Evidence for Convergent Nucleotide Evolution and High Allelic Turnover Rates at the *complementary sex determiner* Gene of Western and Asian Honeybees

Martin Hasselmann, Xavier Vekemans,† Jochen Pflugfelder,‡ Nikolaus Koeniger,‡ Gudrun Koeniger,‡ Salim Tingek,§ and Martin Beye**

*Heinrich-Heine Universitaet Duesseldorf, Institut fuer Genetik, Universitaetsstr, Duesseldorf, Germany; †Laboratoire de Génétique et Evolution des Populations Végétales, Unité Mixte de Recherche Centre National de Recherche Scientifique 8016, Université de Lille, Villeneuve d'Ascq, France; ‡Institut für Bienenkunde, Johann-Wolfgang-Goethe Universität Frankfurt/M, Oberursel, Germany; and §Agricultural Research Station Tenom, Tenom, Sabah, Malaysia

Our understanding of the impact of recombination, mutation, genetic drift, and selection on the evolution of a single gene is still limited. Here we investigate the impact of all these evolutionary forces at the *complementary sex determiner* (*csd*) gene that evolves under a balancing mode of selection. Females are heterozygous at the *csd* gene and males are hemizygous; diploid males are lethal and occur when *csd* is homozygous. Rare alleles thus have a selective advantage, are seldom lost by the effect of genetic drift, and are maintained over extended periods of time when compared with neutral polymorphisms. Here, we report on the analysis of 17, 19, and 15 *csd* alleles of *Apis cerana*, *Apis dorsata*, and *Apis mellifera* honeybees, respectively. We observed great heterogeneity of synonymous (πS) and nonsynonymous (πN) polymorphisms across the gene, with a consistent peak in exons 6 and 7. We propose that exons 6 and 7 encode the potential specifying domain (*csd*-PSD) that has accumulated elevated nucleotide polymorphisms over time by balancing selection. We observed no direct evidence that balancing selection favors the accumulation of nonsynonymous changes at *csd*-PSD (πN/πS ratios are all <1, ranging from 0.6 to 0.95). We observed an excess of shared nonsynonymous changes, which suggest that strong evolutionary constraints are operating at *csd*-PSD resulting in the independent accumulation of the same nonsynonymous changes in different alleles across species (convergent evolution). Analysis of *csd*-PSD genealogy revealed relatively short average coalescence times (~6 Myr), low average synonymous nucleotide diversity ($\pi S < 0.09$), and a lack of trans-specific alleles that substantially contrasts with previously analyzed loci under strong balancing selection. We excluded the possibility of a burst of diversification after population bottlenecking and intragenic recombination as explanatory factors, leaving high turnover rates as the explanation for this observation. By comparing observed allele richness and average coalescence times with a simplified model of csd-coalescence, we found that small long-term population sizes (i.e., $N_e < 10^4$), but not high mutation rates, can explain short maintenance times, implicating a strong historical impact of genetic drift on the molecular evolution of highly social honeybees.

Introduction

The *complementary sex determiner* (*csd*) gene offers an informative example to study the molecular evolution of a single gene evolving under a well-documented mode of selection (balancing selection) that maintains multiple functional alleles (Beye et al. 2003; Hasselmann and Beye 2004). The *csd* determines the sexual fate of the honeybee (*Apis mellifera*), with the combination of 2 *csd* alleles in the zygote resulting in a female and 1 unique allele resulting in a male (Beye 2004). The *csd* gene segregates in more than 15 allelic forms in populations (Adams et al. 1977; Hasselmann and Beye 2004). Bees with a heterozygous *csd* combination are females and those with a hemizygous copy (haploid, unfertilized eggs) or with a homozygous combination are males. Diploid males, however, are eaten in the larval stage by worker bees. Fertile males are produced from haploid, unfertilized eggs. The strong selection against homozygotes at *csd* results in a frequency-dependent selection regime. Rare or newly evolved alleles increase in frequency in a population as they will rarely combine to form homozygotes and initiate diploid male development. Very frequent alleles, in contrast, will often form homozygotes. These diploid males are lethal and do not contribute alleles to the next generation, which results in a decrease in the frequency of common alleles. As a general consequence of the rare allele advantage or negative frequency dependence (balancing selection), these alleles are seldom lost by the effect of genetic drift and have a much longer persistence times in populations than neutral alleles (Takahata 1990). Consequently, we expect that regions that are under balancing selection have much longer persistence time and have accumulated substantially more nucleotide differences than regions that are unlinked to these sites. Our previous analyses have shown that *csd* of *A. mellifera* is evolving under balancing selection and that several regions of the sequence are possible targets of selection (Hasselmann and Beye 2004). In addition to finding that recombination has operated among some *csd* alleles, we have shown as a further sign of balancing selection that the gene has accumulated 10–13 times more variation than the genome-wide average (Hasselmann and Beye 2006).

The prolonged persistence time and the accumulation of substantial nucleotide differences within the *csd* gene offer the possibility to study 1) long-term population processes by the coalescence process and 2) molecular evolution of a single gene under a well-documented mode of selection. The coalescence process of functional alleles with equal fitness evolving under strong balancing selection has been described analytically in models of overdominant selection acting on the major histocompatibility complex (MHC) (Adams et al. 1977; Takahata 1990) and negative frequency-dependent selection acting on the self-incompatibility *S* locus of plants (Vekemans and Slatkin 1994; Uyenoyama 2003). The persistence time of an allele depends on the strength of selection, the rate of origination (mutation to novel

Key words: sex determination, balancing selection, genetic drift, social insects, convergent adaptive evolution, molecular evolution, nucleotide polymorphism.

E-mail: martin.hasselmann@uni-duesseldorf.de.

Mol. Biol. Evol. 25(4):696–708. 2008
doi:10.1093/molbev/msn011
Advance Access publication January 12, 2008

alleles), and the loss of alleles by genetic drift, which is a function of population size (Wright 1939, 1960; Yokoyama and Nei 1979). *S* alleles and *MHC* alleles have long overall coalescence times and have been maintained for more than 40 Myr (Ioerger et al. 1990; Uyenoyama 1995) and 30 Myr (Takahata 1993), respectively, as suggested by the observation of trans-specific polymorphisms (some alleles are more closely related to an allele from another species than to other alleles from the same species). Differences in nucleotide divergence among *S* alleles have been used to infer relative coalescence times and population history (Takahata 1990; Richman et al. 1996; Richman and Kohn 1999) that reflect long-term effective population size. In small populations or in those having experienced a bottleneck, the overall coalescence times are expected to be shorter and the average sequence divergence lower because alleles are more likely to be lost by genetic drift.

In this study, we analyze *csd* gene sequences of 3 related honeybee species, *Apis cerana*, *Apis dorsata* (both Asian honeybee species), and *A. mellifera* (the western honeybee) that diverged over the last 10 Myr (Sheppard and Berlocher 1989; Garnery et al. 1992, see Materials and Methods). These species share very similar life histories, have a highly social organization, and show substantial division of labor and reproduction (i.e., they are eusocial).

A survey of nucleotide polymorphisms of genomic fragments supports the notion that balancing selection is also operating among *csd* of other honeybee species: *A. cerana* and *A. dorsata* (Cho et al. 2006). However, the latter study relied on fragmented open reading frames (ORFs) in which the relation of ORFs to single alleles was unknown and furthermore included only a very limited set of alleles (i.e., ~8 alleles for *A. dorsata* of the most variable genomic fragment, excluding neutral variants with low divergence).

Here we analyze for the first time 1) the full-coding sequence of *csd* from 3 related *Apis* species (*A. cerana*, *A. dorsata*, and *A. mellifera*) and 2) a significant number of *csd* alleles (>14). These comprehensive sequence data provide us with the unique power to study the molecular evolution of nucleotide polymorphisms and the coalescence process across different species. Thanks to consistent patterns of nonsynonymous and synonymous polymorphisms across species, we could narrow down the target of balancing selection. This is of importance as the molecular nature of the *csd* specifying domain is still unknown. By further exploring the distribution of synonymous and nonsynonymous changes, we shed light onto the evolutionary forces that have shaped these specificities. By comparing our empirical results with expectations from an analytic model of the coalescence process, we obtained insights into the long-term population history of these highly social bees.

