Global allele polymorphism indicates a high rate of allele genesis at a locus under balancing selection

Guiling Ding¹ · Martin Hasselmann² · Jiaying Huang¹ · John Roberts³ · Benjamin P. Oldroyd⁴ · Rosalyn Gloag⁴

Received: 11 May 2020 / Revised: 13 August 2020 / Accepted: 16 August 2020 / Published online: 27 August 2020
© The Author(s), under exclusive licence to The Genetics Society 2020

Abstract

When selection favours rare alleles over common ones (balancing selection in the form of negative frequency-dependent selection), a locus may maintain a large number of alleles, each at similar frequency. To better understand how allelic richness is generated and maintained at such loci, we assessed 201 sequences of the *complementary sex determiner* (*csd*) of the Asian honeybee (*Apis cerana*), sampled from across its range. Honeybees are haplodiploid; hemizygotes at *csd* develop as males and heterozygotes as females, while homozygosity is lethal. Thus, *csd* is under strong negative frequency-dependent selection because rare alleles are less likely to end up in the lethal homozygous form. We find that in *A. cerana*, as in other *Apis*, just a few amino acid differences between *csd* alleles in the hypervariable region are sufficient to trigger female development. We then show that while allelic lineages are spread across geographical regions, allelic differentiation is high between populations, with most *csd* alleles (86.3%) detected in only one sample location. Furthermore, nucleotide diversity in the hypervariable region indicates an excess of recently arisen alleles, possibly associated with population expansion across Asia since the last glacial maximum. Only the newly invasive populations of the Austral-Pacific share most of their *csd* alleles. In all, the geographic patterns of *csd* diversity in *A. cerana* indicate that high mutation rates and balancing selection act together to produce high rates of allele genesis and turnover at the honeybee sex locus, which in turn leads to its exceptionally high local and global polymorphism.

Associate editor: Darren Obbard

Supplementary information The online version of this article (<https://doi.org/10.1038/s41437-020-00358-w>) contains supplementary material, which is available to authorized users.

✉ Rosalyn Gloag
ros.gloag@sydney.edu.au

- ¹ Key Laboratory of Pollinating Insect Biology of the Ministry of Agriculture and Rural Affairs, Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing 100093, China
- ² Department of Livestock Population Genomics, Institute of Animal Science, University of Hohenheim, Stuttgart, Germany
- ³ Commonwealth Scientific and Industrial Research Organisation, Canberra, ACT 2601, Australia
- ⁴ Behaviour, Ecology and Evolution Laboratory, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

Introduction

Natural populations harbour persistent polymorphism, despite the fact that directional natural selection and genetic drift tend to erode genetic diversity over time (Clarke 1979). At some loci, this evolutionary puzzle is resolved by negative frequency-dependent selection. Under this form of selection, the fitness benefit conferred by an allele depends on its frequency relative to other alleles in the population, with rare gene variants favoured over common ones (Wright 1939; Clarke 1979; Levin 1988). As a result, loci that are subject to negative frequency-dependent selection retain allelic diversity, often with alleles at equal frequency. Together with other forms of balancing selection (e.g. heterozygote advantage), negative frequency-dependent selection plays an important role in maintaining adaptive genetic diversity in diverse organisms (Wu et al. 2017). For example, the self-incompatibility locus (*S* locus) in plants (Wright 1939; Nasrallah et al. 1987; Nasrallah 1997), mating-type genes in fungi (Raper 1966; May and Matzke 1995) and mitochondrial haplotypes in fruit flies (Kurbalija Novičić et al. 2020) are all subject to negative frequency-dependent selection.

The sex locus, or complementary sex determiner (*csd*), of the honeybee is also under strong negative frequency-dependent selection, and offers an opportunity to investigate how polymorphism at such loci is generated, maintained and constrained (Beye et al. 2003; Hasselmann and Beye 2004). The *csd* gene underpins the honeybee sex-determination pathway (Beye et al. 2003). Diploid individuals heterozygous at *csd* develop into females (workers or queens) and haploid individuals become males, while diploid individuals homozygous at *csd* are infertile males that are killed as larva by the colony's workers (Mackensen 1951; Woyke 1963). Thus, negative frequency-dependent selection in this case stems from an extreme form of over-dominant selection (the selection coefficient against homozygotes = 1; Yokoyama and Nei 1979). In natural populations, the lethal homozygous condition is uncommon because an extraordinarily large number of *csd* alleles exist, each at low frequency. For example, 23–28 alleles were observed in wild African populations of the Western honeybee, *Apis mellifera* (Lechner et al. 2014). The total number of *csd* alleles in the global *A. mellifera* population is unknown, but certainly exceeds 100 (Lechner et al. 2014; Zareba et al. 2017).

The *csd* gene consists of nine exons spanning 9 kb that produce a single transcript encoding a protein of about 400 amino acid residues (Beye et al. 2003). Nucleotide polymorphism is concentrated in exons 6–7 (the potential-specifying domain, PSD; Hasselmann et al. 2008), and especially so in the domain's hypervariable region (HVR), which consists of repeated amino acid motifs that differ in arrangement and number between alleles (Hasselmann et al. 2008). Previous studies suggest that the HVR is key to understanding *csd*'s allelic diversity. The *csd* gene arose via gene duplication of an ancestor gene in the sex-determination system of hymenopteran insect, *fem* (Hasselmann et al. 2008; Koch et al. 2014). As ancestral *fem* lacks a repeat region, the *csd* HVR evidently evolved during the gene's subsequent neofunctionalization as the primary signal of sex determination (Hasselmann et al. 2008). Functional *csd* allele pairs (i.e. those that trigger female development in heterozygotes) typically vary in the length and sequence of the HVR. At least in *A. mellifera*, a large number of motif combinations can produce functional allele pairs, and only a few amino acid changes in the HVR may be sufficient to give rise to novel *csd* alleles (*A. mellifera*; Lechner et al. 2014).

The *csd* HVR has all the hallmarks of a mutation hotspot (Hasselmann et al. 2008). Heterozygosity is intrinsically mutagenic (Amos 2009; Yang et al. 2015), with mutation rates particularly elevated when alleles differ in their length (Amos et al. 2015). At loci under balancing selection, this leads to a positive feedback loop between mutation and selection, in which mutations are common and mutants are

also more likely to persist and spread, maintaining the heterozygosity that drives further mutation (Lynch 2015). The mutational mechanisms that generate polymorphism in the HVR remain poorly understood. Single-nucleotide polymorphisms and indels contribute to *csd* allelic variation, throughout the specifying domain, but diversity is many times higher in the HVR than elsewhere in the PSD (Hasselmann et al. 2008; Biewer et al. 2016). Here, slipped strand mispairing during replication is presumed to play an important role in allele genesis, adding or removing repeat units in a manner similar to that of the simple sequence repeats (SSRs) within other genes (Li et al. 2004), and microsatellites (Chapuis et al. 2015).

