HAL open science

Complex population structure and haplotype patterns in the Western European honey bee from sequencing a large panel of haploid drones

David Wragg, Sonia Eynard, Benjamin B. Basso, Kamila Canale-Tabet, Emmanuelle Labarthe, Olivier Bouchez, Kaspar Bienefeld, Malgorzata Bieńkowska, Cecilia Costa, Aleš Gregorc, et al.

To cite this version:

David Wragg, Sonia Eynard, Benjamin B. Basso, Kamila Canale-Tabet, Emmanuelle Labarthe, et al.. Complex population structure and haplotype patterns in the Western European honey bee from sequencing a large panel of haploid drones. Molecular Ecology Resources, 2022, 22 (8), pp.3068-3086. 10.1111/1755-0998.13665 . hal-03807491

HAL Id: hal-03807491

https://hal.inrae.fr/hal-03807491v1

Submitted on 4 Apr 2024

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

Check for updates

Received: 24 November 2021 | Revised: 26 May 2022 | Accepted: 1 June 2022

DOI: 10.1111/1755-0998.13665

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Complex population structure and haplotype patterns in the Western European honey bee from sequencing a large panel of haploid drones

David Wragg[1] | Sonia E. Eynard[1] | Benjamin Basso[2] | Kamila Canale-Tabet[1] | Emmanuelle Labarthe[1] | Olivier Bouchez[3] | Kaspar Bienefeld[4] | Małgorzata Bieńkowska[5] | Cecilia Costa[6] | Aleš Gregorc[7] | Per Kryger[8] | Melanie Parejo[9] | M. Alice Pinto[10] | Jean-Pierre Bidanel[11] | Bertrand Servin[1] | Yves Le Conte[12] | Alain Vignal[1]

[1]GenPhySE, Université de Toulouse, INRAE, INPT, INP-ENVT, Castanet Tolosan, France
[2]Institut de l'abeille (ITSAP), UMT PrADE, Avignon, France
[3]GeT-PlaGe, Genotoul, INRAE, Castanet Tolosan, France
[4]Bee Research Institute, Hohen Neuendorf, Germany
[5]Apiculture Division, National Research Institute of Horticulture, Puławy, Poland
[6]CREA Research Centre for Agriculture and Environment, Bologna, Italy
[7]Faculty of Agriculture and Life Sciences, University of Maribor, Pivola, Slovenia
[8]Department of Agroecology, Science and Technology, Aarhus University, Slagelse, Denmark
[9]Agroscope, Swiss Bee Research Centre, Bern, Switzerland
[10]Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Bragança, Portugal
[11]GABI, INRAE, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France
[12]INRAE, UR 406 Abeilles et Environment, UMT PrADE, Avignon, France

Correspondence
Alain Vignal, GenPhySE, Université de Toulouse, INRAE, INPT, INP-ENVT, Castanet Tolosan, France.
Email: alain.vignal@inrae.fr

Present address
David Wragg, Roslin Institute, University of Edinburgh, Midlothian, UK

Benjamin Basso, INRAE, UR 406 Abeilles et Environment, UMT PrADE, Avignon, France

Melanie Parejo, Applied Genomics and Bioinformatics, Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, Leioa, Spain

Funding information
FranceAgriMer, Grant/Award Number: 14-21-AT

Handling Editor: Jeremy B. Yoder

Abstract

Honey bee subspecies originate from specific geographical areas in Africa, Europe and the Middle East, and beekeepers interested in specific phenotypes have imported genetic material to regions outside of the bees' original range for use either in pure lines or controlled crosses. Moreover, imported drones are present in the environment and mate naturally with queens from the local subspecies. The resulting admixture complicates population genetics analyses, and population stratification can be a major problem for association studies. To better understand Western European honey bee populations, we produced a whole genome sequence and single nucleotide polymorphism (SNP) genotype data set from 870 haploid drones and demonstrate its utility for the identification of nine genetic backgrounds and various degrees of admixture in a subset of 629 samples. Five backgrounds identified correspond to subspecies, two to isolated populations on islands and two to managed populations. We also highlight several large haplotype blocks, some of which coincide with the position of centromeres. The largest is 3.6 Mb long and represents 21% of chromosome 11,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with two major haplotypes corresponding to the two dominant genetic backgrounds identified. This large naturally phased data set is available as a single vcf file that can now serve as a reference for subsequent populations genomics studies in the honey bee, such as (i) selecting individuals of verified homogeneous genetic backgrounds as references, (ii) imputing genotypes from a lower-density data set generated by an SNP-chip or by low-pass sequencing, or (iii) selecting SNPs compatible with the requirements of genotyping chips.

KEYWORDS
genome, haplotype, honey bee, population genetics, SNP

1 | INTRODUCTION

The honey bee *Apis mellifera* comprises more than 30 subspecies, each of which is defined according to morphological, behavioural, physiological and ecological characteristics suited to their local habitat (Chen et al., 2016; Ilyasov et al., 2020; Meixner et al., 2013; Ruttner, 1988). European subspecies broadly group into two evolutionary lineages representing on one side western and northern Europe (M lineage), and on the other eastern and southern Europe (C lineage) (Ruttner, 1988). The two European M lineage subspecies are the Dark European or "black" honey bee *A. m. mellifera* and the Iberian honey bee *A. m. iberiensis*, while the C lineage subspecies include, amongst others, the Italian honey bee *A. m. ligustica* and the Carniolan honey bee *A. m. carnica* (Ilyasov et al., 2020). Prior to the involvement of apiarists, the Alps are thought to have presented a natural barrier between *A. m. mellifera* to the north, *A. m. carnica* to the southeast and *A. m. ligustica* to the southwest (Rinderer, 2013).

Before the turn of the 19th century, French honey bee populations were solely represented by the native *A. m. mellifera*, for which regional ecotypes have previously been described (Cornuet et al., 1978; Cornuet et al., 1982). However, during the 20th century, much interest arose amongst apiarists in developing hybrids between the native *A. m. mellifera* and other subspecies including *A. m. ligustica*, *A. m. carnica* and the Caucasian *A. m. caucasia* from Georgia (Cornuet et al., 1979; Fresnaye et al., 1974; Ruttner, 1988). Apiarists found the hybrids performed better than the native *A. m. mellifera* with regard to the production of honey and royal jelly, spurring further interest in the imported subspecies, which were also reported to be more docile and easier to manage (Ruttner, 1988). *A. m. ligustica* is a very popular subspecies amongst apiarists because of its adaptability to a wide range of climatic conditions, its ability to store large quantities of honey without swarming and its docile nature if disturbed (Franck et al., 2000). *A. m. ligustica* queens are frequently exported worldwide, and after the first introductions of *A. m. mellifera* and *A. m. iberica*, most of the honey bees imported over recent centuries into the New World were of Italian origin (Carpenter & Harpur, 2021; Franck et al., 2000). Apiculture involving *A. m. carnica* is also very popular amongst apiarists (Puškadija et al., 2021), in particular for further selection throughout central and western Europe (Gregorc et al., 2008; Gregorc & Lokar, 2010) given their calm temperament and higher honey yield compared to *A. m. mellifera* (Ruttner, 1988). *A. m. caucasia* is a subspecies that was also imported to France, to generate *A. m. ligustica* × *A. m. caucasia* hybrids, that were themselves crossed naturally to the *A. m. mellifera* present in the local environment. Another popular hybrid used in apiculture is the so-called Buckfast, created and bred by Brother Adam of Buckfast Abbey in England (Adam, 1986).

Following the extensive imports of queens from "exotic" subspecies, the genetic makeup of honey bee populations in France became complex, and genetic pollution of local populations had clear phenotypic consequences such as changes in the colour of the cuticle (Cornuet et al., 1986). The increasing admixture of divergent honey bee subspecies has fostered conservationists to protect the native genetic diversity of regional ecotypes, such as *A. m. iberiensis* in Spain and Portugal, *A. m. ligustica* and *A. m. siciliana* in Italy (Fontana et al., 2018), and *A. m. mellifera* in France, Scotland and Switzerland amongst other places (De la Rúa et al., 2009; Fontana et al., 2018; Hassett et al., 2018; Parejo et al., 2016; Pinto et al., 2014). As a result of the different breeding practices, the necessity arose for a study targeted towards *A. m. mellifera* conservatories and French bee breeders specialized in rearing and selling queens. In this context, the genomic diversity project "SeqApiPop" emerged. Within this project, samples from French conservatories, from individual French breeders and from breeders' organizations were analysed, including Buckfast samples.

Traditionally, wide diversity studies have been performed using a small number of molecular markers such as microsatellites (Techer et al., 2015) or limited sets of single-nucleotide polymorphisms (SNPs) (Henriques et al., 2018a; Parejo et al., 2018; Whitfield et al., 2006; Zayed & Whitfield, 2008), enabling population stratification, introgression and admixture levels to be characterized. However, to understand complex population admixture events, as has occurred for the managed honey bee populations in France and elsewhere, or to identify signatures of natural (Harpur et al., 2014; Henriques et al., 2018b; Parejo et al., 2020; Zayed & Whitfield, 2008) or artificial (Parejo et al., 2017; Wragg et al., 2016) selection in the genome, a much higher density of markers is required. As no high-density SNP chip was available for the honey bee at the onset of the project, and as the honey bee genome is very small compared to most animal genomes, being only 226.5 Mb long (Wallberg et al., 2019), we used a

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

whole-genome sequencing approach (Harpur et al., 2014; Wallberg et al., 2014). Although the sequencing of honey bee workers has proven successful for detecting signatures of selection or admixture events (Christmas et al., 2018; Dogantzis et al., 2021; Harpur et al., 2014; Wallberg et al., 2014; Wragg et al., 2018), analysing haploid drones allows sequencing at a lower depth and with greater accuracy in variant detection, as demonstrated by studies on a limited number of samples (Henriques, et al., 2018b; Parejo et al., 2016; Wragg et al., 2016). An additional advantage of sequencing haploids is that the alleles are phased, which is invaluable for studies investigating genome dynamics such as recombination hotspots and haplotype structure. Although some insights into recombination patterns in the honey bee have been made through the analysis of drones from individual colonies (Kawakami et al., 2019; Liu et al., 2015) and linkage disequilibrium (LD)-based approaches (Jones et al., 2019; Wallberg et al., 2015), a deep understanding of the recombination landscape, essential for fine-scale genetic analyses, requires hundreds of phased genomes. Such "HapMap" projects have been conducted in humans and cattle, initially using SNP arrays (Bansal et al., 2007; Bovine HapMap Consortium et al., 2009) and more recently by whole-genome sequencing as in the "1000 genome" projects (Chaisson et al., 2015; Sudmant et al., 2015).