Materials and Methods

Each of the 3 species *A. mellifera*, *A. cerana*, and *A. dorsata* were collected in 2 sampling locations. Geographic locations for *A. cerana* and *A. dorsata* were Tenom (Borneo) and Thailand (Samut Songkram for *A. cerana* and Wanmanaow for *A. dorsata*) and for *A. mellifera* were Litija (Slovenia) and Pretoria (South Africa). In each location, 150–300 embryos were collected from 2 colonies. Because of multiple mating of queens, these eggs have 16–28 different sources of chromosomes derived from as many different fathers.

csd sequences were identified from cDNA as described previously (Hasselmann and Beye 2004). For *A. cerana* and *A. dorsata*, a new set of primers were designed based on sequence information that was obtained from 10 different 5′ and 11 different 3′ rapid amplification of cDNA ends (RACE) sequences. Templates for RACE experiments (First choice RACE kit: Ambion, Austin, TX) were derived from different geographic locations. The following primers were developed: csd_rev4CIII: TCTCATTATTCAATAC-GTTGGCATC, csd_forCer2: CTCTAAGCGTGGATTA-CAGGTT, csd_IIIfor: GTTAAATTTCATWRATATACA TATAC, csd_IIIrev3: ATTCAGTTCATTATTCATTATT-TGCA, cons2csd_for: TCATAAAAATGAAACGAAATA TATC, and the ones used previously (Hasselmann and Beye 2004). Primer combinations were as follows: csd_forCer2/ csd_rev4CIII/; csd_IIIfor/csd_IIIrev3, and the combinations as described previously (Hasselmann and Beye 2004). The *csd* sequence variants (and *fem* sequences) were identified from 60 to 100 cloned fragments for each primer pair and sampled by restriction fragment length analysis (*Apo* I, *Mbo* I, and *Taq* I [Fermentas]). The quality of the primer sets was tested on a sample of 15 haploid drones. The *csd* alleles of all drones were successfully amplified by polymerase chain reaction (PCR). The PCR-induced mutations are lower than 2 in 1,000 nt (sequence data of identical sequences obtained from different high-fidelity PCR amplifications), indicating that PCR-based amplifications do not influence the conclusions of potential specifying domain (*csd*-PSD) allele analyses that have nucleotide of diversity of ~5% between alleles.

The ORFs of 89 *csd* sequences were aligned and edited as described elsewhere (Hasselmann and Beye 2004). Sequence variants that we classified as neutral variants of a single specificity/allele were excluded from analysis on the basis of clusters of low divergence that form bush-like structures in the genealogy (Hasselmann and Beye 2004). These clusters were identified by π estimates and statistical tests (*Z*-test) on full-coding sequences (Hasselmann and Beye 2004). Only 1 sequence for each of these neutral sequence clusters was included in the analysis. This resulted in a final data set of 51 *csd* alleles ($n = 15$ for *A. mellifera*, $n = 17$ for *A. cerana*, and $n = 19$ for *A. dorsata*). Our classification is supported by the finding that alleles/specificities differ in the structure of the hypervariable domain, whereas the sequences classified as neutral variants have the same repeat structure. Nucleotide polymorphism parameters were calculated using MEGA version 3.1 program (Kumar et al. 2004). Trees and genealogies of sequences were constructed using the minimum evolution (ME) method and Kimura's 2-parameter distances. The number of shared polymorphisms and fixed differences between species pairs was calculated by using the Sites program (Hey and Wakeley 1997). Linear regression and correlation analyses were performed with the SPSS program (version 12.0). The McDonald–Kreitman test was performed using DnaSP version 4.0 (Rozas J and Rozas R 1999).

An estimate of the genomic mutation rate (7×10^{-9} per site per year) is derived from the average pairwise synonymous divergence per site (dS = 0.1) of *fem* and *elongation factor-alpha 1* (*EF-*α *1*) sequences of *A. mellifera* and *A. cerana* and estimates of their divergence times (~7 Myr) (Sheppard and Berlocher 1989; Garnery et al. 1992) according to the following relationship: number of polymorphisms at synonymous sites that accumulate $=2g\mu$, where g is the number of generations, assuming 1 generation per year for the honeybee, and μ is the mutation rate. From the mutation rate and average synonymous divergence between *A. mellifera* and *A. dorsata* for *fem* and *EF-*α *1* sequences (dS = 0.135), a rough estimate of their divergence is obtained (10 Myr).

We derive expectations for the coalescence times of functional alleles at *csd* following Yokoyama and Nei's (1979) model which assumes that sex-determining alleles in honeybees are selectively equivalent and are evolving under overdominant selection with lethal homozygotes. The derivation assumes that new functional alleles are continuously generated under an infinite-allele model of mutation. It also assumes a finite population of $N_{\rm f}$ and $N_{\rm m}$ reproductive females and males, respectively, where females are always heterozygotes at *csd* and males are strictly haploids. Using diffusion approximations, Yokoyama and Nei (1979) showed that the homozygosity F at the sex-determining locus is given by

$$F=\sqrt{\frac{-3}{4N}\ln(u\sqrt{12\pi.N})}, \tag{1}$$

where u is the mutation rate to new alleles per generation and N is the effective population size computed as $N = 9 N_{\rm m} N_{\rm f}/(4 N_{\rm m} + 2 N_{\rm f})$. The effective number of alleles maintained within a population, $*n$, can then be computed as $1/F$. Takahata (1990) showed that the coalescence process among functionally distinct alleles at a locus subject to strong overdominant selection is similar to a neutral coalescent but with a timescale expanded by a factor $f_{\rm s}$, with $f_{\rm s}$ determined by the parameters of the selection model. Takahata's scaling factor $f_{\rm s}$ can be computed, according to Uyenoyama (2003) as

$$f_{\rm s}=\frac{*n^2}{4N\lambda}, \tag{2}$$

where λ is the turnover rate of alleles that can be computed based on the formula given by Uyenoyama (2003):

$$\lambda=2Nu(a-2b), \tag{3}$$

where $a=\frac{2}{3(1-F)}$ and $b=\frac{1-2F}{3(1-F)}+u$ (Yokoyama and Nei 1979). From (2) and (3) we obtain

$$f_{\rm s}=\frac{3(1-F)}{32N^2uF^3}. \tag{4}$$

Finally, the pairwise coalescence time of alleles ($\rm T_d$) at *csd* is given by (Takahata 1990)

$$T_{\rm d}=2.N.f_s., \tag{5}$$

and replacing with (4) we finally get

$$T_{\rm d}=\frac{3(1-\mathrm{F})^2}{16NuF^3}. \tag{6}$$

We computed expected values of $*n$ and $T_{\rm d}$ for a large range of parameter values u and N and compared them with the actual number of alleles observed in the 3 species and with the total coalescence times of alleles estimated based on the application of a molecular clock to the maximum synonymous nucleotide divergence among extant alleles.

Results

We used reverse transcription–polymerase chain reaction to survey allelic diversity of *A. mellifera* (western honeybee), *A. cerana* and *A. dorsata* (the 2 Asian honeybees) from distant geographical locations. We identified 17, 19, and 15 alleles for *A. cerana*, *A. dorsata*, and *A. mellifera*, respectively. The sequences obtained span the full ORF of *csd* sequences.

Despite substantial variation within and among species in the deduced amino acid sequences, all sequences share the same domains: an arginine/serine-rich domain, a terminal proline-rich domain, and a hypervariable region in between (Beye et al. 2003). This observation is consistent with a homologous sex-determining function of *csd* in the Asian species.

Previous analysis had identified 2 major types (type I and type II [Hasselmann and Beye 2004]) of sex-determining alleles. We have confined our analysis to type I that corresponds to the actual *csd* gene and is the target of balancing selection. With the help of the honeybee genome project, we characterized type II (*feminizer*, *fem*) to be a paralog of *csd* with sex differentiation function downstream of *csd* (Hasselmann M, Gempe T, Schioett M, Nunes-Silva C, Otte M, Beye M, unpublished data); thus, it is neither the initial signal of sexual regulation nor the target of balancing selection. Consistent with this observation, type II sequences show a near-absence of intraspecies divergence and cluster into a single clade, suggesting that gene duplication occurred before the split of these species (supplementary fig. 1, Supplementary Material online). This finding is inconsistent with a previous report (Cho et al. 2006) of partial type I and type II genomic sequences (their fig. 2 and genomic region 1) that were considered to represent trans-specific alleles at a single locus. Even if we confine our analysis to the region that corresponds to the genomic fragment in which Cho et al. (2006) detected trans-specific polymorphism, we detected no trans-specific polymorphism (data not shown). An explanation of their finding is that they evidently have not isolated a type II (*fem*) female sequence from *A. cerana* (the presented type 2 sequence is not a type II sequence but probably a pseudogene or female intronic sequences that were amplified by the genomic DNA approach) resulting in a misinterpretation of the data set. In this study, we excluded paralog type II (*fem*) and confined our analysis to *csd* (former type I) sequences (Hasselmann and Beye 2004, 2006). Using cDNA sequences, we were able to exclude sequence information of pseudogenes.

Common Patterns of Heterogeneity of Nucleotide Diversity

Patterns of average pairwise nonsynonymous (πN) and synonymous (πS) differences per site across exons are reported for each species separately (fig. 1). A striking feature is wide heterogeneity of polymorphism across exons, with some displaying extremely high mean synonymous or nonsynonymous diversity. There is also substantial variation around the mean pairwise nucleotide differences, suggesting that similar and divergent alleles coexist within species representing different divergence times.

In *A. mellifera* and *A. dorsata*, the nonsynonymous differences (πN) are consistently the highest in exons 6 and 7 ($\pi N > 0.05$), whereas πN is lower for exons 1–5 ($\pi N < 0.05$). This suggests that at least exons 6 and 7 are targets of balancing selection that have extended maintenance times and thus have accumulated more nucleotide differences. Assuming exons 6 and 7 are the only targets of selection, the decline of πN among other exons could be due to the strength of genetic hitchhiking to the target sites of selection or to relaxed evolutionary constraints. Because sites that are genetically linked to the target of selection are also maintained over extended periods of time, they also accumulate more nucleotide differences. Recombination, however, decouples and relaxes the impact of balancing selection, leading to a decline of polymorphism at this locus (Hasselmann and Beye 2006). Considerable higher πN in exon 8 than in exons 2–5 can thus presumably be the result of stronger genetic linkage and smaller genomic distances of exon 8 to the target of selection exons 6 and 7 (the genetic distance of exon 8–exon 7 is 60 bp, whereas that between exons 5 and 6 is 3.9 kb). In *A. cerana*, exons 4 and 5 have the highest πN. However, this enhanced polymorphism results from divergence between 2 sequence clusters (O and T type), which are characterized by low intracluster differences (supplementary fig. 2, Supplementary Material online), suggesting that this part of the sequence is not a general target of balancing selection across species. Polymorphisms of the T type for exon 5 and part of exon 4 are more similar to *A. dorsata* polymorphisms (data not shown), suggesting that this restricted region harbors trans-specific polymorphisms.

