



HAL
open science

Genetic diversity and population genetic structure analysis of *Apis mellifera* subspecies in Algeria and Europe based on complementary sex determiner (CSD) gene

Riad Fridi, Nacera Tabet Aoul, Guillau Catays, Benjamin B. Basso, Kaspar Bienefeld, Aleš Gregorc, Alain Vignal, Kamila Canale-Tabet

► To cite this version:

Riad Fridi, Nacera Tabet Aoul, Guillau Catays, Benjamin B. Basso, Kaspar Bienefeld, et al.. Genetic diversity and population genetic structure analysis of *Apis mellifera* subspecies in Algeria and Europe based on complementary sex determiner (CSD) gene. *Apidologie*, 2022, 53 (1), pp.4. 10.1007/s13592-022-00920-x . hal-03626711

HAL Id: hal-03626711

<https://hal.inrae.fr/hal-03626711>

Submitted on 31 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Genetic diversity and population genetic structure analysis of *Apis mellifera* subspecies in Algeria and Europe based on complementary sex determiner (*CSD*) gene

Riad FRIDI¹, Nacera TABEL AOUL^{1,2} , Guillaume CATAYS³, Benjamin BASSO^{4,5}, Kaspar BIENEFELD⁶ , Aleš GREGORC⁷ , Alain VIGNAL³ , and Kamila CANALE-TABET³

¹ Molecular and Cellular Genetics Laboratory, Department of Genetics Applied, Faculty of SNV, Sciences and Technology University of Oran Mohamed Boudiaf, BP 1505, El M'naouer, Oran, Algeria

² Department of Biotechnology, Faculty of SNV, University of Oran, 1 Ahmed Benbella, Oran, Algeria

³ INRAE Occitanie-Toulouse: Institut National de Recherche pour l'Agriculture l'Alimentation et l'Environnement Centre Occitanie, Toulouse, France

⁴ Institut de l'abeille (ITSAP), UMT PrADE, 8914 Avignon, France

⁵ INRA, UR 406 Abeilles et Environnement, UMT PrADE, 84914 Avignon, France

⁶ Institut for Bee Research Hohen Neuendorf, F.-Engels-Straße 32, 16540 Hohen Neuendorf, Germany

⁷ Agricultural Institute of Slovenia, Ljubljana, Slovenia and University of Maribor, Faculty of Agriculture and Life Sciences, Maribor, Slovenia

Received 3 August 2021 – Revised 14 January 2022 – Accepted 27 January 2022

Abstract – In honeybees, the mechanism of sex determination depends on genetic variation at the *complementary sex determiner* (*CSD*) locus, which has a large allelic diversity. In this study, we examined the population genetic structure and genetic diversity within the highly variable region (*HVR*) of *CSD* in five *Apis mellifera* subspecies, in addition to Buckfast and unknown mixed ancestry bees. We sequenced *CSD* in 329 drones, 146 from Algeria (*A. m. intermissa* and *A. m. sahariensis* subspecies) and 183 from Europe (*A. m. ligustica*, *A. m. carnica*, *A. m. mellifera* subspecies, Buckfast samples, and individuals of unknown mixed ancestry). A total of 119 nucleotide haplotypes were detected. These corresponded to 119 protein haplotypes, of which 81 were new. The analysis of these haplotypes showed that *HVR* diversity levels were comparable with those in other populations of honeybee worldwide. Paradoxically, this high level of diversity at the locus did not allow for a separation of the samples according to their subspecies origin, which suggested either an evolutionary convergence or a conservation of alleles across subspecies, and an absence of genetic drift. Our results can be used to provide more information about the *CSD* diversity to include in breeding programs of honeybee populations.

Apis mellifera / complementary sex determiner / polymorphism / DNA / genetic diversity / *HVR*

1. INTRODUCTION

Hymenoptera is one of 11 orders of holometabolous insects, which includes 200,000 species of sawflies, wasps, ants, and honeybees (Wilson and Holldobler 2005). Hymenoptera lack sex chromosomes, and sex is usually determined by a single locus. These species are characterized by

Corresponding author: K. CANALE-TABET,
kamila.tabet@inrae.fr
Manuscript editor: Klaus Hartfelder

the principle of haplodiploid sex determination system found in 12% of all animal species (Beye et al. 2003). In many haplodiploid hymenopteran species, the molecular mechanism underlying female development depends on heterozygosity at the complementary sex determiner (*CSD*) locus (Heimpel and de Boer 2008). In honeybees, females (either queens or workers) develop from diploid embryos (two chromosome sets) and are heterozygous for the *CSD* locus, whereas individuals homozygous for this locus give diploid males. The colony destroys any occurring diploid male during the early development stages, whereas males (drones) developed from unfertilized haploid embryos (one chromosome set) survive (Woyke 1963). *CSD* which is located in chromosome 3 is 11,734 bp long and consists of 9 exons (Kaskinova and Nikolenko 2017). According to Hasselmann and Beye (2004), these 9 exons are distributed in three regions: 1 to 3, 4 to 5, and 6 to 9 and are separated by two prolonged introns. The polymorphism level in the *CSD* coding regions is seven times higher than in the noncoding regions (Beye et al. 2003). The genomic region from exons 6 to 9 has the highest polymorphism level compared to the other two regions (Hasselmann and Beye 2004). Nucleotide polymorphisms accumulate within a confined part of the *CSD* gene, characterized as the Potential Specifying Domain (*PSD*) (Hasselmann et al. 2008). The *PSD* region which includes a hypervariable region (*HVR*) is located on exon 7 which consists of a variable number of repeats, mainly containing A/T rich motifs, and encodes large tyrosine (Y) and asparagine (N) residues. It is flanked by the domains encoding the arginine/serine and proline-rich regions, both of which likely take part in protein–protein interactions. The gene encodes an SR-type protein and is a potential splicing factor (Beye et al. 2003). The *CSD* locus variability and the allelic composition can be done by only amplifying and sequencing exons 6, 7, and 8 (Lechner et al. 2014).

