Research article for *Molecular Biology and Evolution*

Evidence for convergent nucleotide evolution and high allelic turnover rates at the *complementary sex determiner* (*csd*) gene of western and Asian honey bees

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

* Heinrich-Heine Universitaet Duesseldorf, Institut fuer Genetik, Universitaetsstr. 1, 40225 Duesseldorf, Germany

† Laboratoire de Génétique et Evolution des Populations Végétales
UMR CNRS 8016, Bâtiment SN2, Université de Lille 1, F-59655 Villeneuve d'Ascq, France

‡ Institut für Bienenkunde, Johann-Wolfgang-Goethe Universität Frankfurt/M,
Karl-von-Frisch-Weg 2, 61440 Oberursel, Germany

§ Agricultural Research Station Tenom, P.O. Box 197, 89908, Tenom, Sabah, Malaysia

to whom correspondence should be addressed:
Martin Hasselmann, Martin Beye
Heinrich-Heine Universitaet Duesseldorf, Institut fuer Genetik, Universitaetsstr. 1, 40225 Duesseldorf, Germany,
telephone:++49 0211 8114808 ; fax: ++490211 8112279

martin.hasselmann@uni-duesseldorf.de, martin.beye@uni-duesseldorf.de

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

Running title: evolutionary forces operating among *csd* alleles

Abstract

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 of these evolutionary forces at the *complementary sex determiner* (*csd*) gene which 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 to neutral polymorphisms. Here, we report on the analysis of 17, 19 and 15 *csd* alleles of *Apis cerana*, *Apis dorsata* and *Apis mellifera* honey bees respectively. We observed great heterogeneity of synonymous (πS) and nonsynonymous (πN) polymorphisms across the gene, with a consistent peak in exon 6 and 7. We propose that exons 6 and 7 encode the potential specifying domain (*csd*-PSD) which 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 ($\pi N/\pi S$ ratios are all < 1, ranging from 0.6 to 0.95). We observed an excess of shared nonsynonymous changes, which suggests 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 a *csd*-PSD genealogy revealed relatively short average coalescence times (~6 million years), low average synonymous nucleotide diversity ($\pi S < 0.09$) and a lack of trans-specific alleles which 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 turn-over 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 impact of genetic drift on the molecular evolution of highly social honey bees.

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). *csd* determines the sexual fate of the honey bee (*Apis mellifera*), with the combination of two *csd* alleles in the zygote resulting in a female and one unique allele resulting in a male (Beye 2004). The *csd* gene segregates in more than 15 allelic forms in populations (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 offers the possibility to study (i) long-term population processes by the coalescence process; and (ii) 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 MHC complex (Takahata 1990) and of 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 alleles) and the loss of alleles by genetic drift, which is a function of population size (Wright 1960; Wright 1939; Yokoyama and Nei 1979). *S*-alleles and *MHC*-alleles have long overall coalescence times and have been maintained for more than 40 MY (million years) (Ioerger et al. 1990; Uyenoyama 1995) and 30 MY (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 (Richman et al. 1996; Richman and Kohn 1999; Takahata 1990) 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 three related honey bee species, *Apis cerana*, *Apis dorsata* (both Asian honey bee species) and *Apis mellifera* (the western honey bee) which diverged over the last 10 MY (million years) (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 honey bee species: *A. cerana* and *A. dorsata* (Cho et al. 2006). However, the latter study relied on fragmented ORFs (open reading frame) 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 (i) the full coding sequence of *csd* from three related *Apis* species (*A. cerana*, *A. dorsata*, *A. mellifera*) and (ii) 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 three species *Apis mellifera, A. cerana*, and *A. dorsata* were collected in two sampling locations. Geographic locations for *A. cerana* and *A. dorsata* were Tenom (Borneo) and Thailand (Samut Songkram for *A. cerana*; Wanmanaow for *A. dorsata*) and for *A. mellifera* were Litija (Slovenia) and Pretoria (South Africa). In each location, 150 – 300 embryos were collected from two 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´ 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: TCTCATTATTCAATACGTTGGCATC; csd_forCer2: CTCTAAGCGTGGATTACAGGTT; csd_IIIfor: GTTAAATTTCATWRATATACATATAC; csd_IIIrev3: ATTCAGTTCATTATTCATTATTTGCA; cons2csd_for: TCATAAAAATGAAACGAAATATATC 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).
csd sequence variants (and *fem* sequences) were identified from 60-100 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. *csd*-alleles of all drones were successfully amplified by PCR. The PCR-induced mutations are lower than 2 in 1000 nucleotides (sequence data of identical sequences obtained from different high-fidelity PCR amplifications), indicating that PCR based amplifications do not influence the conclusions of *csd*-PSD allele analyses which 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 one sequence for each of these neutral sequence clusters was included in the analysis. This resulted in a final dataset of 51 *csd* alleles (n = 15 for *A. mellifera*, n = 17 for *A. cerana*, n = 19 for *A. dorsata*). Our classification is supported by the finding that alleles/specificities differ in the structure of the hypervariable domain while 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 two-parameter distances. The number of shared polymorphisms and fixed differences between species pairs were 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 and Rozas 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 (~ 7MY) (Sheppard and Berlocher 1989; Garnery et al. 1992) according to the following relationship: number of polymorphisms at synonymous sites that accumulate = $2 g \mu$, where g is the number of generations, assuming one generation per year for the honey bee, 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 MY).