The synonymous differences follow the heterogeneous patterns of nonsynonymous differences across exons, with maximum πS values in exons 6 and 7 in *A. mellifera* and *A. dorsata* ($\pi S > 0.05$). This observation is consistent with the prediction that synonymous differences accumulate in regions of tight genetic linkage to sites under balancing selection (Schierup et al. 2000). When *A. cerana* sequences are separated into 2 clusters, O and T (supplementary fig. 2, Supplementary Material online), intracluster πS values follow the pattern in the other species with low πS in exon 5 but high πS in exons 6 and 7 and that extends to exon 8.

Based on the comparative analysis of nonsynonymous and synonymous nucleotide polymorphisms across species, we propose that exons 6 and 7 are a common target of selection across species. Targets of selection in exons 6 and 7 are nonsynonymous sites that encode amino acids determining the specificity of an allele in the sex determination process. We suggest that exons 6 and 7 encode the *csd*-PSD.

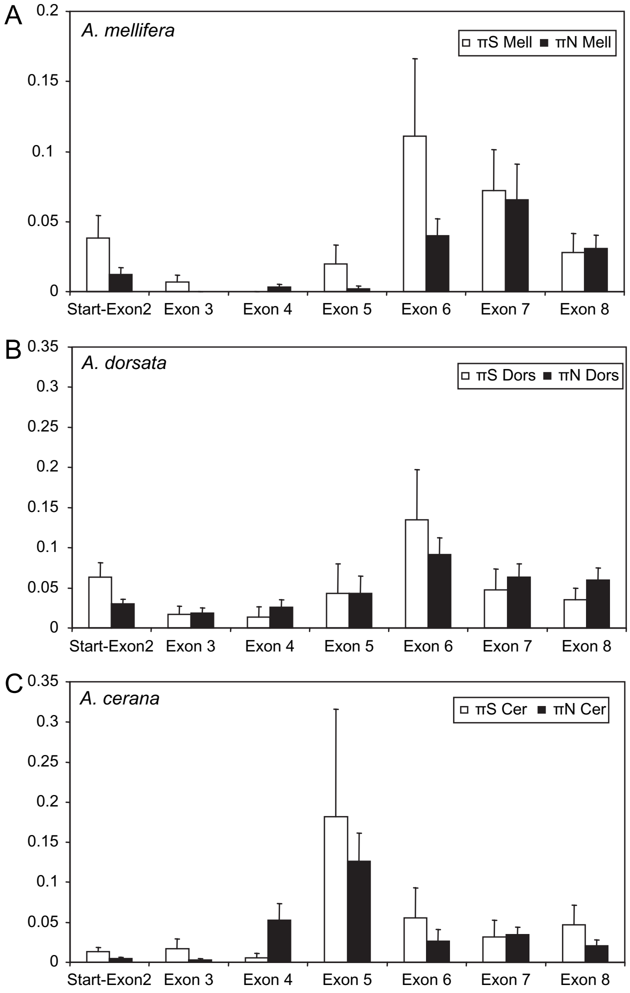

FIG. 1.—Average pairwise nonsynonymous (πN) and synonymous (πS) differences per site of exons for 3 honeybee species (*A*) *Apis mellifera*, (*B*) *Apis dorsata*, and (*C*) *Apis cerana*. Average nonsynonymous (πN) and synonymous (πS) differences per site per exon are presented with their standard errors. Sequence assignment to exons is based on *A. mellifera csd* gene structure (Beye et al. 2003). Exon 9, which encodes only 9 amino acids, was excluded from the analysis. Number of alleles included are 15, 19, and 17 for *A. mellifera*, *A. dorsata*, and *A. cerana*, respectively.

Short Coalescence Times in *csd* Genealogies

The relationship among *csd*-PSD sequences is reflected by their genealogy (fig. 2). The alleles fall into 3 clades, supported by bootstrap values >90%, such that each clade represents a single species (see also supplementary fig. 1 [Supplementary Material online] for the whole sequence). Hence, no trans-specific allelic pattern of the common target of selection, *csd*-PSD, is observed in these honeybee species that have been separated by maximally 10 Myr (see Materials and Methods). This contrasts remarkably with the general pattern observed in other loci under balancing selection such as the *S* locus and the *MHC*

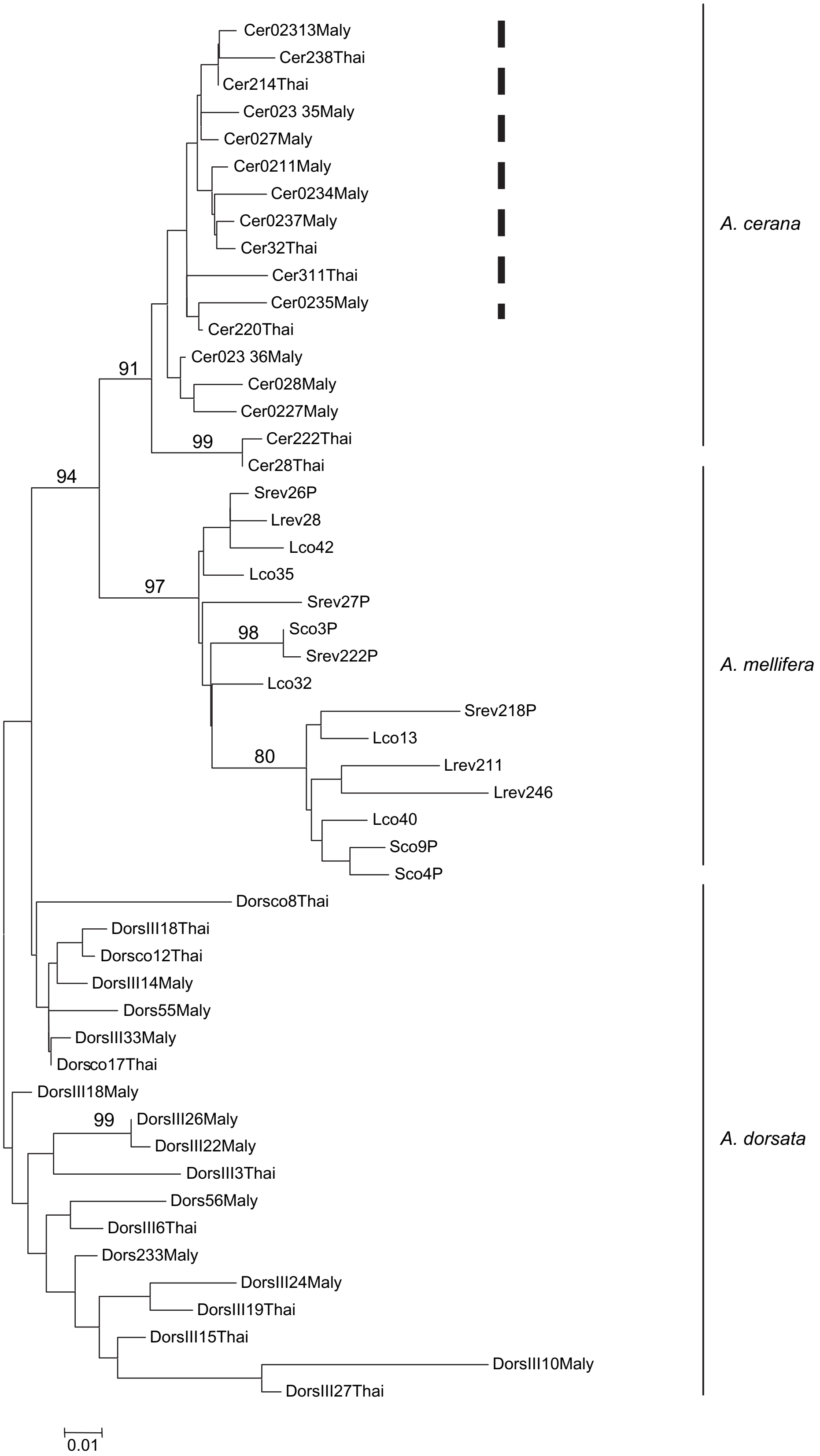

FIG. 2.—Genealogy of *csd*-PSD nucleotide sequences of 3 honeybee species based on the ME method of genetic differences (Kimura's 2-parameter distances). The alleles cluster into 3 clades representing the species *Apis mellifera*, *Apis dorsata*, and *Apis cerana*. The scale on the left represents nucleotide differences per site, whereas the stippled bar marks a burst of diversification among *A. cerana* alleles (for further information see text). Numbers are bootstrap values exceeding 80%.

complex in which several trans-specific alleles have been observed even after 30–40 Myr of species separation but is consistent with the relatively low average synonymous nucleotide diversity in each bee species ($\pi S = 0.089$ for *A. mellifera*, $\pi S = 0.032$ for *A. cerana*, and $\pi S = 0.069$ for *A. dorsata*). A lack of trans-specific lineages could be due 1) to bursts of diversification of *csd* alleles that occurred after the speciation events, 2) to the process of recombination in which trans-specific allelic patterns are lost, or 3) to high allelic turnover rates leading to short coalescence times among *csd* alleles.

