Heredity (2016) 116, 12–22

www.nature.com/hdy

ORIGINAL ARTICLE

Similar but not the same: insights into the evolutionary history of paralogous sex-determining genes of the dwarf honey bee *Apis florea*

M Biewer[1], S Lechner[2] and M Hasselmann[1]

Studying the fate of duplicated genes provides informative insight into the evolutionary plasticity of biological pathways to which they belong. In the paralogous sex-determining genes *complementary sex determiner* (*csd*) and *feminizer* (*fem*) of honey bee species (genus *Apis*), only heterozygous *csd* initiates female development. Here, the full-length coding sequences of the genes *csd* and *fem* of the phylogenetically basal dwarf honey bee *Apis florea* are characterized. Compared with other *Apis* species, remarkable evolutionary changes in the formation and localization of a protein-interacting (coiled-coil) motif and in the amino acids coding for the *csd* characteristic hypervariable region (HVR) are observed. Furthermore, functionally different csd alleles were isolated as genomic fragments from a random population sample. In the predicted potential specifying domain (PSD), a high ratio of $\pi_N/\pi_S = 1.6$ indicated positive selection, whereas signs of balancing selection, commonly found in other *Apis* species, are missing. Low nucleotide diversity on synonymous and genome-wide, non-coding sites as well as site frequency analyses indicated a strong impact of genetic drift in *A. florea*, likely linked to its biology. Along the evolutionary trajectory of ~30 million years of *csd* evolution, episodic diversifying selection seems to have acted differently among distinct *Apis* branches. Consistently low amino-acid differences within the PSD among pairs of functional heterozygous *csd* alleles indicate that the HVR is the most important region for determining allele specificity. We propose that in the early history of the lineage-specific *fem* duplication giving rise to *csd* in *Apis*, *A. florea csd* stands as a remarkable example for the plasticity of initial sex-determining signals.
Heredity (2016) **116,** 12–22; doi:10.1038/hdy.2015.60; published online 8 July 2015

INTRODUCTION

Duplicated genes may represent a rich source for novel gene function, although our current understanding of the different evolutionary fates leading to duplicated genes with modified gene functions remains incomplete (Innan and Kondrashov, 2010). The single-locus mechanism of complementary sex determination in honey bees provides a revealing example of how new gene functions can arise by duplication and limit the evolution of other genes (Hasselmann *et al.*, 2008a, 2010). The primary signal of the sex-determining pathway in honey bees, the gene *complementary sex determiner* (*csd*), originated by tandem gene duplication from its paralog gene *feminizer* (*fem*) within the honey bee (*Apis*) lineage. The *csd* gene encodes an SR-type protein, harboring an arginine/serine-rich and a proline-rich domain. In addition to these domains and in contrast to *fem*, *csd* is characterized by the hypervariable region (HVR), forming repeated structures of species-specific amino acid motifs that are highly variable in length (Hasselmann *et al*, 2008b). Heterozygous *csd* is required to induce the female pathway by interacting with *transformer 2* (Nissen *et al*, 2012), leading to a female-spliced *fem* transcript; this decision is maintained throughout the development by a positive feedback loop. Homozygous or hemizygous *csd* induces the male pathway, mediated by a truncated Fem protein, which results from an early stop codon in the male *fem* mRNA (Beye *et al*, 2003; Gempe *et al*, 2009). Fertilized eggs, homozygous at *csd*, develop into diploid males that have zero fitness, as they are consumed by worker bees at early embryonic stages (Woyke, 1963). Consequently, rare or newly evolved *csd* alleles that increase in frequency within a population rarely contribute to the formation of diploid males (Yokoyama and Nei, 1979). This rare-allele advantage or negative frequency-dependent selection as one form of balancing selection results in a prolonged persistence time for *csd* alleles that segregate in honey bee populations (Hasselmann and Beye, 2004; Hasselmann *et al*, 2008b).

The *csd* alleles of the different *Apis* species analyzed so far are on an average not older than the corresponding species split, although several *csd* alleles of *Apis dorsata* and *A. mellifera* could have predated speciation, as the divergence of allele pairs exceeds the mean interspecies divergence ($d_S > 0.14$; Hasselmann *et al.*, 2008b). Interestingly, in this previous study, no trans-specific alleles (alleles that are more closely related to an allele from another species than to other alleles from the same species) were detected (Hasselmann *et al.*, 2008b). Trans-species alleles that can be maintained for >30 million years are frequently observed in other systems under balancing selection, such as in two well-studied cases: the major histocompatibility complex of vertebrates (Takahata, 1990) and the self-incompatibility S-locus of plants (Vekemans and Slatkin, 1994). In *Apis*, the relatively short average coalescence time of functional *csd*

[1]Department of Livestock Population Genomics, Institute of Animal Science, University of Hohenheim, Stuttgart, Germany and [2]CeGaT GmbH - Center for Genomics and Transcriptomics, Tübingen, Germany
Correspondence: M Biewer, Department of Livestock Population Genomics, Institute of Animal Science, University of Hohenheim, Garbenstrasse 17, Stuttgart 70599, Germany.
E-mail: matthias.biewer@uni-hohenheim.de
Received 11 December 2014; revised 1 May 2015; accepted 6 May 2015; published online 8 July 2015

alleles (5–7 million years) can be best explained by a high allelic turnover rate, indicating small, long-term effective population sizes. Consequently, a strong historical impact of genetic drift as an evolutionary force might have affected *csd* allele evolution (Hasselmann *et al*, 2008b). A recent comprehensive evaluation of the dynamics of *csd* alleles in *A. mellifera* subpopulations showed that (i) the evolutionary rate giving rise to a new specificity is high and particularly driven by a HVR and that (ii) only few (4–5) amino acid substitutions are sufficient to give rise to a novel functional heterozygotic *csd* specificity (Lechner *et al*, 2014). In another study (Beye *et al*, 2013), the function and molecular evolution of 14 natural variants among 76 genotypes of *csd* were analyzed. The authors found evidence that at least five amino acid differences and length variation of the HVR are sufficient to induce female development. Interestingly, the rise of new specificities of *csd* likely evolves through a series of single mutations leading to incomplete penetrance of the advantageous phenotype (femaleness), supported by an evolutionary intermediate with only three amino-acid length differences.

Our knowledge about the evolutionary history of these paralogous genes is, however, still incomplete. Thus far, the *csd* gene has been intensively studied in three species of the genus *Apis*:*A. mellifera*, *Apis cerana* and *A. dorsata*. The *csd* alleles for each of the three *Apis* species are, on an average, 3% different at the amino acid level, excluding the HVR. In all three *Apis* species, support is given for a common target of balancing selection in *csd*, the potential specifying domain (PSD) located in exons 6 and 7, based on high average diversity at synonymous (π_S) and nonsynonymous (π_N) sites (Hasselmann *et al*, 2008b). For *A. florea* and other Hymenoptera (for example, bumble bees), genome predictions have been used to describe *fem* and its orthologs (Schmieder *et al*, 2012). A putative *csd* ortholog has been reported in *A. florea*, showing exceptionally high-nucleotide diversity exceeding those described for any other *Apis* (Liu *et al*, 2011). However, their results are unexpected and may result from pseudogenic fragments with similarity to *csd* occurring in the genome, which has already given misleading assignments to putative csd alleles in other *Apis* species (Cho *et al*, 2006; Hasselmann *et al*, 2008b).

Previous evolutionary analyses showed that strong positive selection has acted on *csd* after the duplication event, leading to the divergence of *A. mellifera*, *A. cerana* and *A. dorsata*, accompanied by purifying selection for *fem* (Hasselmann *et al*, 2008a). Six amino acid residues associated with the fixation of nonsynonymous substitutions in the phylogenetic branch of *csd* prior to *Apis* species divergence gave rise to a predicted coiled-coil motif in the *csd* protein encoding protein-binding properties (Lupas *et al*, 1991). In close sister lineages of corbiculate bees (for example, bumble and stingless bees), taking available genomic resources into account, no *csd* ortholog could be detected, indicating that the gene duplication of honey bee *csd* and *fem* occurred after the split of stingless, bumble and honey bees (~80 million years ago), but before honey bee divergence, as an independent event (Koch *et al.*, 2014). Recent phylogenetic analyses of molecular and morphological data place the open-nesting dwarf honey bee, *A. florea*, as the most basal clade in the phylogeny of the genus *Apis* (Lo *et al*, 2010). The estimated divergence time of this species from the rest of the genus *Apis* ranges from 29 to 33 million years, establishing *A. florea* as an ideal *Apis* species to gain broader insight into the evolutionary history and the early stage of the paralogous genes *fem* and *csd* (Ramirez *et al*, 2010).