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 honey bees 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_f and N_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

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

where u is the mutation rate to new alleles per generation, and N is the effective population size computed as $N = 9\ N_m\ N_f/(4\ N_m + 2\ N_f)$. 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 time scale expanded by a factor f_s, with f_s determined by the parameters of the selection model. Takahata's scaling factor f_s can be computed, according to Uyenoyama (2003) as

$$f_s = \frac{*n^2}{4N\lambda} \quad (2),$$

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

$$\lambda = 2Nu(a - 2b) \quad (3);$$

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:

$$f_s = \frac{3(1-F)}{32N^2uF^3} \quad (4).$$

Finally, the pairwise coalescence time of alleles at *csd* is given by (Takahata, 1990):

$$T_d = 2.N.f_s. \quad (5),$$

and replacing with (4) we finally get:

$$T_d = \frac{3(1-F)^2}{16NuF^3} \quad (6).$$

We computed expected values of $*n$ and T_d for a large range of parameter values u and N, and compared them to the actual number of alleles observed in the three species, and to the total coalescence times of alleles estimated based on application of a molecular clock to the maximum synonymous nucleotide divergence among extant alleles.

Results

We used reverse transcription polymerase chain reaction (RT-PCR) to survey allelic diversity of *A. mellifera* (western honey bee), *A. cerana* and *A. dorsata* (the two Asian honey bees) 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 two major types (type I and type II (Hasselmann and Beye 2004)) of sex-determining alleles. We have confined our analysis to type I which corresponds to the actual *csd* gene and is the target of balancing selection. With the help of the honey bee genome project we characterized type II (*feminizer*, *fem*) to be a paralog of *csd* with sex differentiation function downstream of *csd* (Hasselmann, Gempe, Schioett, Nunes-Silva, Otte, Beye, submitted) and thus it is not the initial signal of sexual regulation nor the target of balancing selection. Consistent with this observation, type II sequences show a near-absence of intra-species divergence and cluster into a single clade suggesting that gene duplication occurred before the split of these species (supplemental Figure 1). This finding is inconsistent with a previous report (Cho et al. 2006) of partial type I and type II genomic sequences (their figure 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 and co-workers 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; Hasselmann and Beye 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 co-exist within species representing different divergence times.

In *A. mellifera* and *A. dorsata*, the non-synonymous differences (πN) are consistently the highest in exons 6 and 7 ($\pi N > 0.05$), whereas πN is lower for exons 1 to 5 ($\pi N < 0.05$). This suggests that at least exons 6 and 7 are targets of balancing selection which 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 to exon 7 is 60 bp, whereas that between exon 5 and exon 6 is 3.9 kb). In *A. cerana* exons 4 and 5 have the highest πN. However, this enhanced polymorphism results from divergence between two sequence clusters (O and T type), that are characterized by low intra-cluster differences (supplementary Figure 2), 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 non-synonymous 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 two clusters, O and T (supplementary Figure 2), intra-cluster πS values follow the pattern in the other species with low πS in exon 5 but high πS in exons 6 and 7 and which extends to exon 8.

Based on the comparative analysis of nonsynonymous and synonymous nucleotide

polymorphisms across species and the identification of a conserved coiled coil motif, 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 potential specifying domain (*csd*-PSD).

Short coalescence times in csd genealogies

The relationship among *csd*-PSD sequences is reflected by their genealogy (fig 2). The alleles fall into three clades, supported by bootstrap values > 90%, such that each clade represents a single species (see also supplementary Figure 1 for the whole sequence). Hence, no trans-specific allelic pattern of the common target of selection, *csd*-PSD, is observed in these honey bee species that have been separated by maximally 10 MY (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* complex in which several trans-specific alleles have been observed even after 30 to 40 MY 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*, $\pi S = 0.069$ for *A. dorsata*). A lack of trans-specific lineages could be due (i) to bursts of diversification of *csd* alleles that occurred after the speciation events, (ii) to the process of recombination in which trans-specific allelic patterns are lost, or (iii) to high allelic turnover rates leading to short coalescence times among *csd* alleles.

(i) 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 to 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 to

mean number of synonymous and nonsynonymous differences which 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).

(ii) The process of intragenic recombination at *csd*-PSD within each species could re-assemble 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 figure 3 A-C) shows that the two 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 is consistent with a tight linkage of synonymous to nonsynonymous sites (Spearman rank correlation, two-tailed, *A. dorsata*: rho = 0.55; $P < 0.001$, *A. cerana*: rho = 0.45; $P < 0.001$, *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, 0.32 for *A. mellifera*, *A. cerana*, *A. dorsata* respectively), we assumed a binomial distribution from which the 95% confidence interval 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.

(iii) 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 to 15×10^{-9} per site per year for *Drosophila*), as estimates from divergence of gene sequences and speciation events suggest (see Material and Methods), it must

have taken about 14.6 million years (MY) or generations in *A. mellifera*, 16.7 MY in *A. dorsata* and 7.9 MY in *A. cerana* for these synonymous differences to accumulate (number of synonymous sites that accumulate = 2 $g\,\mu$ with g number of generations and one generation per year for the honey bee 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 to 8 MY) as the largest allelic divergence (dS = 0.11) is in the order of mean inter-species divergence (dS = 0.1, see Material 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 inter-species divergence (dS > 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 i.e. that the two 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 supplementary Material and Methods) to compare with our empirical data. The coalescence process generating allelic genealogies at *csd* can be approximated using Takahata's coalescent approach (Takahata 1990) 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 ($u = \mu$ x 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 pairwise coalescence times of alleles (T_d)

expected 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 honey bees ($n \approx 20$) enforces a strong constraint on the effective population size, with compatible N_e values ranging between 1500 and 4000 individuals assuming a rate of origin of new alleles of 10^{-8} to 10^{-7} (fig. 3a). In contrast, the inferred average pairwise coalescence time of alleles of about 6 MY years 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 ($N_e < 10^4$) for *A. mellifera* and *A. dorsata*. Hypothesised N_e is even lower for *A. cerana* harbouring less nucleotide diversity. In the above calculations, we assumed a long-term generation time of one 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 one year will result in an even lower population size estimate.