Therefore, as the first step towards a deeper understanding of French and Western European managed honey bee populations and of their genome dynamics, and also to provide a large data set of phased genotypes from sequences aligned to the latest Amel_HAv3.1 genome assembly (Wallberg et al., 2019), we undertook the extensive sequencing of haploid drones. These data comprised samples from French conservatories and commercial breeders in addition to samples from several European countries each representing potentially pure *A. m. ligustica*, *A. m. carnica*, *A. m. mellifera* and *A. m. caucasia* populations typically imported by French breeders. Finally, *A. m. iberiensis*, the Iberian subspecies only separated from the native French *A. m. mellifera* by the natural barrier of the Pyrenees was also studied. In total, 870 samples were sequenced for SNP detection and 629 were used for a detailed genetic analysis of present-day honey bee populations in France. The results are publicly available as a phased vcf file for high-quality SNP markers carefully filtered against sequencing and mapping artefacts, which can be used for imputation or for selecting already genotyped reference individuals to be used in future studies.

2 | MATERIALS AND METHODS

2.1 | Sampling and sequencing

For the population genomics analyses, one individual drone per colony was sampled before emergence, from colonies throughout France, Spain, Germany, Switzerland, Italy, the UK, Slovenia, Poland, Denmark and China, and from a French beekeeper having imported queens from Georgia, amounting to a total of 642 samples (Figure S1, Table S1; Table 1). The robustness of the primary SNP detection by GENOTYPEGVCFS (see below) and of the filtering steps on mapping and genotype quality metrics estimated across samples were improved by increasing the size of the data set for these technical steps: a further 30 colony replicate samples, which had been collected from colonies already sampled for this study, in addition to 198 samples of similar genetic backgrounds from two other ongoing projects. Thus, although 642 colonies were included for population genomics analyses, in total 870 samples were used for SNP detection (Table S1).

DNA was extracted from the thorax of adult bees or from pupae as described in Wragg et al. (2016). Briefly, drones were sampled at either the pupae/nymph or larval stage and stored in absolute ethanol at −20°C. DNA was extracted from the thorax or from diced whole larvae. Tissue fragments were first incubated for 3 h at 56°C in 1 ml of a solution containing 4 M urea, 10 mM Tris–HCl pH 8, 300 mM NaCl, 1% SDS, 10 mM EDTA and 0.25 mg proteinase K, after which 0.25 mg proteinase K was added and incubated overnight at 37°C. Four hundred microlitres of a saturated NaCl solution was added to the incubation, which was then gently mixed and centrifuged for 30 min at 15,000 *g*. The supernatant was treated for 5 min at room temperature with RNAse (Qiagen) and then centrifuged again, after which the DNA in the supernatant was precipitated with absolute ethanol and resuspended in 100 µl TE 10/0.1. Pair-end sequencing was performed on Illumina HiSeq 2000, 2500 and 3000 sequencing machines with 20 samples per lane, or on a NovaSeq machine with 96 samples per lane, following the manufacturers' protocols for library reparations.

2.2 | Mapping and genotype calling

Sequencing reads were mapped to the reference genome Amel_HAv3.1 (Wallberg et al., 2019) using BWA-MEM (version 0.7.15) (Li, 2013), and duplicates marked with PICARD (version 2.18.2; http://broadinstitute.github.io/picard/). Libraries that were sequenced in multiple runs were merged with SAMTOOLS (version 1.8) (Li et al., 2009) prior to marking duplicates. Local realignment and base quality score recalibration (BQSR) were performed using GATK (version 4.1.2.) (McKenna et al., 2010), using SNPs called with GATK HAPLOTYPECALLER as covariates for BQSR. Each drone was independently processed with the pipeline and genotyped independently with HAPLOTYPECALLER. Although the drones sequenced are haploid, variant calling was performed using a diploid model to allow the detection and removal of SNPs for which heterozygous genotypes are called in >1% of samples, and that might have arisen for example as a result of short-tandem repeats (STRs) or could highlight copy number variants (CNVs) in the genome. Individual gVCF files were combined with COMBINEGVCFS and then jointly genotyped with GENOTYPEGVCFS, resulting in a single VCF file for the 870 samples containing 14,990,574 raw variants. After removing Indels with GATK SELECTVARIANTS, 10,601,454 SNPs remained. Sequencing depth was estimated using MOSDEPTH (Pedersen & Quinlan, 2018). Further details are given in https://github.com/avignal5/SeqApiPop/blob/v1.5/SeqApiPop_1_MappingCalling.md.

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Samples used for the diversity study

Genetic type	Geographical origin[a]	Samples	Status
A. m. carnica	France	13	Reference (breeders)
A. m. carnica	Germany	18	Reference (breeders)
A. m. carnica	Poland	19	Reference (breeders)
A. m. carnica	Slovenia	20	Reference (breeders)
A. m. carnica	Switzerland	31	Reference (breeders)
A. m. caucasia	France	15	Reference (breeder)
A. m. iberiensis	Spain	30	Reference (beekeepers)
A. m. ligustica	Italy	30	Reference (breeders)
A. m. mellifera	Ariège, France	8	Reference (conservatory)
A. m. mellifera	Brittany, France	4	Reference (conservatory)
A. m. mellifera	Colonsay, UK	28	Reference (conservatory)
A. m. mellifera	Ouessant, France	41	Reference (conservatory)
A. m. mellifera	Porquerolles, France	15	Reference (conservatory)
A. m. mellifera	Savoy, France	31	Reference (conservatory)
A. m. mellifera	Solliès, France	14	Reference (conservatory)
Buckfast	Haut-Rhin, France	6	Reference (breeder)
Buckfast	Switzerland	17	Reference (breeders)
Royal Jelly[b]	Breeder organization	65	Reference
Unknown	Ariège, France	12	Queen breeder
Unknown	Brittany, France	3	Breeder
Unknown	Corsica, France	44	Breeder organization
Unknown	Hautes Pyrénées, France	19	Queen breeder
Unknown	Hérault, France	7	Breeder
Unknown	Isère1, France	6	Breeder
Unknown	Isère2, France	11	Breeder
Unknown	Sarthe, France	17	Unselected apiary
Unknown	Tarn1, France	44	Queen breeder
Unknown	Tarn2, France	31	Queen breeder
Unknown	Vaucluse, France	20	Breeder
Unknown	China	10	Breeder
Unknown	Unknown	13	Breeder

Note: Four hundred and five samples were used as references for the genetic types commonly used in breeding in Western Europe. These include 317 samples representing five subspecies (*A. m. carnica* [$n = 101$], *A. m. ligustica* [$n = 30$], *A. m. caucasia* [$n = 15$], *A. m. iberiensis* [$n = 30$], *A. m. mellifera* [$n = 141$]), 23 samples from the Buckfast strain and 65 samples selected for royal jelly production. The remaining 237 samples were from French breeders or breeders' organizations, except 10 samples from China. All samples were collected in 2014 and 2015.

[a]Regional geographical origins in France are indicated by their administrative "département" (see Figure S1 for locations on the map and further information on the populations).

[b]The royal jelly samples come from several French beekeepers in Moselle, Alpes maritimes, Bouches-du-Rhône and Isère exchanging genetic material within the GPGR breeders' organization.

2.3 | Quality filters on SNPs

The first run of filters concerns technical issues related to the sequencing and alignment steps and was therefore used for the total data set of 870 samples, to benefit from its larger size for SNP detection and validation (Figure S2). These filters included (i) strand biases and mapping quality metrics (SOR ≥3; FS ≤60 and MQ ≥40), (ii) genotyping quality metrics (QUAL >200 and QD <20), and (iii) individual SNP genotyping metrics (heterozygote calls <1%; missing genotypes <5%, allele number <4 and less than 20% of the samples with genotypes having individual GQ <10). Distribution and ECDF plots of values for all the filters used on the data set were used to select thresholds and are shown in https://github.com/avignal5/SeqApiPop/blob/v1.5/SeqApiPop_2_VcfCleanup.md.

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.4 | Haplotype block detection, LD pruning, PCA, ADMIXTURE, TREEMIX, RFMIX

Haplotype blocks were detected with PLINK (version 1.9) (Chang et al., 2015) using the blocks function, "--blocks no-pheno-req no-small-max-span," with the parameter "--blocks-max-kb 5000." LD pruning was performed with PLINK using the indep-pairwise function, with a window of 1749 SNPs, corresponding to a mean chromosome coverage of 100 kb (see Section 3), 10% overlap between windows and an LD value of 0.3. Principal component analyses (PCAs) were performed with PLINK and the contributions of individual SNPs to the principal components were estimated using smartpca from the EIGENSOFT package version 7.2.1 (Patterson et al., 2006). The significance of the contribution of the SNPs to PC1 was evaluated with the R package PCADAPT version 4.3.3 (Privé et al., 2020) and qvalues estimated with the R package QVALUES version 2.26.0 (Storey et al., 2022). Further details are given in https://github.com/avignal5/SeqApiPop/blob/v1.5/SeqApiPop_3_LDfilterAndPCAs.md. Admixture analysis was performed with the program ADMIXTURE version 1.3.0 (Alexander & Lange, 2011), with values of *K* ranging from 2 to 16. Fifty runs were performed each time using a unique random seed. The PONG software (Behr et al., 2016) was used for aligning runs with different *K* values and for grouping results from runs into clustering modes, setting the similarity threshold to 0.98. Further details are given in https://github.com/avignal5/SeqApiPop/blob/v1.5/SeqApiPop_4_Admixture.md. Analyses of population migration was performed with TREEMIX (Pickrell & Pritchard, 2012), with the option for grouping SNPs set to $-k = 500$, testing between 0 and 9 migrations and performing 100 runs per migration with a unique random seed. The optimum number of migrations was estimated with the R package OPTM (Fitak, R. R.: https://CRAN.R-project.org/package=OptM) using the Evanno method provided (Evanno et al., 2005). Tree summaries for the 100 runs per migration tested were performed with DENDROPY (Sukumaran & Holder, 2010) and drawn with FIGTREE version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Further details are given in https://github.com/avignal5/SeqApiPop/blob/v1.5/SeqApiPop_5_TreeMix.md. Local ancestry inference and positioning of haplotype switches were performed with RFMIX version 2.03-r0 (Maples et al., 2013). Three main genetic backgrounds were considered for this analysis, corresponding to the three major groups highlighted in the PCA. Reference samples were selected as having >95% pure background. Although most diploid data were removed and the data were already phased, SHAPEIT Version 2.904 (Delaneau et al., 2012) was run to format the vcf files for RFMIX. RFMIX was run using genetic maps generated from the data of Liu et al. (2015). Briefly, reads from the project SRP043350 were retrieved from the Short Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra), aligned to the reference genome for SNP detection and recombinants were detected with the custom script find_crossing_overs.py to produce a genetic map. Further details on the RFMIX analysis are given in https://github.com/avignal5/SeqApiPop/blob/v1.5/SeqApiPop_6_RFMix.md.