In this study, we analyzed *csd* and *fem* gene sequences of *A. florea* to obtain a comprehensive view on the early evolutionary history of these important genes in bees. We identified the full-length coding sequence and genomic structure of both genes, obtained nucleotide polymorphism data from *A. florea csd* sequences, including nucleotide divergence of functional *csd* allele pairs in *A. florea*, and compared it against the genomic background. Using maximum-likelihood-based evolutionary models, we tested for evidence of positive selection within *csd* of the *Apis* phylogeny. Our data provide comparative insight into the evolutionary processes that have shaped the *csd/fem* complex in the past and show that modification of gene structures may have contributed to the astonishing diversity of primary signals in sex determination pathways.

MATERIALS AND METHODS

Sequence data

A. florea eggs (0–48 h, ~150 eggs) and 20 adult female individuals were initially sampled in February 2009 from a colony in Thailand (Samut Songkram). Because of multiple mating of the queen, these samples have up to 15 different sources of chromosomes derived from as many different fathers (Palmer and Oldroyd, 2000). In a second sampling in February 2012, adult females were collected from 12 colonies in Chom Bueng District (Thailand) within a range of 1.6 km (Supplementary Table 1), and we used two individuals per colony to isolate the maximum-possible number of *csd* alleles (48 chromosomes). As shown in previous studies, high *csd* allelic variability is found within single colonies and localities of *A. mellifera* (Hasselmann and Beye, 2004; Lechner *et al.*, 2014). Therefore, our sampling provides a substantial basis for the evaluation of *csd* allele diversity within *A. florea*. Sets of oligonucleotides were designed based on sequence information that was obtained from eight different 5′ and 3′ rapid amplification of complementary DNA (cDNA) ends (RACE) sequences. The following primers were developed: S-166 5′-CGGTTTCT CTAAGCATATAGGTGA-3′ and S-158 5′-GTCAAGGCTGAGTAATAGTAT TAA-3′ (used for full-length *csd* amplification); consfor 5′-GGTGATT TATACATTTGCAGGT-3′ and A_rev3III 5′-ATTCAGTTCATTATTCA TTATTTGCA-3′ (used for full-length *fem* amplification); S-156 5′-CTCC CGTTCTTCTTTTGATTATCACATT-3′ and S-143 5′-CAGAAGAACGAT TACGACGGA GACGCG-3′ (genomic fragment of *csd* exons 2 and 3); and S-176 5′-CATTTGACCCGCTAGTTGTCCAATCTCG-3′ and S-177 5′-GTTGC AGTAGAGATAGAAATAGAGG-3′ (genomic fragment of *csd* exons 6–9). Full-length *csd* and *fem* sequences were obtained from cDNA from pooled eggs of the 2009 sample. We noticed that our established protocol to identify different csd alleles (Hasselmann and Beye, 2004; Hasselmann *et al*, 2008b) using restriction patterns of full-length cDNA clones out of the pooled eggs failed for technical reasons (no detectable variation in restriction pattern). Samples of 20 female individuals (2009) were used to amplify genomic fragments of *csd* covering the corresponding regions of exons 2+3 and 6–9 by PCR using high-fidelity proofreading DNA polymerase (Phusion: Fermentas, Schwerte, Germany; Q5: New England Biolabs, Frankfurt a.M., Germany) according to the manufacturer's instructions. *csd* PCR fragments were cloned into pGEM-T vectors (Promega, Mannheim, Germany), and the positive clones were subjected to double-strand sequencing (GATC Biotech, Konstanz, Germany). From the 2012 samples, we obtained 32 *csd* exon 6–9 sequences, of which four were duplicates (identical sequences), which we removed in all subsequent analysis.

The genomic structures of *csd* and *fem* were identified through comparisons of full-length open reading frame cDNA sequences and amplified partial gene fragments from genomic DNA (this study) against the genomic resources available for *A. florea* (NCBI genome sequence accession numbers AEKZ01000001-AEKZ01019341, assembled as scaffolds GL575021-GL582965) using BLAST algorithms (blastn, tblastx) implemented in BeeBase (http://hymenopteragenome.org/beebase/?q=blast). Amino acid sequences of *csd* and *fem* were deduced from corresponding cDNA sequences using EMBOSS transeq (http://www.ebi.uk/Tools/st/emboss_transeq).

Seven sets of oligonucleotides were developed to obtain nucleotide polymorphism from the genomic background of *A. florea* using seven presumably neutral evolving loci. Initially, we used pairs of oligonucleotides known to amplify neutral loci in *A. mellifera* and *A. cerana* (Beye *et al*, 2006). In those cases where amplification failed, we used alternative loci deduced from the genome sequences (GeneBank numbers AEKZ01002059.1, AEKZ01002102.1,

AEKZ01002230.1; Supplementary Table 2). All loci were generated with high-fidelity proofreading DNA polymerases and were subjected to direct sequencing.

Sequence data of *csd* alleles from *A. mellifera, A. cerana* and *A. dorsata* obtained in a previous study were used for comparative evolutionary analysis (Hasselmann *et al*, 2008b). For outgroup analyses, we used *fem* sequence data from *Bombus terrestris* (NCBI: NM_001280924.1), *B. impatiens* (NCBI: XM_003493748.1) and *Melipona compressipes* (NCBI: EU139305.1).

Molecular evolutionary analysis of nucleotide and amino acid substitutions

The *csd* sequences were aligned and edited as described elsewhere (Hasselmann and Beye, 2004). Sequence summary statistics in terms of haplotype diversity, number of segregating sites, estimates of nucleotide diversity and neutrality tests (Tajima's D (Tajima, 1989), Fu's F_s (Fu, 1997) and Fay and Wu's H (Fay and Wu, 2000)) were computed in DnaSP 5.10 (Librado and Rozas, 2009). Combining tests for deviation from neutral evolution using the allele frequency spectrum can provide insight into the evolutionary processes within populations. A negative Tajima's D indicates population size expansion or selective sweeps, whereas a positive Tajima's D is associated with a recent bottleneck or overdominant selection. A negative Fu's F_s statistic is evidence for an excess number of alleles, as expected from a population expansion or genetic hitchhiking. Positive values of F_s are evidence for a deficiency of alleles and would indicate recent population bottleneck or overdominant selection. Fay and Wu's H statistic is most sensitive to a signal of a selective sweep, as reflected by the excess of high-frequency polymorphism that results in negative values. DnaSP 5.10 was also used to perform the Hudson–Kreitman–Aguadé test (Hudson *et al*, 1987) for neutral molecular evolution and test for recombination (four gamete test (Hudson and Kaplan, 1985)).

We used the best-fit substitution model application implemented in MEGA 5.1 (Tamura *et al*, 2011) to compare the available nucleotide substitution models and obtain the best description of the substitution pattern by maximum likelihood. The model with the lowest BIC scores (Bayesian Information Criterion) is considered to describe the substitution pattern the best. Non-uniformity of evolutionary rates among sites was modeled by using a discrete Gamma distribution (+G) with five rate categories. By this approach, the Kimura-2 parameter model (with two categories and Gamma = 1.63) was assigned as the best model, and trees were inferred from evolutionary distances by maximum likelihood.

Evolutionary sequence analyses to detect signs of selection in target genes were performed using the datamonkey web server interface (http://www.datamonkey.org), which accesses the HyPhy package (Kosakovsky Pond *et al*, 2005). The HyPhy package performs tests on evolutionary rate differences and signatures of selection using the Branch-Site REL model approach, which uses both branches and sites simultaneously for increased power and accuracy (Kosakovsky Pond *et al*, 2011).

Structural analysis of amino acid composition and motifs

The COIL program (Lupas *et al*, 1991; www.ch.embnet.org/software/ COILS_form.html) was used to search the deduced full-length amino acid sequences of *csd* for predicted coiled-coil regions. The COIL program compares input sequences to a database of known coiled coils and derives a similarity score. The probability that the sequence will form a coiled-coil motif is obtained within the program by comparing the similarity score against the distribution of scores in lobular and coiled-coil proteins. InterPro scan (http://www.ebi.ac.uk/interpro/) was used for additional sequence pattern analyses including coil motifs, domains and signatures in DNA and protein sequences.

RESULTS

fem and *csd* gene analyses in *A. florea*

We identified and isolated the sex-determining genes, *fem* and *csd*, from *A. florea* (*Af-fem* and *Af-csd*) through reverse transcription PCR experiments using RNA from a pool of early (0–48 h) embryonic-stage eggs (~150 eggs). The exon/intron structure of both genes is shown in Figure 1a and was deduced using the genome sequence information of *A. florea*. For the *fem* gene of *A. mellifera*, a female (encoding a full-length protein) and a male transcript (encoding a truncated protein) are known (Hasselmann *et al*, 2008a). We isolated the female splice form of *Af-fem*, consisting of 10 exons and encoding an open reading frame of 401 amino acids. *Af-csd* consists of nine exons, and the deduced open reading frame encodes ~408 amino acids (depending on length variation within the HVR). In *A. florea*, the genomic region covered by exons encoding the *fem* gene is 7 kb long, whereas *csd* spans a genomic region of 8.7 kb located 3.5 kb downstream of *fem*. This is in contrast to *A. mellifera*, for which the distance between *fem* and *csd* is approximately four times as large (~12 kb). The amino acid sequences of *Af-fem* and *Af-csd* deduced from the full-length cDNA indicated a similar domain structure as described previously for Csd and Fem protein in other *Apis* species (Hasselmann *et al*, 2008a, Figure 1b). Csd protein harbors a arginine-serine (RS)-rich (red) and proline (P)-rich (blue) region corresponding to exons 6–9, which are flanking a HVR (green). Fem protein is characterized by two RS-rich domains and a P-rich domain and lacks a HVR.