Excess of shared nonysynonymous but not synonymous changes

csd-PSD has on average less nonsynonymous than synonymous polymorphisms per nonsynonymous and synonymous site ($\pi N/\pi S$: 0.6, 0.95 and 0.73 for *A. mellifera*, *A. dorsata* and *A. cerana*, respectively). $\pi N/\pi S > 1$ 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 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 two 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 which removes deleterious mutations. Alternatively, evolutionary constraints may be a consequence of selection favoring new allelic variants which 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 $\pi N/\pi S$ 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, 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 $\pi N/\pi S$ change with different values of πS. 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 $\pi N/\pi S$ ratio is in all three 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 downwards bias of $\pi N/\pi S$ estimate due to πS signal saturation. 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 two 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 not included in the former polymorphism analyses. The repeat encodes a basic $((N)_{1\text{-}5}/Y)$ sequence motif of

asparagine/tyrosine residues. This basic sequence motif is modified at the end of each reiteration by either a set of two amino acids (KK less often with KP, KQ) or with the more complex $((KHYN)_{1-4})KH$ motif. This combination is reiterated three to seven times, encoding peptides of varying length. The former repeat ending is observed in all three species (with only one representative in *A. dorsata*), while the latter one 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 suggests functional significance of this domain in specifying sex determination alleles.

Discussion

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 three 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 which have comparable πS values (supplementary figure 2). 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* 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 HLA genes within the MHC system (Takahata and Satta 1998).

Short coalescence times relates to long term small population sizes

When compared to other genetic systems under strong balancing selection, *csd*-PSD alleles have relatively short average coalescence times (~6MY) and low average synonymous nucleotide diversity (πS = 0.032 to 0.089). In addition, we have not identified trans-specific *csd*-PSD alleles among species although the separation times of these species is very short (~7 and 10 MY). For plant self-incompatibility systems, diversity of synonymous sites in the recognition domain within *Arabidopsis lyrata* is more than four times larger (Charlesworth et al. 2003), and several trans-specific allelic lineages have been observed in a variety of species (Castric and Vekemans 2004; Ioerger et al. 1990; Uyenoyama 1995) that correspond to separation times of up to 40 MY. 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 non-synonymous 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 to 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 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, i.e. $N_e < 10^4$, 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 honey bees. 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 to 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 honey bees 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 $\pi = 0.0049 +/- 0.0013$ SE of neutral polymorphisms (Beye et al. 2006) to $\pi = N_e \times \mu$, N_e becomes 70,000 (with a C.I. of 30,000 – 110,000). The greater short-term over long-term effective population size would indicate a recent expansion and/or sub-structuring of European honey bee populations (*A. mellifera*). This is supported by biogeographic molecular studies which have shown an expansion and differentiation of *A. mellifera* in Europe and Africa over the last 0.7-1.3 MY (Arias and Sheppard 2005; Garnery et al. 1992). 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 ($\pi N/\pi S = 0.6$ to 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 two basic peptide motifs across species, suggesting their evolutionary origin in a common ancestor. These regions further diversify in the different honey bee 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 non-coding 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 and tables are provided on this journal's web site.

Acknowledgements

We thank very much Ross Crozier, George Heimpel, Robert Paxton and two 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,E.D., Kerr,W.E., Paulino,Z.L. 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,M.C. and Sheppard,W.S. 2005. Phylogenetic relationships of honey bees (Hymenoptera:Apinae:Apini) inferred from nuclear and mitochondrial DNA sequence data. Mol. Phylogenet. Evol. 37:25-35.

Bechsgaard,J.S., Castric,V., Charlesworth,D., Vekemans,X., Schierup,M.H. 2006. The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. Mol. Biol. Evol. 23:1741-1750.

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

Biesmeijer,J.C., Roberts,S.P., Reemer,M., *et al.* (12 co authors). (2006). Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. Science 313:351-354.

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

Castric,V. and 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,M.H., Mable,B.K. 2003. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. Mol. Biol. Evol. 20:1741-1753.

Cho,S., Huang,Z.Y., Green,D.R., Smith,D.R., 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,R.H. 1979. Genetics of Sociality. In: H.R.Hermann, editor. Social Insects. New York: Academic Press. p. 223-286.

Eyre-Walker,A., Keightley,P.D., Smith,N.G., 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. and Beye,M. 2004. Signatures of selection among sex-determining alleles of the honey bee. Proc Natl Acad Sci U S A 101:4888-4893.

Hasselmann,M. and 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. and Wakeley,J. 1997. A coalescent estimator of the population recombination rate. Genetics 145:833-846.

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

Hudson,R.R. and Kaplan,N.L. 1988. The coalescent process in models with selection and recombination. Genetics 120:831-840.

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.

Hughes,A.L. and Yeager,M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. 32:415-435.

Ioerger,T.R., Clark,A.G., Kao,T.H. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. U. S. A 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. and Crozier,R.H. 1997. Population biology of social insect conservation. Memoirs of the Museum of Victoria 56:411-419.

Richman,A.D. and Kohn,J.R. 1999. Self-incompatibility alleles from Physalis: implications for historical inference from balanced genetic polymorphisms. Proc Natl Acad Sci U S A 96:168-172.