3 | RESULTS

3.1 | Sequencing and genotyping

Sequencing of the honey bee drones for the SeqApiPop diversity project began in 2014 on Illumina HiSeq instruments and some of the first samples had such low coverage that a second run (or even three in the case of OUE8) of sequencing was performed. For these samples, the resulting BAM files were merged prior to variant calling. Only four samples of the diversity project were sequenced on Novaseq instruments, for which higher sequencing depths were achieved. Therefore, to improve the robustness of the SNP detection pipeline, we included drone genome sequences from other ongoing projects using the same genetic types, which had the advantage of all being produced with a Novaseq instruments and at a higher depth. Samples sequenced with the HiSeq and NovaSeq instruments had mean sequencing depths of 12.5 ± 6.1 and 33.5 ± 10.2 respectively (Table S1, Figures S3 and S4).

Genotyping the whole data set of 870 drones with the GATK pipeline allowed the detection of 10,601,454 raw SNPs (Figure S2). Results of the subsequent filtering steps are shown on the Venn diagrams in Figures S5–S7. A total of 7,023,976 high-quality SNPs remained after filtering. The 198 samples from the other projects and 30 within-colony replicate samples from the present diversity project were removed from the data set for downstream analyses. Although a filter on genotyping rate ≥95% was applied in the primary filtering steps, the final filter on heterozygote calls was set to keep SNPs with up to 1% of heterozygote samples, and these remaining heterozygous genotypes were set to missing (Figure S2). After this, a final filter on missing data in samples was applied and 15 samples were removed due to the fraction of missing genotypes exceeding 10%. The final diversity data set comprised 629 drones (Table S1) and 7,012,891 SNPs, and was used for all subsequent analyses unless stated otherwise.

3.2 | Contribution of SNPs to the variance in PCAs: detection of large haplotype blocks

Principal component analyses, performed on the 629 samples and 7 million SNPs, resulted in a clear differentiation of three groups of samples. The first principal component, representing 10.8% of the total variance, broadly differentiates M lineage bees, *A. m. mellifera* and *A. m. iberiensis*, from the *A. m. ligustica*, *A. m. carnica* and *A. m. caucasia* bees. The second principal component, representing 3.1% of the variance, separates the O lineage *A. m. caucasia* bees from the C lineage *A. m. ligustica* and *A. m. carnica* bees (Figure S8). PC3 represents 1.2% of the variance and the remaining 626 principal components each represent 0.7% or less. When looking at the individual contributions of SNPs to the variance, we can see that only a very small proportion of the ~7 million markers contribute significantly to PC1 (red lines on Figure S9) and that this proportion is even much smaller for PCs 2 and 3. Two reasons for such a limited

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

contribution to the variance of the majority of markers is the low informativity of markers of low minor allele frequency (MAF) and the redundancy of markers that are in strong LD. Therefore, to thin the data set, we tested the effect of several MAF filters and chose the most pertinent one for subsequent testing of various LD pruning values. The effects of these filters were estimated by inspecting the contributions of the SNPs to the principal components. The MAF filters tested showed clearly that data sets containing only SNPs with MAF >0.01 or MAF >0.05 are sufficient to allow a higher proportion of markers contributing to the PCs, with a notable increase of SNPs contributing to PC2 and PC3 (Figure S9). To avoid losing too many potential population-specific markers present at low frequency in the data, we chose to use the lowest MAF threshold tested, leaving a data set of 3,285,296 SNPs having MAF >0.01 for subsequent analyses. On inspecting the contributions of individual SNPs to principal components along the genome, a striking feature we observe is that for several large chromosomal regions, five of which are larger than 1 Mb, a high proportion of SNPs have a significant contribution to PC1 (Figure 1a,b; and Figures S10 and S11). Such observations suggest the existence of large haplotype blocks driving differentiation along principal components, in particular PC1. To explore this further we compared these genomic regions to the haplotype blocks detected with PLINK (Table S2) revealing significant overlap by visual inspection (Figure S12). The largest of these blocks spans 3.6 Mb, representing 21% of chromosome 11 and close to 1.6% of the honey bee genome size (Figure 1c). Four other blocks on chromosomes 4, 7 and 9 are larger than 1 Mb (Figures S10–S12, Table S2).

3.3 | LD filtering

Population structure and admixture analyses rely largely on the assumption that markers along the genome are independent. Indeed, markers in strong LD such as those in haplotype blocks can influence genetic structure. Therefore, we sought to investigate the impact of LD pruning on inferences of population structure. The number of SNPs used in a window for LD pruning was determined such that most windows would correspond to a physical size of 100 kb. To achieve this, we used the mode of the distribution of the number of SNPs in 100-kb bins, which is 1749 for the data set of 3,285,296 SNPs with MAF >0.01 (Figures S13 and S14). LD pruning was thus performed with a window size of 1749 SNPs and 175-bp (10%) overlap and various values were tested, spanning between LD $.1 < r^2 \leq .9$. These various thresholds show that with LD $r^2 \leq .3$ the global structure of the data set is altered, with the *A. m. iberiensis* population as the major contributor to PC2 and *A. m. caucasia* being a separate population only in PC3 (Figures S15A,B), whereas with LD $r^2 > .3$, the contributions to the variance in PC1 is not as widely distributed amongst subspecies (Figure S16). The effect of LD pruning on the haplotype blocks is drastic, with the few SNPs retained having a distribution of their contributions to the variance in PC1 and PC2 similar to that of the rest of the genome (Figure S17). After pruning for LD $r^2 < .3$, 601,945 SNPs were left in the data set, which were subsequently used in the analysis of population structure.

3.4 | Analysis of population structure

The PCA revealed a distinct population structure within the data. For instance, some populations from French breeding organizations, such as the royal jelly breeders' organization (GPGR: Groupement des Producteurs de Gelée Royale), and the Corsican breeder's organization (AOP Corse), appear quite homogenous (Figure 2), with GPGR samples clustering close to the *A. m. ligustica* and *A. m. carnica* reference populations, while AOP Corse samples appear as a distinct group between the C lineage *A. m. ligustica* and *A. m. carnica* on one side and the M lineage *A. m. mellifera* and *A. m. iberiensis* on the other side. Other populations from French breeders appear much less homogeneous, with individuals scattered across the whole graph (e.g. Tarn 2 on Figure 2) suggesting various degrees of admixture between the three principal genetic groups (Figure S18). To further investigate the genetic structure and the effects of human-mediated breeding, we performed admixture analyses. Our data set consists of reference samples from 13 origins, including two islands, in addition to samples from several commercial breeders and conservatories. The genetic makeup is therefore expected to be complex and the first task was to estimate the optimal number of genetic backgrounds (K). We performed 50 independent runs with the ADMIXTURE software for each value of $2 \leq K \leq 16$ on the LD-pruned data set, totalling 750 independent analyses. Cross-validation (CV) error estimates of the results computed by the software are shown in Figure 3a. Results suggest that the most likely number of genetic backgrounds is 8 or 9, with $K = 8$ having runs with the lowest CV values overall, and $K = 9$ having the lowest median CV value over its 50 runs. The resulting Q matrices were jointly analysed using PONG (Behr et al., 2016), where runs of each value of K are grouped together by similarity into modes and the mode containing the largest number of similar runs is defined as the major mode. As PONG failed to find disjoint modes with the default similarity threshold of 0.97, we increased the stringency of this value to 0.98. Naturally, for low values of K, such as 2 or 3, most of the Q matrices are very similar and the major modes contain most runs, if not all. Typically, for $K = 2$, all 50 runs are in a single mode and for $K = 3$, the major mode contains 49 out of all 50 runs and reflects the three main groups from the PCA. Amongst the values of K having the lowest CV values (Figure 3a), $K = 9$ stands out as having a major mode containing 33 runs out of 50. While $K = 8$ had the lowest overall mean CV value, its major mode contained only 12 runs, indicating $K = 9$ to be a better model (Figure 3b; Table S3). As expected, the pattern observed when considering only $K = 3$ genetic backgrounds recapitulates the general pattern observed in the PCAs, in which the reference populations separate into three groups.