The *CSD* gene has evolved by a gene duplication event from an ancestral copy of its paralog gene *feminizer* (*FEM*) within the honeybee *Apis* genus. Heterozygosity at *CSD* is required to induce the female pathway by interacting with

transformer 2 (a protein that contains the RNA-binding domain) leading to a female-spliced *FEM* transcript. This creates the protein product that is responsible for the female-specific splicing of the *Am-DSX* transcript (Gempe et al. 2009). 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. Consequently, the translation is terminated prematurely, and *Am-DSX* is spliced in a male-specific manner, producing a protein with a male-specific carboxy-terminal end (Nissen et al. 2012; Biewer et al. 2015; Hasselmann et al. 2010).

Low allelic diversity for *CSD* leads to a loss of colony strength due to the emergence of non-viable diploid drones; therefore, the identification of alleles is an important task for breeding programs, particularly for those employing artificial insemination of (Kaskinova et al. 2019). The *CSD* gene in honeybees was first found as having 20 alleles (Beye et al. 2003). Later, Lechner et al. (2014) raised this number to 87 and extrapolated to 116–145 existing alleles. In 2017, 121 different *CSD* alleles were identified within two Polish honeybee populations (Zareba et al. 2017). In 2020, Bilodeau et al. identified 83 and 62 alleles in two studied populations. *CSD* evaluation in breeding programs in New Zealand (Hyink et al. 2013), Russia (Kaskinova et al. 2019), and the USA (Bilodeau et al. 2020) found *HVR* diversity levels comparable to those found in populations of honeybees worldwide (Lechner et al. 2014; Zareba et al. 2017).

In this study, we assessed the haplotypic diversity of the *CSD* locus, focusing on the *HVR* of honeybees from breeding apiaries in Europe: *A. m. ligustica*, *A. m. carnica*, *A. m. mellifera*, Buckfast, and unknown population, and from Algeria: *A. m. intermissa*, *A. m. sahariensis*, which represents the North Africa. Determination of genetic diversity and phylogenetic analysis were done in order to examine if samples could be separated according to their subspecies and identify the presence or not of haplotypes that had not been previously described. Since some samples are collected from regions never studied before, we expect to find new haplotypes.

stage, we used head and thorax, which was the case for all the Algerian samples. At a nymphal stage, all the body was used for *DNA* extraction (European samples).

After cutting with a scalpel, the material was placed in tubes containing 490 μL cell lysis buffer (TNES- Urea: 1 M Tris HCl, pH 8, 3 M NaCl, 0.5 M EDTA, 10% SDS). Proteinase K (eurobio GEXPRO01) 10 mg/mL was added in two times: (i) 12.5 μL at 56 °C incubation under constant agitation during 3 h. (ii) 5 μL at 37 °C with an overnight incubation. Precipitation of cell debris was performed by adding 200 μL 3 M NaCl to each sample then stirring the tubes by inversion followed by centrifugation at 12,500 g for 30 min at 4 °C. Each supernatant was carefully transferred to a new tube and 2.5 μL 100 mg/mL RNase were added, and then transferred to a new tube with 100% ethanol to allow *DNA* precipitation. *DNA* was pelleted then washed with 70% ethanol, air dried for 5 min, then resuspended in 100 μL TE buffer 10/0.1 (10 mM Tris HCl, 0.1 mM EDTA). When the *DNA* did not appear after adding 100% ethanol, a centrifugation at 8000 g for 10 min at 4 °C was performed. The supernatant was eliminated and the pellet was resuspended in the same way as described before. *DNA* solutions were incubated at 37 °C overnight with constant agitation. *DNA* quality and quantity controls were measured using a NanoDrop 8000 (Thermo Scientific, Wilmington, DE, USA) spectrophotometer and the PicoGreen fluorescence assay. All *DNA* were submitted to an electrophoresis in 0.8% agarose gel.

2.3. PCR amplification

Two primers pairs (set 1 and 2) were designed in exons 6 and 8 with primer 3 0.4.0 (Koressaar and Remm 2007; Untergasser et al. 2012) using the *Amel_4.5* reference genome. These primers sets allowed the analysis of introns 6, 7 and exon 7 including *HVR*. Primers set 1 (F1 *CSD10984*, 5'-AGACGATATGAAAATTACACAATGA-3', R1 *CSD8657*, 5'-CCTAAATCTTGGTATTTGTTCTT-3') theoretical size according

to *Amel_4.5*: 566pb. Primer set 2 (F2 *CSD8110*, 5'-ACAATGAAAAAGAAAACTTTTAGA-3', R2 *CSD8695*, 5'-GGAATGAATCGTGGAATGG-3') theoretical size according to *Amel_4.5*: 588 pb.

A final volume of 25 μL amplified sample consisted of 50 ng *DNA* [10 ng/ μL], 0.5 U TaqTM (Promega ref M8295), 1 \times Buffer, 0.2 mmol/L dNTP, 1.5 mmol/L MgCl₂, 0.4 mmol/L of each primer and ultrapure water. PCR amplification initiated by 5 min initial denaturation at 94 °C, followed by 30 cycles of 30 s denaturation at 94 °C, 30 s annealing at 58 °C, and 30 s elongation at 72 °C, ultimately completed by 20 min terminal elongation at 72 °C. Positive (*DNA*) and negative (water) control were systematically used in each PCR run. Five microliters of each PCR product was loaded on 1.5% agarose gel. Whenever no PCR product was observed with the primer set 1, an attempt was done with set 2. If unsuccessful, the sample was removed from the study.

2.4. Sanger sequencing

The PCR products were sequenced in both directions by an automated Sanger sequencer (ThermoFisher ABI3730). According to the intensity of the band observed on agarose gel, 3 to 8 μL of PCR product was digested with 10 U exonuclease I (Promega M9910) and 0.5 U Shrimp Alkaline Phosphatase (OzymeM0293L) for 45 min at 37 °C. After enzyme inactivation (incubation for 30 min at 80 °C), the sequencing reaction was immediately carried out as follows: 5 \times sequencing buffer, BigDye Terminator Cycle Sequencing mix v 3.1, and 5 μM *CSD* primer (1 cycle of 5 min denaturation at 95 °C, followed by 25 cycles elongation of 5 min denaturation at 95 °C, 15 s at 55 °C, and 4 min at 60 °C). Excess fluorescent ddNTP was removed by sephadex G50 (GE Healthcare) gel filtration chromatography and centrifugation at 755 g for 3 min at 4 °C without break. The purified sequencing products were subjected to capillary electrophoresis (ThermoFisher ABI3730).