In other well-studied loci under balancing selection, alleles or allele families are so long-lived that they transcend species boundaries (e.g. the *trans*-specific polymorphisms of genes of the major histocompatibility complex in vertebrates, Azevedo et al. 2015; the *S* locus in plants, Richman and Kohn 1999; the *het-c* locus in fungi, Muirhead et al. 2002). Interestingly, allele lineages at *csd* are, in contrast, not shared between *Apis* species (Hasselmann et al. 2008). Only a few codons have been conserved across the 20 million or so years of *Apis* speciation (Lechner et al. 2014). The relatively short average coalescence time of 6 million years in the *csd*-PSD genealogy supports a high allelic turnover rate (Hasselmann et al. 2008). If local populations maintain sufficiently large number of alleles, and readily generate new alleles, the selective advantage of any particular rare allele is eroded, making it more likely that old alleles are lost over time to drift and replaced by new ones (Schierup et al. 2008; Hasselmann et al. 2008).

The distribution of *csd* diversity between natural honeybee populations can provide new insights into how, and how quickly, new alleles arise and spread. A recent study of two *A. mellifera* populations in Poland found that they contained many private alleles, including alleles at lower-than-expected frequencies (i.e. unequal frequencies, Zareba et al. 2017). While this observation is consistent with a high turnover of alleles, it is difficult to rule out the effects of recent allele migrations arising from the movement of commercial *A. mellifera* colonies. In this study, we assess the sequence polymorphism of *csd* within and between populations of the Asian honeybee *Apis cerana*. In particular, we determine whether patterns of *csd* polymorphism are consistent with high rates of allele genesis and turnover proposed for this locus. Here we use "allele genesis rate" to refer to the rate at which novel alleles arise (via mutation) and become established in a population (via selection). *A. cerana* has a large native range spanning Pakistan, India, China, Japan and South-East Asia (Oldroyd and Wongsiri 2006). In the last 40 years, it has also established an invasive range in the Austral-Pacific (Koetz 2013; Gloag et al. 2017); Fig. 1. Although *A. cerana* has

Fig. 1 The distribution of analyzed samples of *A. cerana* across its native range (yellow) and invasive range (red). Our dataset included samples analyzed for this study (stars) and those represented in published sequences (circles). Populations with less than 5 sequences are indicated with a white cross. The colour of the sample locations

corresponds to that of their alleles indicated in Fig. 3: China (orange), India (green), Thailand and Malaysia (purple), Japan (grey), The Philippines (blue) and Sundaland (including invasive Pacific, black). Information on sample locations and number of *csd* alleles per location is provided in Table 1.

been used in Asia for beekeeping and honey hunting for thousands of years (Oldroyd and Wongsiri 2006), it has not been the subject of extensive anthropogenic migration in the same way as *A. mellifera* has been (Moritz et al. 2005). Thus, while today's populations of *A. cerana* have undoubtedly been impacted by the small-scale movements of colonies by beekeepers, at regional scales within its native range, we expect patterns of genetic diversity in this species to reflect the bee's evolutionary history rather than human interference.

Two previous studies that sequenced *A. cerana*'s *csd* focused on interspecific comparisons with other *Apis* (Cho et al. 2006, $n = 34$ sequences; Hasselmann et al. 2008, $n = 17$ sequences). Here we combine those sequences with new data to assess the sequence polymorphism of 201 *csd* fragments, collected from populations across *A. cerana*'s native and invasive range (Fig. 1, Table 1). Specifically, we describe the nucleotide diversity of the *A. cerana* HVR, assess the distribution of *csd* allele lineages across populations and estimate the proportion of shared *csd* alleles between

Table 1 *Apis cerana* alleles at the complementary sex determiner (*csd*) used in this study.

Sample region	Sample location	<i>N</i> <i>csd</i> alleles	Reference
Populations with >5 <i>csd</i> alleles available for analysis			
South-East Asia	Samut Songkhram, Thailand	22	This study; Hasselmann et al. (2008)
	Maha Sarakham, Thailand	15	This study
	Flores, Indonesia	15	This study
	Tenom, Borneo	10	Hasselmann et al. (2008)
	Sagada, Luzon, Philippines	7	Cho et al. (2006)
India	Mysore, India	25	This study
	Bangalore, India	8	Cho et al. (2006)
China	Beijing, China	19	This study
	Shandong, China	27	This study
	Sichuan, China	22	Gloag et al. (2017)
Austral-Pacific (invasive)	Cairns, Australia	7	Gloag et al. (2017)
	Vanimo, Papua New Guinea	14	This study
	Makira Island, Solomon Islands	8	This study
Populations with <5 <i>csd</i> alleles available for analysis			
South-East Asia	Los Baños, Luzon, Philippines	4	Cho et al. (2006)
	Chiang Mai, Thailand	1	
	Po Thong, Thailand	1	
	Prachuap Khiri Khan, Thailand ^a	2	
	Phet Buri, Thailand ^a	1	
	Southern China ^b	2	
Japan	Morioka, Japan	3	
	Kumamoto, Japan	1	
	Nasushiobara, Japan	1	
	Tokyo, Japan	1	
Total number of alleles		201	

Alleles were nominated based on amino acid differences in the gene's potential-specifying domain (exons 6–7).

^aClose locations shown with one circle in Fig. 1.

^bSpecific location not given in the source publication.

populations (i.e. differentiation) compared with microsatellite loci. We further use our dataset to explore the mutational mechanisms that drive allele genesis at *csd* by establishing a criterion to determine combinations of *csd* alleles that generate the female phenotype, estimating the rate of *csd*-length changes relative to those of microsatellites and identifying the

type of polymorphisms that characterise recently diverged allele pairs between and within populations. We also discuss *csd* diversity of three invasive populations in the Austral-Pacific, in light of their invasion history.

Materials and methods

Sampling

To assess inter- and intra-population polymorphism at *csd*, we obtained sequences and genotypes from *A. cerana* collected from seven native-range locations (three in China, one in Indonesia, one in India and two in Thailand) and three invasive-range locations (Papua New Guinea, Solomon Islands and Australia). Of these, *csd* sequences for two locations (Sichuan, China and Cairns, Australia) were previously reported in Gloag et al. 2017; Fig. 1, Table 1. We then enriched our dataset with published *A. cerana* *csd* sequences obtained from additional sampling locations in Japan, China, Thailand, Malaysia, India and the Philippines (32 “Type 1” alleles, excluding pseudogenes and duplicate sequences from the same population; Cho et al. 2006), Borneo and Thailand (17 alleles, Hasselmann et al. 2008). In total, we analysed 201 *csd* sequences from 23 sampling locations; 13 locations were sufficiently well sampled to provide at least 7 *csd* sequences (range: 7–27), while the remaining 10 locations (all from Cho et al. 2006) were represented with just one or a few sequences each; Fig. 1, Table 1. We hereafter refer to these sample locations as local populations. We defined alleles as sequences with unique amino acid sequences (though in all but five cases, alleles also had unique nucleotide sequences).