The small physical distance of 3.5 kb between *Af-fem* and *Af-csd* located on a single scaffold segment provides strong support for the existence of a single sex determination locus in *A. florea*, which is consistent with the genome structure described for *A. mellifera*. Nucleotide sequence analyses on full-length open reading frames of both genes using BLAST against available genome sequence resources confirmed the instances as single-copy genes of each, verified by highly significant single matches (score: 166, *e*-value 10^{-169}).

Differences between the *A. florea* csd protein and those of other *Apis* species

To elucidate the functional characteristics of *A. florea csd* alleles, we further analyzed the Csd protein sequence. A striking characteristic of all *csd* genes is the HVR. This region has been shown to be highly variable in length and composition among *csd* alleles within the same *Apis* species (Beye *et al*, 2003; Hasselmann *et al*, 2008b). The amino acid residues flanking the HVR are highly conserved within a single *Apis* species. A cross-species comparison reveals that several of those amino acids are also conserved, allowing an unambiguous alignment shown in Figure 1b. Remarkably, the HVR of *A. florea* consists only of asparagines together with some single serine residues (Figures 1b and 2a). This structure contrasts with the recurrence of characteristic motifs found in other *Apis* species ($N_{1–5}Y$ for *A. mellifera*, $KHYN_{1–4}KH$ for *A. cerana* and *A. dorsata*; Figure 1b, Hasselmann *et al*, 2008b). By isolating both *csd* alleles from *A. florea* females, we were able to compare pairs of functional alleles forming a heterozygous *csd* gene. In all the pairs of heterozygous *csd* alleles analyzed, the HVR between the two chromosomes differed in length (1–18 amino acids of length difference) and composition (serine and asparagine residues; Figure 2a). To gain additional insight into putative functional differences of *A. florea* Csd compared with other known Csd proteins in *Apis*, we searched for coiled-coil motifs in *A. florea*. Coiled-coil motifs are known to be involved in protein–protein interactions (Lupas *et al*, 1991) and have been proposed to contribute to the interaction between Csd proteins (Hasselmann *et al*, 2008a). First, we focused on the region in which the coiled-coil motif is predicted for other *Apis* species (Figure 1a, marked with an asterisk). In this particular region, we found among the six amino acids in Csd that are conserved within the motif of *A. mellifera, cerana* and *dorsata* (Figure 2b, black box) one amino acid residue that was different in *A. florea* (Valin (V) instead of glutamate (E)), leading to a strongly decreased probability for a coiled-coil formation in *A. florea* (>98% for *A. mellifera, cerana* and *dorsata* vs <92% for *A. florea*). Interestingly, a highly significant prediction of a coiled-coil motif was found in

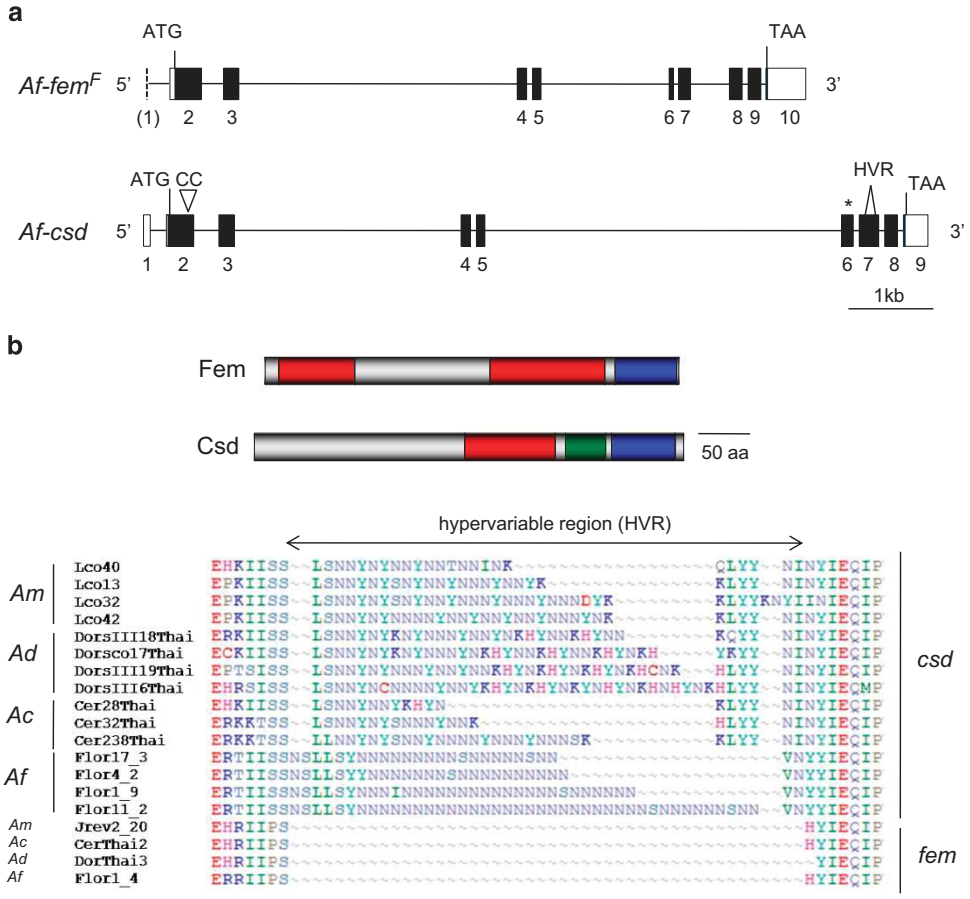

Figure 1 Genomic structure of the *fem* (*Af-fem*^{F}, female splice variant) and *csd* (*Af-csd*) genes in *Apis florea*, domain diagram of Csd and Fem and conceptual translations spanning the hypervariable region (HVR) of four honey bee species. (**a**) Coding exons are marked in black; untranslated regions are in white. Translational start and stop sites are indicated. In the *csd* gene, the position of the HVR is marked (exon 7). * indicates the position of the predicted coiled-coil motif in *A. mellifera, A. cerana, A. dorsata*, but missing in *A. florea* (see Figure 2b). In *A. florea*, a coiled-coil motif is predicted in exon 2 (marked as CC). (**b**) Protein domains of Csd and Fem (above): arginine-serine domains are indicated in red, the HVR in green and the proline-rich region in blue. Below: conceptual alignment of deduced amino acids residues of the HVR in *csd* and the corresponding region of *fem*. Am, *A. mellifera*; Ad, *A. dorsata*; Ac, *A. cerana*; Af, *A. florea*.

the N-terminal region (position aa 46–70; Figure 1a, marked with a triangle; Figure 2c) located ~240 amino acids upstream compared with the region analyzed first. No coiled-coil motif has been predicted within the N-terminal region of Csd in any other *Apis* species studied, likely due to the absence of two amino acids, displayed as gaps in the alignment, and a single amino acid substitution of leucine (L) for arginine (R; Figure 2c, black box). In addition to this finding, one N-terminal coiled-coil motif at the corresponding position to Csd (aa 46–70) was found for Fem in *A. florea*, but not in the other three *Apis* species (98.9% probability for *A. florea* vs <65% for *A. mellifera/A. cerana/A. dorsata*; Figure 2c, black box), accompanied by a lineage-specific deletion of two amino acids. No such deletion is found in any other Hymenopteran non-*Apis* outgroup species we tested (*B. terrestris*, *B. impatiens*, *M. compressipes*), and no coiled-coil motif is predicted for these species (data not shown). Consequently, the additional coiled-coil motifs for *Af*-Csd and *Af*-Fem in *A. florea* suggest an altered position for protein–protein interaction among these molecules in the process of sex determination compared with other *Apis* species. In the early phase of the evolution of paralogous genes within the *Apis* lineage, we observed a transition in primary structure resulting in modifications of protein structure and function.

Reduced variability and evolutionarily young *csd* alleles in *A. florea*

We obtained nucleotide polymorphism data from exons 6–9 of *Af-csd* from females of 1 colony sampled in 2009 and 12 colonies sampled in 2012 to analyze the molecular evolutionary history of the *csd* alleles in *A. florea* (Table 1). We compared these data with the nucleotide polymorphisms of seven non-coding, unlinked loci obtained from 12 individuals of different colonies examined in 2012 (Table 2), as well as with data on *Af-csd* exons 2+3 and *Af-fem* exons 2–7 obtained from a subsample of these colonies (Supplementary Table 3).