2.5. Data analysis

The agreement between forward and reverse sequence chromatograms was checked by visualization with Chromas 2.6 (Technelysium Pty Ltd, Brisbane, Australia) and the sequences were exported as FASTA files. The sequence alignment was done with Clustal Omega 1.2.4 with default parameters (Madeira et al. 2019). EMBOSS-Transeq was used to translate DNA sequences to protein in silico (Madeira et al. 2019). Among the 6 reading frames, we selected the one giving the conserved amino acid sequence IEQIP situated after the hypervariable region as described by Lechner et al. (2014). Alignment visualization and edition were done with Jalview 2.10.5 using the default alignment parameters (Waterhouse et al. 2009).

2.5.1. Genetic diversity analysis

Haploid DNA sequence data were analyzed in DnaSP (DNA Sequence Polymorphism) 6.12.01 (Rozas et al. 2017) for diversity measurements: haplotype diversity (H_d), nucleotide diversity (π), average number of nucleotide differences between haplotypes (k), number of variable sites (S), number of haplotypes excluding and including indel sites h_1 , h_2 respectively, number of non-synonymous substitutions per non-synonymous site (K_a), and number of synonymous substitutions per synonymous site (K_s).

2.5.2. Phylogenetic analysis

Molecular phylogenetic analyses were conducted using MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms (Kumar et al. 2018) to perform the maximum likelihood (ML) tree searches and bootstrap estimation of node support. We used the best-fit substitution model application implemented in MEGA X to compare the available amino acid substitution models and obtain the

best description of the substitution pattern by maximum likelihood. Using this approach, the GTR + G + I (General Time Reversible, (+G) non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution, (+I) assuming that a certain fraction of sites is evolutionarily invariable) model was assigned to be the best-fit evolutionary model for the data set of this present study by MEGA X and tree was inferred from evolutionary distances by maximum likelihood. Trust in nodes was estimated by 1000 bootstrap repeats. The median joining network (MJN) analysis of the haplotypes based on the variable characters of the complete alignment was constructed to infer evolutionary relationships between the Algerian and European haplotypes of this study using the NETWORK 10 software (Bandelt, Forster, and Rohl 1999). The principal component analysis (PCA) was carried out by RStudio v1.4 (R Core Team 2018) using the following packages: the Adegenet 2.1.3 (Jombart and Ahmed 2011), the Ape 5.5 (Paradis and Schliep 2019), and Ade4 1.7–17 (Bougeard and Dray 2018). In this analysis, two outlier samples (Buckfast_ITSAP5 and Sahariensis_Bayadh26) were removed to improve the resolution of the figure.

3. RESULTS

Out of 381 samples tested, 47 from Algeria and 5 from Europe could not amplify, as evaluated by agarose gel electrophoresis, with either of the two primer sets tested (Supplementary Table I). Consequently, we obtained 329 amplified DNA samples from Algeria ($N = 146$) and Europe ($N = 183$). For some individuals, the length polymorphism of the *HVR* could be visually observed on agarose gel (Fig. 2A). We obtained complete sequence coverage and agreement between forward and reverse reactions from the start of intron 6 to the end of intron 7 for all remaining 329 samples: 146 and 183 from Algerian and European populations respectively (Fig. 2B; Supplementary Fig. 1; Supplementary Table II).

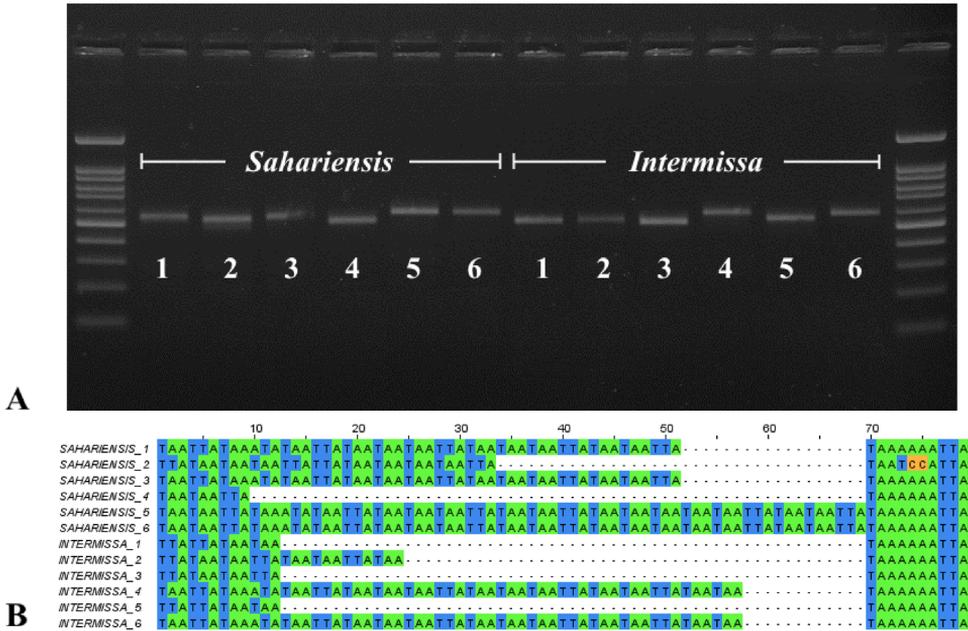


Figure 2. Examples of *HVR-CSD* polymorphism in Algerian adult drones from two Algerian subspecies (*sahariensis* and *intermissa*), **A** on agarose gel, **B** after Sanger sequencing.