Extraction, amplification and sequencing

We obtained *csd* sequences from each of our sampled populations ($N = 31$ –105 per population; sample sizes in Table S1). We extracted DNA from one hind leg per bee (males or workers) using a 5% Chelex solution protocol (Walsh et al. 1991), and then amplified a fragment of *csd* spanning the PSD and exon 8 (Fig. 2a) using primers from Gloag et al. (2017). This region of *csd* is the target of balancing selection and the location of most polymorphism (Hasselmann et al. 2008). We used high-fidelity KAPA2G Robust DNA Polymerase (Kapa Biosystems, Wilmington, USA) and standard PCR conditions (94 °C for 5 min, 38 cycles of 94 °C for 30 s, 60 °C for 30 s and 72 °C for 45 s, with a final extension at 72 °C for 15 min). For males (i.e. hemizygotes), PCR products were sequenced directly (SangonBiotech Co., Ltd., Shanghai, China). For workers (i.e. heterozygotes), we cloned PCR products into TOPO® vectors (ThermoFisher Scientific, Waltham, USA). We then

estimate the extent of differentiation between populations at neutral loci (Ellegren 2004). We amplified DNA in 5- μ l reactions (1 $\times$ reaction buffer, 2.5 mM MgCl₂, 0.16 mM dNTP mixture, 0.32–0.8 μ M of fluorescent dye-labelled primers, Sigma-Aldrich, USA, 0.4 units of Taq polymerase and 1 μ l of extracted DNA) and resolved PCR products using an ABI3130xl genetic analyser (Applied Biosystems, USA).

Sequence analyses

Nucleotide diversity

We determined the position of exons and introns in our *csd* nucleotide sequences by consulting the *A. cerana* cDNA sequences of Hasselmann et al. (2008). We aligned nucleotide sequences in MEGA X (Kumar et al. 2018) and translated them to amino acid sequences. We used DnaSP v.6.10.01 (Rozas et al. 2017) to calculate summary statistics for our total sample of 201 sequences, and per population. Because the extreme polymorphism of the HVR prevents reliable alignment of this region, we used only the HVR-flanking regions for these summary statistics (flanking region defined in Fig. 2a). We calculated the number of segregating sites, haplotype diversity, nucleotide diversity, Watterson's theta and three tests of selective neutrality: Tajima's D (Tajima 1989), Fu's F_s (Fu 1997) and Fay and Wu's H_n (Fay and Wu 2000). Statistical significance was obtained using coalescence simulation against the standard-neutral model, implemented in DnaSP v.6.10.01 with 1000 replicates, an intermediate level of recombination (recombination per site: 0.0294) and input parameters (theta, number of sites and sample size) provided by the data. In combination, these tests for deviation from neutral evolution based on the allele frequency spectrum provide insight into the evolutionary processes driving nucleotide diversity within populations. A negative Tajima's D is indicative of population-size expansion or selective sweeps, whereas a positive Tajima's D suggests a recent bottleneck or overdominant selection (Tajima 1989). A negative Fu's F_s statistic is evidence for an excess number of alleles, as expected from a population expansion or genetic hitchhiking, whereas positive values of F_s indicate a deficiency of alleles, and would indicate recent population bottlenecks or overdominant selection (Fu 1997). Finally, positive values of Fay and Wu's H_n are a sensitive indicator of a selective sweep, while negative values indicate an excess of high-frequency polymorphism (Fay and Wu 2000). We then looked for putative recently diverged alleles pairs, based on low divergence of the HVR-flanking regions, and identified the number and type of sequence changes that such pairs had accumulated in their HVRs.

Evolutionary rate of length change in the HVR relative to microsatellites

Length variations in the repetitive region of the HVR are one of several types of mutation that may drive allelic novelty at *csd*, and presumably occur via replication slippage in a manner similar to microsatellites. Lechner et al. (2014) devised a factor that estimates the rate at which the repetitive region of the *csd* HVR evolves length variation relative to selectively neutral microsatellites (the evolutionary rate factor of differences, or F_{erd}). This estimation takes observed length variation at a locus as a proxy for the rate at which length variations evolve. Microsatellite alleles are often highly variable in length because they are prone to length-change mutations (Ellegren 2004). Greater variation in allele lengths at *csd*, relative to microsatellite loci (i.e. $F_{erd} > 1$), would thus be consistent with length changes evolving often at *csd*, either because slippage rates are similarly high, selection favours many length-change mutations (i.e. they produce new functional alleles) or both.

We calculated F_{erd} for each of the five *A. cerana* populations in our dataset that had sufficient sample sizes (Shandong and Beijing, China; Mysore, India; Samut Songkhram and Maha Sarakham, Thailand, each with 32–63 drones). To do this, we first calculated the deviation in length of each allele from the mean length for each locus j :

$$\alpha_j = x_i - \bar{x}.$$

For *csd*, lengths were calculated for the repetitive region of the HVR only (given in Fig. 2). We then calculated the standard deviation of α_j as a measure of relative variability in allele lengths at each locus:

$$\delta = \sqrt{\frac{1}{n-1} \sum_{i=1}^n [\alpha_j - \bar{\alpha}]^2}$$

F_{erd} was then calculated per population as the ratio of δ in the repetitive region of the HVR (δ_{HVR}) to the average value across microsatellite loci (δ_{msats}):

$$F_{erd} = \delta_{HVR} / \delta_{msats}$$

Functional heterozygosity

We estimated the number of amino acid changes that are required for diverging alleles to become functionally distinct (Lechner et al. 2014). This estimate assumes that all functional polymorphisms are found in our target fragment of exons 6–8 (as for *A. mellifera*; Lechner et al. 2014). First, we determined a minimum difference for functional heterozygosity of *csd* alleles based on 140 pairs of allelic

combinations that are clearly functional because they were detected in phenotypic females (140 unique combinations; Table S3). For each functional pair, we calculated the number of amino acid differences between alleles in (i) the HVR region (d_{HVR}), (ii) the PSD adjacent to HVR (d_{PSD}) and (iii) exon 8 (d_{e8}). We calculated Pearson correlations between d_{HVR} and d_{PSD} , d_{HVR} and d_{e8} and d_{PSD} and d_{e8} to determine which measures of divergence to include in our criteria of functional heterozygosity. A scatterplot of the amino acid mismatches was made for the d_x values that were significantly ($P < 0.05$) correlated (SPSS ver. 20). We then took the convex envelop of the data from the scatterplot as the bounds of our criteria for determining functional heterozygosity (see Fig. S1). Finally, we applied our criteria to all possible pairings of alleles in our dataset (within and between populations) to determine the proportion of combinations that would fail to generate femaleness.