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

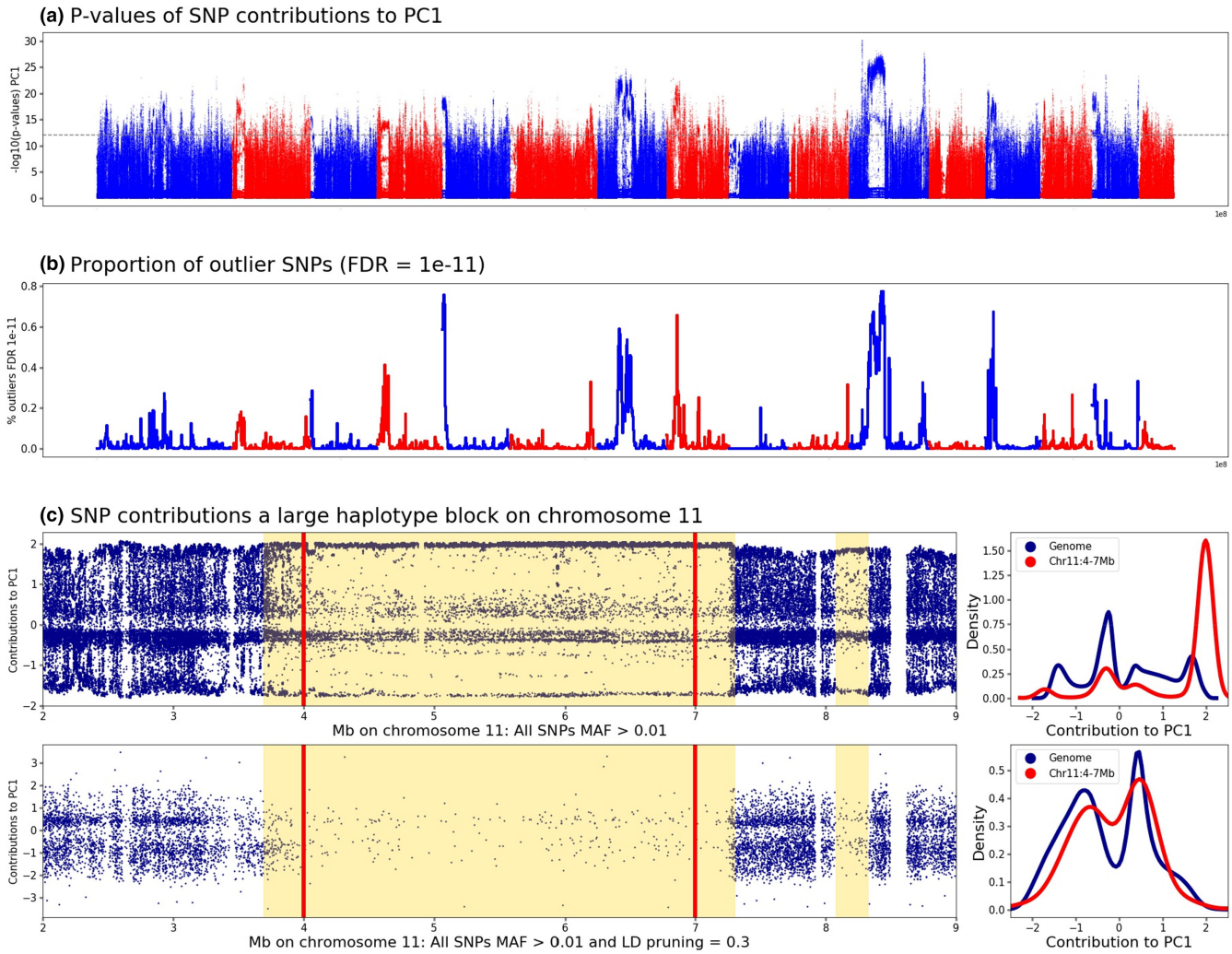


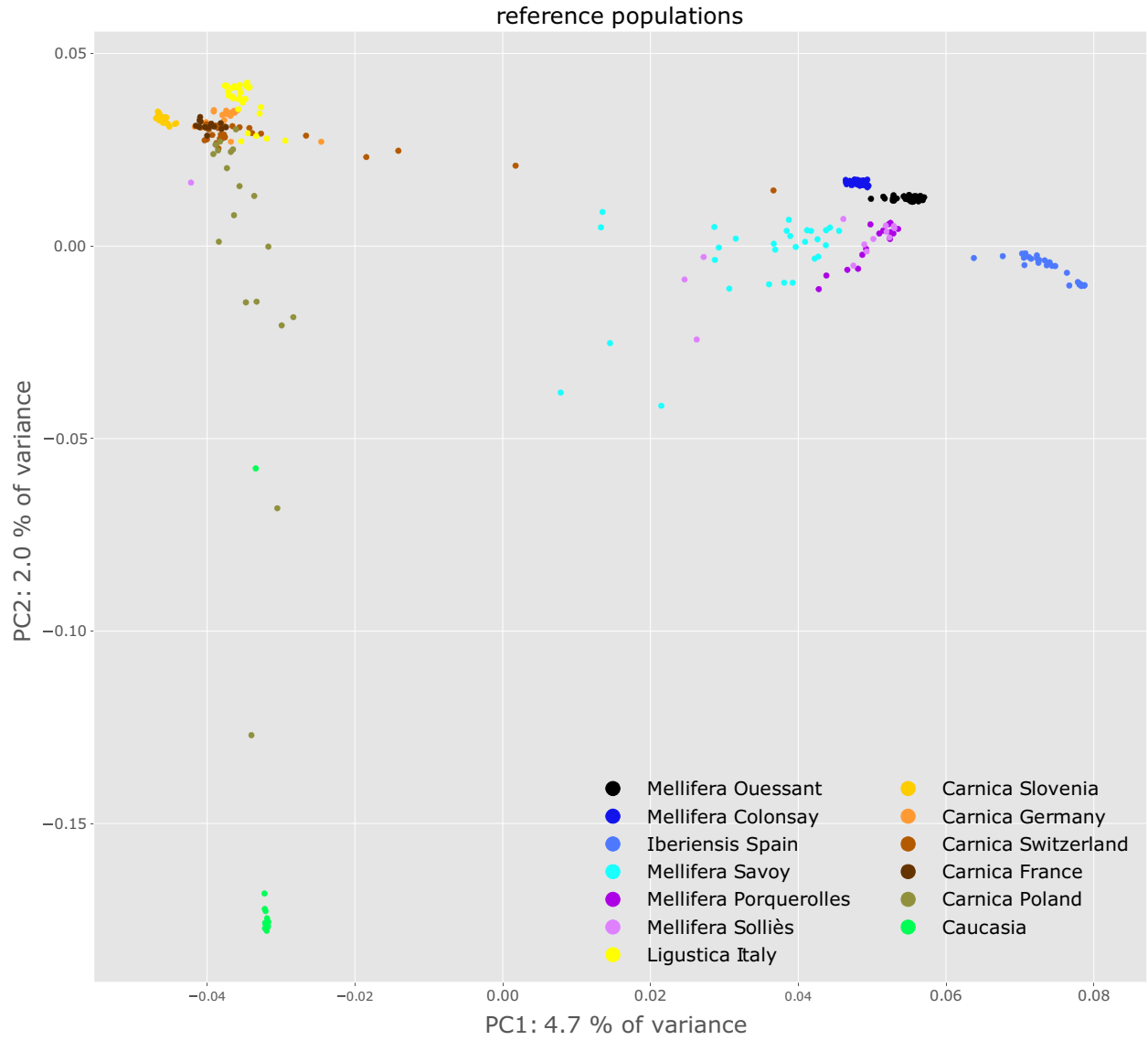
FIGURE 1 Contribution of SNPs to principal component 1: genome-wide and in a large haplotype block on chromosome11. (a) Component-wise *p*-values are plotted for the correlations between PC1 and each of the 3,285,296 SNPs with MAF >1. The dashed line represents the -$\log_{10}$(*p*-value) corresponding to an FDR of 10^{-11}. (b) Proportion of outlier SNPs, as determined by the FDR = 10^{-11} threshold. (c) Blue points show the contribution of individual SNPs to PC1 along a 6-Mb region of chromosome 11 containing two haplotype blocks of >3 Mb and ~200 kb (yellow backgrounds) before (top) and after (bottom) LD pruning at LD = 0.3. The LD pruning successfully eliminates the markers in the haplotype blocks and the distribution of the marker contributions approaches that of the rest of the genome, as shown in the corresponding density plots on the right

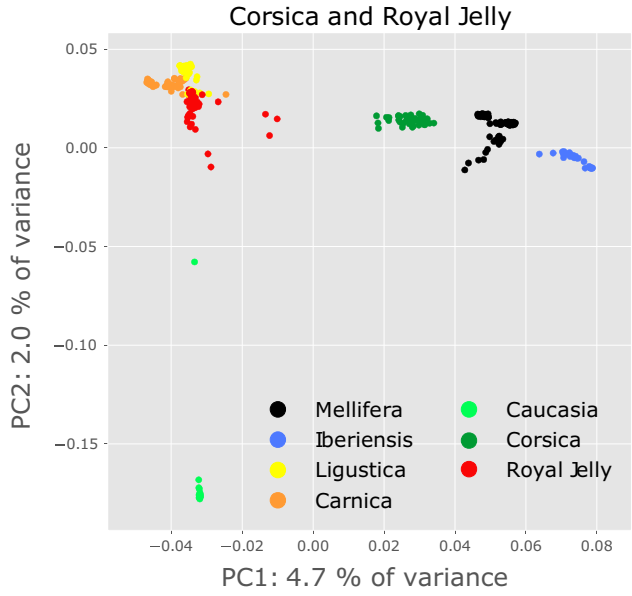
These groups reflect the main evolutionary lineages present in the data set, being the M lineage (*A. m. mellifera* and *A. m. iberiensis*), C lineage (*A. m. ligustica*, *A. m. carnica*) and O lineage (*A. m. caucasia*). For *K* = 2, *A. m. caucasia* bees are considered as having the same genetic background as the *A. m. ligustica* and *A. m. carnica* bees, also reflecting the results from the PCA (Figure 2; Figure S19). Some admixture can be observed for a small proportion of the reference samples. For instance, the reference samples from the Savoy conservatory appear to have a small proportion of genetic background from *A. m. ligustica* and/or *A. m. carnica*, which is consistent with the PCA results (Figure 2). Likewise, the *A. m. carnica* samples from Poland have a small proportion of genetic background from *A. m. caucasia*. Finally, the *A. m. carnica* samples from Switzerland show some proportion of *A. m. mellifera* genetic background.

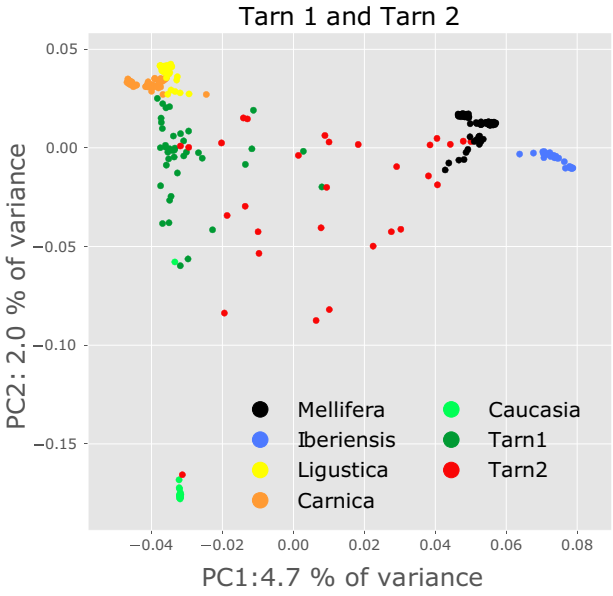
When examining the admixture pattern representing the 33 runs at *K* = 9 genetic backgrounds, the three main groups are now further subdivided. The M lineage group from the *K* = 3 backgrounds is now

FIGURE 2 PCA on the reference populations and on a sample of representative breeder populations. The 601,945 SNPs obtained after MAF filtering and LD pruning were used. Left: reference populations only, with a colouring scheme according to their origin. Middle and left: only the reference populations with a high proportion of pure background individuals, as observed after ADMIXTURE analysis, were kept and coloured according to the five subspecies. Some breeders' populations appear homogeneous, such as the honey bees selected for Royal Jelly or those from Corsica. Others are heterogeneous, such as populations Tarn1 and Tarn2, from breeders