3.1. CSD diversity analysis

3.1.1. Nucleotide polymorphism level

Analysis of exon 7 and introns 6 and 7 with DnaSP showed that the length of the nucleotide sequences for the 329 samples ranges between 351 and 370 pb. Due to the different biological mechanisms leading to the generation of SNPs and Indels, to the fact that ambiguities in the placement of indels can influence the count of SNPs, and to the fact that most analysis softwares do not take indels into account, we counted haplotypes with indels, but also gave a more conservative number without indels. Including indel sites, we identified 124 *DNA* haplotypes: 50 and 72 found in Algerian and European subspecies respectively; and 2 in common between both populations. Excluding indel sites, we identified 108 *DNA* haplotypes (Table I).

When restricting our analysis to exon 7, the length of the sequences ranged between 211 and 301pb. Including indel sites, we identified 119 different *DNA* haplotypes: 46 in Algerian

subspecies while 69 others were found in European ones. And four were common to both Algerian and European subspecies. Excluding indel sites, we identified 104 *DNA* haplotypes: 43 and 55 were found in Algerian and European subspecies respectively; and 6 were common to both of them (Table I).

Population genetic diversity indices were calculated using nucleotide data for exon 7. The genetic diversity indices (*S*, *Hd*, π , and *K*) were calculated without including indels except indice *h1* (number of haplotypes including indel sites) (Table I). As shown in Table I, the *CSD* gene has a high level of polymorphism in all *A. m.* subspecies. A difference of 15 haplotypes was found when comparing the number of haplotype including indels ($h1 = 119$) and excluding indels ($h2 = 104$). The number of segregating (polymorphic) sites was 79, in which tri-allelic ($N = 20$) and tetra-allelic SNPs ($N = 5$) were detected. The average number of nucleotide differences (*k*) varied from 14.04 (for *A. m. carnica*) to 24.50 (for Buckfast). The haplotype diversity level (*Hd*) for the 329 sequences was 0.984 confirming the high

Table I

Genetic variability of exon 7 of *CSD* for all studied subspecies. *N*, number of sequences obtained; *S*, number of variable sites; *h1*, number of haplotypes including indel sites; *h2*, number of haplotypes excluding indel sites; *Hd*, haplotype diversity; π , nucleotide diversity; *k*, average number of nucleotide differences

Subspecies	N	S	h1	h2	Hd	π	k
<i>intermissa</i>	98	72	42	41	0.965	0.087	16.28
<i>sahariensis</i>	48	56	9	9	0.823	0.083	17.55
<i>carnica</i>	50	55	29	27	0.969	0.07	14.04
<i>mellifera</i>	67	63	23	21	0.909	0.08	16.45
<i>ligustica</i>	29	59	25	23	0.985	0.078	16.28
Buckfast	4	43	4	4	1.000	0.106	24.50
unknown	33	64	23	22	0.964	0.088	17.36
All subspecies	329	79	119	104	0.984	0.091	15.14

genetic diversity characteristic, and was substantially similar between Algerian and European subspecies. It varied from 0.823 (*A. m. sahariensis*) to 1 (Buckfast).

In order to compare the variability found in introns 6 and 7 and to compare it to exon 7, the nucleotide diversity (π) was calculated separately for the three regions and was found to be 0.034, 0.091, and 0.026 for intron 6, exon 7, and intron 7 respectively. Thus, exon 7 showed at least a threefold greater diversity than that of its neighboring introns (Fig. 3; Table I; Supplementary Table III), consequently confirming the previous

observations which showed that the polymorphism level in the *CSD* coding regions is much higher than in the noncoding regions (Beye et al. 2003).

We compared our haplotypes with the 854 *CSD* nucleotide sequences available in the GenBank database for the same region. The analysis of exon 7 and introns 6 and 7 shows that 101 of the haplotypes identified here were new: 50 specific to each of the Algerian and European subspecies and one shared by both. The analysis of the exon 7 sequences revealed 85 haplotypes that had not been previously described: 45 were

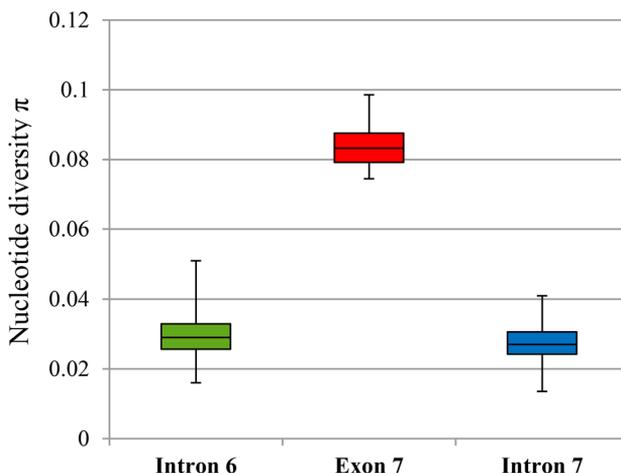


Figure 3. Nucleotide diversity π of introns 6 and 7 and exon 7 of *CSD* for all studied subspecies.

found only in Algeria, and 39 in Europe. One haplotype was common to the Algerian and European subspecies.

3.1.2. Amino acid polymorphism level

The 329 exon 7 nucleotide sequences were translated into amino acid sequences, giving peptides between 70 and 116 amino acids long. The HVR, enriched in N and Y residues and situated between a SLS and a KKL motif, was variable in length, between 10 and 37 amino acids (Supplementary Fig. 2).

The 329 amino acid sequences were assigned to 119 different amino acid haplotypes, among 46 were found in Algeria, and 69 in Europe. Four haplotypes were shared between Algerian and European populations with different frequencies (Supplementary Table IV). Among all 119 haplotypes, 48 haplotypes were present only once in each individual: *A. m. intermissa* = 18, *A. m. sahariensis* = 3, *A. m. carnica* = 8, *A. m. mellifera* = 3, *A. m. ligustica* = 9, Buckfast = 2, unknown = 5. Forty-four haplotypes could be found in more than one individual from the same population. However, 27 haplotypes were shared between individuals from different populations.