Allele genealogy

We constructed a genealogy of alleles to obtain insights into the evolutionary relationships among *csd* sequences within and across populations. We predicted that balancing selection would lead to allele lineages shared between populations; that is, each branch of a genealogy should contain related alleles from multiple populations. We aligned *csd*-coding sequences and used Modeltest, implemented in MEGA X (Kumar et al. 2018), to compare the available amino acid substitution models, and determined the best description of the substitution pattern via the maximum likelihood method. The model with the lowest Bayesian information criterion scores was considered to best describe the substitution pattern. We next used GARD (Kosakovsky Pond et al. 2006) implemented in the datamonkey platform (Weaver et al. 2018), to test for signs of recombination within the dataset that would otherwise impact the reliability of subsequent tree construction methods. This heuristic approach looks for phylogenetic incongruence among different partitions of the data in a stepwise placement of potential breakpoints. The best-fitting model is given based on Akaike information criterion derived from a maximum likelihood model fit to each segment. Further, we used the software RDP 3.44 (Martin et al. 2015) to evaluate potential recombination events via three tests that each search a dataset for recombinant sequences and the parental sequences from which they derived: BootScan (Martin et al. 2005), Chimaera (Posada and Crandall 2001) and SiScan (Gibbs et al. 2000). No recombination signal or event was detected by any of these tests. We therefore proceeded with *csd* genealogy construction via two approaches. In the first and most conservative approach, we removed the HVR region and positions containing gaps to consider only the remaining 109 amino acid positions. Non-uniformity of

evolutionary rates among sites was modelled by using a discrete Gamma distribution (+G, parameter = 0.56) with 5 rate categories, and by assuming that a certain fraction of sites is evolutionarily invariant (+I, 16% of sites) based on the Jones, Taylor and Thornton (JTT) matrix and maximum likelihood method. The second approach took the full spectrum of all sites (183 positions), including those reflecting the length variation within the HVR into account, using the maximum parsimony method. The tree was obtained using the heuristic Subtree-Pruning-Regrafting algorithm with search level 1 in which the initial trees were obtained by the random addition of sequences (ten replicates).

Differentiation between populations

Loci under balancing selection are predicted to show low levels of differentiation between populations, relative to neutral loci, because of strong positive selection for any migrant allele not present in the resident population. This means even if migration is infrequent, migrant alleles are likely to become established (Schierup et al. 2000; Muirhead 2001). Whether this phenomenon should hold for *csd* is unclear, however, because high rates of allele turnover within populations may erase any signal of migration. We therefore investigated between-population differentiation at *csd* and microsatellites for all populations with available data ($n = 8$ populations). As all diploid individuals in a population are heterozygous at *csd*, and diversity at the locus is extreme, the commonly used statistics *Fst* and *Gst* are unreliable indicators of inter-population differentiation at this locus (Hedrick 2005; Wang 2015). Instead, we compared the pattern of allele sharing for the *csd* locus and each of our nine microsatellite loci by calculating h_k , the proportion of alleles at a locus that are shared among k populations (Muirhead 2001). We used a chi-square test to assess whether the proportion of alleles in each h_k category at *csd* differed from the average proportion across microsatellite loci. Allele-sharing patterns reflect the relative magnitude of the effects of migration (m) and mutation (μ) on allelic distributions (Muirhead 2001).

Results

Nucleotide diversity of *csd* in *A. cerana*

We identified 170 unique alleles (i.e. distinct amino acid sequences) from our 201 *csd* sequences. Among these alleles, the HVR varies greatly in sequence, but is flanked by two highly conserved regions (S motif in the N terminus, and NINYI motif downstream). The HVR-repetitive region encodes peptides of varying length (6–34 residues) and is rich

Table 2 Summary statistics of *A. cerana* *csd* sequences, based on the HVR-flanking regions.

Sample location	<i>n</i>	Sites	<i>h</i>	<i>S</i>	<i>Hd</i>	π total	π_s	π_n	π_n / π_s	θ_w	Tajima's D	Fu's F_s	Fay and Wu's H_n
Samut Songkhram, Thailand	22	351	20	49	0.991	0.03199 ± 0.00247	0.04781	0.02837	0.5934	0.04142	-0.900	-8.109**	-1.248
Maha Sarakham, Thailand	15	381	15	45	1.000	0.02397 ± 0.00274	0.03715	0.02054	0.5529	0.03713	-1.521*	-7.958**	-2.057*
Flores, Indonesia	15	369	13	39	0.981	0.03244 ± 0.00323	0.04889	0.02878	0.5887	0.03334	-0.115	-2.628	-0.877
Tenom, Borneo	10	399	10	32	1.000	0.02551 ± 0.00263	0.03361	0.02371	0.7054	0.03012	-0.739	-3.431	-0.462
Sagada, Luzon, Philippines	7	387	7	27	1.000	0.02535 ± 0.00242	0.02857	0.02499	0.8747	0.02848	-0.625	-1.596	0.098
Mysore, India	25	381	25	59	1.000	0.02617 ± 0.00254	0.04657	0.02074	0.4454	0.04310	-1.526**	-18.251***	-1.485*
Bangalore, India	8	387	8	28	1.000	0.02261 ± 0.00306	0.03330	0.01990	0.5976	0.02890	-1.152	-2.466	-1.613*
Beijing, China	19	384	19	46	1.000	0.02668 ± 0.00204	0.03553	0.02471	0.6955	0.03651	-1.094	-11.145**	-1.638*
Shandong, China	27	378	25	47	0.994	0.02398 ± 0.00186	0.03173	0.02224	0.7009	0.03500	-1.197*	-15.486***	-1.749*
Sichuan, China	22	375	20	44	0.991	0.02756 ± 0.00184	0.04812	0.02222	0.4618	0.03365	-0.712	-8.756**	-1.216
Vanimo, Papua New Guinea	14	375	14	44	1.000	0.03053 ± 0.00307	0.05086	0.02545	0.5004	0.03857	-0.911	-5.921*	-0.896
Makira Island, Solomon Islands	8	381	8	31	1.000	0.02709 ± 0.00390	0.04641	0.02206	0.4753	0.03239	-0.870	-2.109	-0.980
Cairns, Australia	7	381	7	31	1.000	0.03112 ± 0.00371	0.05376	0.02538	0.4720	0.03428	-0.527	-1.266	-0.191
Total <i>A. cerana</i>	201	327	146	111	0.996	0.03054 ± 0.00090	0.05042	0.02566	0.5089	0.06503	-1.656**	-33.153***	-1.731*

Calculations were performed using the total dataset and each sample location separately: number of sequences per location (*n*), number of nucleotides (sites), number of haplotypes (*h*), number of variable sites (*S*), haplotype diversity (*Hd*), average pairwise nucleotide diversity (π) for all (π total), synonymous (π_s) and nonsynonymous (π_n), Watterson's theta (θ_w), Tajima's D, Fu's F_s and Fay and Wu's H_n .

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

in asparagine (N) and tyrosine (Y) motifs typically in the order [(N)₁₋₈(Y)₁₋₂]_n. Some sequences have an additional motif of (KHYN)₁₋₃ (Fig. 2b).

Our measure of relative allelic length divergence in *csd* versus microsatellites, F_{erd} , ranged from 4.43 to 7.28 across our sampled populations, with an average of 5.5 (Table S4). This indicates that in *A. cerana*, length variation in *csd*'s repetitive region is 4–7 times greater than that at selectively neutral microsatellite loci. The relative contribution of mutation rate and selection to this difference in length variation is unknown.