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License







17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

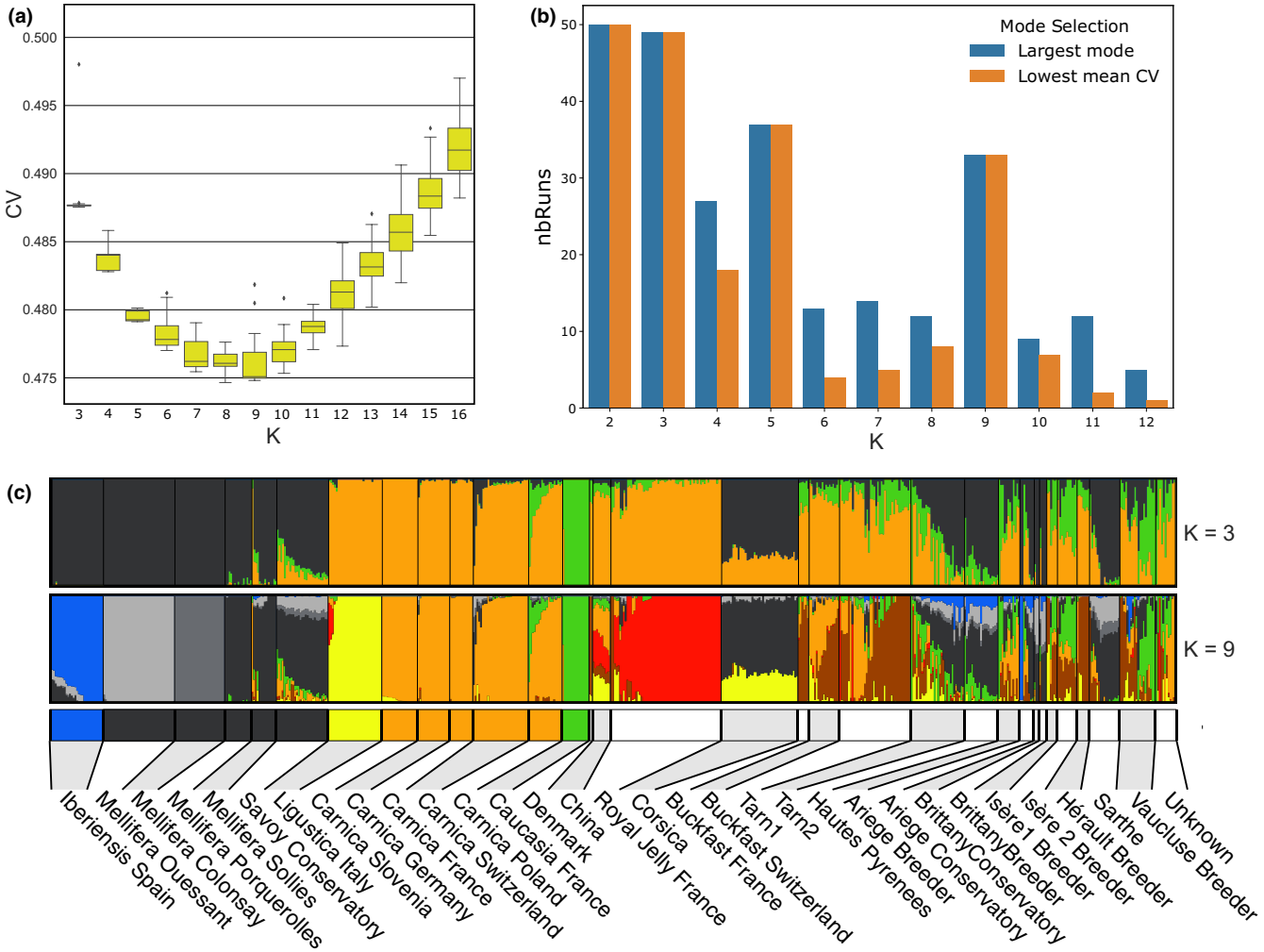


FIGURE 3 Admixture analysis. (a) Estimation of cross-validation (CV) error for 50 runs of ADMIXTURE for $3 \leq K \leq 16$. (b) Major modes and modes with the lowest mean CV error for ADMIXTURE runs. For each value of *K* ranging between 2 and 12, Q matrices from ADMIXTURE runs were grouped by similarity in modes by using the PONG software (Behr et al., 2016). Blue: number of runs in the major mode; orange: number of runs in the major or minor mode having the lowest mean CV value. Amongst the values of *K* having the lowest CV values from ADMIXTURE runs, $K = 9$ stands out as having a major mode containing 33 runs out of 50 (Figure S19), which is also the mode having the lowest mean CV value from the ADMIXTURE runs. For other values of *K*, such as 4, 6, 7 and 8, the major modes do not have the lowest mean CV values. (c) Admixture plots for all 629 samples for $K = 3$ (major mode containing 49 out of 50 runs) and $K = 9$ (major mode containing 33 out of 50 runs). Reference populations on the left have a colour code under the admixture plot that recapitulates their colour on the PCA plots of Figure 2; other populations are indicated with alternating grey and white colours

composed of four genetic backgrounds: *A. m. iberiensis* is separated from *A. m. mellifera*, and the *A. m. mellifera* bees are separated into three groups from mainland France, and the two islands of Ouessant and Colonsay. The other three subspecies *A. m. ligustica*, *A. m. carnica* and *A. m. caucasia* each have their own genetic background. An eighth background corresponds to the samples from the bees selected for the production of royal jelly and a ninth appears in the two populations that were noted as Buckfast bees. Although it is a major background in these two populations, a majority of samples have also a large proportion of *A. m. carnica* and, to a lesser extent, of *A. m. ligustica* backgrounds. This ninth background can also be found in other breeders' populations, principally in Hérault and Tarn1 (Figure 3c). Apart from the royal jelly population, all honey bees from breeders show high levels of admixture. Moreover, there is great variability in the genetic origins and proportions of backgrounds, even for samples coming from the same location (Figure 4). The exception is the population from Corsica, for which all samples show proportions close to 75% of *A. m. mellifera* and 25% of *A. m. ligustica* backgrounds.

3.5 | Migrations between populations

Due to the commercial interest expressed by beekeepers for the Buckfast bees and the peculiar genetic composition observed in the Corsican population, we performed a population migration analysis with TREEMIX (Pickrell & Pritchard, 2012). All samples having more than 80% ancestry from one of the nine backgrounds detected in

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the ADMIXTURE analysis were selected from one of the $K = 9$ major mode Q-matrices (Table S4), and the list supplemented with the 43 Corsican samples, making our data set composed of 10 representative groups for the European populations.

Estimations on the number of migrations (m) between the populations in the data set, based on the Evanno method (Evanno et al., 2005), return a mode of $m = 1$, strongly suggesting a single migration, and a relatively high Δm value for $m = 2$ supports the existence of a second migration. The Δm values for three or more migrations are close to zero, suggesting that more than two migrations between populations in the data set are unlikely (Figure 5a). For $m = 1$ the 100 TREEMIX runs indicated a migration from *A. m. ligustica* to the Corsican population. For $m = 2$ the 100 TREEMIX runs show the two migrations as being from *A. m. ligustica* to the Corsica population, and from *A. m. caucasia* to the Buckfast bees (Figure 5b). Summaries of the resulting trees with DENDROPY (Sukumaran & Holder, 2010) are shown in Figure 5c, indicating that when the two migrations are taken into account, the Corsican samples are grouped with the *A. m. mellifera* M lineage bees, and the Buckfast bees group with the *A. m. ligustica* and *A. m. carnica* C lineage bees.

3.6 | Haplotype conservation in the admixed populations

To investigate further the haplotype blocks detected, we performed a local ancestry inference on our data set with RFMIX. Reference samples were selected as bees having >95% ancestry for a given background following the ADMIXTURE analysis at $K = 3$ (Figure 3c), resulting in 131 samples for group 1, 148 for group 2 and 17 for group 3, while the remaining 333 samples formed the query data set. To perform the local ancestry inference, we constructed a genetic map from crossovers identified in the sequence data of 43 males from three colonies (Liu et al., 2015) aligned to the Amel_HAv3.1 reference genome. The results indicate that few historical recombination events have occurred in the large haplotype blocks since admixture between the subspecies. The most notable example is that of the 3.6-Mb haplotype block between positions 3.7 and 7.3 Mb on chromosome 11, in which almost all 333 samples from the query data set show one continuous stretch for one of the three backgrounds. Only one of the 43 samples from Corsica presents two different ancestral haplotypes within this interval, with a switch from a group 1 to a group 2 haplotype at position ~4.5 Mb on chromosome 11, within the 3.6-Mb haplotype block, whereas numerous switches can be observed on the rest of the chromosome (Figure 6a). When counting the haplotype switches detected in all 333 query samples, only 28 are located within the 3.6-Mb haplotype block on chromosome 11, whereas other regions of the chromosome can have more than 50 switches per 100 kb (Figure 6b; and see Figure S20 for the other chromosomes). Interestingly, *LOC724287*, which is the largest gene described in the Gnomon annotation set for the Amel_HAv3.1 genome assembly, is found in this block at position 11:5,292,072–6,161,805. This gene is 869,734 bp long and encodes protein rhomboid transcript variant X2, its large size being due to intron 4, which

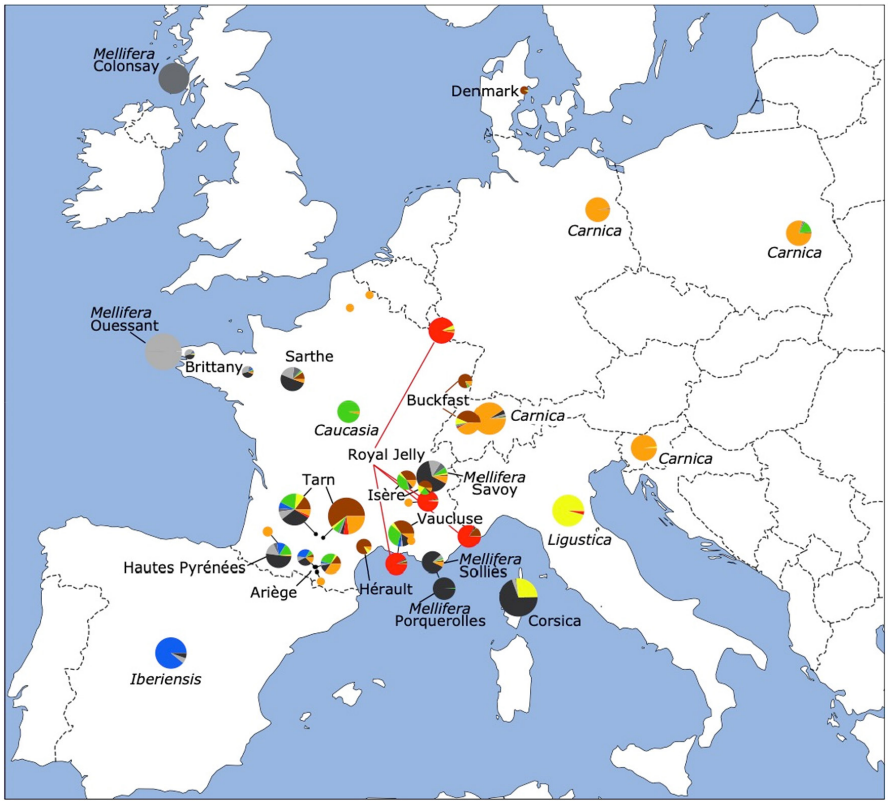


FIGURE 4 Admixture proportions and location of sample populations used in the diversity study. The size of the pie charts indicates the number of samples from a given location, with the number ranging from two samples (e.g., Denmark) to 43 samples (Corsica). Positions in France indicate the coordinates of the breeder or honey bee conservatory sampled. In other countries, reference samples are all grouped together, unless two genetic types were sampled (e.g., Switzerland). Colours in the pie charts correspond to the backgrounds found in the admixture analysis for $K = 9$, as presented in Figure 3. Reference populations for the five subspecies are indicated in italics. Two Buckfast populations in France and Switzerland are indicated, as the four breeders from the Royal Jelly breeders' organization (GPGR: Groupement des Producteurs de Gelée Royale) having provided samples