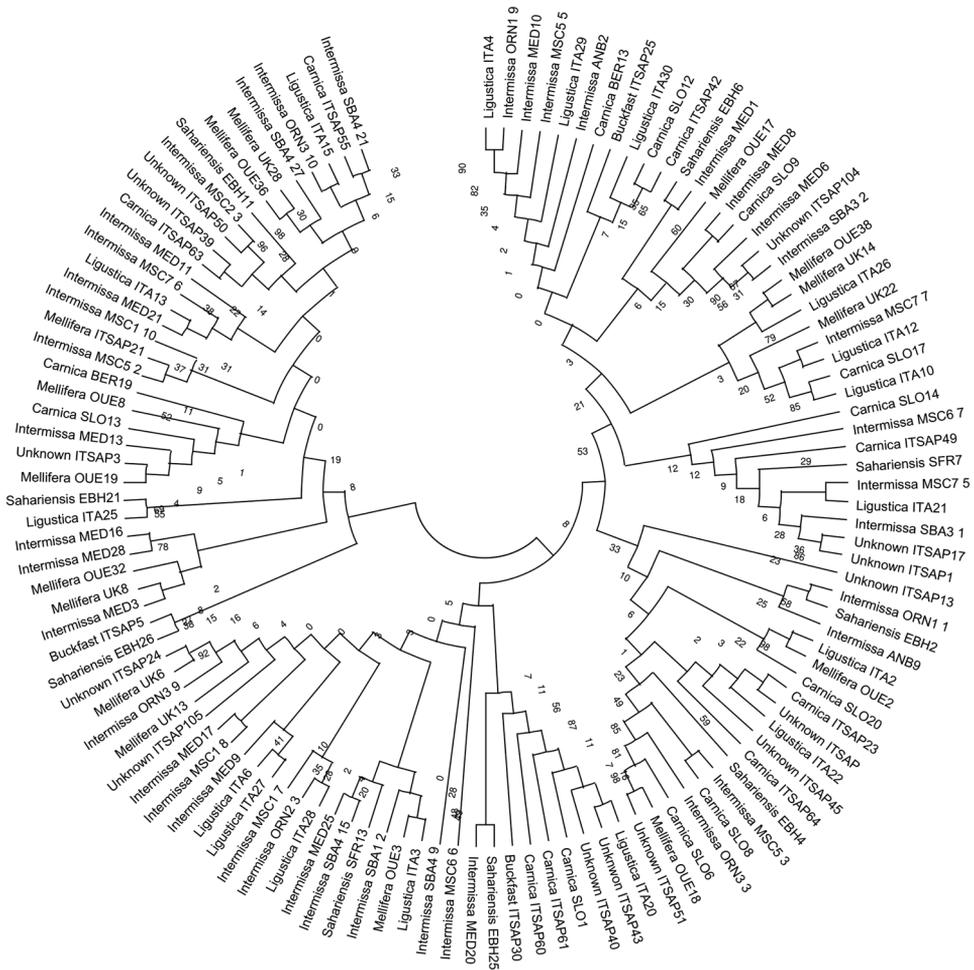


Figure 4. Maximum likelihood phylogenetic tree of *CSD* protein haplotypes. The figure outlines phylogenetic relationships inferred from protein sequence alignments between *CSD* haplotypes. Each tip of the tree is labeled with the haplotype name. The percent bootstrap value is represented by the numbers at the node after 1000 replication.

We compared the obtained amino acid sequences with 650 *CSD* sequences available in GenBank database. Among the 119 haplotypes, 81 haplotypes were newly described, of which 45 found only in Algeria, 36 only in Europe. The remaining 38 haplotypes were identical to sequences previously found in GenBank. The set of new nucleotide sequences identified in this study have been deposited to GenBank (www.ncbi.nlm.nih.gov/genbank) nucleotide sequence database and have been assigned the accession numbers (MZ674091-MZ674175).

In exon 7, the number of non-synonymous substitutions ($N=80$) was much higher than the number of synonymous substitutions ($N=27$). Testing for positive selection was done by

calculating the Ka/Ks ratio. The number of non-synonymous substitutions per non-synonymous site (Ka) was equal to 0.10, whereas the number of synonymous substitutions per synonymous site (Ks) was equal to 0.09, giving a ratio of 1.1, indicative of positive selection.

3.2. Phylogenetic and network analysis

As part of our investigation to check if the relationships between the 119 protein *CSD* haplotypes identified in the Algerian and European subspecies follow the known genetic structure of the populations studied, we performed a maximum likelihood phylogenetic tree, a median

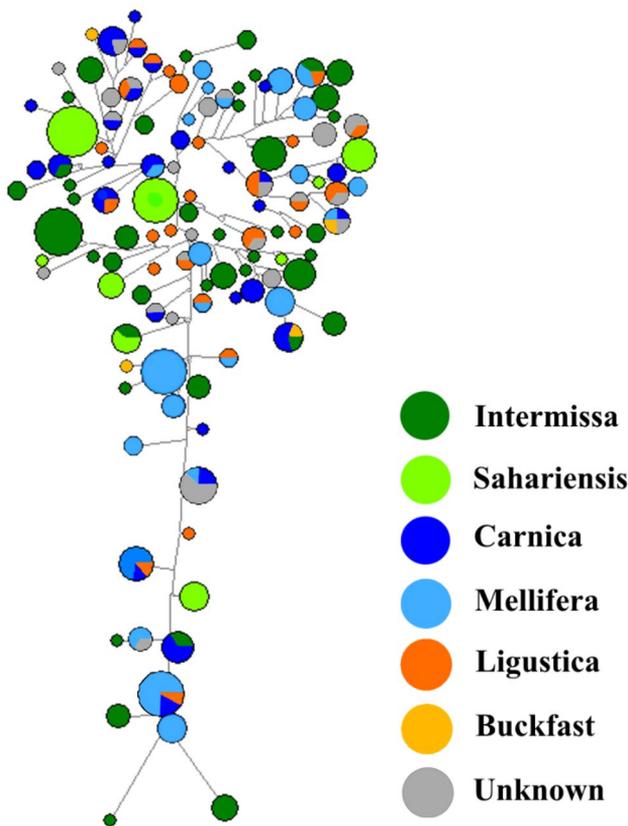


Figure 5. Median-joining network and distribution of unique *CSD* nucleotide haplotypes for exon 7. Each circle represents one unique sequence in the data set. Circle size is proportional to the total number of individuals with the haplotype. Colors correspond to each one of subspecies.

joining network, and a principal component analysis (PCA). All three analyses showed that the haplotypes were not grouped according to their *A. m.* subspecies origin, but that these genetic origins were well mixed with each other (Figs. 4, 5, 6).