Population-genetic statistics indicated skewed allele frequency distributions with an excess of recently arisen haplotypes, and a history of strongly expanding population size (Table 2). Population growth is associated with negative values of Tajima's D, and all populations showed negative values of Tajima's D; three significantly so (Maha Sarakham, Thailand, $P < 0.05$, Mysore, India, $P < 0.01$, Shandong, China, $P < 0.05$). Fu's F_s statistic, which is particularly sensitive to recent population expansion based on the number of haplotypes, was also consistent with population expansion shaping *A. cerana* *csd* polymorphism across much of our sampled range (Table 2). Although selective sweeps can produce similar patterns, this scenario is unlikely based on the significant negative values of F_{ay} and Wu's H statistic (Table 2).

We identified 40 *csd* allele pairs or sets that had low synonymous divergence in sequence flanking the HVR ($\pi_s = 0$ –0.0055, equivalent to 0–1 synonymous SNPs in this region; Table S5). Consistent with an excess of recently arisen alleles, 73% of all alleles in our sample (128 alleles) belonged to at least one such pair/set. To better understand the role of the HVR in the genesis of *csd* alleles, we used pairwise alignments to further consider only those allele pairs with identical nucleotide sequences in the flanking regions (24 pairs). We found that the HVRs of these pairs differed on average by 0.54 SNPs and 1.3 indels in the repetitive region (where indels varied in length from 1 to 15 amino acids; Table S6). That is, diverging alleles that had not acquired polymorphisms in HVR-flanking regions had accumulated one to several polymorphisms in the HVR. To estimate the timescale needed to accumulate such changes in the HVR, we assumed that synonymous site diversity in the conserved flanking regions of the HVR (π_s) was generated at approximately neutral rates. Assuming a neutral substitution rate of $\mu = 5.27 \times 10^{-9}$ mutations/base/generation in honeybees (Wallberg et al. 2014), and two generations per year for *A. cerana* (Oldroyd and Wongsiri 2006), these alleles arose 130,000 years ago or less ($\pi_s = 4 N\mu$).

Spatial distribution of allelic lineages and alleles

Genealogical trees grouped *csd* alleles into 31 lineages, where most lineages contained alleles from multiple

diversification following past bottlenecks in these locations (Fig. 3; Table S8).

Although allelic lineages were dispersed across populations, particular alleles (i.e. identical amino acid sequences) were not. Consistent with a high rate of allele turnover at *csd*, just 14% of alleles (22 of 160) were detected in more than one population (Fig. 4a). The majority of alleles in each population were thus private (86%, $h_k = 1$, Fig. 4b). Furthermore, populations were far less likely to share alleles at *csd* than at microsatellite loci, consistent with the allele genesis rate at *csd* being greater than the rate of effective migration (median h_k across all microsatellites = 4, and $h_k > 1$ for 89% of microsatellite alleles; chi-square test: $\chi = 788$, $df = 7$, $p < 0.001$, Fig. 4b). Where alleles were shared between populations, there was no clear effect of distance; rather, shared alleles were as likely to be found in two populations from non-adjacent geographic regions ($n = 11$ alleles; regions = South-East Asia, India, China and Austral-Pacific) as they were from populations in adjacent regions ($n = 11$ alleles) (Table S8).

Only the invasive populations of the Austral-Pacific share most or all of their *csd* alleles, with all alleles present in Australia ($n = 7$ alleles) and Solomon Islands ($n = 8$) also being present in New Guinea ($n = 14$) (Table S8). This is consistent with New Guinea, which harbours the earliest-established invasive population, being the source population from which Australia and Solomon Islands were subsequently colonised. All three invasive populations had lower allelic richness than native-range populations, despite much greater sampling effort, indicating that the invasive populations have a reduced complement of alleles as a result of founding bottlenecks (Table 1, Table S1). Despite such bottlenecks, however, *csd* diversity in all three invasive populations exceeded that of neutral microsatellite loci, just as it did in the native populations (16.7 ± 6.6 alleles for *csd* per population, compared to 7.4 ± 1.8 alleles per population for Ac32, the most polymorphic of our microsatellite loci; Table S9; $F = 24.252$, $P = 0.000$, d.f. = 9, 90).

Functional heterozygosity of *csd* in *A. cerana*

We evaluated unique pairs of *csd* alleles present in 140 workers. We focused on the length and amino acid differences in exons 6–8 between each worker's pair of alleles. Representative *csd* pairs are shown in Fig. 2b. For most functional pairs, the repetitive region within the HVR differed in the number of amino acids (1–21 amino acid-length differences; Table S3). Only 5 out of 140 *csd* pairs comprise alleles with a repetitive region of the same length; in these cases, alleles had at least three different amino acids within the HVR. In one allele pair, differences occurred only in the HVR, indicating that HVR variation is sufficient for functional heterozygosity.

We assessed the similarity of pairs of *csd* alleles based on three parameters: d_{HVR} , d_{PSD} and d_{e8} . The Pearson correlation between d_{HVR} and d_{PSD} was significant ($r = 0.293$, $P < 0.001$, $n = 140$), while it was small and non-significant between d_{HVR} and d_{e8} ($r = -0.048$, $P = 0.571$, $n = 140$) and between d_{PSD} and d_{e8} ($r = 0.041$, $P = 0.629$, $n = 140$). A scatterplot of amino acid mismatches between the d_{HVR} and d_{PSD} and a curve representing possible conditions for functional heterozygotes illustrate the convex association between these parameters (Fig. S1). The convex envelope gave $d_{HVR} \geq 2$, $d_{HVR} + 4d_{PSD} \geq 7$ as the criteria for *csd* specificities (i.e. alleles that in combination would produce femaleness).

We then looked for evidence that populations may harbour recently arisen alleles that are currently below our threshold for new specificities relative to their parent allele. For our global set of 201 *csd* alleles, there were 20,100 possible heterozygous combinations. Among them, just 132 pairings (0.66%) failed to meet our criteria for functional heterozygosity; that is, the great majority of alleles in our dataset (as determined by unique amino acid sequence) would trigger female development in combination with all other alleles in our sample. Among the pairs that failed to meet the threshold, most were pairings of alleles detected in different populations, though eight pairs occurred in the same population. We examined these eight pairs to see if their sequences were indicative of recent divergence. In five of these pairs, alleles had identical repetitive regions, 0–1 amino acid differences within the HVR and 1–6 additional non-synonymous SNPs in the PSD or Exon 8 ($n = 3$, Sagada, Luzon, The Philippines; $n = 1$, Los Baños, Luzon, The Philippines; $n = 1$, Samut Songkhram, Thailand). In the remaining three pairs ($n = 1$, Bangalore, India; $n = 1$, Flores, Indonesia; $n = 1$, Samut Songkhram, Thailand), allele pairs differed either by just one amino acid or a single-motif repeat (N or YNN) in the repetitive region (Fig. 2c). In all, about half of our native-range populations ($N = 4$ of 10, Table 1) contained at least one allele pair that was consistent with alleles having not yet diverged sufficiently from each other to produce the female phenotype in heterozygotes.

