A. dorsata*. In *A. mellifera* and *A. cerana*, only type 1 alleles were used in the analysis. Five data points in *A. cerana* and one in *A. dorsata* could not be plotted because their d_S values are 0, making d_N/d_S infinite.

is that trans-species polymorphisms are found in region 1, but not region 3 of the *csd*. This may be due to two reasons. First, there are more insertions and deletions in region 3 than in region 1, and they reduce the number of informative sites in region 3 for the phylogenetic reconstruction. Consequently, the reliability of the gene tree is compromised, indicated by relatively low bootstrap percentages (Fig. 4). It is possible that trans-species polymorphisms do exist in region 3, but the phylogenetic signal is too weak to be revealed. Second, it is known that the recombination rate is high at the *csd* locus (Beye et al. 1999), and recombination among alleles may have led to a reduced resolution in the reconstructed gene tree. Thus far, occurrences of trans-species polymorphisms by balancing selection have been best studied at the DNA sequence level in two biological systems: the major histocompatibility complex (*MHC*) of jawed vertebrates (Klein 1987; Hughes and Nei 1988; Takahata and Nei 1990; Takahata et al. 1992) and the self-incompatibility (SI) system of angiosperms (Ioerger et al. 1990; Dwyer et al. 1991; Richman et al. 1996; Charlesworth and Awadalla 1998) and fungi (Wu et al. 1998; May et al. 1999). Our results mark the first report of DNA-level trans-species polymorphisms caused by balancing selection in invertebrates and also the first in sex-determination systems.

It is interesting to determine the age of the *csd* alleles. The average number of nucleotide substitutions per site (*d*) between *A. mellifera* and *A. cerana* is 0.07 in the six neutral genomic regions sequenced (Table 1), and the two species diverged ~7 Mya (Sheppard and Berlocher 1989; Garnery et al. 1991). Therefore, the honey bee molecular clock ticks at a rate of ~10 substitutions per kilobase per million years, which is close to that in fruit flies (11.1) (Tamura et al. 2004). The d_S between the most divergent pair of *csd* alleles in *A. mellifera* is 0.142 (between a type 1 allele and a type 2 allele). Thus, these alleles are ~14 million yr old, indeed, older than the species.

The level of polymorphism is much higher in *csd* than in the neutral regions examined (Fig. 6). Without this comparison, we would not be able to rule out explanations, such as a large population size or a high genomic mutation rate, for the observation of the high intraspecific polymorphism at *csd*. Furthermore, our HKA test results (Table 1) indicate that it cannot be due to an elevated mutation rate at the *csd* locus. Thus, balancing selection remains as the only explanation of the high levels of polymorphism. If the *csd* locus is, indeed, under long-term balancing selection, the introns that connect the exons are also expected to have high diversity by hitchhiking (Maynard-Smith and Haigh 1974; Charlesworth 2004). Our results indicate that this is, indeed, the case (Fig. 6; Table 1). Interestingly, we observed similar levels of polymorphisms in exons and introns, but much higher levels of divergence in exons than introns, leading to a more significant HKA result for introns than exons (Table 1). This is probably because the hitchhiking effect is transient and gradually decays with evolutionary time. In the future, it would be interesting to determine the extent of this effect around the *csd* locus. Because the recombination is unusually high around *csd* (Beye et al. 1999), we expect that this effect is limited to a small area surrounding the gene.

To our knowledge, our study is the first to characterize the level of polymorphism in neutral regions of the honey bee nuclear genome. Our results, based on six randomly chosen noncoding regions totaling ~6000 nucleotides, show that the mean nucleotide diversity (π) is ~0.0084 per site in both *A. mellifera* and *A. cerana* (Table 1), slightly lower than the mean π of non-coding regions in *Drosophila melanogaster* (0.01082) (Moriyama and Powell 1996).

One of the most striking observations about *csd* is that many alleles segregate in a population. Using the frequency of homozygous males produced, Adams et al. (1977) estimated the number of *csd* alleles in an *A. mellifera* population of ~500 hives in Sao

Paulo, Brazil to be ~19, which is close to a theoretical prediction by Yokoyama and Nei (1979). In our study, we identified 18 distinct alleles from 27 workers sampled from a single hive in East Lansing, Michigan, USA. Because seven of these 18 alleles have a single occurrence in our data set, and our data are from only one hive, it is likely that the total number of *csd* alleles that segregate in an entire population is much higher than 18. Direct determination of the allele number in a population by sequencing *csd* alleles is now feasible and will clarify this issue. For *A. cerana*, we identified 48 alleles from 45 workers sampled from 19 different locations distributed in five different countries (Supplemental Table S1). This result suggests that the total number of *csd* alleles distributed in the entire species is much higher than that in a population.