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

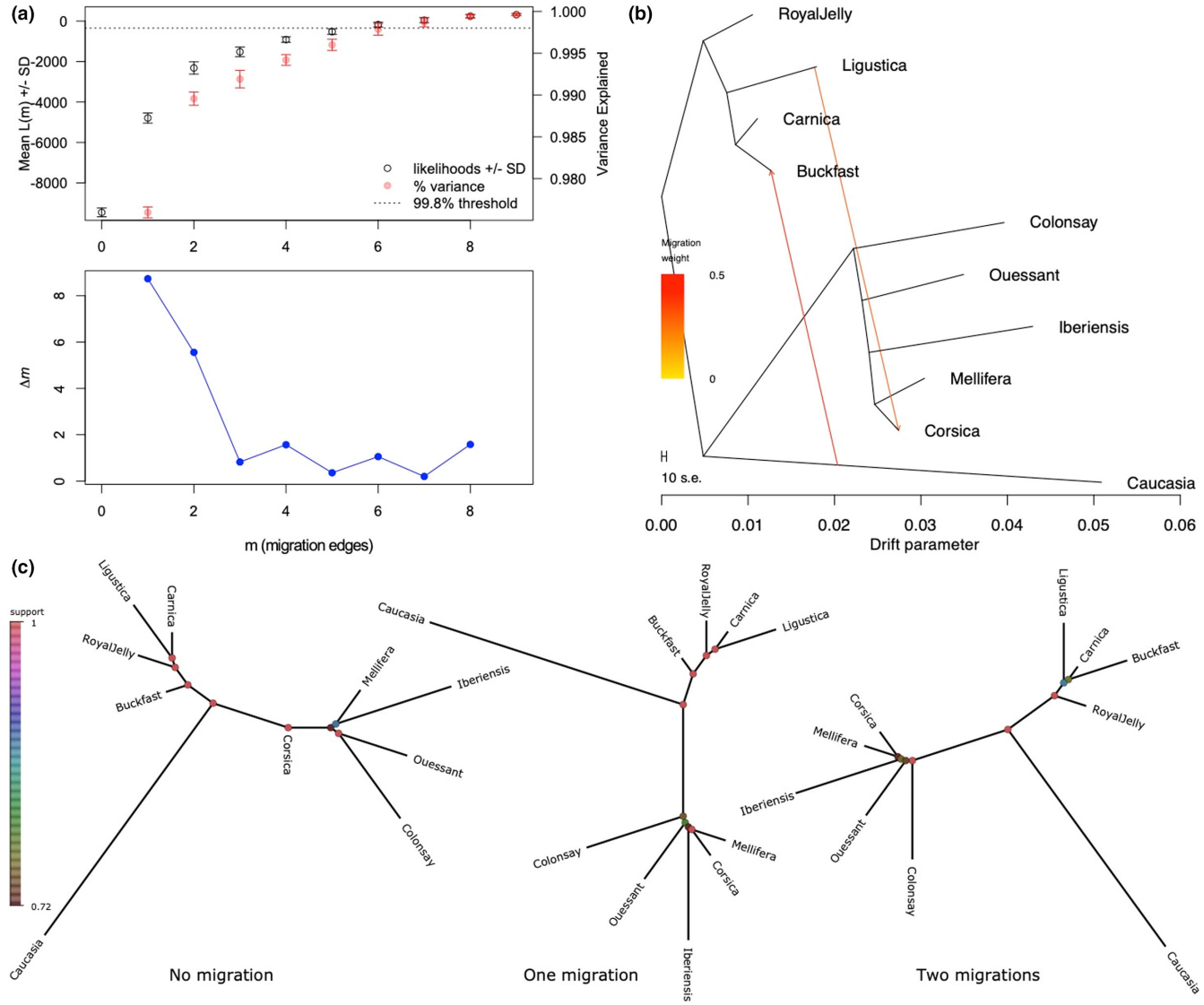


FIGURE 5 Analysis of migrations with TREEMIX. (a) The OptM package was used to determine the optimal number of migrations between populations and backgrounds. The Δ*m* values suggest one or two migrations. (b) TREEMIX graph selected amongst the 100 runs showing the two migrations identified. (c) Summaries of 100 trees from TREEMIX, estimated from the 100 runs per migration with DENDROPY. The topologies correspond to the hypothesis of no migration (left), the Corsican population as being of the *Apis mellifera mellifera* subspecies, with a migration from an *A. m. ligustica* population (middle) or with an additional migration of *A. m. caucasia* in the Buckfast population (right). Support for the nodes is indicated by the colour, with values indicated by the scale on the left. Results suggest the original Corsican population is very similar to *A. m. mellifera* from mainland France

is 596,047 bp long. However, on investigating a possible relationship between haplotype blocks and gene sizes in the honey bee genome no obvious association could be found (data not shown).

4 | DISCUSSION

4.1 | SNP detection in a haploid data set

Our complete data set of haploid drones is composed of 870 samples sequenced using Illumina's HiSeq and NovaSeq technologies. The results clearly show that although a few of the early sequences produced on the HiSeq are of lower depth, only 15 samples were eliminated due to the fraction of missing genotypes exceeding 10%. By contrast, the fraction of missing genotypes over the ~7 million SNPs detected was considerably lower in samples sequenced on the NovaSeq sequencing platform. Having sequenced haploids, the removal of heterozygote SNPs in individual samples is recommended to reduce the likelihood of "pseudo SNPs," as we have shown previously that heterozygote SNPs tend to cluster together (Wragg et al., 2016) and colocate with repetitive elements (data not shown). This set of ~7 million markers can now be used as a basis for the realization of high-density SNP chips, allowing selections of markers according to optimized spacing and to defined MAF values in the main subspecies of interest. Indeed, an important technical issue in SNP chip design is that very high SNP

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

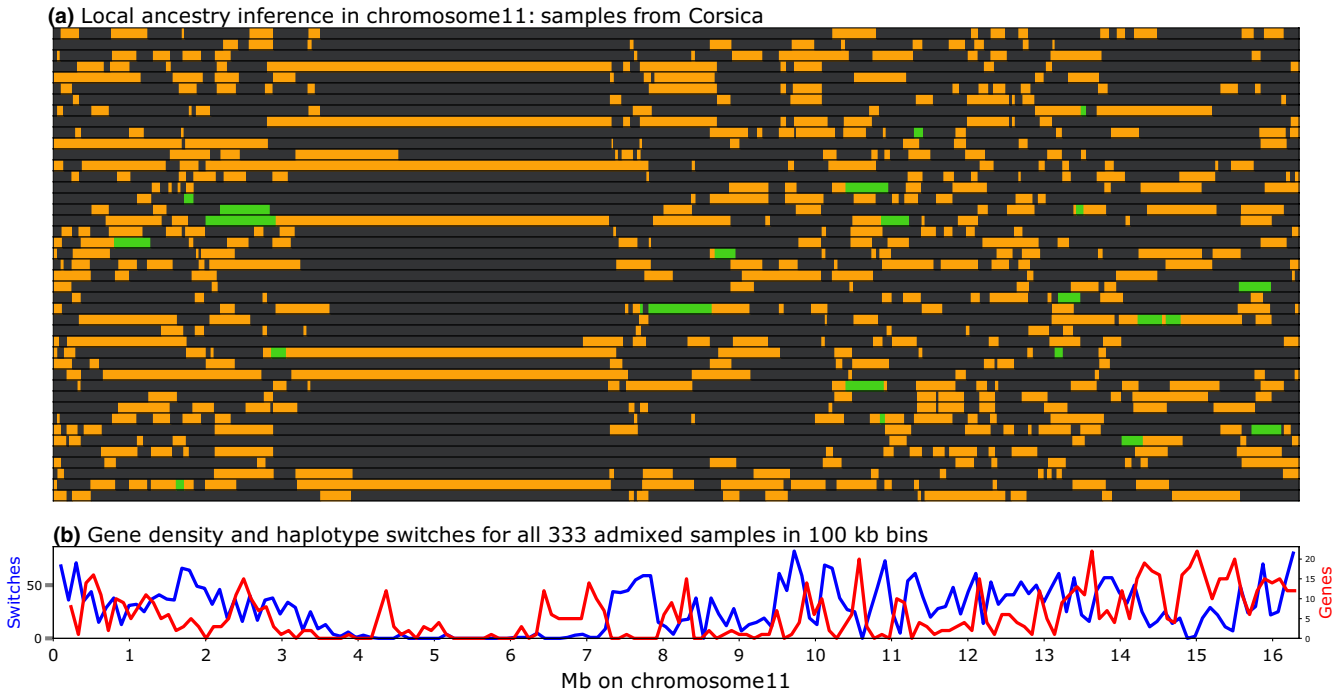


FIGURE 6 Local ancestry inference on chromosome 11 in the admixed samples from Corsica. (a) Each horizontal line represents the ancestry inference on one of the 43 individual samples from Corsica. Grey: *Apis mellifera mellifera* and *A. m. iberiensis* backgrounds; yellow: *A. m. ligustica* and *A. m. carnica* backgrounds; green: *A. m. caucasia* background. (b) Haplotype switches in all 412 admixed samples analysed. The 3-Mb haplotype block at positions 4–7 Mb on chromosome 11 shows very little historical recombination