Nucleotide diversity (π) in exons 6+7 of *Af-csd* significantly exceeded the diversity in exons 2+3 of *Af-csd* for both data sets ($\pi_{ex2+3} = 0.00248$ vs $\pi_{ex6+7} = 0.00579$ (2009), *Z*-test, $P < 0.001$ and $\pi_{ex6+7} = 0.00675$ (2012), $P < 0.001$). Nucleotide diversity in exons 6+7 of *Af-csd* also exceeded the diversity in *Af-fem* ($\pi_{ex6+7fem} = 0.00135$; $P < 0.001$ (2009) and $P < 0.001$ (2012)). Remarkably, intron diversity within *Af-csd* exceeded those for synonymous sites ($\pi_{intron\_combined} = 0.0094$ vs $\pi_{s\_combinedPSD} = 0.0035$, $P < 0.05$), which are both surprisingly low when compared with the corresponding values from other *Apis* species (for example, *A. mellifera* $\pi_{syno\_PSD} = 0.07$–0.09, Lechner *et al.* (2014)).

A comparison of nonsynonymous nucleotide diversity (π_n) between the two exonic regions of the *csd* gene showed a higher π_n for exons 6

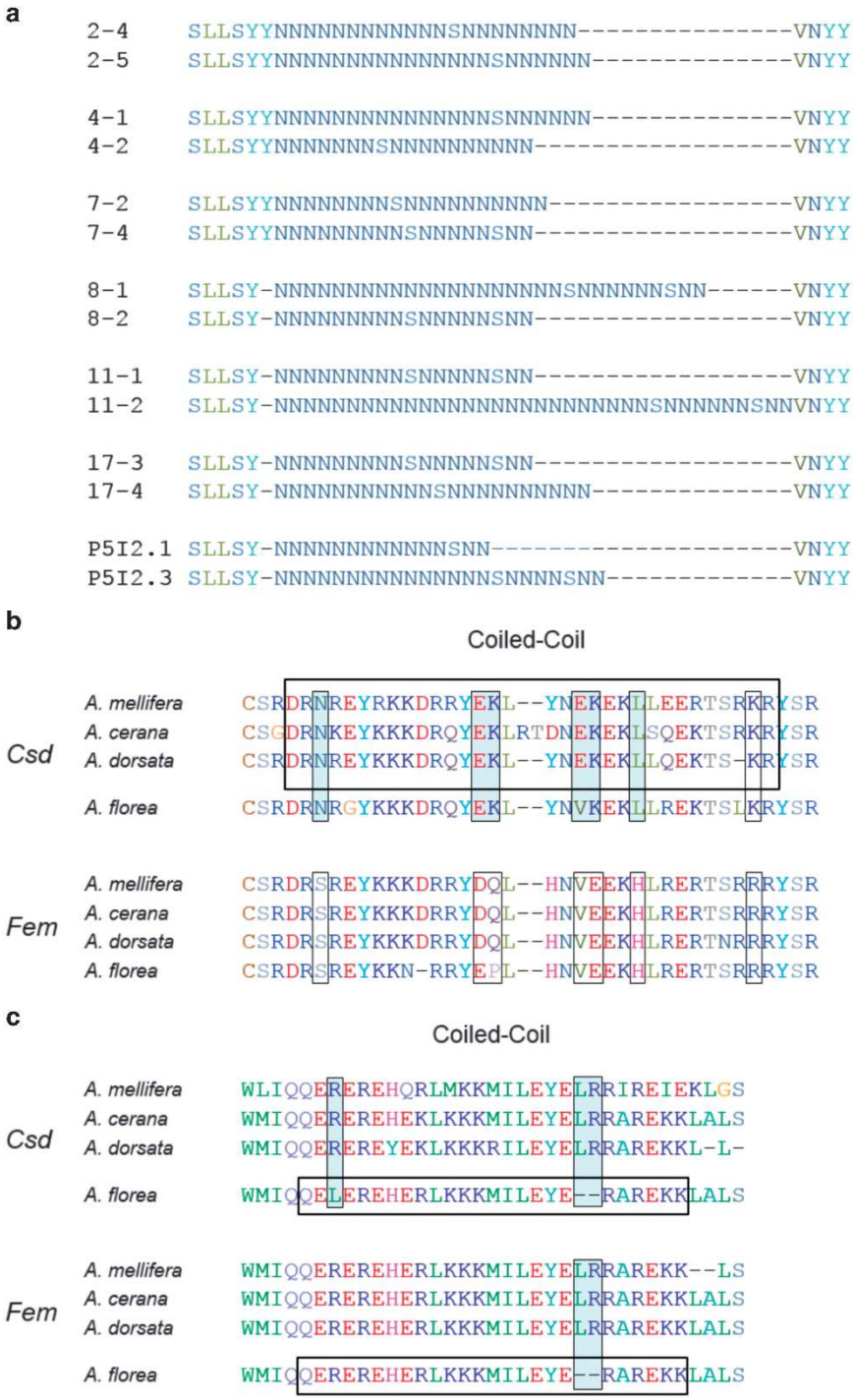

Figure 2 Characteristics of Csd and Fem protein in *A. florea*. (**a**) The aligned conceptual translation of the hypervariable region (HVR) between seven pairs of functionally different csd alleles of females is shown. The HVR encodes a succession of asparagine residues (N) and single serine residues (S). (**b**) Deduced amino acid sequence encoding the region of the predicted coiled-coil motif, framed in black. For *A. mellifera*, *A. cerana* and *A. dorsata csd,* the prediction is based on the six amino acids in gray shaded boxes. *A. florea* differs in one position compared to the other species and does not form a coiled-coil motif (unframed). The homologous region for *fem* is placed below for comparison, with the corresponding amino acids marked. (**c**) Deduced amino acid sequence encoding the region of the predicted N-terminal coiled-coil motif for *A. florea* framed in black. Only one amino acid change and a two amino acid difference shown as a gap (shaded boxes) seem sufficient for the formation in Csd and Fem in *A. florea*, whereas this is not found in the other *Apis* species (unframed).

+7 than exons 2+3 ($P<0.05$) in both the 2009 and 2012 samples. We noted that the diversity of nonsynonymous sites exceed those of the synonymous sites (Fisher's exact test, $P<0.01$) in the region of exons 6 +7, which is more pronounced in the 2009 data than in the 2012 data (2009: $\pi_n/\pi_s=5.71$; 2012: $\pi_n/\pi_s=1.17$). These findings are in agreement with our previous results, establishing the region of exon 6+7 as a common target of balancing selection and predicting the specifying domain (*csd*-PSD) in the *Apis* species. The ratio $\pi_n/\pi_s=1.6$ of the

Table 1 Summary statistic of csd sequences of *Apis florea*

	N	H_d	n	S	π	*csd*		π_n	π_n/π_s	D	D_s	D_n	F_s	H
						Θ_W	π_s							
Sample 2009														
Exon 6–9	23	0.68	438	15	4.78±0.1	9.28±3.7	2.42	5.61	2.32	−1.73*	−1.73	−1.51	−3.81*	−3.04*
PSD (6+7)	23	0.58	273	10	5.79±1.4	9.92±4.4	1.28	7.31	5.71	−1.4	−1.16	−1.28	−1.66	−1.38
Intron	23	0.79	201	7	9.56±1.6	9.44±4.5	—	—	—	0.04	—	—	−1.99	—
Sample 2012														
Exon 6–9	28	0.88	444	9	4.90±0.5	5.21±2.3	2.83	5.71	2.02	−0.14	0.29	−0.23	−6.57**	−3.15*
PSD (6+7)	28	0.85	279	7	6.75±0.7	6.45±3.1	4.52	5.29	1.17	0.14	0.29	−0.39	−3.65*	−1.3
Intron	28	0.69	201	5	8.1±0.9	6.39±3.4	—			0.75	—	—	0.61	—
comb. data														
Exon 6–9	51	0.84	423	19	5.15±0.6	9.98±3.5	2.87	5.95	2.07	−1.53*	−1.50*	−1.33	−11.97**	−5.66**
PSD (6+7)	51	0.78	258	13	6.87±0.8	11.20±4.3	3.45	5.46	1.58	−1.16	−0.9	−1.50*	−3.34*	−1.9
Intron	51	0.77	200	8	9.36±0.9	8.89±3.9	—	—	—	0.14	—	—	−1.39	—

Abbreviations: D, Tajima´s D; F_s, Fu' s F_s; H, Fay and Wu's H statistic; H_d, haplotype diversity; N, sequence number; n, number of nucleotides; S, number of segregating sites; π, average pairwise nucleotide diversity for all; s and n, synonymous and nonsynonymous sites; Θ_W, Watterson's Theta.
Values for π and Θ_W are given ×1000. PSD: potential specifying domain. Intron sequences are from csd region exon 6–9.
*$P<0.05$, **$P<0.01$.