To initiate female development in a diploid honey bee, her two *csd* alleles need to be recognized as distinct from each other. How are the allelic specificities of *csd* established? One potential mechanism is single amino acid substitutions. Our detection of positive selection promoting amino acid substitutions in young *csd* alleles (Fig. 7) supports this idea. Another potential mechanism is the use of short repetitive sequences, which usually have a high rate of mutation and therefore a high level of polymorphism (Fondon and Garner 2004). In fact, there is a hypervariable region between the SR and P-rich domains, where apparent differences between alleles can be found. Variations in the HVR have been suggested to mediate allelic specificity (Beye et al. 2003; Beye 2004). To our surprise, three different kinds of repeating units are identified in the HVRs of the three honey bee species (Fig. 5). Among them, only $[RRERSRN]_n$ repeats are specific to one species (*A. cerana*), and the other two kinds of repeats are shared by more than one species. Notably, the $[N_{1-4}Y]_n$ repeats are found in all three species, and thus were likely present in the common ancestor of the three species. However, we find that $[N_{1-4}Y]_n$ is neither required for csd function nor for allelic specification. The reasons are as follows. First, because $[N_{1-4}Y]_n$ is absent in type 2 alleles of *A. mellifera* and *A. cerana* (Fig. 5), the repeats must be unnecessary for the function of csd. Second, because there are seven *A. cerana* workers that contain two different copies of type 2 alleles that lack the repeats (Supplemental Table S3), the $[N_{1-4}Y]_n$ repeats must be unnecessary for determining allelic specificities at least in *A. cerana*. In *A. mellifera*, however, no workers have two copies of type 2 alleles. But, because type 2 alleles are rare (~10%) in *A. mellifera*, this result may simply be due to chance. In other words, it is still possible that the $[N_{1-4}Y]_n$ repeats are unnecessary for allelic specificity in *A. mellifera*. These analyses suggest that while the HVR enhances the allelic diversity, it is neither necessary for the function of csd nor for the determination of allelic specificity. It is likely that the combination of single amino acid substitutions and repeat variations in HVR contribute to allelic specificities.

In this study, we took population genetic and molecular evolutionary approaches to elucidate the evolutionary forces acting on the complementary sex-determination locus in honey bees and the molecular mechanisms determining allelic specificities. Honey bees have a prime economic importance not only for their honey production but also for their being the major pollinator for agriculturally important plants. Understanding the mode and mechanism of honey bee sex determination is instrumental to developing bee-breeding technology and designing successful mating. In this respect, our data and analysis of the *A. mellifera* and *A. cerana csd* sequences provide useful resources for breeding these economically important honey bees.

Methods

Bee sampling

Adult workers were sampled by hand or using a vacuum bee collector directly from colonies with the presence of sedative smoke. Sampled bees were frozen at −70°C until used, or stored in 70% ethanol solution when a freezer was not available. For this study, we collected samples of *A. mellifera*, *A. cerana*, and *A. dorsata*, whose localities are listed in Supplemental Table S1.

Genomic DNA purification

We purified genomic DNA from either head or thorax of the sampled bees using the PUREGENE genomic DNA purification kit manufactured by Gentra Systems, following the manufacturer's instruction for insect samples. Bee samples stored in 70% ethanol were dried in an oven at 60°C for 2 h before being used for DNA purification. We used disposable plastic pestles and microcentrifuge tubes produced by Kontes Glass Company to break up the exoskeleton of frozen bee tissue. The final concentration of genomic DNA purified was adjusted to ~50 ng/μL.

PCR and sequencing

We used the Expand High Fidelity PCR Systems (Roche Diagnostics Corporation) for all of the PCR reactions, with the presence of bovine serum albumin at a final concentration of 0.4 mg/mL. Because all bee workers are heterozygous for *csd*, identifying haploid sequences by directly sequencing PCR products was impossible. Instead, we had to clone each PCR product into the pCR2.1 vector (Invitrogen) for sequencing. For region 1 of *csd* and the six neutral regions (see below), a single allele was sequenced for each worker. This is because it is unknown whether a worker is homozygous for the genomic region sequenced, and a large number of colonies have to be sequenced in order to ensure the recovery of both alleles. To study the role of the hypervariable region in determining allelic specificity, we tried to obtain both alleles of each worker for region 3 of *csd* by increasing the number of clones subject to sequencing. We distinguished the two alleles obtained from the same individual by labeling them with "A" and "B" at the end of the individual name. However, for 10 of 42 *A. mellifera*, seven of 45 *A. cerana*, and three of 13 *A. dorsata* workers, we could identify only one allele even though we sequenced up to 12 clones for each worker. This is probably because one allele is preferentially amplified over the other, which has primer mismatches owing to high levels of divergence between alleles. We used some primers from Beye et al. (2003) along with the primers we designed for amplifying region 1 and region 3 of *csd*. The primers used in this study are listed in Supplemental Table S4. All PCRs were performed at 50°C for primer annealing.