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

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

Satta,Y. 1993. Balancing selection at *HLA* loci. In: Takahata N, Clark, AG editors. Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology. Tokyo: Sunderland, Japan Scientific Societies Press, Sinnauer Associates Inc. p. 129-148.

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

Sheppard,W.S. and Berlocher,S.H. 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 paleo-populations. In: Takahata N, Clark, AG editors. Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology. Tokyo: Sunderland, Japan Scientific Societies Press, Sinnauer Associates Inc. p. 1-21.

Takahata,N. and 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,M.K. 2003. Genealogy-dependent variation in viability among self-incompatibility genotypes. Theor. Popul. Biol. 63:281-293.

Uyenoyama,M.K. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatiblity. Genetics 139:975-992.

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

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

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

Zayed,A. and Packer,L. 2005. Complementary sex determination substantially increases extinction proneness of haplodiploid populations. Proc. Natl. Acad. Sci. U. S. A 102:10742-10746.

Fig. 1. - Average pairwise nonsynonymous (πN) and synonymous (πS) differences per site of exons for three honey bee species (A) *A. mellifera;* (B) *A. dorsata;* (C) *Apis cerana.* Average nonsynonymous (πN) and synonymous (πS) differences per site per exon are presented with their standard errors. The genomic structure 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.

Fig. 2. - Genealogy of *csd*-PSD nucleotide sequences of three honey bee species based on the minimum evolution method of genetic differences (Kimura´s two-parameter distances). The alleles cluster into three clades representing the species *A. mellifera*, *A. dorsata* and *A. cerana*. The scale on the left represents nucleotide differences per site while the stippled bar marks a burst of diversification among *A. cerana* alleles (for further information see text). Numbers are bootstrap values exceeding 80 %.

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.

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 *A. mellifera*/*A. cerana* is shown by triangles, *A. mellifera*/*A. dorsata* by squares and *A. dorsata*/*cerana* by diamonds. We approximate the total number of possible nonsynonymous changes to 501 based on the numbers of 167 nonsynonymous sites.

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

Fig. 6. - Aligned conceptual translations spanning the hypervariable region of three 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 which is modified at the end by either a set of two amino acids (KK less often with KP, KQ) or with the more complex $((KHYN)_{1-4})$KH motif. This hypervariable region reiterates three to seven times, giving rise to peptides of varying length. Two alleles of *A. cerana* have the same repeat structure.

Table 1 - The pairwise largest and mean synonymous (S) and nonsynonymous (N) differences of *csd*-PSD

			S	N
A. 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
A. dorsata	sample size (*i*)	19		
	mean		3.5 ± 0.97	11.6 ± 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
A. 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 one 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).

Table 2 - Numbers and shared numbers of synonymous (S) and nonsynonymous (N) polymorphisms.

	S	N
A. cerana	7	22
A. mellifera	16	27
shared between species	1	5
P	0.4	0.004
expected number of shared		1.2
A. 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. 1