4. DISCUSSION

Previous studies have shown that the *CSD* gene in *A. mellifera*, *A. cerana*, and *A. dorsata* has a very high level of polymorphism in these three honeybee species (Hasselmann and Beye 2004; Cho et al. 2006; Hasselmann and Beye 2006; Hasselmann et al. 2008; Wang et al. 2012). In this study, we confirmed that the *CSD* gene also shows a high level of within subspecies polymorphism, and presented the first results for Algerian subspecies.

In the present work, we sequenced 329 honeybees from different subspecies collected from Algeria in addition to five different European countries. Analysis of these sequences showed that genetic diversity indices such as total nucleotide diversity per site and average number of nucleotide differences between any two sequences agreed with values previously observed in other studies (Hasselmann and Beye 2006; Hasselmann et al. 2008; Wang et al. 2012; Lechner et al. 2014; Bilodeau et al. 2020). The estimate of nucleotide diversity was three times higher in exon 7, as compared to introns 6 and 7. Our results reveal that exon 7 of the

CSD locus has a much higher genetic diversity than the mean value in the *A. m.* genome. This is supported by the lower overall genome average nucleotide diversity found in modern samples ($\pi \cong 0.001$) and historical ones ($\pi \cong 0.003$) (Themudo et al. 2020), as compared to the exon 7 of the *CSD* locus ($\pi \cong 0.090$) studied here (Table I).

The haplotype diversity of the *CSD* protein in the populations sampled is not independent on the sampling size. In the case of *A. m. sahariensis*, we sequenced in one of the locations (the Ain Sefra site) 26 drones from a single hive, which gave us only two haplotypes, corresponding as expected to the two alleles present in the queen's genome. In the remaining 22 *A. m. sahariensis* drones sequenced, seven more haplotypes were found. For all the other subspecies or populations with more than 20 drones sequenced, the number of haplotypes ranges between 21 and 41 (Table I). This is well within the number of protein haplotypes found in other studies, in which values vary from 16 to over 100 haplotypes (Hyink et al. 2013; Zareba et al. 2017; Kaskinova et al. 2019; Bilodeau et al. 2020). The diversity could represent even more haplotypes if the 52 samples which were not amplified could be analyzed. This was due probably to the high divergence in the flanking regions where the primers were designed. Additional data on flanking sequence for these samples, such as obtained in whole genome sequencing data, for which there is no specific amplification step, would allow designing new primers.

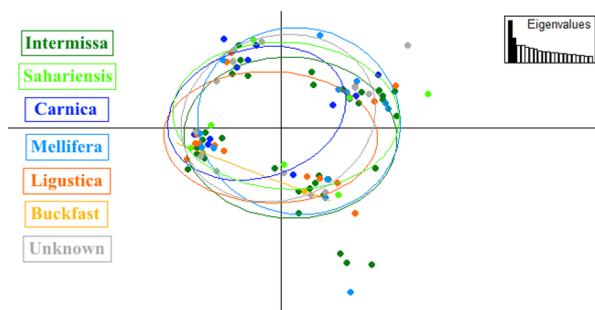


Figure 6. Principal component analysis (PCA) of unique *CSD* nucleotide haplotypes for exon 7. Each point represents one unique sequence in the data set. Colors correspond to subspecies.

The length of the *HVR* shows considerable variation across populations worldwide. The range of 10 to 37 amino acids for the length of the *HVR* we found in our study was well within that found for other bee populations (Hasselmann and Beye 2006; Hasselmann et al. 2008; Wang et al. 2012; Hyink et al. 2013; Lechner et al. 2014; Zareba et al. 2017; Kaskinova et al. 2019; Bilodeau et al. 2020). Typically large ranges can be found, such as 6 to 33 amino acids in Kenyan populations (Lechner et al. 2014) or more restricted ones: between 21 and 38 amino acids in a smaller sampling of a closed breeding population (Hyink et al. 2013).

We found a total of 119 nucleotide and amino acid haplotypes in exon 7 (introns excluded), 48 of which being unique and present in a single individual. Conversely, 27 haplotypes were present in sequences from different subspecies, suggesting either evolutionary convergence or the possibility of a high pressure for haplotype conservation across subspecies. A striking fact is that all 119 nucleotide haplotypes correspond to 119 protein haplotypes, meaning that each nucleotide haplotype has at least one non-synonymous mutation.

When comparing our data to the *CSD* haplotype sequences found in the international databases, forty-five new haplotypes were found only in the 122 samples from Algeria (excluding *A. m. sahariensis* drones coming from the same hive), which represent 37%, and 36 new European haplotypes from 183 samples, which represent 19%. Therefore, sequencing new subspecies such as *A. m. intermissa* and *A. m. sahariensis* contributed to the identification of new haplotypes that had not been studied with European subspecies.

However, although it appears that each subspecies has its own set of specific haplotypes, the phylogenetic tree, the haplotype network, and the PCA analyses of the sequences obtained failed to reveal any structure following the expected subspecies pattern. This is due to the fact that closely related sequences differing by only very few SNPs can be found in very different subspecies. One hypothesis for the presence of identical haplotypes in two completely unrelated subspecies could be an evolutionary convergence, as

already suggested (Hasselmann et al. 2008). The other is the rate of positive selection incredibly high as Hasselmann and Beye (2004) showed. The hypothesis of positive selection acting at the *CSD* locus was confirmed by the ratio of non-synonymous to synonymous substitutions ($dN/dS = 1.10$) (Jeffares et al. 2015) and supported by the number of non-synonymous substitutions ($N = 80$) that was much higher than the number of synonymous substitutions ($N = 27$), in addition to the presence of tri-allelic and tetra-allelic SNPs (usually a rare event), coding for three different amino acids.