Table 2 Summary statistic of unlinked, neutral loci sequences of *Apis florea*

	N	H_d	n	S	*Neutral loci*		D	F_s	H
					π	Θ_W			
Locus 2	12	0.17	346	1	0.48±0.4	0.96±0.9	−1.14	−0.48	0.15
Locus 4	12	0.62	223	2	3.19±0.8	2.97±2.3	0.22	0.03	0.51
Locus 8	11	0.82	263	7	8.16±1.6	9.09±4.8	−0.42	−2.65*	0.09
Locus 9	12	0.41	195	1	2.10±0.7	1.70±1.7	0.54	0.74	0.27
Locus 11	12	0.82	316	5	4.22±0.9	5.24±2.9	−0.72	−2.46*	—
Locus 12	12	0.8	317	5	4.35±0.9	5.22±2.9	−0.62	−1.11	—
Locus 13	12	0.65	544	3	1.42±0.4	1.83±1.2	−0.73	−1.17	—
Average		0.61		3.4	3.42	—	—	—	—

Abbreviations: D, Tajima´s D; F_s, Fu' s F_s; H, Fay and Wu's H statistic; H_d, haplotype diversity; N, sequence number; n, number of nucleotides; S, number of segregating sites; π, average pairwise nucleotide diversity for all; Θ_W, Watterson's Theta.
Values for π and Θ_W are given ×1000.
*$P<0.05$.

combined data set is still remarkably higher in *A. florea*-PSD when compared with the three other *Apis* species with ratios of $\pi_n/\pi_s<1$, ranging from 0.6 to 0.95 for *csd*-PSD, due to an accumulation of synonymous polymorphism over time in obviously older alleles (Hasselmann *et al*, 2008b). Consequently, as we observe π_n/π_s ratios >1 for *A. florea*, we conclude that csd alleles in *A. florea* populations are comparatively young.

We used polymorphism data from seven non-coding, presumably neutrally evolving loci, to gain insight into the genome-wide nucleotide diversity of *A. florea*. On an average, π_{genome} is 0.00341 (ranging from 0.00048 to 0.00816, Table 2), which is almost identical to the diversity of synonymous sites in *Aflor*-csd-PSD ($\pi_{syno\_combinedPSD}=$ 0.00345), indicating the genetic drift as the dominant evolutionary force affecting *csd* allele evolution in *A. florea*.

To gain insight into population dynamics, we determined whether the observed nucleotide polymorphisms might be affected by demographic effects by analyzing their frequency spectrum. Population growth is known to result in a biased frequency spectrum of polymorphism, which would be reflected in negative values of Tajima's D. For all data sets, except *csd* intron sequences, Tajima's D-values were negative or close to zero. The values for synonymous sites at *csd* exons 6–9 (combined data: $D_s=-1.50$), nonsynonymous sites at *csd*-PSD ($D_n=-1.50$, Table 1) and overall ($D=-1.53$) were significant ($P<0.05$), which might indicate an expanding population or enhanced fixation of nonsynonymous substitutions within the *csd* gene (selective sweep). We applied Fu's F_s statistic, which is known to be particularly sensitive for recent population expansions, based on the number of haplotypes observed in the sample (Fu, 1997). Fu's F_s was significant for *csd* exons 6–9 ($P<0.01$) and *csd*-PSD ($P<0.05$, combined data). These values support Tajima's test and led us to discriminate the potential effect of population expansion from an alternative scenario, which would be a selective sweep, by applying Fay and Wu's H-statistic (Fay and Wu, 2000). Fay and Wu's H-statistic was significant for *csd* exons 6–9 ($H=-5.66$, $P<0.01$), but marginally insignificant for csd-PSD ($H=-1.9$). Combined, our results suggest that the observed *csd* alleles belong to an expanding population after a bottleneck event, with an advantage for new alleles, and a fixation of haplotypes that have recently arisen.

The analyses of the non-coding loci revealed that Tajima's D values were close to zero in all loci, which is expected from neutral evolution

following mutation–drift equilibrium. To test for the signs of population expansion in neutral loci, we applied Fu's F_s-test. In two of the seven loci, we observed significant values for F_s ($P<0.05$), providing weak indications of an expanding population after a bottleneck in these populations, as was found in the *csd* samples. Fay and Wu's H-statistic was close to zero, as expected for these unlinked neutral loci.

To relate our findings to the evolutionary time, which would be necessary for the rise of new *csd* specificities within *A. florea*, we followed the approach based on the neutral mutation rate per site per generation ($\mu = 2.16 \times 10^{-9}$; Lechner *et al*, 2014). The number of generations per year in *A. florea* can vary from one to three, depending on the reproduction cycles linked to swarming activities that can be induced by variation in food resources and climatic conditions (Hepburn, 2011). Using a model for neutral evolving nucleotides, $\pi_s = 4N\mu$ (using $\pi_s = 0.00345$), with the above average neutral mutation rate per site per generation, we obtain $N = 4 \times 10^5$ generations to accumulate these differences. Under this assumption, the expected number of nonsynonymous substitutions within the PSD region is 1 ($N\mu \times$ the number of nonsynonymous sites (209)), which equals the number of amino acid changes (1–3) found among *A. florea csd* alleles using the corresponding π_n values ranging from $\pi_n = 0.0048$ to 0.0145. Based on this result, we calculated an origination rate (u) of new *csd* alleles per gene per generation of about $u = 4.5 \times 10^{-7}$. If we assume that two amino acid differences within the PSD are needed to generate a new specificity, our results suggest the formation of a new *csd* allele in *A. florea* every 400 000–800 000 generations. As mentioned before, *A. florea* may swarm (and reproduce) up to three times a year. Consequently, the time necessary for a new *csd* specificity to arise may range from 130 000 (three swarmings) to 800 000 (one swarming) years.

To exclude the alternative scenario of an elevated mutation rate in *A. florea csd*-PSD, indicated by the high ratios found for π_n/π_s, we performed the Hudson–Kreitman–Aguadé test. We compared the pattern of polymorphism of *csd*-PSD and the unlinked neutral locus 4 within *A. florea* to the divergence data from *A. dorsata*. The Hudson–Kreitman–Aguadé test detected no significant deviation in the number of synonymous sites from a constant ratio of polymorphism to divergence in *csd*-PSD and locus 4. However, we are aware of the fact that the Hudson–Kreitman–Aguadé test is only valid in the absence of recombination within and free recombination between the loci, which we cannot completely rule out to be the case here.

To obtain further insight into the divergence of the PSD among *csd* alleles, we calculated the nucleotide differences for 14 pairs of functionally different *csd* alleles separately. The pairwise calculations of synonymous and nonsynonymous substitutions between functional *csd*-PSD alleles are given in Table 3. Notably, when the numbers of amino acid differences excluding the HVR were compared among these pairs, we found five (out of 14) pairs with no amino acid differences in the HVR flanking region, but with changes within the HVR ranging from 2 to 17, providing evidence for the importance of the HVR in determining *csd* specificities.

Combined, our comparative sequence analyses of *csd* exons 6–9, *csd*-PSD and the neutral loci suggest that very few (1–2) nonsynonymous amino acid changes, basically located within the HVR, are sufficient for generating new functional csd alleles in *A. florea*.

Evolutionary history of sex-determining genes in *A. florea*

To explore the evolutionary history of *Af-csd* and *Af-fem* genes, we compared their genealogical relationship to other *Apis* species using a subset of the examined alleles (Figure 3). The *csd* alleles fall into separate clades according to their species origin, which is supported by high bootstrap values (>99). No trans-specific allelic pattern appeared in the genealogy as a result of adding the *A. florea* data, which is consistent with the findings of the previous studies from 2008. Next we applied the Branch-Site REL model (see Materials and methods section) to test for an excess of nonsynonymous over synonymous substitutions along phylogenetic branches. We used this model to circumvent the problem, where with one foreground branch and treating other branches as background leads to an exclusion of important sites, reducing the statistical power of the likelihood ratio tests, as would happen in branch-site two tests.

Table 3 Synonymous (d_S) and nonsynonymous (d_N) differences per site and number of amino acid differences in functional pairs of csd-PSD alleles

Allele 1	*Allele 2*	d_s	*s.e.*	d_n	*s.e.*	*Amino acid differences*
P5I2.1	P5I2.3	0.009	0.009	0.003	0.003	1 (10)
P4I1.1	P4I1.2	0.000	0.000	0.000	0.000	0 (13)
P5I1.1	P5I1.3	0.000	0.000	0.000	0.000	0 (2)
P1I1.1	P1I1.3	0.000	0.000	0.006	0.004	2 (5)
P12I1.1	P12I1.4	0.009	0.009	0.003	0.003	1 (8)
P3I1.3	P3I1.4	0.000	0.000	0.006	0.004	2 (8)
P1I2.1	P1I2.2	0.000	0.000	0.003	0.003	1 (1)
P7I1.1	P7I1.2	0.000	0.000	0.003	0.003	1 (1)
P7I2.1	P7I2.6	0.000	0.000	0.006	0.004	0 (5)
P9I1.3	P9I1.4	0.000	0.000	0.000	0.000	0 (6)
P10I1.1	P10I1.2	0.000	0.000	0.003	0.003	1 (1)
P3I2.1	P3I2.3	0.000	0.000	0.003	0.003	1 (1)
P6I2.1	P6I2.2	0.000	0.000	0.006	0.004	0 (17)
P8I1.1	P8I1.1	0.000	0.000	0.003	0.003	1 (3)

The hypervariable region (HVR) was excluded in the analysis. Number of amino acid differences including the HVR are given in parenthesis.