Neutral region selection

We randomly chose six genomic sequences of ~1 kb from the *A. mellifera* genome sequence (http://www.ncbi.nlm.nih.gov/genome/guide/bee/) produced by the Human Genome Sequencing Center at Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/projects/honeybee/). No annotated or predicted open reading frames (ORFs) exist within a 5-kb range from the sequence in both directions. We used the GENESCAN server (http://genes.mit.edu/GENSCAN.html) to detect any ORFs. The locations of these six sequences are listed in Supplemental Table S5.

Sequence analysis

Protein and nucleotide sequence alignments were made by CLUSTAL X (Thompson et al. 1997) with manual adjustments.

MEGA3 (Kumar et al. 2004) was used for sequence alignment and evolutionary analyses. Evolutionary trees were reconstructed using the neighbor-joining method (Saitou and Nei 1987) based on Kimura's two-parameter distances, with 2000 bootstrap replications (Felsenstein 1985). We used the complete deletion option for all trees. Numbers of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions were computed by the modified Nei-Gojobori method (Zhang et al. 1998). Nucleotide diversity (π) and Watterson's θ were computed as described by (Tajima 1989). More specifically, π is the number of nucleotide differences per site between two randomly picked alleles within a population or species and θ is the number of polymorphic sites per site divided by

$$\sum_{i=1}^{n-1} \frac{1}{i} \quad ,$$

where n is the number of alleles sampled. The Hudson-Kreitman-Aguade (HKA) test (Hudson et al. 1987) was used to compare *csd* with neutral sequences. DnaSP (Rozas et al. 2003) was used for all population genetic analyses.

Acknowledgments

We thank Xiaoxia Wang for technical assistance and Wendy Grus and Peng Shi for valuable comments. We also thank the Honey Bee Genome Project at the Baylor Human Genome Sequencing Center for making the honey bee genome sequence available. This work was supported by grants from the Office of Vice President for Research of the University of Michigan (to J.Z.), the National Institutes of Health (to J.Z.), and the University of Kansas General Research Fund (to D.R.S.).

References

Adams, J., Rothman, E.D., Kerr, W.E., and Paulino, Z.L. 1977. Estimation of number of sex alleles and queen matings from diploid male frequencies in a population of *Apis mellifera*. *Genetics* **86:** 583–596.

Alexander, B. 1991. *A cladistic analysis of the genus* Apis. Westview Press, New Delhi, India.

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

Beye, M. 2004. The dice of fate: The *csd* gene and how its allelic composition regulates sexual development in the honey bee, *Apis mellifera*. *Bioessays* **26:** 1131–1139.

Beye, M., Hunt, G., Page, R., Fondrk, M., Grohmann, L., and Moritz, R. 1999. Unusually high recombination rate detected in the sex locus region of the honey bee (*Apis mellifera*). *Genetics* **153:** 1701–1708.

Beye, M., Hasselmann, M., Fondrk, M., Page, R., and Omholt, S. 2003. The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* **114:** 419–429.

Bull, J. 1983. *The evolution of sex-determination mechanisms*. Benjamin/Cummings, Menlo Park, CA.

Charlesworth, D. 2004. Sex determination: Balancing selection in the honey bee. *Curr. Biol.* **14:** R568–R569.

Charlesworth, D. and Awadalla, P. 1998. Flowering plant self-incompatibility: The molecular population genetics of *Brassica* S-loci. *Heredity* **81:** 1–9.

Cook, J. 1993. Sex determination in the Hymenoptera—A review of models and evidence. *Heredity* **71:** 421–435.

Crozier, R. 1971. Heterozygosity and sex determination in haplodiploidy. *Am. Nat.* **105:** 399–412.

Dwyer, K.G., Balent, M.A., Nasrallah, J.B., and Nasrallah, M.E. 1991. DNA-sequences of self-incompatibility genes from *Brassica campestris* and *Brassica oleracea*—Polymorphism predating speciation. *Plant Mol. Biol.* **16:** 481–486.

Dzierzon, J. 1845. Gutachten ueber die von Herrn Direktor Stoehr in ersten und zweiten Kapitel des General-Gutachtens aufgestellten Fragen. *Bienenzeitung* **1:** 109–113, 119–121.