The sequencing of the *HVR-CSD* in exon 7 and its flanking introns in Algerian and European populations contributed to the identification of new haplotypes, especially in the Algerian subspecies that was investigated for the first time for *CSD*. Thus, it is important to continue to include African subspecies that are underrepresented on *CSD* variability studies, in addition to the fact that *Apis mellifera* originated in Africa (Whitfield et al. 2006). The high level of diversity at the locus does not allow for a separation of the samples according to their subspecies, suggesting either evolutionary convergence or a conservation of alleles across subspecies since their separation, and an absence of genetic drift. According to our results, this region will certainly be enriched by new haplotypes as new subspecies are analyzed. This study may help inform bee-breeding programs about *CSD* haplotype diversity in their populations and help make breeding decisions that minimize the loss in brood viability caused by inbreeding.

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1007/s13592-022-00920-x>.

ACKNOWLEDGEMENTS

We thank Hamzaoui Mohamed and the "Association nationale des apiculteurs professionnels" (ANAP) for providing honeybee samples from Medea, Annaba, and El Bayadh (Algeria); Saci Soumia for samples from Ain Sefra (Algeria); Andrew Abrahams for samples from Colonsay (Scotland); Cecilia Costa for samples from

Italy and the “Association Conservatoire de l’Abeille Noire Bretonne” (ACANB) for samples from Oues-sant (France). Colonsay and Ouessant samples were collected for the SeqApiPop program. We would also like to thank all the beekeepers for allowing access to their apiaries and their disinterested assistance in sample collection. Sequencing was performed in collaboration with the GeT platform, Toulouse (France), a partner of the National Infrastructure France Génomique, thanks to the support by the Commissariat aux Grands Investissements (ANR-10-INBS-0009). The mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation.

AUTHOR CONTRIBUTION

NTA, KT, AV, and FR designed the project, FR provided samples for Algerian population, and BB, KB, and AG provided samples for the European populations. FR, GC, and KT performed the laboratory experiments and data analyses. FR and KT deposited sequences in the database. NTA, KT, AV, GC, and FR discuss the results. FR drafted the manuscript. AV, KT, and NTA revised the manuscript. All authors reviewed and accepted the final draft of the manuscript.

FUNDING

This research was funded by a grant from the INRA Département de Génétique Animale (INRA Animal Genetics division), by FranceAgriMer grant 14–21-AT and by the PHC Tassili 21MDU312 (Partenariat Hubert Curien Tassili) program.

AVAILABILITY OF DATA AND MATERIAL

The set of new nucleotide sequences identified in this study have been deposited to GenBank (www.ncbi.nlm.nih.gov/genbank) nucleotide sequence database and have been assigned the accession numbers (MZ674091–MZ674175).

CODE AVAILABILITY

Not applicable.

DECLARATIONS

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

REFERENCES

- Baldensperger, P.J. (1932) Variétés d’abeilles en afrique du nord. proceedings of the five international congress on entomology, 1932.
- Bandelt, H. J., Forster P., Rohl A. (1999) Median-Joining Networks for inferring intraspecific phylogenies. *Molecular biology and evolution* 16 (1): 37–48. <https://doi.org/10.1093/oxfordjournals.molbev.a026036>.
- Beye, M., Hasselmann M., Fondrk M.K., Page R. E., Omholt S. W. (2003) The gene CSD Is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* 114 (4): 419–29. [https://doi.org/10.1016/S0092-8674\(03\)00606-8](https://doi.org/10.1016/S0092-8674(03)00606-8).
- Biewer, M., Schlesinger F., Hasselmann M. (2015) The evolutionary dynamics of major regulators for sexual development among hymenoptera species. *Frontiers in Genetics* 6 (124). <https://doi.org/10.3389/fgene.2015.00124>.
- Bilodeau, L., Avalos A., Danka R.G. (2020) Genetic diversity of the complementary sex-determiner (Csd) gene in two closed breeding stocks of varroa-resistant honey bees. *Apidologie* 51 (6): 1125–32. <https://doi.org/10.1007/s13592-020-00790-1>.
- Bougeard, S., Dray S. (2018) Supervised multiblock analysis in R with the Ade4 package. *Journal of Statistical Software* 86 (1). <https://doi.org/10.18637/jss.v086.i01>.
- Buttel-Reepen H. (1906) Contributions to the systematics biology as well as the historical and geographical distribution of honeybees (*Apis Mellifica* L.) Their variability and other apis species. *Mitteilungen aus dem zoologischen museum in berlin* 7: 117–201.
- Cho, S., Huang Z. Y., Green D. R., Smith D. R., Zhang J. (2006) Evolution of the complementary sex-determination gene of honey bees: balancing