Discussion

Sequence polymorphism at the honeybee sex locus

We find that the distribution of *csd* polymorphism across the native range of *A. cerana* has two defining characteristics: (i) allele lineages are cosmopolitan across populations (consistent with balancing selection), yet (ii) alleles themselves (i.e. sequences identical at the amino acid level) are mostly private to one population. This pattern

Fig. 4 Allele differentiation between populations.

a Proportion of total alleles at *csd* that were unique to that population for populations with sample sizes >5 alleles. **b** The distribution of allele-sharing level for the *csd* locus and nine microsatellite loci among our dataset of eight populations: Beijing, Shandong and Sichuan of China, Samut Songkhram and Maha Sarakham of Thailand, Mysore of India, Flores of Indonesia and Vanimo of Papua New Guinea. h_k is the probability of choosing an allele from a population that is shared among k populations.

supports the view that high rates of allele genesis and turnover help to explain the very high local and global allelic diversity at the honeybee sex locus.

At first glance, a high rate of allele turnover is not what we would predict for a locus under negative frequency-dependent selection, given that such selection reduces the

probability that alleles are lost due to genetic drift, relative to neutral loci (Takahata and Nei 1990). As local allele number becomes high, however, selection for rare alleles is weakened. Thus, the high allelic richness at *csd* facilitates allele turnover, by increasing the chance that any given allele is eventually lost from a population via drift. In the case of *A. cerana*, historical changes in range and population sizes may have further contributed to high allele turnover at *csd*. While it is difficult to infer population history from *csd* alone, our frequency spectrum analysis of *csd*, and the observed excess of *csd* alleles with little or no synonymous diversity, is consistent with a burst of young alleles that arose in the last 100,000 years during *A. cerana*'s post-glaciation expansions. Population-genetic studies of *A. mellifera* point to one or more large-scale expansions in the past 1 million year, punctuated by occasional regional contractions during glacial maxima (Wallberg et al. 2014). Although the past range shifts of *A. cerana* are less well understood, mitochondrial diversity suggests that similar fluxes are likely, with Indochina serving as a refuge during Pleistocene glaciation, and *A. cerana* then expanding both north and west as regions became more hospitable for honeybees (Smith 2011). Likewise, population expansions eastwards into the Sunderland region may have occurred in the last few hundred-thousand years, coinciding with repeated sea-level changes (Smith 2011). Historical expansions are also proposed to have shaped *csd* diversity in another Asian honeybee species, *A. florea* (Biewer et al. 2016). Interestingly, both *A. cerana* (Koetz 2013) and *A. florea* (Bezabih et al. 2014) are successful invasive species today, each having undergone range expansions of hundreds of thousands of square kilometres in the past 50 years (Moritz et al. 2010; Gloag et al. 2019).

Balancing selection is generally predicted to lower allelic differentiation of a locus between populations, relative to neutral loci, because even very low migration rates are likely to successfully transfer alleles between populations (Muirhead 2001). Such a homogenisation of allelic diversity across populations has been observed at some other loci under balancing selection (e.g. *S* locus, Glémin et al. 2005; Schierup et al. 2008). At *csd*, the distribution of allele lineages across populations is consistent with historical gene flow via migration that was enhanced by selection for immigrant alleles. Rarely though, did our sampled *A. cerana* populations share alleles with identical nucleotide sequence, and we found that differentiation of populations was always higher at *csd* than at neutral microsatellite loci. This is best explained by a mutation rate at *csd* that exceeds the rate of realized migration at the scale of our sampling (Schierup et al. 2008). Recent migration may account, however, for the small number of identical *csd* nucleotide sequences detected in distant populations. For example, human-assisted movements of *A. cerana* in recent centuries

could have led to the occasional transfer of alleles between regions. Alternatively, these shared alleles may be identical by descent (having evaded change and replacement by chance) or have originated independently from a common lineage.

What is the limit of *csd* polymorphism? Lechner et al. (2014) estimated 145 functional alleles in *A. mellifera* worldwide, while a subsequent study of managed European *A. mellifera* identified at least 121 alleles (Zareba et al. 2017). Based on our dataset of 170 unique *csd* alleles from 201 sequences, it seems certain that the global *csd* allele number in *A. cerana* far exceeds even the large previous estimates from *A. mellifera*. Given that we find native *A. cerana* populations separated by as little as 500-km carry sets of alleles that are mostly distinct, our dataset of 170 unique alleles has presumably captured only a small fraction of the total *csd* diversity present across the species' native range of 30 M km² (Radloff et al. 2010).

The HVR of *csd* as a mutation hotspot

Polymorphism in the HVR is proposed to be key to allelic functionality and specificity at *csd* (Hasselman et al. 2008; Lechner et al. 2014). In *A. cerana*, we find that polymorphism of the HVR alone can generate specificity even where flanking regions are identical, although this was rare (just 1% of heterozygotes). More often, functional heterozygosity is generated by pairs of alleles that differ in amino acid sequence of both flanking regions (1–15 differences) and the HVR itself (3–22 differences), with the HVRs usually showing length differences as well. These criteria for functional pairs in *A. cerana* are similar to those reported for two other *Apis* species (*A. mellifera*: Beye et al. 2013; Lechner et al. 2014; *A. florea*: Biewer et al. 2016). One result of this wide functional envelope for nucleotide polymorphism in *csd* is that almost all alleles are also specificities in a global sense. In > 99% of cases, any two *A. cerana* alleles picked from the global population would generate femaleness in a heterozygote.

Heterozygosity is mutagenic, and loci under balancing selection are thus likely to be mutational hotspots (Yang et al. 2015). The mechanism by which heterozygosity stimulates mutation is not yet known. Heterozygous sites might stimulate an extra round of DNA replication during the initial stages of crossing over, increasing the opportunity for copying errors (Amos 2009), or heterozygosity may lead to poor pairing of homologous chromosomes during meiosis and thus to mutation, with pairing most impaired when the length difference between alleles is large (Yang et al. 2015; Amos 2016). For *csd*, the result may be a positive feedback loop (Lynch 2015) in which polymorphism (concentrated in the HVR) leads to mutation and thus more polymorphism.

Several types of mutation appear to contribute to allele genesis at *csd*, including point mutations, insertions and deletions. In the HVR's repetitive region, duplications or deletions of amino acid motifs presumably occur via slippage during replication (Lechner et al. 2014). This mutational mechanism is best known from microsatellites, but SSRs are common in coding regions, where their mutability is proposed to make them important drivers of adaptation (Li et al. 2004). The *csd* repeat region is not nearly as uniform as a typical SSR. Nevertheless, the presence of some allele pairs in our dataset that differ in their HVR by one or two repeated units supports replication slippage as one pathway contributing to the generation of new alleles. Mutation rate in microsatellites increases as the discrepancy between allele sizes increases (Amos et al. 2015), and a similar outcome at *csd* would further accelerate the mutation rate, given that local *csd* alleles can vary in length by as much as 100 bp. Indeed, we estimate length variation in the *A. cerana* HVR-repetitive region to be 4–7 times greater than length variation at microsatellites, which is broadly consistent with an equivalent estimate for *A. mellifera* (2.4-fold length variation, relative to microsatellites; Lechner et al. 2014). Length variation at microsatellites is the result of a high mutation rate. This rate varies between loci and species, but is typically in the order of one per 10^4 – 10^5 meioses (Ellegren 2004; Chapuis et al. 2015). Length variation at *csd* is the result of both selection and mutation, and the relative contribution of each to length polymorphism at this locus remains unclear.