1) A burst of diversification of alleles after each speciation event is not a general explanation for the lack of trans-specific alleles at *csd*. Such diversification of *csd* alleles over short evolutionary time (e.g., after a bottleneck) would be consistent with a star-like structure in the genealogy. However, this is only observed in parts of the *A. cerana* genealogy (alleles marked by a stippled bar in fig. 2) in which the genealogy has substantially reduced internal branches when compared with other internal branches of the same species (Mann–Whitney *U* test. $P < 0.01$) or other species ($P < 0.002$). Another argument against a burst of diversification is that the observed nucleotide diversity at the *csd*-PSD is largely compatible with the expected coalescence process at equilibrium. Under a conservative assumption, all sex determination alleles have equal fitness and equal probability of extinction because alleles with lower fitness will have a higher probability of extinction. Thus, the expected genealogical relationships will be similar to sampled neutral gene genealogies (Takahata 1990). A way to test this is to compare ratios of the largest with mean number of synonymous and nonsynonymous differences that is close to $2(1-1/i)$ for i sampled alleles (Takahata et al. 1992). The ratios for nonsynonymous differences are close to the expected values for *A. mellifera* and *A. cerana* (table 1) and slightly larger for *A. dorsata*. We detected, however, no significant differences between the observed and expected ratios of the largest and the mean nucleotide differences (Fisher's exact test, $P > 0.05$ for all comparisons applied to the absolute numbers of synonymous and nonsynonymous differences).

2) The process of intragenic recombination at *csd*-PSD within each species could reassemble ancestral polymorphisms in independent combinations resulting in a loss of trans-specific alleles. Such a process involving recombination and genetic drift will slow down the accumulation of synonymous differences because synonymous polymorphisms but not nonsynonymous polymorphisms maintained by balancing selection would become sensitive to the process of genetic drift. The plot of mean synonymous and nonsynonymous differences for each species (supplementary fig. 3A–C, Supplementary Material online) shows that the 2 types of polymorphism are highly correlated; the more synonymous differences that have accumulated, the older an allele, the more nonsynonymous differences that have accumulated over time. This is consistent with a tight linkage of synonymous to nonsynonymous sites (Spearman rank correlation, 2-tailed, *A. dorsata*: rho = 0.55; $P < 0.001$, *A. cerana*: rho = 0.45; $P < 0.001$, and *A. mellifera*: rho = 0.27; $P < 0.01$). To investigate whether some alleles deviate from this general pattern of tight linkage of synonymous (S) and nonsynonymous (N) polymorphisms given the mean S/N ratios (0.32, 0.29, and 0.32 for *A. mellifera*, *A. cerana*, and *A. dorsata*, respectively), we assumed a binomial distribution from which the 95% confidence interval (CI) was set. The distribution is so broad as to be largely compatible with the data. Some allelic pairs, however, exhibit unusual values of S and N. This approach identifies 6 of 171 *A. dorsata* and 2 of 120 of *A. mellifera* and none of 153 *A. cerana* allele pairs that have less synonymous changes than expected.

Table 1
The Pairwise Largest and Mean Synonymous (S) and Nonsynonymous (N) Differences of *csd*-PSD

			S	N
Apis mellifera	Sample size (*i*)	15		
	Mean		3.9 ± 0.99	8.1 ± 1.9
	Largest		8.5 ± 2.6	15.5 ± 4.2
	Ratio		2.2	1.9
	Expected ratio:			
	2(1−1/*i*)		1.87	1.87
Apis dorsata	Sample size (*i*)	19		
	Mean		3.5 ± 0.97	116 ± 1.6
	Largest		9.2 ± 2.9	28 ± 5.8
	Ratio		2.6	2.4
	Expected ratio:			
	2(1−1/*i*)		1.89	1.89
Apis cerana	Sample size (*i*)	17		
	Mean		1.9 ± 0.7	4.8 ± 1.2
	Largest		5 ± 2.1	10 ± 3.2
	Ratio		2.6	2.1
	Expected ratio:			
	2(1−1/*i*)		1.88	1.88

NOTE.—Sex determination alleles have equal fitness and equal probability of extinction; thus, the expected genealogical relationships are similar to sampled neutral gene genealogies (Takahata 1990) in which the ratio of largest to mean number of synonymous and nonsynonymous differences is close to $2(1-1/i)$ for i sampled alleles (Takahata et al. 1992). Each sample consists of alleles (sample size i) that differ at least by 1 nonsynonymous difference. The ratio is defined as the number of the largest to the mean number of nonsynonymous and synonymous changes. Variation around the mean is presented as standard errors. The expected ratio of the largest to the mean under the coalescence model is given by $2(1-1/i)$ (Takahata et al. 1992). Actual and expected ratios are not significantly different ($P > 0.05$).

3) High allelic turnover rates leading to short coalescence times among *csd*-PSD alleles could also explain the lack of trans-specific lineages. We estimate coalescence times of alleles based on the application of a molecular clock to the maximum synonymous nucleotide divergence among alleles that have accumulated by mutation. The largest synonymous differences per corresponding site (dS) between alleles in *csd*-PSD are 0.2, 0.23, and 0.11 for *A. mellifera*, *A. dorsata*, and *A. cerana*, respectively. If the neutral mutation rate is about 7×10^{-9} per site per year (compared with 15×10^{-9} per site per year for *Drosophila*), as estimates from divergence of gene sequences and speciation events suggest (see Materials and Methods), it must have taken about 14.6 Myr or generations in *A. mellifera*, 16.7 Myr in

A. dorsata, and 7.9 Myr in *A. cerana* for these synonymous differences to accumulate (number of synonymous sites that accumulate $=2g\mu$ with g number of generations and 1 generation per year for the honeybee and μ the mutation rate). None of the *A. cerana* alleles are substantially older than the most recent species split between *A. mellifera* and *A. cerana* (estimated at about 6–8 Myr) as the largest allelic divergence ($d\mathrm{S} = 0.11$) is in the order of mean interspecies divergence ($d\mathrm{S} = 0.1$, see Materials and Methods). This is consistent with the absence of trans-specific alleles in this species. However, 4 and 6 allelic lineages of *A. mellifera* and *A. dorsata*, respectively, could have predated speciation as divergence of allele pairs exceeds mean interspecies divergence ($d\mathrm{S} > 0.14$). Thus, based on our rough computations, trans-specific patterns could have been observed among *A. dorsata* and *A. mellifera*. Assuming that the ancestral species possessed 20 alleles, the probability that the historical sample of 4 and 6 allelic lineages did not overlap, that is, that the 2 descendant species have no trans-specific alleles, is as high as $P > 0.2$ when we apply the hypergeometrical distribution.

To explore which factor is most likely responsible for the short coalescence times in bees and the observed low average synonymous nucleotide diversity, we applied a theoretical model of the coalescence process (see Material and Methods) to compare with our empirical data. The coalescence process generating allelic genealogies at *csd* can be approximated using Takahata's (1990) coalescent approach applied to the diffusion analysis of the sex-determining locus in the honeybee (Yokoyama and Nei 1979). Takahata's main result is that the scale of the coalescence process for alleles under balancing selection is determined not only by the effect of genetic drift that is determined by the effective population size N_e but also depends strongly on the rate of origin of new specificities, u. An upper bound of $u = 10^{-6}$ (per gene per year) can be computed based on the genomic mutation rate and the assumption that any change among the 167 nonsynonymous sites at *csd*-PSD would give rise to a new allelic specificity ($u = \mu \times d$ with u the rate of origin of new alleles per gene per year, μ genomic mutation rate, and d number of nonsynonymous sites within *csd*-PSD). This estimate is, however, unrealistically high because of functional constraints in the protein sequence (see below). The lower bound of $u = 10^{-8}$ is close to the mutation rate per site when a single-specific site change within *csd*-PSD is required for a new allele to originate ($u = \mu$; rate of origin of new alleles equals the mutation rate). Given these broad bounds of the rate of origin of new alleles, we computed the number of expected alleles ($*n$) and expected pairwise coalescence times of alleles (T_d) as a function of the long-term effective population size N_e (fig. 3). It can be seen that the observed number of alleles in honeybees ($n \approx 20$) enforces a strong constraint on the effective population size, with compatible N_e values ranging between 1,500 and 4,000 individuals assuming a rate of origin of new alleles of 10^{-8} to 10^{-7} (fig. 3*A*). In contrast, the inferred average pairwise coalescence time of alleles of about 6 Myr or generations (as estimated from the average synonymous divergence of *A. dorsata* and *A. mellifera* alleles) is compatible with a wider range of N_e values (250–10^6 individuals, fig. 3*B*) given a rate of origin of new alleles of 10^{-8} to 10^{-7}. Putting together theoretical expectations and empirical data, we conclude that the observed coalescence times are most likely explained by a moderate rate of origin of new alleles and a strikingly small long-term effective population size ($N_\mathrm{e} < 10^4$) for *A. mellifera* and *A. dorsata*. Hypothesized N_e is even lower for *A. cerana* harboring less nucleotide diversity. In the above calculations, we assumed a long-term generation time of 1 year for these social bees. This assumption is not critical for our upper bound estimate of long-term effective population sizes as generation times of more than 1 year will result in an even lower population size estimate.