As the Branch-Site REL method uses unrestricted combinations of selective regimes across sites and branches, we detected different branches with significant signs of episodic diversifying selection for *A. mellifera* (a*), *A. dorsata* (b*) and *A. cerana* (c*), whereas for *A. florea*, these signs were not detected (d*, $P=0.5$; Figure 4 and Table 4). The branch leading to the *fem* sequences (e*) displays signs of purifying selection, as known from previous studies. Interestingly, the most basal *csd* branch (g*) exhibits a rather weak ($d_n/d_s = 0.6$) but significant ($P<0.001$) sign of diversifying selection, likely influenced by the contrasting evolutionary forces acting on *csd* alleles in *A. florea* and the other *Apis* species.

DISCUSSION

Evolutionary differences and functional conservation of the *csd/fem* gene complex in *Apis*

The identification and analyses of the sex-determining genes *csd* and *fem* in *A. florea* broaden our understanding of the evolutionary dynamic of sex determination in the *Apis* lineage. We found marked differences in the number of polymorphisms within csd alleles (Table 1) and in Csd protein regions in *A. florea* (Figure 2) compared with the three bee species *A. mellifera*, *A. cerana* and *A. dorsata*. Our results suggest that alongside various evolutionary forces, the underlying biology of the open-nesting *A. florea* may explain the structural and sequence-based differences that exist when compared with other *Apis* species.

We found evidences that only few amino acid differences between *A. florea csd* alleles seem to be necessary and sufficient to encode functional heterozygotes (Table 3, Figure 2). Considering the five allelic pairs that show no amino acid differences except for those located within the HVR (Table 3), we suggest that the HVR alone may

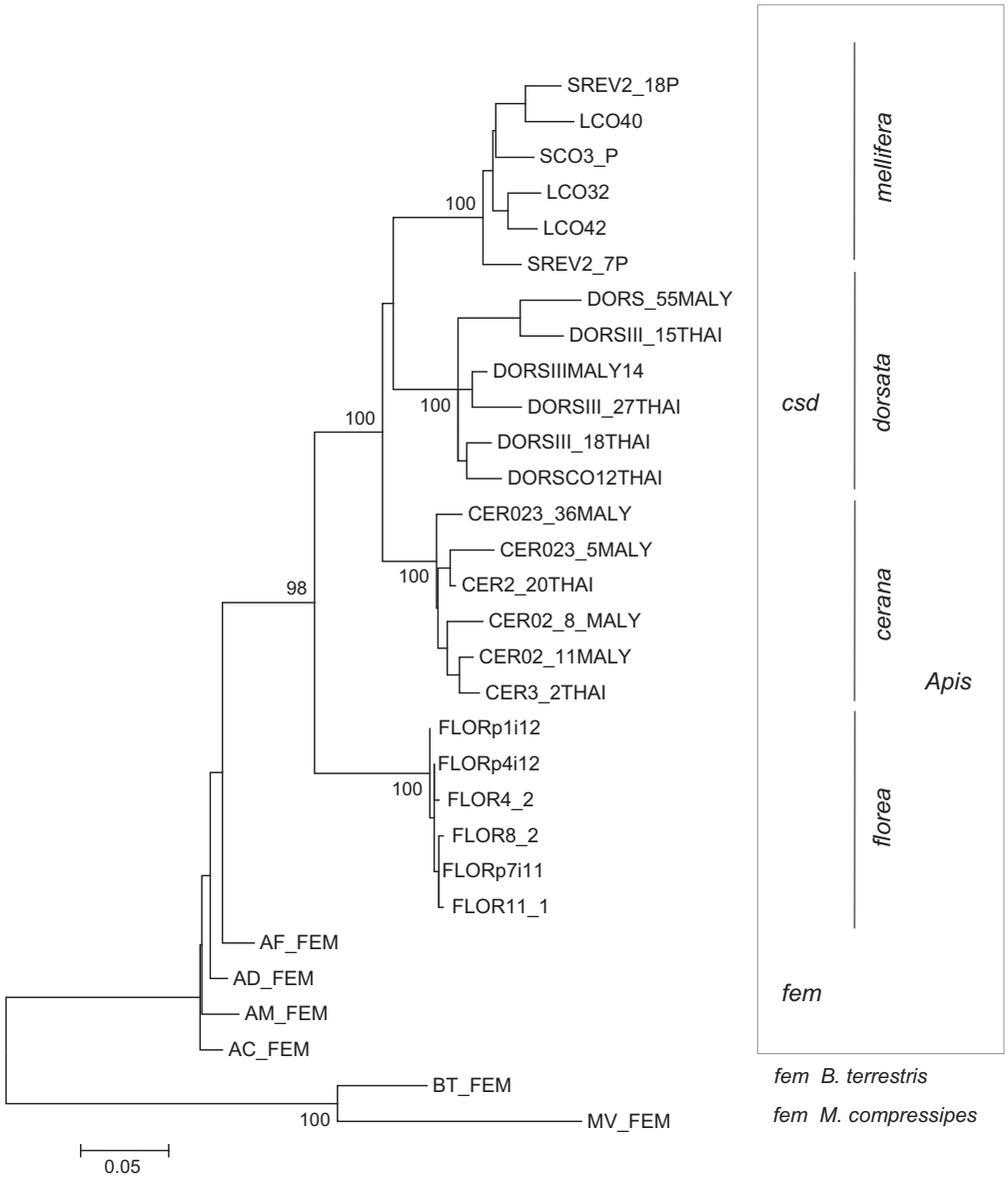

Figure 3 Phylogenetic relationship and substitutions along branches for *csd* and *fem* sequences. The evolutionary history was inferred using maximum likelihood based on the Kimura-2 parameter model. The tree with the highest log-likelihood (−1631.25) is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The evolutionary distances are given in units of the number of nucleotide substitutions per site. A discrete Gamma distribution was used to model the evolutionary rate differences between sites (two categories (+*G*, parameter = 1.63)). The analysis involved 30 nucleotide sequences. The codon positions included were first+second. All positions containing gaps were eliminated. There were a total of 232 positions in the final data set.

encode allelic specificity to initiate female development. Although we have not analyzed the full-length coding sequence of *csd* for these specific allele pairs, based on the low diversity in the other exons (Table 1), it seems unlikely that substantially more amino acid changes have accumulated in the regions other than the PSD region of Csd. The analysis of functionally different *A. mellifera csd* alleles indicates that on an average, no reduced variability to the extent found for *A. florea* exists within *A. mellifera csd*, as they are an order of magnitude more diverse ($\pi_{\text{Amell\_csd}} = 0.07$ vs $\pi_{\text{Aflor\_csd}} = 0.007$). Therefore, the appearance of low-diversity alleles in *A. florea* seems more the rule than the exception when compared with other *Apis* species.

Based on the amino acid composition within the HVR, asparagine (N) is the predominating residue. No repeated motifs composed of other amino acid residues, as have been observed for the HVR of other *Apis* species (for example, $((N)_{1-5}/Y)$ or $(KHYN)_{1-4})KH$), were found in *A. florea csd*. However, single serine residues (S) are consistently found at variable positions in pairs of functionally different *csd*–PSD alleles obtained from female individuals, indicating an as yet unknown pattern of these serines within the HVR. Serine is commonly phosphorylated and exhibits binding properties, including the capacity to establish stable protein structures (Betts and Russell, 2003). Thus, it may be speculated that the formation of the Csd protein leads to structures in the HVR that can directly mediate *csd* allelic specificities in *A. florea*. Support for this hypothesis is given by a *csd* allele pair in the study of *A. mellifera* (Beye *et al.*, 2013) containing three amino acid differences within the HVR only that trigger female development. In