densities, such as found in the honey bee, can potentially cause allele dropout when genotyping, due to interference in the probe designs. Deep knowledge of SNP and indel positions will help select candidates flanked by monomorphic sequences. As an example, we have studied the overlap between our data set and the SNPs present on the 100 k SNP chip developed recently by Jones et al. (2020). The results show that of 103,270 markers originally present on this chip, 61,320 on chromosomes 1–16 are in common with our ~7 million high-quality panel (Table S5). To investigate the effects of the high SNP density found on the honey bee genome on the potential quality of the genotyping results using the chip, we looked for additional SNPs close to each target SNP that could interfere with the genotyping process and cause allele dropout. Indeed, additional variation in the probe sequence within 50 bp flanking the target SNP will interfere with the hybridization-based genotyping (Gershoni et al., 2022) and it is therefore common practice to have at most one extra variant within 50 bp either side of the target SNP. For SNPs with an MAF >0.2, we find that 51,178 out of 61,320 meet this criterion (16.5% of markers lost). If rare variants are considered (MAF >0.01), this number drops to 29,828 (71% of the markers lost). This suggests that SNP chips may have to be tailored towards specific populations and/or designed with prior knowledge of SNP data from population genomics studies. Conversely, for lower density chips, the spacing of markers can be optimized by taking the haplotype structure into account, thus avoiding redundancy while maintaining the highest possible level of genetic information. Another advantage of sequencing haploid samples is that the whole data set represents phased chromosomes. Notably, the present data set will be invaluable for genotype imputation in future studies using lower density genotyping, such as SNP chips or low-pass sequencing (Gilly et al., 2019; Li et al., 2021; Wasik et al., 2021).

Although other studies have used whole genome sequencing for honey bee population genomics, these either concerned workers, thus complicating phasing (Dogantzis et al., 2021; Wallberg et al., 2014), and/or were more limited in terms of the number of samples studied (Henriques, Wallberg, et al., 2018b; Parejo et al., 2016, 2020). Moreover, to make practical use of these published results, the raw reads need to be downloaded from the short-read archive (SRA) and aligned to the genome reference prior to detection of variants. This is a highly demanding task, in terms of both labour and computing time. In contrast, our data set is much larger, is naturally phased, and genotypes for the samples and markers can be directly selected from the vcf file provided. Moreover, our data are based on the latest version AMEL_HAV3.1 of the genome assembly (Wallberg et al., 2019), whereas most other data sets are not, including one of the latest published, which is based on the older and less complete Amel4.5 assembly (Dogantzis et al., 2021). We believe our careful filtering steps on sequencing and alignment metrics provide a reliable set of markers and that the selection of reference samples can be done on the basis of the uniformity of genetic backgrounds, for instance by filtering on the admixture Q matrixes provided in Tables S6 and S7 on a user-defined basis.

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.2 | Population structure in managed honey bees

The deep understanding of European honey bee populations and of their recent admixture via imports of genetic stocks by breeders is not a simple task. Analyses of admixture events in complex population structures can be sensitive to a number of parameters and sometimes yield misleading results, especially if one or several populations have gone through a recent bottleneck (Lawson et al., 2018). PCA on all ~7 million markers indicate that our data set is structured into three main genetic types (Figure S8). The first principal component, representing 10.8% of the variance, separates two major groups corresponding respectively to subspecies from northwestern (M lineage) and southeastern Europe (C lineage). These two groups are represented by several populations, including the Savoy and Porquerolles conservatories from South-East France on one side, and bees that are not geographically far from Italy or Slovenia on the other. This large genetic distance, despite relatively close geographical proximity of the populations, supports the hypothesis of the colonization of Europe by honey bees via distinct western and eastern routes (Estoup et al., 1995; Han et al., 2012; Ruttner, 1988; Whitfield et al., 2006), and the separation between subspecies due to the Alps forming a natural barrier preventing genetic exchange (Rinderer, 2013). Along the second principal component, representing 3.1% of the variance, the population originating from *Apis mellifera caucasia* separates from the southeastern European populations (Figure S8). Prior to investigating admixture, we pruned SNPs in LD taking care to maximize the removal of redundancy while maintaining the general structure of the data (Figure 2; Figures S15–S17).

We explored a range of *K* numbers of genetic backgrounds, running multiple iterations of each, to determine the most likely admixture pattern (Figure 3). Our results indicate that this approach is necessary to ensure the results from each *K* model are stable prior to interpretation. We observe from our ADMIXTURE analyses that CV outliers within a *K* model are common. For instance, at $K = 8$, the mode with the lowest CV is only represented by eight out of 50 ADMIXTURE runs, whereas the major mode has 12 runs. On examining the admixture patterns from these two modes, the major mode suggests the *A. m. mellifera* bees from conservatories on mainland France to be hybrids between bees from Ouessant and Spain, and moreover with roughly 50% of each genetic background on the same mode, the *A. m. iberiensis* background represents also 50% of the M lineage background in the bees from Corsica (Figure S19). This is unlikely given the geography of Western Europe and our knowledge of the history of the bees of Ouessant. Indeed, Ouessant is a very small island (15.6 km^2) off the coast of western Brittany, isolated from the rest of the French honey bee population since its installation in 1987 and the prohibition of imports since 1991 mostly for sanitary reasons. In contrast, the mode with the eight runs and lowest CV presents a better separation of *A. m. mellifera* and *A. m. iberiensis*, which is also found in the major mode at $K = 9$ backgrounds. A smaller level of admixture can still be found between *A. m. mellifera* and *A. m. iberiensis*, which is quite likely to be due to the shared ancestry between these two subspecies.

The major mode at $K = 9$, represented by 33 out of 50 runs, returned the lowest mean CV value. This mode identifies mainland France *A. m. mellifera* samples as having a distinct genetic background and suggests that honey bees from Ouessant may have been re-introduced in the mainland conservatories. This mode also identifies a distinct genetic background in French and Swiss Buckfast bees. Buckfast bees were developed by Brother Adam, and are described in page 14 of "Beekeeping at the Buckfast Abbey" as a cross performed around 1915 between "the leather-coloured Italian bee and the old native English variety" (Brother Adam, 1986). Brother Adam also notes that the Italian bees that were imported in later years were distinct from those used in the development of the Buckfast strain. Our analysis of migrations between populations with TREEMIX suggests that the Buckfasts in our data set were subject to introgression with genetic material from *A. m. caucasia* (Figure 5b), although the timing of this potential admixture event could not be determined. When the two migrations of *A. m. ligustica* into Corsica and *A. m. caucasia* into the Buckfast are considered, which is a likely scenario suggested by the Evanno analysis, the latter is close to *A. m. carnica*, as seen in (Figure 5b,c). Interestingly, a whole genome sequence study of Italian honey bees also suggest that the Buckfast bees are closer to *A. m. carnica* than to *A. m. ligustica* (Minozzi et al., 2021) and no proximity of the Buckfast bees with M lineage bees was found either in their study or in ours, despite the cross at the origin of the Buckfast including an old native variety. Further investigations including more Buckfast samples and additional honey bee subspecies will be needed to fully elucidate this question. The *A. m. carnica* samples from Slovenia, Germany, France, Switzerland and Poland all share the same genetic background, reflecting their identical origin, probably recent imports, relative to the history of honey bee populations.

The population of bees from Corsica comes from a breeders' organization on the island, where importation has been prohibited since the 1980s. The results show that this population has the distinct characteristic of being homogeneous in composition, despite being admixed, with all samples showing mean proportions of 75% and 25% of *A. m. mellifera* and *A. m. ligustica* backgrounds, respectively (Figures 2 and 3). The introgression of Italian bees is confirmed by the TREEMIX migration analysis, and when this is accounted for, the Corsican samples group with *A. m. mellifera* bees from mainland France instead of being situated between the two main genetic subgroups of western and eastern European bees (Figure 2b,c). This result probably reflects the fact that Italian bees may have been imported on the Island before the ban on imports and that the population has been homogenized since then, at least within the breeders' organization. As beekeepers generally prefer the *A. m. ligustica* Italian bees over *A. m. mellifera*, it is very likely that the latter is the original population, as also suggested by Ruttner (1988). Although the hypothesis of the separation of the two subspecies on the mainland by the Alps seems appropriate (Rinderer, 2013), the situation of the

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Mediterranean islands in the region is not as clear. Based on physical geography alone, Corsica being at a closer distance to Italy than to France, the chances would have been greater to have originally C lineage rather than M lineage bees. Moreover, Corsica was under the control of Pisa, then fell to Genoa in 1284 and was only purchased by France in 1768. Further studies including samples from Sardinia would certainly help define the Mediterranean boundaries between the M lineage and C lineage honey bees and confirm observations based on morphology (Ruttner, 1988).

Apart from the subspecies references and the royal jelly populations, the honey bees provided by breeders are largely admixed, exhibiting high variability in background proportions—even for samples sourced from the same region. A typical example is that of the Tarn1 and Tarn2 populations, reflecting the fact that these two breeders, although situated very close to one another (<100 km), have very different genetic management strategies. The Tarn1 breeder produces Buckfast and *A. m. carnica* × *A. m. caucasia* queens by selecting within his own lines, and this is reflected in our results, in which the samples are mainly composed of Buckfast and *A. m. carnica* backgrounds, with a small amount of *A. m. caucasia*. By contrast, the Tarn2 breeder focuses principally on selecting for resistance to *Varroa destructor* and not treating his colonies with acaricides. A large proportion of *A. m. mellifera* background is present and the population is far less homogenous (Figures 2–4). This exemplifies the heterogeneity of the managed populations that can be found in France. A question that needs further investigation is the influence of the mating strategies used by the breeders, such as artificial insemination, mating stations, with drone-producing hives to saturate the environment with the desired genetic strains, or open mating (Cao et al., 2016). Interestingly, in our data set, only three *A. m. ligustica* and all of the bees from China have some royal jelly genetic background.

Previous analyses on worldwide data sets were published, either by whole genome sequencing of workers (Chen et al., 2022; Dogantzis et al., 2021; Wallberg et al., 2014) or by sequencing the mitochondrial DNA (Tihelka et al., 2020). These were intended to understand worldwide populations, the geographical origins and migration routes of *Apis mellifera*, a topic still under debate. Our intentions here are different, being targeted towards managed populations, for which detailed knowledge of genetic makeup is essential for further work concerning traits of interest to queen breeders and beekeepers and the interaction between different subspecies present within a territory. Typically, a refined description of admixture and of its distribution along chromosomes is essential to avoid confounding effects in genome-wide association studies (GWAS).