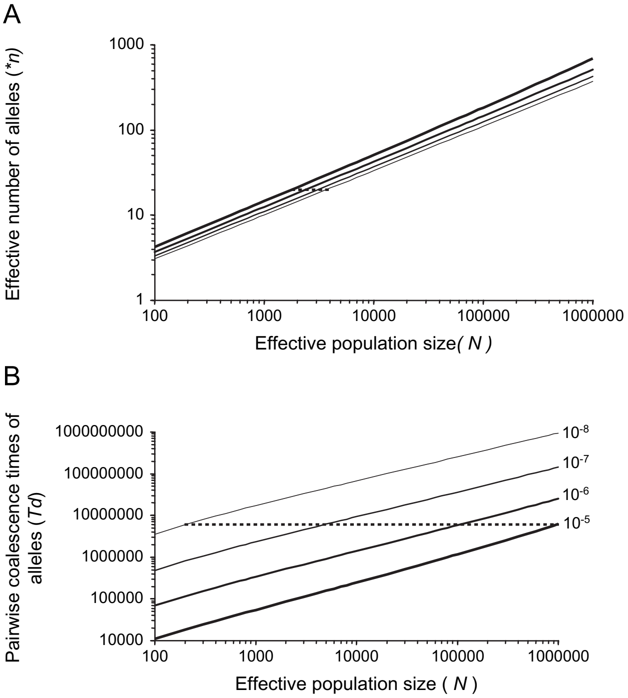

FIG. 3.—Expected values of (*A*) the effective number of alleles ($*n$) and (*B*) the average pairwise coalescence times of alleles (T_d) for a wide range of values of the parameters N (effective population size) and u (mutation rate to new functional alleles) under a model of a sex determination locus subject to overdominant selection with lethal homozygotes. Increasing values of u (from 10^{-8} to 10^{-5} mutations per generation) are represented by increasingly thicker solid lines. The broken horizontal lines indicate the range of N values that are consistent with the values of $*n$ and T_d observed in the present study.

Excess of Shared Nonysynonymous but Not Synonymous Changes

The *csd*-PSD has on average less nonsynonymous than synonymous polymorphisms per nonsynonymous and synonymous site ($\pi\mathrm{N}/\pi\mathrm{S}$: 0.6, 0.95, and 0.73 for *A. mellifera*, *A. dorsata*, and *A. cerana*, respectively). The $\pi\mathrm{N}/\pi\mathrm{S} > 1$ would constitute direct evidence that balancing selection favors the accumulation of nonsynonymous changes and thus the origin of new allelic variants. To test for adaptive nonsynonymous changes, we first applied the

Table 2
Numbers and Shared Numbers of Synonymous (S) and Nonsynonymous (N) Polymorphisms

	S	N
Apis cerana	7	22
Apis mellifera	16	27
Shared between species	1	5
P	0.4	0.004
Expected number of shared		1.2
Apis dorsata	16	59
A. mellifera	16	27
Shared between species	4	12
P	0.08	9×10^{-6}
Expected number of shared		3
A. dorsata	16	59
A. cerana	7	22
Shared between species	2	11
P	0.17	6×10^{-6}
Expected number of shared		2.6

NOTE.—The hypergeometrical distribution was used to obtain the probabilities of observing these shared numbers. We assume 129 synonymous possible changes for 43 synonymous sites and 501 nonsynonymous possible changes for 167 nonsynonymous sites.

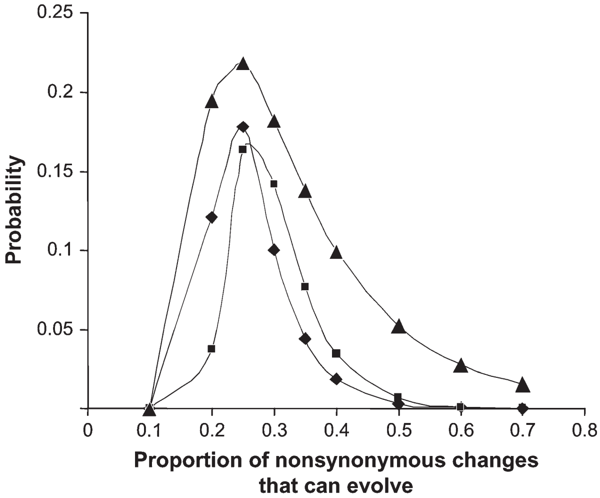

FIG. 4.—The probabilities (*P*) of observing the detected number of shared nonsynonymous changes (see table 2) given the different proportions of nonsynonymous changes that can evolve among *csd*-PSD alleles. Probability estimates were obtained by applying the hypergeometrical distribution to the data of table 2. The probability distribution of the species pair *Apis mellifera*/*Apis cerana* is shown by triangles, *A. mellifera*/*Apis dorsata* by squares, and *A. dorsata*/*A. cerana* by diamonds. We approximate the total number of possible nonsynonymous changes to 501 based on the numbers of 167 nonsynonymous sites.

McDonald–Kreitman test on *csd*-PSD alleles. The test detected no significant excess of nonsynonymous substitutions or polymorphisms in all species comparisons ($P > 0.3$). Notably, we identified a substantial excess of shared nonsynonymous but not synonymous changes at *csd*-PSD (table 2). These changes are shared as they have at least 2 of the same nucleotides in common between alleles across species. These nonsynonymous changes, however, have no common evolutionary origin. They accumulated independently in different alleles across species (homoplasy) because we found no trans-specific alleles in the phylogenetic analysis (fig. 2). Furthermore, under trans-specific evolution, we would expect to observe an excess of shared synonymous sites as well, as these sites are tightly linked to nonsynonymous sites (the impact of recombination is negligibly low). As an explanation for the substantial excess of shared nonsynonymous changes, we propose that only a limited number of nonsynonymous mutations evolved and accumulated across species. These evolutionary constraints of the CSD protein may be the result of purifying selection that removes deleterious mutations. Alternatively, evolutionary constraints may be a consequence of selection favoring new allelic variants that are restricted to only a subset of nonsynonymous changes. We next asked what portion of all possible nonsynonymous changes can accumulate and evolve in a way that supports the observed pattern of homoplasy when the hypergeometrical distribution is applied (fig. 4). In this simplified model, we assume that selection operates equally across species. The highest probability of obtaining these shared numbers is found when ~25% of all possible changes can evolve (fig. 4). This estimate is very robust over all species pairs (similar probability distributions), suggesting that comparable numbers of nonsynonymous changes are evolutionary constrained across species. When we correct the πN/πS ratios for the portion of changes that can possibly evolve, the 25% estimate we calculated from the number of homoplasic changes across species, $\pi N_{sp}/\pi S$ becomes 2.4, 3.8, and 2.9 ($\pi N_{sp} = 4\pi N$) for *A. mellifera*, *A. dorsata*, and *A. cerana*, respectively. These estimates suggest that a substantial fraction of nonsynonymous changes have been favored by balancing selection.

We further analyzed the relation of πN and πS and asked whether ratios of πN/πS change with different values of πS. The πS accumulates over evolutionary time; thus, low πS indicates alleles that have newly diverged from each other (newly diverged allele pairs), whereas high πS points to alleles that have diverged a long time ago (anciently diverged allele pairs). Figure 5 shows the scatter blot of πS versus πN of allele pairs and the linear regressions (fig. 5). The πN/πS ratio is in all 3 species not constant for low and high πS (newly and anciently diverged allele pairs) as shown by the regression lines. For newly diverged alleles, the regression line is above the $\pi N/\pi S = 1$ ratio (represented by the dotted line in fig. 5), whereas for anciently diverged alleles, the regression line drops below $\pi N/\pi S = 1$. This higher evolutionary rate of newly diverged alleles could indicate balancing selection enhancing the accumulation of nonsynonymous changes, or alternatively, could suggest a downward bias of πN/πS estimate due to πS signal saturation. The πS values <0.2, however, are far from signal saturation, suggesting that selection for new allelic variants played a role in the accelerated evolutionary rate among newly diverged alleles.