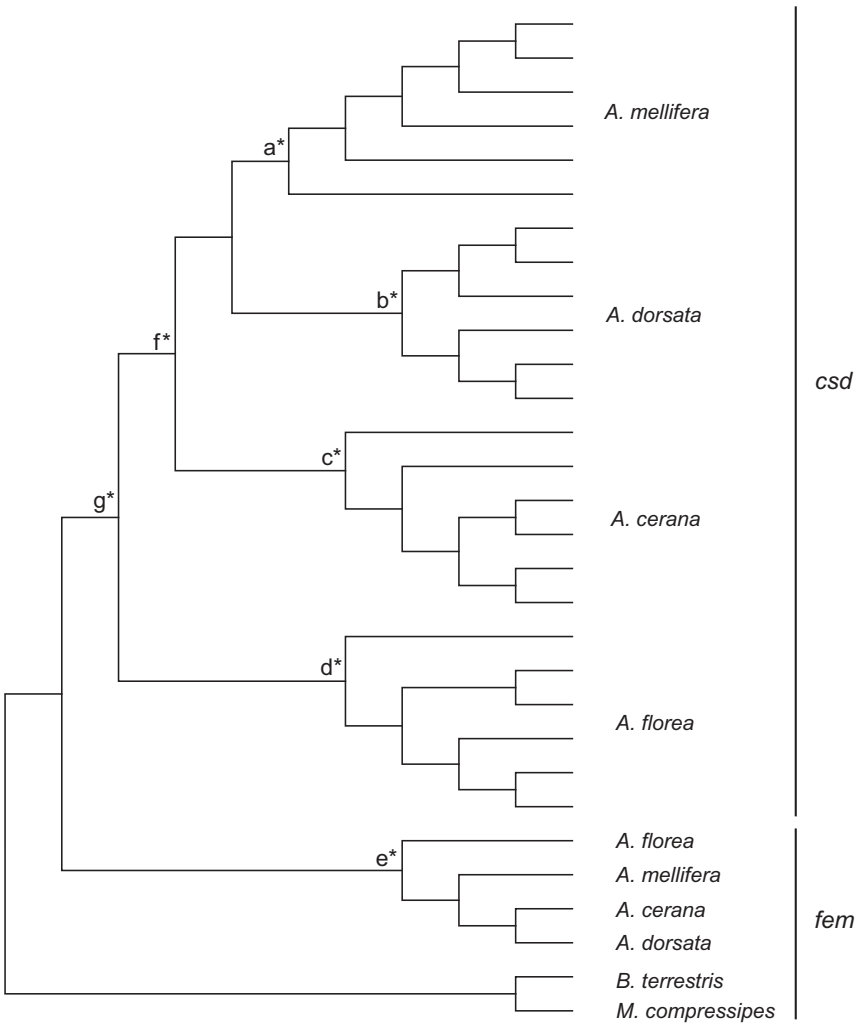

Figure 4 Topology representing *fem* and *csd* evolution in *Apis*. Marks a*–g* represent the branches leading to the different clades of *Apis* (a*–d*) and the internal branches (f* and g*) within *csd*. e* marks the branch of *Apis-fem.* Strong signs of diversifying selection are given for a*–c*but not for d* (see Table 4 for corresponding results of the Branch-Site REL model. *Fem* (*e) shows strong signs of purifying selection.

Table 4 Summary of branches (as marked in Figure 4) under episodic diversifying selection identified by the random effects Branch-Site REL model

Branch	*Mean* d_N/d_S	*LRT*	P-*value*
a*	4.758	5.73	0.008
b*	3.662	10.85	4.92E−04
c*	10	15.55	4.02E−05
d*	0.201	−0.73	0.5
e*	0	0	1
f*	1.46	0.42	0.26
g*	0.61	10.98	4.61E−04

Likelihood Ratio test (LRT) and statistical significance were obtained by using the HyPhy package.

addition, the interaction of proteins produced from different *csd* alleles in *A. florea* may be altered by modification of the coiled-coil motif compared with other *Apis* species (Figures 2b and c). The rise of functional protein domains and structures known to be ubiquitous is documented as being rapid and diverse in manner (Ponting and Russell, 2002). Therefore, a coiled-coil motif may likely change in composition and position in orthologs among species but yet maintain its function. We detected a coiled-coil motif in the N-terminal region of *Af*-Csd (Figure 1a, marked with a triangle) and in the corresponding region in *Af*-Fem (up to 98% probability), but not within the *csd*-PSD region of *A. florea* (Figure 2). Our analysis indicates that in the other *Apis* species, no such coiled-coil can be formed in the corresponding regions of *Af*-Csd and *Af*-Fem. This motif follows the principle of a 7-11-7 coiled-coil motif (the formation of helices in which seven residues flank eleven residues), which contrasts with the motif located within the PSD in other *Apis* species, for which a 4-3-4 type (four residues flank a three residue unit in a helical formation) has been described, which could lead to an altered protein-binding behavior like binding strength or association (Lupas *et al*, 1991; Hicks *et al*, 1997).

The analyses of nucleotide polymorphisms and their frequency spectrum indicate various evolutionary forces acting on *A. florea csd*. We detected low-average nucleotide diversity for both nonsynonymous and synonymous sites, whereas the ratio of $\pi_N/\pi_S = 1.6$ indicates adaptive evolution acting within the region of exons 6+7. This is supported by negative values of Tajima's *D* and Fay and Wu's H-statistic, and a high ratio of π_N/π_S, suggesting positive selection in this particular region, known to be a common target of selection and a PSD (Hasselmann *et al*,

2008b). The remarkably low synonymous diversity on *csd* that equals the one calculated for noncoding loci scattered within the genome indicates that the effect of balancing selection observed in *csd*-PSD of other *Apis* (elevated π_s) is not pronounced in *A. florea*. The frequency spectrum of polymorphic sites in our data suggest that demographic effects may have acted in the recent history of *A. florea*, which could have erased possible signs of balancing selection (for example, increased π_s values). Strengthened by our analyses (Table 1), a possible explanation is a recent population growth after a bottleneck.

Bottlenecks might be a frequent occurring phenomenon in *A. florea*, as this species, originally endemic in South East Asia, is now expanding westward, present now in the Middle East, the Arabian Peninsula and invading Israel and the Sudan (Moritz *et al*, 2010; Hepburn, 2011). Thus, one may expect that over the course of this long-range expansion, the populations have passed several bottlenecks in the past, raising an interesting issue for future research regarding our sampling, which was restricted to the south of Thailand. In addition, in *A. florea*, the phenomenon of nest abandonment by swarming and migration is known to be ubiquitous (Hepburn, 2011), as their nests are freely exposed to the environment (open-air nesting). By contrast, the cavity-nesting *A. mellifera* do not show this frequent and strong fluctuation in population size. In addition to this behavior of abandonment and rebuilding of colonies, *A. florea* queens mate with few (6–10) males, which is in strong contrast to *A. mellifera* queens (mating with up to 25 males) or *A. dorsata* queens (>40 matings), also known to be open-air nesting (Oldroyd *et al*, 1995; Kraus *et al*, 2005). This species-specific behavior may act as a major force leading to a repeating reduction in the effective population size, explaining the biased nucleotide-frequency spectrum.

It seems reasonable to assume that due to the extraordinary advantage of heterozygosity at the *csd* locus, single mutations are selectively favored in a newly founded population. Fay and Wu's H-statistic, known to be very sensitive for selective sweeps, detected such signs when applied to *Af-csd* exons 6–9, but we also see indications for population size expansion after a bottleneck in our data (Table 1). Our previous studies (Hasselmann *et al*, 2008b; Lechner *et al*, 2014) examined genetic drift in *Apis* species and high-origination rates, in *A. mellifera* specifically, as the major drivers for *csd* allele evolution, leading to coalescence times shorter than expected under such a selective regime of balancing selection. Interestingly, in the early history of *Apis* evolution, *A. florea* seems to combine both in a most pronounced way.

We found differences in signs of selection along the phylogenetic branches for *csd* in *A. florea* compared with other *Apis* species (Figure 4). Whereas strong and significant signs of positive selection were found for the *csd* branches leading to *A. mellifera/A. cerana* and *A. dorsata* (Figure 4 and Table 4) with d_n/d_s ranging from 3.6 to 10, this value for *csd* in *A. florea* was not significant (0.2). Thus, although a substantial number of nonsynonymous substitutions have been fixed in the lineage of *A. florea*, the low number of polymorphisms in the current *csd* allele data set may limit the detectable signs of selection. To test whether the results for the *A. florea* branches were influenced by the low-sequence diversity, we used a modified data set of *A.mellifera, A. cerana and A.dorsata* containing *csd* sequences more similar to each other and re-ran the Branch-Site REL analyses (Supplementary Table 4). The d_n/d_s ratio of the non-*florea Apis* was thereupon strongly reduced (0.23–0.56) to the level of *A. florea* (0.23), supporting our hypothesis that the limited *csd*-sequence diversity affects the statistical power of the analysis.

Contrasting evolution in the sex determination pathway among *Apis* species

As is known for several holometabolic species, primary signals of sex determination pathways show an enormous diversity, whereas downstream targets remain conserved as a transductional core (*fem/tra - dsx* genes). Furthermore, these primary signals seem to be rapidly evolving, as they derive from different origins such as from gene duplication or allelic variation (Bopp *et al*, 2014). Among closely related species like *Apis* and *Bombus*, gene duplication of *fem* occurred independently (Koch *et al*. 2014) and likely led to different functions. Therefore, it is reasonable to argue that the evolution of *A. florea* and *A. mellifera/A. cerana/A. dorsata csd* headed in different directions, leading to structural and functional modifications between the *csd* genes for which we found evidence in the current study.