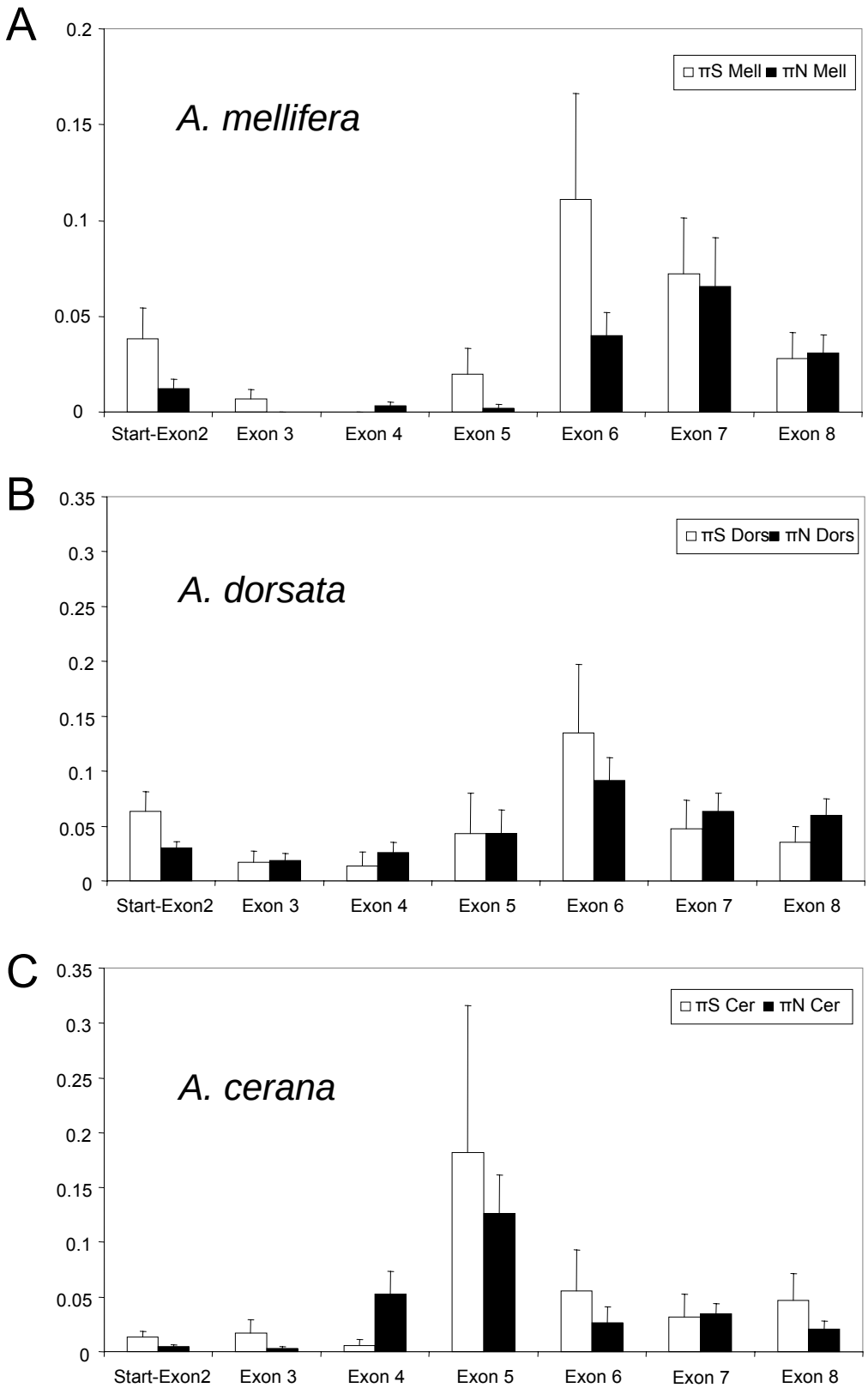

A
0.2
0.15
0.1
0.05
0
□ πS Mell ■ πN Mell
A. mellifera
Start-Exon2
Exon 3
Exon 4
Exon 5
Exon 6
Exon 7
Exon 8
B
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
□ πS Dors ■ πN Dors
A. dorsata
Start-Exon2
Exon 3
Exon 4
Exon 5
Exon 6
Exon 7
Exon 8
C
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
□ πS Cer ■ πN Cer
A. cerana
Start-Exon2
Exon 3
Exon 4
Exon 5
Exon 6
Exon 7
Exon 8

Fig. 2

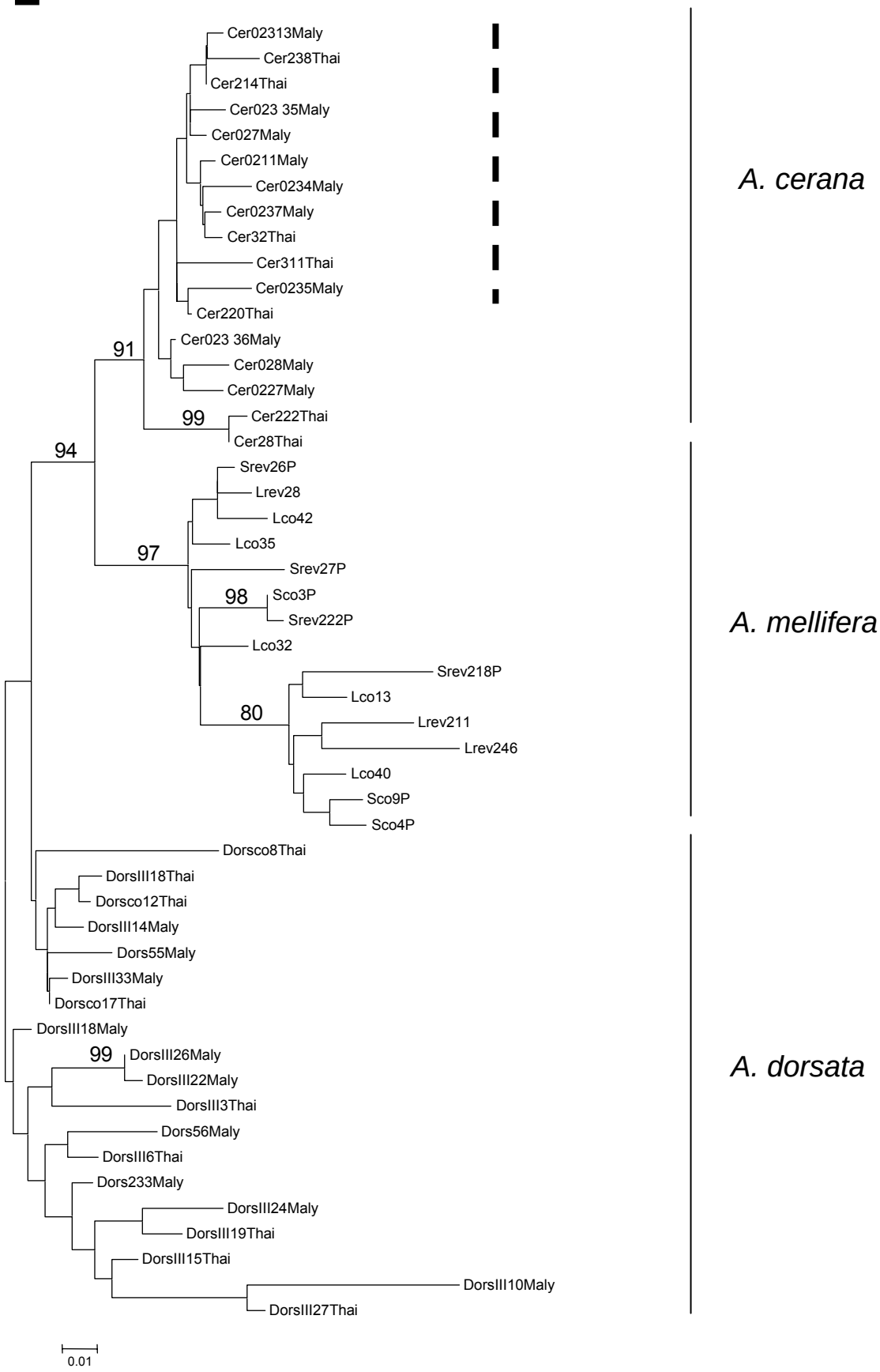
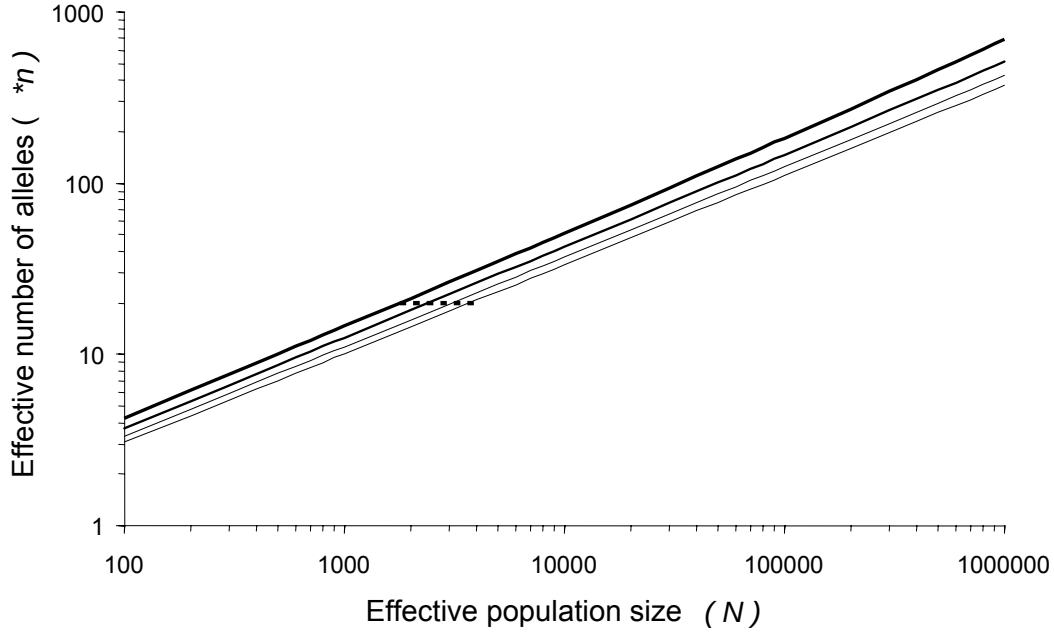