The Hypervariable Domain Encodes 2 Different Asparagine/Tyrosine-Rich Motifs

The hypervariable region consists of a variable number of repeats and is located within the *csd*-PSD (fig. 6). Because of the ambiguous relation of these sites, they were

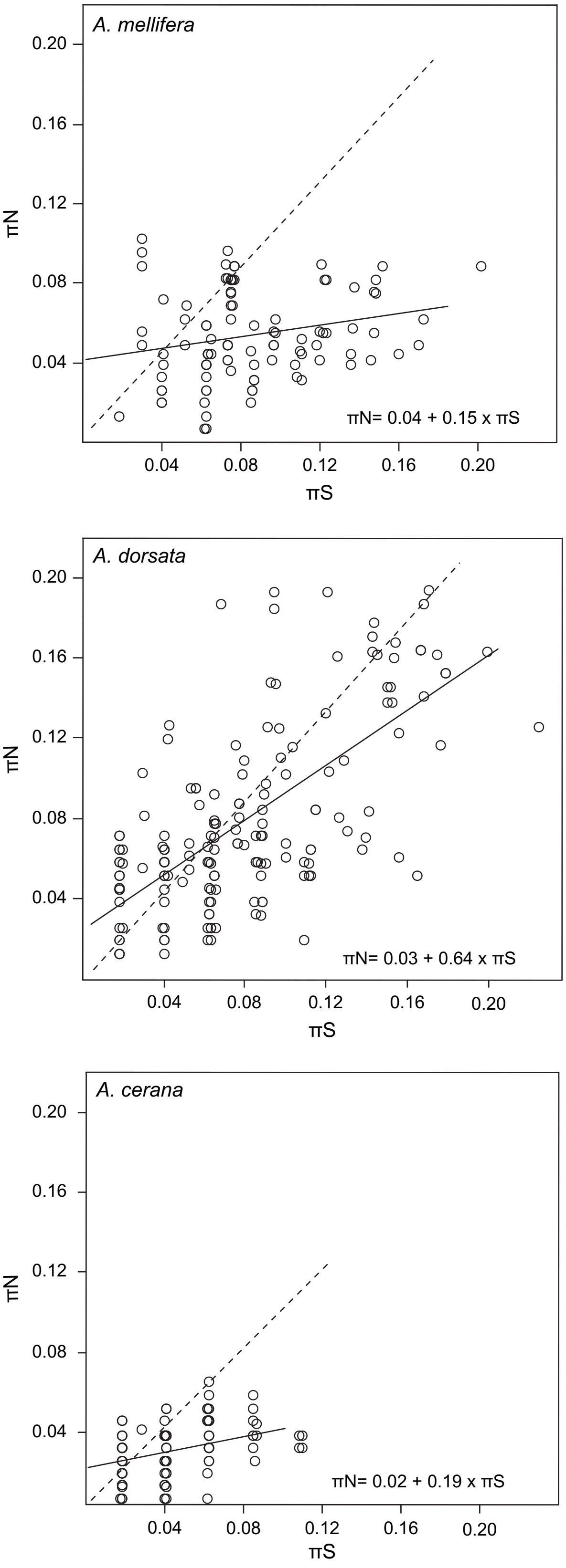

Fig. 5.—Scatter plots and linear regression analysis of nonsynonymous (πN) versus synonymous (πS) differences per site for all pairwise comparisons of (*A*) *Apis mellifera* ($P < 0.05$, $R^2 = 0.06$), (*B*) *Apis dorsata* ($P < 0.001$, $R^2 = 0.46$), and (*C*) *Apis cerana* ($P < 0.001$, $R^2 = 0.18$). The dashed line shows the 1:1 ratio of synonymous to nonsynonymous changes per site.

not included in the former polymorphism analyses. The repeat encodes a basic ($(N)_{1–5}$/Y) sequence motif of asparagine/tyrosine residues. This basic sequence motif is modified at the end of each reiteration by either a set of 2 amino acids (KK less often with KP, KQ) or with the more complex ($(KHYN)_{1–4}$)KH motif. This combination is reiterated 3–7 times, encoding peptides of varying length. The former repeat ending is observed in all 3 species (with only 1 representative in *A. dorsata*), whereas the latter 1 is confined to *A. cerana* and *A. dorsata*. We propose that the ($(KHYN)_{1–4}$)KH motif has most likely been lost in the *A. mellifera* lineage as it was obviously present in the most recent common ancestor of *A. cerana* and *A. mellifera*. The observation that the numbers of repeats and motifs vary between alleles suggest functional significance of this domain in specifying sex determination alleles.

Discussion

The *csd*-PSD Is a General Target of Balancing Selection

The heterogeneity among exons in substitution rates suggests that the combined evolutionary forces of balancing selection, mutation, recombination, and genetic drift have shaped the pattern of synonymous and nonsynonymous polymorphisms at *csd* in the 3 bee species *A. mellifera*, *A. cerana*, and *A. dorsata* (fig. 1). Balancing selection maintains polymorphisms at nonsynonymous sites over extended periods, providing time for synonymous mutations to accumulate as well. Recombination among exons decouples this process and, with loose linkage, these changes are more likely to become lost by the effect of genetic drift. Thus, in regions immediately surrounding the targets of balancing selection, high levels of synonymous polymorphisms are expected (Hudson and Kaplan 1988). The consistently high levels of synonymous differences across species in exons 6 and 7, therefore, serve to narrow down the target of balancing selection to *csd*-PSD. Nonsynonymous differences evolve with both direct and indirect effects of selection (e.g., purifying selection) and are thus less informative to narrow down the target of balancing selection. Regardless of these potential restrictions, nonsynonymous differences follow very well the synonymous pattern, suggesting that balancing selection is a strong selective force operating at *csd*-PSD nonsynonymous sites. Compatible with *csd*-PSD being the target of selection, we have identified a coiled coil domain (Burkhard et al. 2001) that has the capability to be part of the specifying domain by its protein interaction (Beye 2004). An exception of exons 6 and 7 having the highest πS values is found between exons 8 and 7 of *A. cerana* O cluster that have comparable πS values (supplementary fig. 2, Supplementary Material online). We propose that selection may operate differently among sequences of the O cluster, or, alternatively, similar values result from the broad distribution of nucleotide diversity generated by the stochastic nature of the evolutionary process. The decline of πS around *csd*-PSD indicates that recombination events between exons have contributed to the evolution of *csd* sequences. Direct estimates for recombination rates within the *csd*