How many mutations are required to produce a new *csd* specificity? One model of *csd* allele evolution, based on polymorphisms in *A. mellifera*, posits that diverging alleles must accumulate several mutations before they can reliably confer femaleness in combination with their parent allele, and thus new specificities are generated in a stepwise manner (Beye et al. 2013). Consistent with this scenario, we found that four of our studied populations contained at least one pair of alleles that failed to meet the threshold criteria for specificities. These putative still-diverging allele pairs differed either by the insertion of 1–3 amino acid motifs in the HVR, or by SNPs (in the HVR, Exon 8 or both). Indeed, they were equivalent to a reported case of *A. mellifera* alleles with incomplete penetrance (a pair differing by a single YNN motif in the HVR; Beye et al. 2013). Where local *csd* allele number is high, and the incidence of homozygosity low, such new alleles may persist alongside their parent allele even if the two fail to generate femaleness in combination, eventually diverging further so that they become a functional allele pair. Alternatively, such diverging alleles may confer femaleness in some but not all heterozygotes (incomplete penetrance) and thus be under weak positive selection (Beye et al. 2013).

Many mutations at *csd* are likely to be deleterious. Lechner et al. (2014) highlighted conserved codons and

length restrictions among the *csd* alleles of *A. mellifera* that indicate there are limits on *csd* polymorphism, including a fraction of sites with functional constraints (shown in Hasselmann et al. 2008). Nevertheless, the range of mutations at *csd* that will produce a new functional variant is large, a feature that is presumably key to its high rate of allele genesis. In contrast, for example, novel specificities at the *S* locus of plants, another locus under strong balancing selection, require complimentary mutations to occur in the linked genes of the pistil and the style (Chookajorn et al. 2004; Gervais et al. 2011). Such synchronous, multistep mutations presumably arise far less often than those required to generate new specificities at *csd*, resulting in a slower rate of allele genesis at the *S* locus.

Allele diversity at *csd* in invasive honeybee populations

The high polymorphism and local endemism of *csd* make it a useful marker for tracing the origin and routes of invasive honeybee populations. Here we find that diversity at *csd* strongly supports a common origin for the three invasive island populations of *A. cerana* in the Pacific: New Guinea, the Solomon Islands and Australia. As the populations of Australia and the Solomon Islands (established c. 2003 and 2007, respectively; Koetz 2013) each carry distinct subsets of the *csd* alleles present in New Guinea (established 1970s), we conclude that New Guinea is the source of these subsequent invasions.

All three invasive populations in our dataset have fewer *csd* alleles than do native-range populations. Assuming that a typical native-range population maintains at least 25 *csd* alleles (Lechner et al. 2014; Gloag et al. 2017), then these invasive populations have suffered reductions in allele number of 44–72% as a result of founding bottlenecks. Allele number in Australia and Solomon Islands in particular is so low that it was most likely founded by a single multiply-mated queen (Ding et al. 2017; Gloag et al. 2017; Gloag et al. 2019). Yet all populations show significantly higher diversity at *csd* than at microsatellite loci, consistent with balancing selection acting to prevent the loss of initially rare *csd* alleles via genetic drift in the immediate aftermath of a founder event (Gloag et al. 2017).

The recent range expansion of *A. cerana* into the Pacific illustrates the type of large-scale range expansion that may have shaped *csd* evolution throughout *A. cerana*'s evolutionary history. The rapid allele origination rate of *csd* has important implications for bottlenecked honeybee populations because it reduces the time needed to restore a population's full complement of alleles. Indeed, selection for newly diverged alleles in bottlenecked populations will be far stronger than that of native-range populations, as each extra allele confers a significant fitness advantage when allelic diversity is low. Bursts

of allele diversification following ancient bottlenecks may account for the two instances of clustered lineages in our *csd* allele genealogy: sets of alleles from Flores, Indonesia and Luzon, The Philippines. Both these islands have been separated from mainland Asia by ocean even during glacial maxima (i.e. both lie east of Wallace's line; Radloff et al. 2010). We propose that further research of honeybee populations that have experienced past bottlenecks is likely to provide additional insights into how *csd* polymorphisms arise and spread.

Data availability

DNA sequences of *csd* alleles identified in this study have been submitted to GenBank: accession numbers MN334337–MN334481.

Acknowledgements Thanks to Gabriele Buchmann for assistance in the lab, and Nadine Chapman, Greg Good, Jurgen Paar, the Queensland Department of Agriculture and Fisheries and the Australian Government Department of Agriculture and Water Resources for collecting samples. We thank Siriwat Wongsiri, Wandee Wattanachaiyingcharoen Pyamas Sopadawan and Madeleine Beekman for help with collections in Thailand, and M.S. Reddy and Madeleine Beekman for help in India. GD was supported by the National Natural Science Foundation of China (31800463) and the Agricultural Science and Technology Innovation Program (CAAS-ASTIP-2015-IAR). RG was supported by a University of Sydney Postdoctoral Fellowship. Research funding came from Australian Research Council DP190101500 and DP150101985 to BPO and RG