Our results show a modified coiled-coil formation in *A. florea* Csd compared with the other species caused by three amino acid differences in the N-terminal region and one difference in the C-terminal region of Csd. The N-terminal coiled-coil motif in *A. florea* seems to have evolved from specific modifications (one substituted and two deleted amino acids) of its precursor Fem (Figure 2c). In addition, few amino acid changes are needed to give rise to new *csd* specificities, harboring a less complex HVR than in the other *Apis* species (Figure 2). These findings lead us to two alternative scenarios of *fem/csd* evolution. The first scenario involves a single duplication of *fem* at the early stage, prior to the divergence of *Apis* species, with lineage-specific modifications of *csd* governed by evolutionary constraints. This scenario is supported by high homologies among all *Apis-csd*, including protein domains, HVR and coil motifs. Nevertheless, marked differences between *Af-csd* and remaining *Apis-csd* suggest an evolutionary transition, from the early stage of the duplication beginning, which has led to a *fem/csd* complex in *A. florea* compatible with (or influenced by) its biology. The second scenario would be an independent duplication of *fem* within the *Apis* lineage, with one *Apis-csd* lineage present in *A. florea* and the other in *A. mellifera/A. cerana/A. dorsata*, leading to an independent but constrained evolution of *csd* as the primary signal of sex determination. Given the high structural similarity of *csd* within *Apis*, irrespective of the evolutionary differences, we conclude this to be a rather unlikely scenario. Our conclusion is further strengthened by the analysis of the *fem* gene of *A. andreniformis*, a dwarf honey bee closely related to *A. florea*, comprising an identical coiled-coil motif (Hasselmann, unpublished data).

The first scenario is in agreement with the diverse evolutionary fate of gene duplicates during their initial phase. These genes can be altered through randomly occurring mutations, with major consequences for their subsequent evolution (for example, neo-functionalization and pseudogenization; Conant and Wolfe, 2008). Pseudogene sequences and non-coding genomic fragments of *csd* occurring in the *A. mellifera* genome, which were non-functional but still had high similarity to *csd*, were previously identified and interpreted as likely trans-specific alleles (Cho *et al*, 2006; for discussion see Hasselmann *et al*, 2008b). In our present study, none of the evolutionary young *csd* alleles provide indications for being trans-specific. The study of Liu *et al.* (2011) describes variable *csd* HVR repeated motifs amplified from *A. florea* genomic DNA, for which we found no evidence in our RACE and reverse transcription experiments. Thus, the scenario of neo-functionalization of distinct ancestral *csd*-types in *A. florea* and in *A. mellifera/A. cerana/A. dorsata* seems likely. The frequent and independent duplication of *fem* in Hymenopteran lineages (Geuverink and Beukeboom, 2014; Koch *et al*, 2014) (Bee10Genome consortium, Kapheim *et al*, 2015) seems to be a general phenomenon, which

may be of broader relevance to better understand the evolution of sex determination pathways in other insect species.

In summary, our data suggest that the sex-determining function of *csd* in *A. florea* follows the same principle as that proposed for *A. mellifera*, whereas the molecular mechanism of *csd* seems to be encoded by fewer amino acid differences, a less complex HVR and more modified coiled-coil motifs than that described for *A. mellifera*. Moreover, *csd* allele evolution seems to be heavily influenced by the biology of *A. florea*. The resulting strong impact of genetic drift and bottleneck events is also of pronounced interest for the conservation of other (for example, solitary) bee species.

DATA ARCHIVING

Sequence data available from GenBank: accession numbers KS297794–KS297876 (*A. florea fem*), KS297743–KS297793 (*A. florea csd*).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are very thankful to Jochen Pflugfelder for collecting *A. florea* samples and three anonymous reviewers for their helpful comments. This work was supported by grants from the Deutsche Forschungsgemeinschaft (HA 5499/3–1 and HA 5499/3–2 to MH).

Betts MJ, Russell RB (2003). Amino acid properties and consequences of subsitutions. In: Barnes MR, Gray IC (eds). *Bioinformatics for Geneticists*. Wiley: West Sussex.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF *et al.* (2006). Exceptionally high levels of recombination across the honey bee genome. *Genome Res* **16**: 1339–1344.

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

Beye M, Seelmann C, Gempe T, Hasselmann M, Vekemans X, Fondrk MK *et al.* (2013). Gradual Molecular Evolution of a Sex Determination Switch through Incomplete Penetrance of Femaleness. *Current Biology* **23**: 2559–2564.

Bopp D, Saccone G, Beye M (2014). Sex determination in insects: variations on a common theme. *Sex Dev* **8**: 20–28.

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.

Conant GC, Wolfe KH (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.

Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Fu YX (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

Gempe T, Hasselmann M, Schioett M, Hause G, Otte M, Beye M (2009). Sex determination in honeybees: Two separate mechanisms induce and maintain the female pathway. *PLoS Biol* **7**: e1000222.

Geuverink E, Beukeboom LW (2014). Phylogenetic distribution and evolutionary dynamics of the sex determination genes doublesex and transformer in insects. *Sex Dev* **8**: 38–49.

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, Gempe T, Schiott M, Nunes-Silva CG, Otte M, Beye M (2008a). Evidence for the evolutionary nascence of a novel sex determination pathway in honeybees. *Nature* **454**: 519–522.

Hasselmann M, Lechner S, Schulte C, Beye M (2010). Origin of a function by tandem gene duplication limits the evolutionary capability of its sister copy. *Proc Natl Acad Sci USA* **107**: 13378–13383.

Hasselmann M, Vekemans X, Pflugfelder J, Koeniger N, Koeniger G, Tingek S *et al.* (2008b). Evidence for convergent nucleotide evolution and high allelic turnover rates at the complementary sex determiner gene of western and asian honeybees. *Mol Biol Evol* **25**: 696–708.

Hepburn HR Absconding, migration and swarming. Hepburn HR, Radloff S. (2011). *Honeybees of Asia*. Springer: Berlin, Heidelberg, 133–158.

Hicks MR, Holberton DV, Kowalczyk C, Woolfson DN (1997). Coiled-coil assembly by peptides with non-heptad sequence motifs. *Fold Des* **2**: 149–158.

Hudson RR, Kaplan NL (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

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

Innan H, Kondrashov F (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**: 97–108.

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T *et al.* (2015). Genomic signatures of evolutionary transitions from solitary to group living. *Science* **348**: 1139–1143.

Koch V, Nissen I, Schmitt BD, Beye M (2014). Independent evolutionary origin of fem paralogous genes and complementary sex determination in hymenopteran insects. *PLoS One* **9**: e91883.

Kosakovsky Pond SL, Frost SD, Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K (2011). A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol* **28**: 3033–3043.

Kraus FB, Neumann P, Moritz RFA (2005). Genetic variance of mating frequency in the honeybee (*Apis mellifera L.*). *Insectes Soc* **52**: 1–5.

Lechner S, Ferretti L, Schoning C, Kinuthia W, Willemsen D, Hasselmann M (2014). Nucleotide variability at its limit? insights into the number and evolutionary dynamics of the sex-determining specificities of the honey bee *Apis mellifera*. *Mol Biol Evol* **31**: 272–287.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Liu ZY, Wang ZL, Wu XB, Yan WY, Zeng ZJ (2011). csd alleles in the red dwarf honey bee (Apis florea, Hymenoptera: Apidae) show exceptionally high nucleotide diversity. *Insect Sci* **18**: 645–651.

Lo N, Gloag RS, Anderson DL, Oldroyd BP (2010). A molecular phylogeny of the genus Apis suggests that the Giant Honey Bee of the Philippines, A-breviligula Maa, and the Plains Honey Bee of southern India, A-indica Fabricius, are valid species. *Syst Entomol* **35**: 226–233.

Lupas A, Van Dyke M, Stock J (1991). Predicting coiled coils from protein sequences. *Science* **252**: 1162–1164.

Nissen I, Muller M, Beye M (2012). The *Am-tra2* gene is an essential regulator of female splice regulation at two levels of the sex determination hierarchy of the honeybee. *Genetics* **192**: 1015–1026.

Oldroyd BP, Smolenski AJ, Cornuet JM, Wongsiri S, Estoup A, Rinderer TE *et al.* (1995). Levels of polyandry and intracolonial genetic-relationships in *Apis-Florea*. *Behav Ecol Sociobiol* **37**: 329–335.

Palmer KA, Oldroyd BP (2000). Evolution of multiple mating in the genus *Apis*. *Apidologie* **31**: 235–248.

Ponting CP, Russell RR (2002). The natural history of protein domains. *Annu Rev Bioph Biom* **31**: 45–71.

Ramirez SR, Nieh JC, Quental TB, Roubik DW, Imperatriz-Fonseca VL, Pierce NE (2010). A molecular phylogeny of the stingless bee genus Melipona (Hymenoptera: Apidae). *Mol Phylogenet Evol* **56**: 519–525.

Schmieder S, Colinet D, Poirie M (2012). Tracing back the nascence of a new sex-determination pathway to the ancestor of bees and ants. *Nat Commun* **3**: 895.

Tajima F (1989). The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.

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.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739.

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

Woyke J (1963). Drone larvae from fertilized eggs of the honeybee. *Japic Res* **2**: 19–24.

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

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)