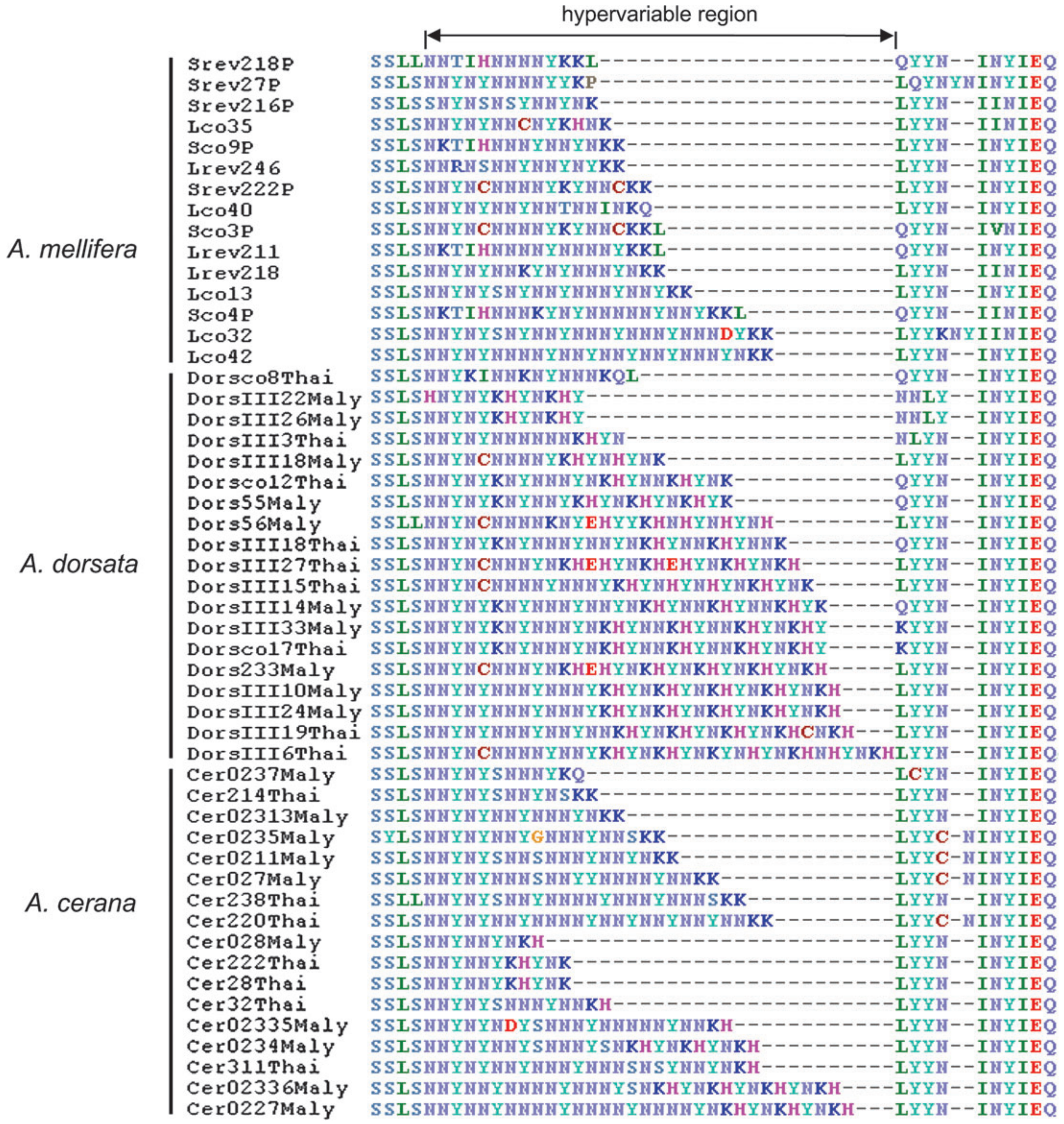
hypervariable region
A. mellifera
Srev218P SSLLNNTIHNNNNYKKL----------------------QYYN--INYIEQ
Srev27P SSLSNNYNYNNNNYYKP----------------------LQYNYNINYIEQ
Srev216P SSLSSNYNSNSYNNYNK----------------------LYYN--IINIEQ
Lco35 SSLSNNYNYNNCNYKHNK---------------------LYYN--IINIEQ
Sco9P SSLSNKTIHNNNYNNYNKK--------------------LYYN--INYIEQ
Lrev246 SSLSNNRNSNNYNNYNYKK--------------------LYYN--INYIEQ
Srev222P SSLSNNYNCNNNNYKYNNCKK------------------LYYN--INYIEQ
Lco40 SSLSNNYNYNNYNNTNNINKQ------------------LYYN--INYIEQ
Sco3P SSLSNNYNCNNNNYKYNNCKKL-----------------QYYN--IVNIEQ
Lrev211 SSLSNKTIHNNNNYNNNNYKKL-----------------QYYN--INYIEQ
Lrev218 SSLSNNYNYNNKYNYNNNYNKK-----------------LYYN--IINIEQ
Lco13 SSLSNNYNYSNYNNYNNNYNNYKK---------------LYYN--INYIEQ
Sco4P SSLSNKTIHNNNKYNYNNNNNYNNYKKL-----------QYYN--IINIEQ
Lco32 SSLSNNYNYSNYNNYNNNYNNNYNNNDYKK---------LYYKNYIINIEQ
Lco42 SSLSNNYNYNNNNYNNYNNYNNYNNNNYNKK--------LYYN--INYIEQ
A. dorsata
Dorsco8Thai SSLSNNYKINNKNYNNNKQL-------------------QYYN--INYIEQ
DorsIII22Maly SSLSHNYNYKHYNKHY-----------------------NNLY--INYIEQ
DorsIII26Maly SSLSNNYNYKHYNKHY-----------------------NNLY--INYIEQ
DorsIII3Thai SSLSNNYNYNNNNNNNKHYN-------------------NLYN--INYIEQ
DorsIII18Maly SSLSNNYNCNNNNYKHYNHYNK-----------------LYYN--INYIEQ
Dorsco12Thai SSLSNNYNYKNYNNNYNKHYNNKHYNK------------QYYN--INYIEQ
Dors55Maly SSLSNNYNYKNYNNYKHYNKHYNKHYK------------QYYN--INYIEQ
Dors56Maly SSLLNNYNCNNNNKNYEHYYKHNHYNHYNH---------LYYN--INYIEQ
DorsIII18Thai SSLSNNYNYKNYNNNYNNYNKHYNNKHYNNK--------QYYN--INYIEQ
DorsIII27Thai SSLSNNYNCNNNYNKHEHYNKHEHYNKHYNKH-------LYYN--INYIEQ
DorsIII15Thai SSLSNNYNCNNNNYNNNYKHYNHYNHYNKHYNK------LYYN--INYIEQ
DorsIII14Maly SSLSNNYNYKNYNNNYNNYNKHYNNKHYNNKHYK-----QYYN--INYIEQ
DorsIII33Maly SSLSNNYNYKNYNNNYNKHYNNKHYNNKHYNKHY-----KYYN--INYIEQ
Dorsco17Thai SSLSNNYNYKNYNNNYNKHYNNKHYNNKHYNKHY-----KYYN--INYIEQ
Dors233Maly SSLSNNYNCNNNYNKHEHYNKHYNKHYNKHYNKH-----LYYN--INYIEQ
DorsIII10Maly SSLSNNYNYNNNYNNNYKHYNKHYNKHYNKHYNKH----LYYN--INYIEQ
DorsIII24Maly SSLSNNYNYNNNYNNNYKHYNKHYNKHYNKHYNKH----LYYN--INYIEQ
DorsIII19Thai SSLSNNYNYNNNYNNYNNKHYNKHYNKHYNKHCNKH---LYYN--INYIEQ
DorsIII6Thai SSLSNNYNCNNNNYNNYKHYNKHYNKYNHYNKHNHYNKHLYYN--INYIEQ
A. cerana
Cer0237Maly SSLSNNYNYSNNNYKQ-----------------------LCYN--INYIEQ
Cer214Thai SSLSNNYNYSNNYNSKK----------------------LYYN--INYIEQ
Cer02313Maly SSLSNNYNYNNYNNNYNKK--------------------LYYN--INYIEQ
Cer0235Maly SYLSNNYNYNNYGNNNYNNSKK-----------------LYYC-NINYIEQ
Cer0211Maly SSLSNNYNYSNNSNNNYNNYNKK----------------LYYC-NINYIEQ
Cer027Maly SSLSNNYNYNNNSNNYYNNNNYNNKK-------------LYYC-NINYIEQ
Cer238Thai SSLLNNYNYSNNYNNNNYNNNYNNNSKK-----------LYYN--INYIEQ
Cer220Thai SSLSNNYNYNNYNNNNYNNYNNYNNYNNKK---------LYYC-NINYIEQ
Cer028Maly SSLSNNYNNYNKH--------------------------LYYN--INYIEQ
Cer222Thai SSLSNNYNNYKHYNK------------------------LYYN--INYIEQ
Cer28Thai SSLSNNYNNYKHYNK------------------------LYYN--INYIEQ
Cer32Thai SSLSNNYNYSNNNYNNKH---------------------LYYN--INYIEQ
Cer02335Maly SSLSNNYNYNDYSNNNYNNNNNYNNKH------------LYYN--INYIEQ
Cer0234Maly SSLSNNYNYNNYSNNNYSNKHYNKHYNKH----------LYYN--INYIEQ
Cer311Thai SSLSNNYNYNNYNNNYNNNSNSYNNYNKH----------LYYN--INYIEQ
Cer02336Maly SSLSNNYNNYNNNNYNNNYSNKHYNKHYNKHYNKH----LYYN--INYIEQ
Cer0227Maly SSLSNNYNNYNNNNYNNNNYNNNNYNKHYNKHYNKH---LYYN--INYIEQ

FIG. 6.—Aligned conceptual translations spanning the hypervariable region of 3 honeybee species. This region was excluded in the *csd*-PSD nucleotide polymorphism and divergence analysis because of the ambiguous relation of these sites. The region encodes a basic $((N)_{1–5}/Y)$ sequence motif that is modified at the end by either a set of 2 amino acids (KK less often with KP, KQ) or with the more complex $((KHYN)_{1–4})$KH motif. This hypervariable region reiterates 3–7 times, giving rise to peptides of varying length. Two alleles of *Apis cerana* have the same repeat structure.

gene, however, have shown so far a drastic suppression of recombination activity in the *csd* gene, at least for *A. mellifera* (Hasselmann and Beye 2006). We propose that even very rare recombination events leave their signatures in the sequence because alleles have extended maintenance times. A similar pattern of a decline in synonymous diversity around the target of balancing selection, the peptide-binding site, has been observed in the human leukocyte antigen genes within the MHC system (Takahata and Satta 1998).

Short Coalescence Times Relates to Long-Term Small Population Sizes

When compared with other genetic systems under strong balancing selection, *csd*-PSD alleles have relatively short average coalescence times (~6 Myr) and low average synonymous nucleotide diversity ($\pi S = 0.032–0.089$). In addition, we have not identified trans-specific *csd*-PSD alleles among species although the separation times of these species are very short (~7 and 10 Myr). For plant self-incompatibility systems, diversity of synonymous sites in the recognition domain within *Arabidopsis lyrata* is more than 4 times larger (Charlesworth et al. 2003), and several trans-specific allelic lineages have been observed in a variety of species (Ioerger et al. 1990; Uyenoyama 1995; Castric and Vekemans 2004) that correspond to separation times of up to 40 Myr. As general explanations for the relatively short average coalescence times, we discarded the effect of intragenic recombination as we observed a strong correlation between numbers of synonymous and nonsynonymous differences. We also excluded the possibility of major bottlenecks as the *csd*-PSD genealogy is largely compatible with the expected ratios of the largest to mean diversity under the assumption of equilibrium, although some size restriction has affected the *A. cerana* population (see below). We suggest that relative low nucleotide diversity is mostly compatible with the occurrence of higher allelic turnover rates of *csd* alleles when compared with other loci that are under balancing selections. The relative high allelic turnover rates of the *csd*-PSD locus under strong balancing selection could result from either a high rate of origin of new alleles or a small long-term effective population sizes. By comparing our data with a *csd*-coalescence model, we discarded the former explanation as this would generate substantially greater allelic richness than that observed. Hence, we conclude that small long-term population sizes, that is, $N_e < 10^4$, are the most likely explanation for the observed patterns of divergence. As a general consequence, this finding would implicate genetic drift as a very strong, long-term evolutionary force in highly social honeybees. The even lower divergence and overall coalescence times of *A. cerana* alleles are suggestive of an additional, but temporal, restriction in population size as seen by a burst of diversification in some allelic lineages. This is seen by the significantly shorter internal branches when compared with other branches of the phylogeny (fig. 2). Such a burst of diversification after a population bottleneck has been previously described for self-incompatibility alleles of *Physalis crassifolia* (Richman et al. 1996).