Cer02313Maly
Cer238Thai
Cer214Thai
Cer023 35Maly
Cer027Maly
Cer0211Maly
Cer0234Maly
Cer0237Maly
Cer32Thai
Cer311Thai
Cer0235Maly
Cer220Thai
Cer023 36Maly
Cer028Maly
Cer0227Maly
Cer222Thai
Cer28Thai
Srev26P
Lrev28
Lco42
Lco35
Srev27P
Sco3P
Srev222P
Lco32
Srev218P
Lco13
Lrev211
Lrev246
Lco40
Sco9P
Sco4P
Dorsco8Thai
DorsIII18Thai
Dorsco12Thai
DorsIII14Maly
Dors55Maly
DorsIII33Maly
Dorsco17Thai
DorsIII18Maly
DorsIII26Maly
DorsIII22Maly
DorsIII3Thai
Dors56Maly
DorsIII6Thai
Dors233Maly
DorsIII24Maly
DorsIII19Thai
DorsIII15Thai
DorsIII10Maly
DorsIII27Thai
91
99
94
97
98
80
99
0.01
A. cerana
A. mellifera
A. dorsata

Fig. 3

A


1000
100
10
1
Effective number of alleles (*n)
100
1000
10000
100000
1000000
Effective population size (N)

B

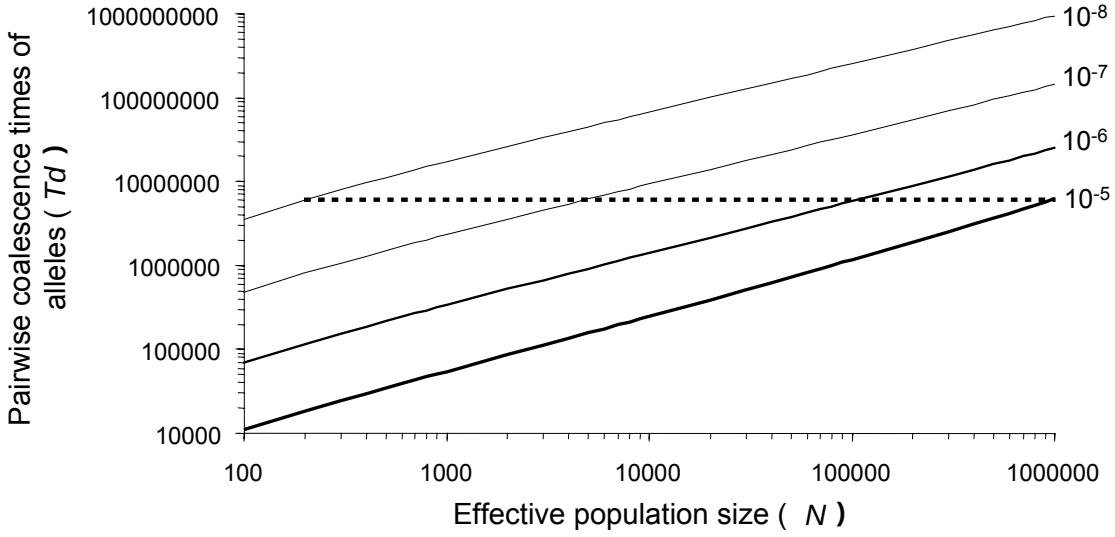

1000000000
100000000
10000000
1000000
100000
10000
Pairwise coalescence times of alleles (Td)
100
1000
10000
100000
1000000
Effective population size (N)
10-8
10-7
10-6
10-5