Felsenstein, J. 1985. Confidence-limits on phylogenies—An approach using the bootstrap. *Evolution Int. J. Org. Evolution* **39:** 783–791.

Fondon, J.W. and Garner, H.R. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci.* **101:** 18058–18063.

Garnery, L., Vautrin, D., Cornuet, J.M., and Solignac, M. 1991. Phylogenetic-relationships in the genus *Apis* inferred from mitochondrial-DNA sequence data. *Apidologie (Celle)* **22:** 87–92.

Hasselmann, M. and Beye, M. 2004. Signatures of selection among sex-determining alleles of the honey bee. *Proc. Natl. Acad. Sci.* **101:** 4888–4893.

Hudson, R.R., Kreitman, M., and Aguade, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153–159.

Hughes, A.L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature* **335:** 167–170.

Ioerger, T.R., Clark, A.G., and Kao, T.H. 1990. Polymorphism at the self-incompatibility locus in *solanaceae* predates speciation. *Proc. Natl. Acad. Sci.* **87:** 9732–9735.

Klein, J. 1987. Origin of major histocompatibility complex polymorphism—The trans-species hypothesis. *Hum. Immunol.* **19:** 155–162.

Klein, J., Sato, A., Nagl, S., and O'Huigin, C. 1998. Molecular trans-species polymorphism. *Annu. Rev. Ecol. Syst.* **29:** 1–21.

Koeniger, N. and Koeniger, G. 2000. Reproductive isolation among species of the genus *Apis*. *Apidologie (Celle)* **31:** 313–339.

Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5:** 150–163.

Mackensen, O. 1951. Viability and sex determination in the honey bee (*Apis mellifera* L.). *Genetics* **36:** 500–509.

Marin, I. and Baker, B.S. 1998. The evolutionary dynamics of sex determination. *Science* **281:** 1990–1994.

May, G., Shaw, F., Badrane, H., and Vekemans, X. 1999. The signature of balancing selection: Fungal mating compatibility gene evolution. *Proc. Natl. Acad. Sci.* **96:** 9172–9177.

Maynard-Smith, J. and Haigh, J. 1974. Hitch-hiking effect of a favorable gene. *Genet. Res.* **23:** 23–35.

Moriyama, E.N. and Powell, J.R. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13:** 261–277.

Nachtsheim, H. 1916. Zytologische Studien ueber die Geschlechtsbestimmung bei der Honigbiene (*Apis mellifera*). *Arch. Zellforsch.* **11:** 169–241.

Nagl, S., Tichy, H., Mayer, W.E., Takahata, N., and Klein, J. 1998. Persistence of neutral polymorphisms in Lake Victoria cichlid fish. *Proc. Natl. Acad. Sci.* **95:** 14238–14243.

Richman, A.D., Uyenoyama, M.K., and Kohn, J.R. 1996. Allelic diversity and gene genealogy at the self-incompatibility locus in the *solanaceae*. *Science* **273:** 1212–1216.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19:** 2496–2497.

Ruttner, F. 1987. *Biogeography and taxonomy of honeybees*. Springer-Verlag, New York.

Saitou, N. and Nei, M. 1987. The neighbor-joining method—A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

Sheppard, W.S. and Berlocher, S.H. 1989. Allozyme variation and differentiation among 4 *Apis* species. *Apidologie (Celle)* **20:** 419–431.

Tajima, F. 1989. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.

Takahata, N. and Nei, M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124:** 967–978.

Takahata, N., Satta, Y., and Klein, J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* **130:** 925–938.

Tamura, K., Subramanian, S., and Kumar, S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21:** 36–44.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

Whiting, P. 1943. Multiple alleles in complementary sex determination of *Habrobracon*. *Genetics* **28:** 365–382.

Woyke, J. 1963. What happens to diploid drone larvae in a honeybee colony. *J. Apic. Res.* **2:** 73–76.
Woyke, J. 1986. Sex determination. In *Bee genetics and breeding* (ed. T. Rinderer), pp. 91–119. Academic Press, Orlando, FL.
Wu, J., Saupe, S.J., and Glass, N.L. 1998. Evidence for balancing selection operating at the het-c heterokaryon incompatibility locus in a group of filamentous fungi. *Proc. Natl. Acad. Sci.* **95:** 12398–12403.
Yokoyama, S. and Nei, M. 1979. Population dynamics of sex-determining alleles in honey bees and self-incompatibility alleles in plants. *Genetics* **91:** 609–626.
Zhang, J., Rosenberg, H.F., and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci.* **95:** 3708–3713.

Received September 23, 2005; accepted in revised form December 7, 2005.