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Amos W (2009) Heterozygosity and mutation rate: evidence for an interaction and its implications. *Bioessays* 32:82–90
- Amos W, Kosanović D, Eriksson A (2015) Inter-allelic interactions play a major role in microsatellite evolution. *Proc R Soc B* 282:20152125
- Amos W (2016) Heterozygosity increases microsatellite mutation rate. *Biol Lett* 12:20150929
- Azevedo L, Serrano C, Amorim A, Cooper DN (2015) Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. *Hum Genomics* 9:21
- Beye M, Hasselmann M, Fondrk MK, Page RE, Omholt SW (2003) The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* 114:419–429
- 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. *Curr Biol* 23:2559–2564
- Bezabih G, Adgaba N, Hepburn HR, Pirk CWW (2014) The territorial invasion of *Apis florea* in Africa. *Afr Entomol* 22:888–890
- Biewer M, Lechner S, Hasselmann M (2016) Similar but not the same: insights into the evolutionary history of paralogous sex-determining genes of the dwarf honey bee *Apis florea*. *Heredity* 116:12–22
- Chapuis MP, Plantamp C, Streiff R, Blondin L, Piou C (2015) Microsatellite evolutionary rate and pattern in *Schistocerca gregaria* inferred from direct observation of germline mutations. *Mol Ecol* 24:6107–6119
- Chookajorn T, Kachroo A, Ripoll DR, Clark AG, Nasrallah JB (2004) Specificity determinants and diversification of the *Brassica* self-incompatibility pollen ligand. *Proc Natl Acad Sci USA* 101:911–917
- 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
- Clarke BC (1979) The evolution of genetic diversity. *Proc R Soc B* 205:453–474
- Ding G, Xu H, Oldroyd BP, Gloag RS (2017) Extreme polyandry aids the establishment of invasive populations of a social insect. *Heredity* 119:381–387
- Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5:435–445
- 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
- Glémin S, Gaude T, Guillemin ML, Lourmas M, Olivier I, Mignot A (2005) Balancing selection in the wild: testing population genetics theory of self-incompatibility in the rare species *Brassica insularis*. *Genetics* 171:279–289
- Gervais CE, Castric V, Ressayre A, Billiard S (2011) Origin and diversification dynamics of self-incompatibility haplotypes. *Genetics* 188:625–636
- Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16:573–582
- Gloag R, Christie JR, Ding G, Stephens R, Buchmann G, Oldroyd BP (2019) Workers' sons rescue genetic diversity at the sex locus in an invasive honey bee population. *Mol Ecol* 28:1585–1592
- Gloag R, Ding G, Christie JR, Buchmann G, Beekman M, Oldroyd BP (2017) An invasive social insect overcomes genetic load at the sex locus. *Nat Ecol Evol* 1:1–6
- 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, Vekemans X, Pflugfelder J, Koeniger N, Koeniger G, Tingek S et al. (2008) 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
- Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* 59:1633–1638
- 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
- Koetz AH (2013) Ecology, behaviour and control of *Apis cerana* with a focus on relevance to the Australian incursion. *Insects* 4:558–592
- Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 23:1891–1901
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 35:1547–1549
- Kurbalija Novičić Z, Sayadi A, Jelić M, Arnqvist G (2020) Negative frequency dependent selection contributes to the maintenance of a

- global polymorphism in mitochondrial DNA. *BMC Evol Biol* 20:20. <https://doi.org/10.1186/s12862-020-1581-2>
- Lechner S, Ferretti L, Schöning C, Kinuthia W, Willemsen D, Haselmann 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
- Levin BR (1988) Frequency dependent selection in bacterial populations. *Philos Trans R Soc Lond B* 319:459–472
- Li Y-C, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 21:991–1007
- Lynch M (2015) Feedforward loop for diversity. *Nature* 523:414–416
- Mackensen O (1951) Viability and sex determination in the honeybee (*Apis mellifera* L.). *Genetics* 36:500–509
- Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* 1:vev003
- Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* 21:98–102
- May G, Matzke E (1995) Recombination and variation at the A mating-type of *Coprinus cinereus*. *Mol Biol Evol* 12:794–802
- Moritz RFA, Haddad N, Bataineh A, Shalmon B, Hefetz A (2010) Invasion of the dwarf honeybee *Apis florea* into the near East. *Biol Invasions* 12:1093–1099
- Moritz RFA, Härtel S, Neumann P (2005) Global invasions of the western honeybee (*Apis mellifera*) and the consequences for biodiversity. *Écoscience* 12:289–301
- Muirhead CA (2001) Consequences of population structure on genes under balancing selection. *Evolution* 55:1532–1541
- Muirhead CA, Glass NL, Slatkin M (2002) Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. *Genetics* 161:633–641
- Nasrallah JB (1997) Evolution of the *Brassica* self-incompatibility locus: a look into S-locus gene polymorphisms. *Proc Natl Acad Sci USA* 94:9516–9519
- Nasrallah JB, Kao TH, Chen CH, Goldberg ML, Nasrallah ME (1987) Amino-acid sequence of glycoproteins encoded by three alleles of the S locus of *Brassica oleracea*. *Nature* 326:617–619
- Oldroyd BP, Wongsiri S (2006) Asian honey bees. Biology, conservation and human interactions. Harvard University Press, Cambridge
- Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* 98:13757–13762
- Radloff SE, Hepburn C, Hepburn HR, Fuchs S, Hadisoelilo S, Tan K et al. (2010) Population structure and classification of *Apis cerana*. *Apidologie* 41:589–601
- Raper JR (1966) Sexuality in fungi. (Book reviews: genetics of sexuality in higher fungi). *Science* 154:758
- Richman AD, Kohn JR (1999) Self-incompatibility alleles from *Physalis*: implications for historical inference from balanced genetic polymorphisms. *Proc Natl Acad Sci USA* 96:168–172
- Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE et al. (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol* 34:3299–3302
- Schierup MH, Bechsgaard JS, Christiansen FB (2008) Selection at work in self-incompatible *Arabidopsis lyrata*. II. Spatial distribution of S haplotypes in Iceland. *Genetics* 180:1051–1059
- Schierup MH, Vekemans X, Charlesworth D (2000) The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genet Res* 76:51–62
- Smith DR (2011) Asian Honeybees and Mitochondrial DNA. In: Hepburn HR, Radloff SE (eds) *Honeybees of Asia*. Springer-Verlag Berlin, Heidelberg, p 69–93
- Solignac M, Vautrin D, Loiseau A, Mougel F, Baudry E, Estoup A et al. (2003) Five hundred and fifty microsatellite markers for the study of the honeybee (*Apis mellifera* L.) genome. *Mol Ecol Notes* 3:307–311
- Tajima F (1989) The effect of change in population size on DNA polymorphism. *Genetics* 123:597–601
- Takahashi J, Shimizu S, Koyama S, Kimura K, Shimizu I, Yoshida T (2009) Variable microsatellite loci isolated from the Asian honeybee, *Apis cerana* (Hymenoptera; Apidae). *Mol Ecol Resour* 9:819–821
- Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124:967–978
- Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N et al. (2014) A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat Genet* 46:1081–1088
- Walsh PS, Metzger DA, Higuchi R (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques* 10:506–513
- Wang J (2015) Does GST underestimate genetic differentiation from marker data? *Mol Ecol* 24:3546–3558
- Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL (2018) Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Mol Biol Evol* 35:773–777
- Woyke J (1963) What happens to diploid drone larvae in honeybee colony? *J Apic Res* 2:73–75
- Wright S (1939) The distribution of self-sterility alleles in populations. *Genetics* 24:538–552
- Wu Q, Han TS, Chen X, Chen JF, Zou YP, Li ZW et al. (2017) Long-term balancing selection contributes to adaptation in *Arabidopsis* and its relatives. *Genome Biol* 18:217
- Yang S, Wang L, Huang J, Zang X, Yuan Y, Chen JQ et al. (2015) Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523:463–467
- Yokoyama S, Nei M (1979) Population dynamics of sex-determining alleles in honey bees and self-incompatibility alleles in plants. *Genetics* 91:609–626
- Zareba J, Blazek P, Laszkiewicz A, Sniezewski L, Majkowski M, Janik S et al. (2017) Uneven distribution of complementary sex determiner (csd) alleles in *Apis mellifera* population. *Sci Rep* 7:2317