- selection and trans-species polymorphisms. *Genome research* 16 (11): 1366–75. <https://doi.org/10.1101/gr.4695306>.
- Themudo, E. G., Rey-Iglesia A., Tascón L. R., Jensen A. B., Da Fonseca R. R., Campos P. F. (2020) Declining genetic diversity of european honeybees along the twentieth century. *Scientific reports* 10 (1): 10520. <https://doi.org/10.1038/s41598-020-67370-2>.
- Gempe, T., Hasselmann M., Schiøtt M., Hause G., Otte M., Beye M. (2009) Sex determination in honeybees: two separate mechanisms induce and maintain the female pathway. Edited by bruce stewart baker. *Plos biology* 7 (10): e1000222. <https://doi.org/10.1371/journal.pbio.1000222>.
- Haccour P. (1960) Recherche sur la race d'abeille saharienne au maroc. *Bull soc sci nat phys Maroc* 6: 96–98.
- Hasselmann, M., Beye M. (2004) Signatures of selection among sex-determining alleles of the honey bee. *Proceedings of the national academy of sciences* 101 (14): 4888–93. <https://doi.org/10.1073/pnas.0307147101>.
- Hasselmann, M., Beye M. (2006) Pronounced differences of recombination activity at the sex determination locus of the honeybee, a locus under strong balancing selection. *Genetics* 174 (3): 1469–80. <https://doi.org/10.1534/genetics.106.062018>.
- Hasselmann, M., Vekemans X., Pflugfelder J., Koeniger N., Koeniger G., Tingek S., Beye M. (2008) Evidence for convergent nucleotide evolution and high allelic turnover rates at the complementary sex determiner gene of western and asian honeybees. *Molecular biology and evolution* 25 (4): 696–708. <https://doi.org/10.1093/molbev/msn011>.
- 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. *Proceedings of the national academy of sciences* 107 (30): 13378–83. <https://doi.org/10.1073/pnas.1005617107>.
- Heimpel, G. E., De Boer J. G. (2008) Sex determination in the hymenoptera. *Annual review of entomology* 53 (1): 209–30. <https://doi.org/10.1146/annurev.ento.53.103106.093441>.
- Hyink, O., Laas F., Dearden P. K. (2013) Genetic tests for alleles of complementary-sex-determiner to support honeybee breeding programmes. *Apidologie* 44 (3): 306–13. <https://doi.org/10.1007/s13592-012-0181-6>.
- Jeffares, D. C., Tomiczek B., Sojo V., Dos Reis M. (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-1438-8_4.
- Jombart, T., Ahmed I. (2011) Adegnet 1.3–1: new tools for the analysis of genome-wide snp data. *Bioinformatics* 27 (September): 3070–71. 10.1093.
- Kaskinova, M. D., Nikolenko A. G. (2017) Csd gene of honeybee: genetic structure, functioning, and evolution. *Russian Journal of Genetics* 53 (3): 297–301. <https://doi.org/10.1134/S1022795417010070>.
- Kaskinova, M. D., Gataullin A. R., Saltykova E. S., Gaifullina L. R., Poskryakov A. V., Nikolenko A. G. (2019) Polymorphism of the hypervariable region of the csd gene in the apis mellifera l. population in southern urals. *Russian Journal of Genetics* 55 (2): 267–70. <https://doi.org/10.1134/S102279541902008X>.
- Koressaar, T., Remm M. (2007) Enhancements and modifications of primer design program primer3. *Bioinformatics* 23 (10): 1289–91. <https://doi.org/10.1093/bioinformatics/btm091>.
- Kumar, S., Stecher G., Li M., Knyaz C., Tamura K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Edited by fabia ursula battistuzzi. *Molecular biology and evolution* 35 (6): 1547–49. <https://doi.org/10.1093/molbev/msy096>.
- Lechner, S., Ferretti L., Schöning 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. *Molecular biology and evolution* 31 (2): 272–87. <https://doi.org/10.1093/molbev/mst207>.
- Madeira, F., Park Y. M., Lee J., Buso N., Gur T., Madhusoodanan N., Basutkar P., et al. (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research* 47 (W1): W636–41. <https://doi.org/10.1093/nar/gkz268>.
- Nissen, I., Müller 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 (3): 1015–26. <https://doi.org/10.1534/genetics.112.143925>.
- Paradis, E., Schliep K. (2019) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Edited by Russell Schwartz. *Bioinformatics* 35 (3): 526–28. <https://doi.org/10.1093/bioinformatics/bty633>.
- R Core Team. (2018) R: a language and environment for statistical computing. vienna, austria: R foundation for statistical computing. <https://www.R-project.org/>.
- Rozas, J., Ferrer-Mata A., Sánchez-DelBarrio J. C., Guirao-Rico S., Librado P., Ramos-Onsins S. E., Sánchez-Gracia A. (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular biology and evolution* 34 (12): 3299–3302. <https://doi.org/10.1093/molbev/msx248>.
- Untergasser, A., Cutcutache I., Koressaar T., Ye J., Faircloth B. C., Remm M., Rozen S. G. (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Research* 40 (15): e115–e115. <https://doi.org/10.1093/nar/gks596>.

- Wang, Z., Liu Z., Wu X., Yan W., Zeng Z. (2012) Polymorphism analysis of CSD gene in six *Apis mellifera* subspecies. *Molecular Biology Reports* 39 (3): 3067–71. <https://doi.org/10.1007/s11033-011-1069-7>.
- Waterhouse, A. M., Procter J. B., Martin D. M. A., Clamp M., Barton G. J. (2009) Jalview version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25 (9): 1189–91. <https://doi.org/10.1093/bioinformatics/btp033>.
- Whitfield, C. W., Behura S. K., Berlocher S. H., Clark A. G., Johnston J. S., Sheppard W. S., Smith D. R., Suarez A. V., Weaver D., Tsutsui N. D. (2006) Thrice out of africa: ancient and recent expansions of the honey bee, *Apis Mellifera*. *Science* 314 (5799): 642–45. <https://doi.org/10.1126/science.1132772>.
- Wilson, E. O., Holldobler B. (2005) The rise of the ants: a phylogenetic and ecological explanation. *Proceedings of the National Academy of Sciences* 102 (21): 7411–14. <https://doi.org/10.1073/pnas.0502264102>.
- Woyke, J. (1963) What Happens to Diploid Drone Larvae in a Honeybee *Colony*. *Journal of Apicultural Research* 2 (2): 73–75. <https://doi.org/10.1080/00218839.1963.11100063>.
- Wragg, D., Eynard S. E., Basso B., Canale-Tabet K., Labarthe E., Bouchez O., Bienefeld K., et al. (2021) Complex population structure and haplotype patterns in western Europe honey bee from sequencing a large panel of haploid drones. Preprint. *Genetics*. <https://doi.org/10.1101/2021.09.20.460798>.
- Zareba, J., Blazej P., Laszkiewicz A., Sniezewski L., Majkowski M., Janik S., Cebat M. (2017) Uneven distribution of complementary sex determiner (CSD) alleles in *Apis mellifera* population. *Scientific Reports* 7 (1): 2317. <https://doi.org/10.1038/s41598-017-02629-9>.

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