Fig. 4

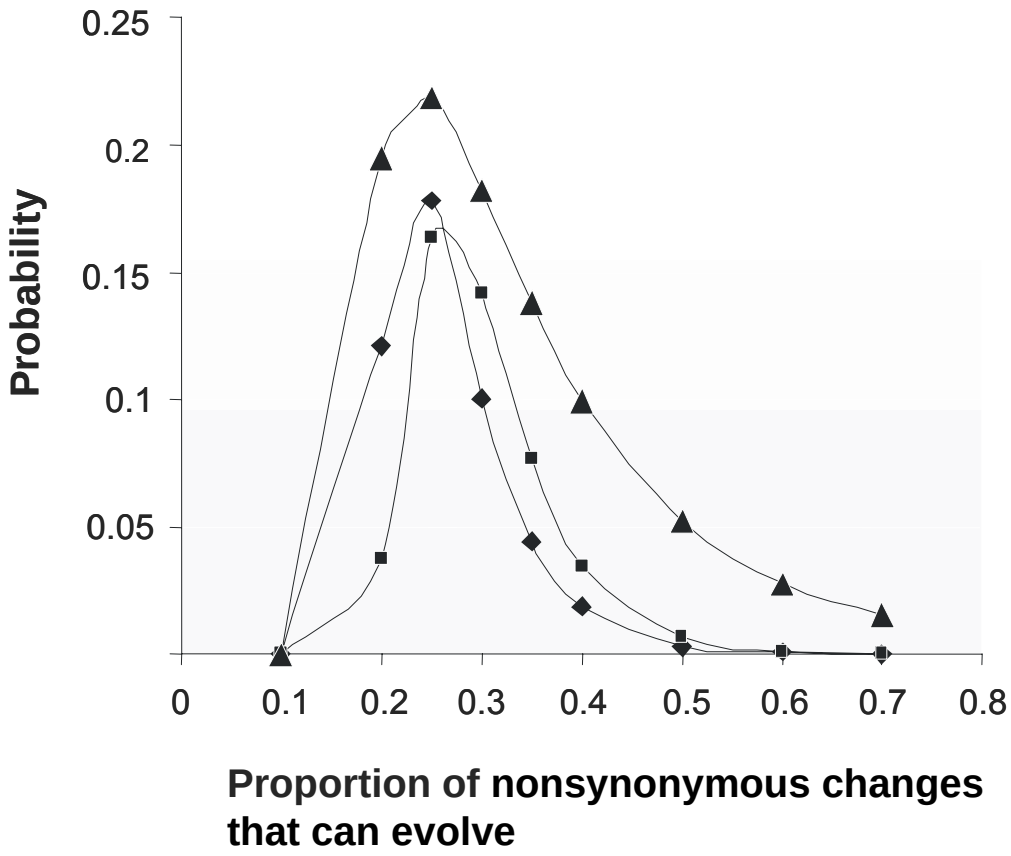

0.25
0.2
0.15
0.1
0.05
Probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Proportion of nonsynonymous changes that can evolve