We hypothesize that small long-term population size in these honeybees is a consequence of highly social organization. The strong division of reproduction among individuals under highly social organization results in a dramatic decrease in long-term effective population size (Wilson 1963; Crozier 1979; Pamilo and Crozier 1997). Thousands of sterile female worker bees are headed by a single reproductive queen, thus substantially reducing the size of the breeding population relative to the total number of individuals that can be sustained by the local environment. Under balancing selection, the estimate of N_e is not strongly affected by population structure (Schierup et al. 2000) so that it truly represents the size of the breeding population at the species level.

The long-term estimates can be related to short-term estimates of effective population size based on neutral polymorphism data for *A. mellifera*. When applying a mutation rate $\mu = 7 \times 10^{-9}$ per site per generation and the mean nucleotide diversity $\pi = 0.0049 +/- 0.0013$ standard error of neutral polymorphisms (Beye et al. 2006) to $\pi = N_e \times \mu$, N_e becomes 70,000 (with a CI of 30,000–110,000). The greater short-term over long-term effective population size would indicate a recent expansion and/or substructuring of European honeybee populations (*A. mellifera*). This is supported by biogeographic molecular studies that have shown an expansion and differentiation of *A. mellifera* in Europe and Africa over the last 0.7–1.3 Myr (Garnery et al. 1992; Arias and Sheppard 2005). In contrast to strongly balanced polymorphisms, these neutral polymorphisms are strongly affected by the substructure of populations thus making both estimates less comparable.

Convergent and Adaptive Evolution of Nonsynonymous Changes at csd-PSD

The observed $\pi N/\pi S$ ratios for *csd*-PSD ($\pi N/\pi S$ = 0.6–0.9) are above average estimates from other genes (0.12 and 0.2 for *Drosophila melanogaster* and *Drosophila simulans*, respectively (Eyre-Walker et al. 2002), which suggests that the *csd* gene is under strong balancing selection favoring new rare allelic variants and nonsynonymous changes or that it has substantially relaxed functional constraints. We identified a substantial excess of shared nonsynonymous changes at *csd*-PSD that accumulated independently in different alleles across species (homoplasy). This provides indirect evidence that positive selection has favored some specific nonsynonymous changes. The lack of trans-specific alleles and the absence of an excess of synonymous changes suggest that the excess of shared nonsynonymous changes have an independent evolutionary origin. In a simplified model, we have approximated that only ~25% of all possible nonsynonymous changes can evolve to become compatible with the observed numbers of shared polymorphisms. Our approximation is constant over all species pairs, suggesting that equivalent evolutionary constraints at *csd*-PSD are operating across species. If we correct for the limited numbers (non constrained changes) of possible nonsynonymous changes that can evolve, $\pi N/\pi S$ becomes substantially >1 (corrected estimates: $\pi N_{sp}/\pi S$ ranges from 2.4 to 3.8). Selection on a confined subset of nonsynonymous changes thus masks the strong effect of selection for new specificities, which is usually revealed by $\pi N/\pi S$ ratios >1. The polymorphisms that evolved convergently at *csd*-PSD could be very informative to identify molecular determinants of allelic specificity. Significantly, these shared sites are scattered across *csd*-PSD, suggesting that functionally relevant polymorphisms are not confined to a specific region.

The hypervariable region most likely adds to the specificity of alleles. We identified 2 basic peptide motifs across species, suggesting their evolutionary origin in a common ancestor. These regions further diversify in the different honeybee lineages essentially by changing the number of repeats. A recent study (Cho et al. 2006) claimed to have identified species-specific repeat structures. Our study, which includes a sophisticated set of alleles relying on transcribed sequences (thus excluding possible noncoding sequences of the genomic fragment approach [Cho et al. 2006]), showed the contrary.

The finding of higher $\pi N/\pi S$ ratios of newly over anciently diverged alleles (fig. 5) is additional evidence that balancing selection has favored nonsynonymous changes at *csd*-PSD. Balancing selection has enhanced the relative rate of accumulation of nonsynonymous changes among newly diverged alleles. These enhanced evolutionary rates of newly diverged alleles may also explain our previous report on longer terminal over internal branches in the *csd* genealogy (Hasselmann and Beye 2004). This longer terminal branch pattern is unexpected from theory (Uyenoyama 1997) but has also been reported for *S* allele genealogies (Richman and Kohn 1999).

Overall, this study has deciphered a complex pattern of evolutionary forces (selection, genetic drift, mutation, and recombination) that have shaped nucleotide polymorphism at *csd*-PSD and linked sites. A conclusive test of these evolutionary genetic evidences awaits detailed functional analysis. This will provide further evidence that the *csd*-PSD corresponds to the specifying domain and that shared polymorphisms are informative to identify the molecular determinants of allele specificity and protein activation (Beye et al. 2003; Beye 2004). This merged approach will give a more detailed understanding of how selection has shaped nucleotide diversity and the function of alleles. Extending studies on allelic richness and nucleotide diversity to other bee species will decipher whether social organization is generally associated with a high turnover rate of sex determination alleles and small long-term effective population sizes. Such a general association will have implications for the understand of natural selection in relation to social organization (Pamilo and Crozier 1997) and may have broad consequences such as genome-wide high recombination rates (Beye et al. 2006).

Supplementary Material

All new *csd* and *fem* sequences presented in this paper are found in GenBank under accession numbers EU100885–EU100941. Supplementary figures 1–3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank very much Ross Crozier, George Heimpel, Robert Paxton, and 2 anonymous reviewers for very helpful comments and corrections on an earlier version of the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft.

Literature Cited

Adams J, Rothman ED, Kerr WE, Paulino ZL. 1977. Estimation of the number of sex alleles and queen matings from diploid male frequencies in a population of *Apis mellifera*. Genetics. 86:583–596.

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

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, Gattermeier I, Hasselmann M, et al. (15 co-authors). 2006. Exceptionally high levels of recombination across the honey bee genome. Genome Res. 16:1339–1344.

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

Burkhard P, Stetefeld J, Strelkov SV. 2001. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol. 11:82–88.

Castric V, Vekemans X. 2004. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Mol Ecol. 13:2873–2889.

Charlesworth D, Bartolome C, Schierup MH, Mable BK. 2003. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. Mol Biol Evol. 20:1741–1753.

Cho S, Huang ZY, Green DR, Smith DR, Zhang J. 2006. Evolution of the complementary sex-determination gene of honey bees: balancing selection and trans-species polymorphisms. Genome Res. 16:1366–1375.

Crozier RH. 1979. Genetics of sociality. In: Hermann HR, editor. Social insects. New York: Academic Press. p. 223–286.

Eyre-Walker A, Keightley PD, Smith NG, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol. 19:2142–2149.

Garnery L, Cornuet J-M, Solignac M. 1992. Evolutionary history of the honeybee *Apis mellifera* inferred from mitochondrial DNA analysis. Mol Ecol. 1:145–154.

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

Hasselmann M, Beye M. 2006. Pronounced differences of recombination activity at the sex determination locus (SDL) of the honey bee, a locus under strong balancing selection. Genetics. 174:1469–1480.

Hey J, Wakeley J. 1997. A coalescent estimator of the population recombination rate. Genetics. 145:833–846.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. Genetics. 120:831–840.

Ioerger TR, Clark AG, Kao TH. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc Natl Acad Sci USA. 87:9732–9735.

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

Pamilo P, Crozier RH. 1997. Population biology of social insect conservation. Mem Mus Vic. 56:411–419.

Richman AD, Kohn JR. 1999. Self-incompatibility alleles from Physalis: implications for historical inference from balanced genetic polymorphisms. Proc Natl Acad Sci USA. 96:168–172.

Richman AD, Uyenoyama MK, Kohn JR. 1996. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science. 273:1212–1216.

Rozas J, Rozas R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics. 15:174–175.

Schierup MH, Vekemans X, Charlesworth D. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genet Res. 76:51–62.

Sheppard WS, Berlocher SH. 1989. Allozyme variation and differentiation among four *Apis* species. Apidologie. 20:419–431.

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc Natl Acad Sci USA. 87:2419–2423.

Takahata N. 1993. Evolutionary Genetics of human paleopopulations. In: Takahata N, Clark AG, editors. Mechanisms of molecular evolution: introduction to molecular paleopopulation biology. Sunderland (MA): Sinnauer Associates Inc. p. 1–21.

Takahata N, Satta Y. 1998. Footprints of intragenic recombination at *HLA* loci. Immunogenetics. 47:430–441.

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

Uyenoyama MK. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. Genetics. 139:975–992.

Uyenoyama MK. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics. 147:1389–1400.

Uyenoyama MK. 2003. Genealogy-dependent variation in viability among self-incompatibility genotypes. Theor Popul Biol. 63:281–293.

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics. 137:1157–1165.

Wilson EO. 1963. Social modifications related to rareness in ant species. Evolution. 17:249–253.

Wright S. 1939. The distribution of self-sterility alleles in populations. Genetics. 24:538–552.

Wright S. 1960. On the number of self-incompatibility alleles maintained in equilibrium by a given size: a re-examination. Biometrics. 16:61–85.

Yokoyama S, Nei M. 1979. Population dynamics of sex-determining alleles in honey bees and self-incompatibility in plants. Genetics. 91:609–626.

Diethard Tautz, Associate Editor

Accepted January 7, 2008