Fig. 5

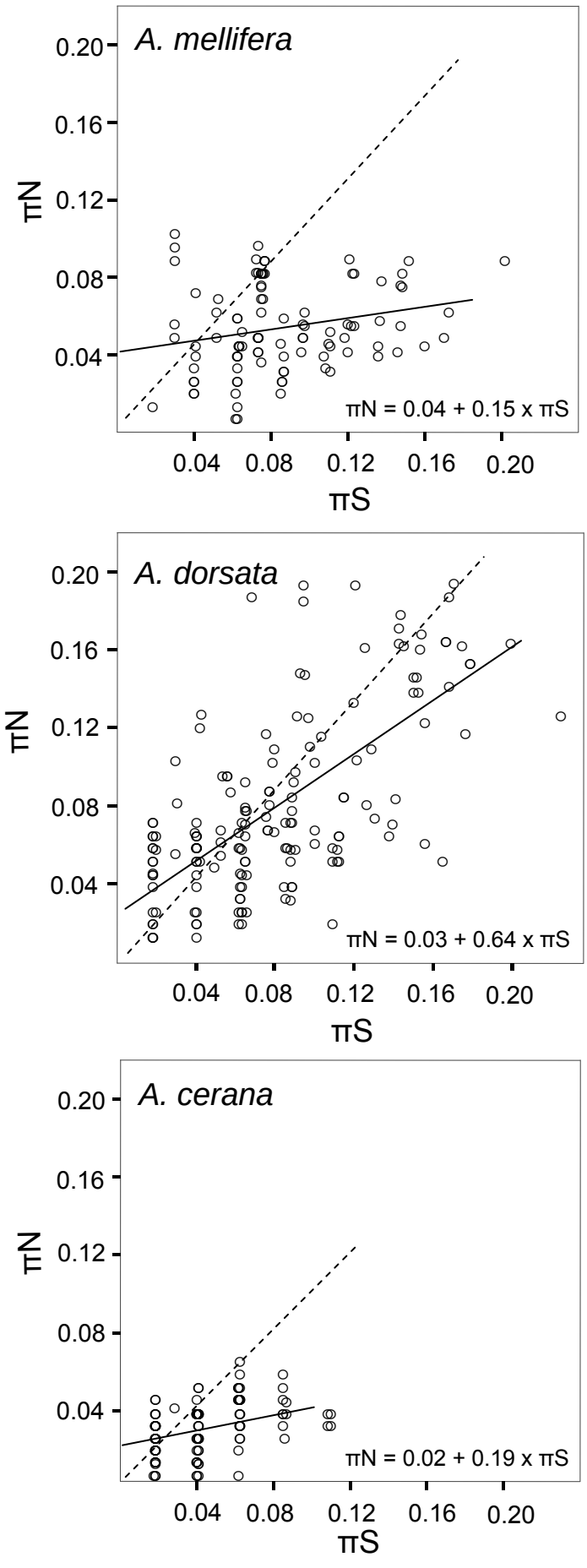
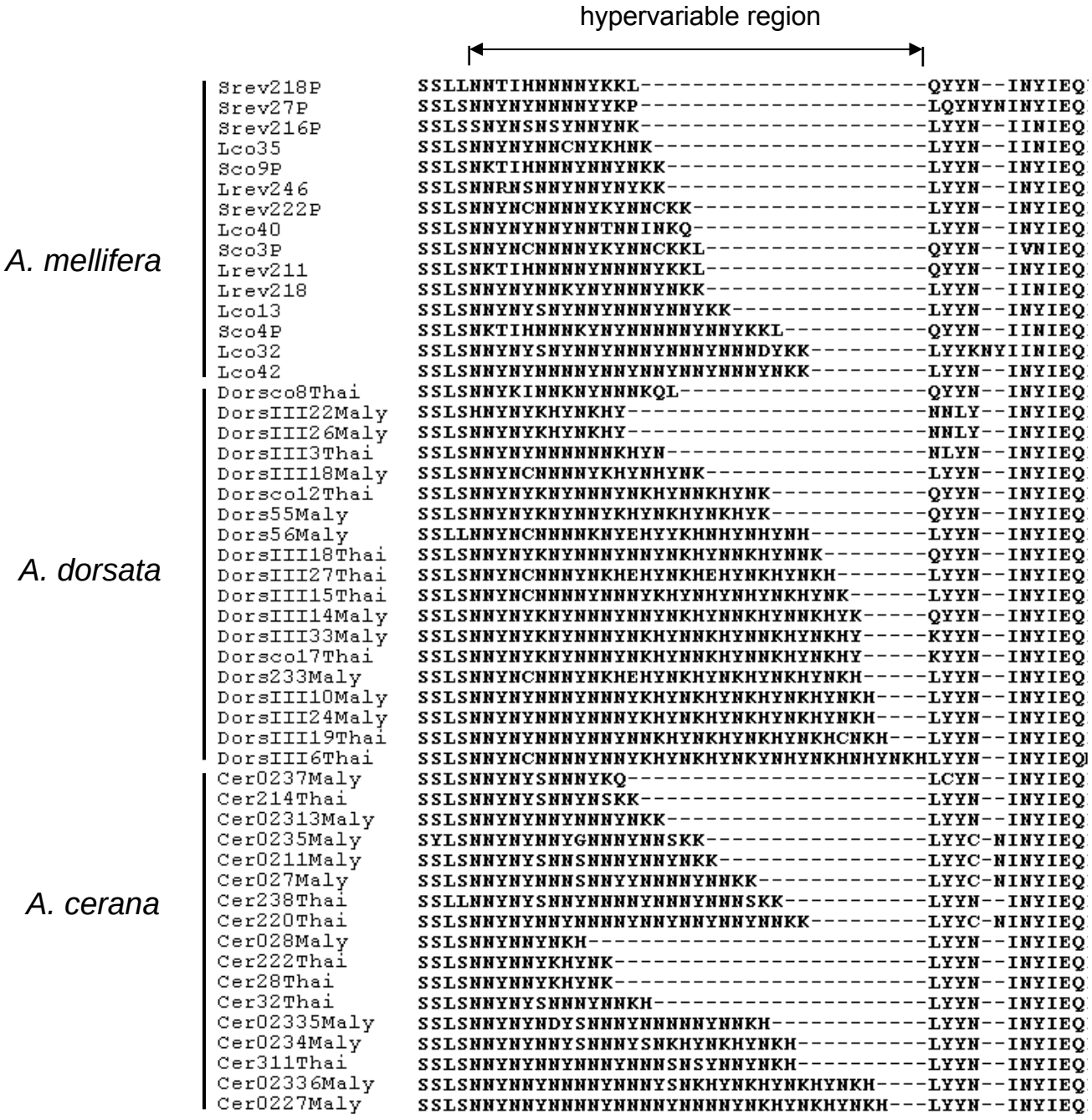

A. mellifera
πN = 0.04 + 0.15 x πS
A. dorsata
πN = 0.03 + 0.64 x πS
A. cerana
πN = 0.02 + 0.19 x πS
πN
πS
0.04
0.08
0.12
0.16
0.20

Fig. 6


hypervariable region
A. mellifera
A. dorsata
A. cerana
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 SSLSNNYNYNNNNYNNYNNYNNYNNNYNKK----------LYYN--INYIEQ
Dorsco8Thai SSLSNNYKINNKNYNNNKQL----------QYYN--INYIEQ
DorsIII22Maly SSLSHNYNYKHYNKHY----------NNLY--INYIEQ
DorsIII26Maly SSLSNNYNYKHYNKHY----------NNLY--INYIEQ
DorsIII3Thai SSLSNNYNYNNNNNNKHYN----------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
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