4.3 | Large haplotype blocks in the honey bee genome, specific to the M and C lineages

When investigating the contribution of SNPs to variance in the PCA, we noted that several large genomic regions, up to 3.5 Mb long, in which almost all markers contributed very strongly to the first principal component, separate bees from northwestern (M lineage) and southeastern Europe (C lineage). These regions were noted to coincide with haplotype blocks detected with PLINK. To investigate the matter further, we performed local ancestry inference in the admixed samples with RFMIX, using samples exhibiting 95% ancestry for the three main genetic backgrounds as references. A low recombination rate is confirmed by the observation of very few switches between the three main genetic backgrounds within these haplotype blocks. Interestingly, some of our regions, including the largest one detected on chromosome 11, coincide with regions of low recombination rate detected in other studies. These include an LD map produced with 30 diploid sequences from African worker bees (Wallberg et al., 2015), ancestry inference in an admixed population (Wragg et al., 2018), low-resolution genetic maps produced by RAD or ddRAD sequencing, with microsatellite or SNP markers, ddRAD sequencing (DeLory et al., 2020; Ross et al., 2015), or higher resolution genetic maps produced by whole genome sequencing of European (Liu et al., 2015) and African subspecies (Kawakami et al., 2019).

Most of these regions coincide with the position of the centromeres such as described in the reference genome assembly, which is based primarily on the combination of the location of *Ava*I repeats, which were previously assigned to centromeres by cytogenetic analysis, and of a low GC content (Beye & Moritz, 1995; Wallberg et al., 2019). However, the *Ava*I repeats only represent a very small fraction of the centromeric regions described, with the largest one only covering 14 kb (Wallberg et al., 2019), whereas the estimate of the extent of the centromeres, based on a GC content lower than the genome average, is much larger although imprecise and supposes a similar organization as for the AT-rich alpha-satellite repeats in vertebrates, such as human (Altemose et al., 2022). Although in some cases the boundaries of our regions of low recombination rate coincide with the actual positioning of the centromere on the genome assembly (Wallberg et al., 2019), such as in chromosomes 5 or 8, in other instances, such as in chromosome 12, the region we define is much narrower. Due to the difficulties in interpreting banding patterns in honey bee chromosomes, the position of the centromeres is not well defined. Some evidence based on G- and C-banding suggests there are four metacentric and 12 submetacentric or subtelocentric chromosomes (Hoshiba, 1984), whereas other evidence based on fluorescence in situ hybridization of a centromere probe suggests there are two metacentric, four submetacentric, two subtelocentric and eight telocentric chromosomes (Beye & Moritz, 1994). Our evidence suggests at least six chromosomes that could be telocentric or acrocentric: chromosomes 3, 5, 6, 9, 14 and 15.

Some of the haplotype blocks/regions of low recombination are large, such as representing up to 21% in the case of chromosome 11 (Figure 6). This may seem a lot, but recent findings in a complete sequencing of the human genome give a similar proportion for chromosomes 9, in which 40 Mb of satellite arrays represent 20% of the chromosome (Nurk et al., 2022). One important difference, however, is that the block on honey bee chromosome 11 contains some genes, except in the central region, whereas the satellite array described on human chromosome 9 does not. This

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reaffirms that our understanding of the centromere positions in the honey bee chromosomes requires refinement. The specific case of the acrocentric chromosomes in terms of gene content (Figure S20) seems to compare better to the situation described in humans, as the sequencing of the p-arm of the five human acrocentric chromosomes has allowed the discovery of novel genes within the satellite repeat-containing regions (Altemose et al., 2022). Centromeric DNA evolves rapidly, suggesting it goes through a genetic conflict known as the centromere drive hypothesis. This is due to the fact that in female meiosis, only one of the resulting chromosomes is included in the oocyte, leading to a competition between centromeres (Rosin & Mellone, 2017). The differentiation of centromeric regions between subspecies we observe here could be in line with this hypothesis. However, although in most species the rapidly evolving DNA sequence of centromeres is typically composed of highly repeated elements, this is not the case of the haplotype blocks found here, which seem to be associated with their respective centromeres due to a lack of recombination.

Some haplotype blocks may have another origin than centromeric DNA. For instance, genetic divergence could have been maintained by limiting recombination via the presence of structural variants such as inversions. Indeed, two of the blocks described here, between positions 4.0–5.1 and 5.8–6.9 Mb on chromosome 7, seem to coincide at least partially with two regions of haplotype divergence possibly due to inversions, detected between positions 3.9–4.3 and 6.3–7.3 Mb on the same chromosome, in a highland vs. lowland study of East African bees (Christmas et al., 2018). The slight differences in coordinates found between the two studies could be due to the fact that different version of the Amel_HAv3 assembly were used. However, if confirmed, this finding suggests that haplotype blocks differing between M lineage and C lineage bees such as found here might coincide with blocks found in other subspecies in Africa. Another study identifying the thelytoky locus (*Th*) in the South African Cape honey bee *Apis mellifera capensis* showed it was in a nonrecombining region over 100 kb long on chromosome 1, although long-read mapping failed to detect any inversion (Aumer et al., 2019).

Given the current hypotheses on the colonization of Europe by honey bees via distinct western and eastern routes (Estoup et al., 1995; Han et al., 2012; Ruttner, 1988; Whitfield et al., 2006), it is not surprising that the haplotype blocks described here, whether or not representing centromeric regions, tend to separate mainly the M and C lineage bees. Further analyses will be necessary to define the centromeric regions more precisely and study their implication, together with the other haplotype blocks, in the subspecies structure of honey bee populations.

5 | CONCLUSIONS

The sequencing of 870 haploid honey bee drones was shown here to be an invaluable approach for variant detection and for understanding the fine genetic makeup of a complex population having gone through multiple events of admixture. In addition, the extent of regions of extremely low recombination rate could be defined with higher precision than previously. The data set generated, based on the latest genome assembly, is a solid base for future research involving other honey bee populations and for any analyses requiring a reference set for simulations (Eynard et al., 2021), phasing or imputation.

AUTHOR CONTRIBUTIONS

YLC, J-PB, BB and AV designed the experiment. BB, YLC and AV coordinated colony selection, and sampling and samples were provided by KB, MB, CC, AG, PK, MP and AP. KC-T, EL and OB performed DNA extraction, library preparation and sequencing. DW, AV, SE and BS performed the bioinformatic analyses and cowrote the manuscript. All authors read and commented on the final manuscript.

ACKNOWLEDGEMENTS

This work was performed in collaboration with the GeT platform, Toulouse (France), a partner of the National Infrastructure France Génomique, thanks to support by the Commissariat aux Grands Invetissements (ANR-10-INBS-0009). Bioinformatics analyses were performed on the GenoToul Bioinfo computer cluster. This work was funded by a grant from the INRA Département de Génétique Animale (INRA Animal Genetics division) and by the SeqApiPop programme, funded by the FranceAgriMer grant 14-21-AT. We thank John Kefuss for helpful discussions. We thank Andrew Abrahams for providing honey bee samples from Colonsay (Scotland), the Association Conservatoire de l'Abeille Noire Bretonne (ACANB) for samples from Ouessant (France), CETA de Savoie for sample from Savoie, ADAPI for samples from Porquerolles and all beekeepers and bee breeders who kindly participated in this study by providing samples from their colonies.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge, for making publicly available the digitally-shareable data necessary to reproduce the reported results. The sequence data is available at from the SequenceRead Archive (SRA) at www.ncbi.nlm.nih.gov/sra under the BioProject accessionsPRJNA311274 as part of the SeqApiPop French honey bee diversity project dataset and PRJEB16533 as part of the Swiss honey bee population and conservationgenomics project dataset. A vcf file with the filtered 7 million SNP and 870 samples is available at (https://doi.org/10.52581/zenodo.5592452) for download, together with the list of the 629 unique samples used for the diversity analysis. Scripts and supplementary description of bioinformatic analyses are available in GitHub: (https://github.com/avignal5/SeqApiPop/tree/v1.5 and a

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

version of record of these is deposited in https://zenodo.org/record/6346402

DATA AVAILABILITY STATEMNT

DNA Sequences and samples metadata for this project have been deposited in the Sequence Read Archive (SRA) at www.ncbi.nlm.nih.gov/sra under the BioProject accessions PRJNA311274 as part of the SeqApiPop French honey bee diversity project dataset and PRJEB16533 as part of the Swiss honey bee population and conservation genomics project dataset. Individual SRA run and BioSample accessions for all samples are given in Table S1. A vcf file with the filtered 7 million SNP and 870 samples is available at https://doi.org/10.5281/zenodo.5592452 for download, together with the list of the 629 unique samples used for the diversity analysis. Scripts and supplementary description of bioinformatic analyses are available in GitHub: https://github.com/avignal5/SeqApiPop/tree/v1.5 and a version of record of these is deposited in https://zenodo.org/record/6346402.

ORCID

David Wragg https://orcid.org/0000-0002-4007-953X
Sonia E. Eynard https://orcid.org/0000-0002-8609-5869
Kaspar Bienefeld https://orcid.org/0000-0002-5201-475X
M. Alice Pinto https://orcid.org/0000-0001-9663-8399
Bertrand Servin https://orcid.org/0000-0001-5141-0913
Alain Vignal https://orcid.org/0000-0002-6797-2125

REFERENCES

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246. https://doi.org/10.1186/1471-2105-12-246

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A., Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science*, *376*(6588), eabl4178. https://doi.org/10.1126/science.abl4178

Aumer, D., Stolle, E., Allsopp, M., Mumoki, F., Pirk, C. W. W., & Moritz, R. F. A. (2019). A single SNP turns a social honey bee (*Apis mellifera*) worker into a selfish parasite. *Molecular Biology and Evolution*, *36*(3), 516–526. https://doi.org/10.1093/molbev/msy232

Bansal, V., Bashir, A., & Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Research*, *17*(2), 219–230. https://doi.org/10.1101/gr.5774507

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817–2823. https://doi.org/10.1093/bioinformatics/btw327

Beye, M., & Moritz, R. F. (1995). Characterization of honeybee (*Apis mellifera* L.) chromosomes using repetitive DNA probes and fluorescence in situ hybridization. *The Journal of Heredity*, *86*(2), 145–150.

Beye, M., & Moritz, R. F. A. (1994). A centromere-specific probe for fluorescence in-situ hybridization on chromosomes of *Apis mellifera* L. *Apidologie*, *25*(3), 322–326.

Bovine HapMap Consortium, Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., Green, R. D., Hamernik, D. L., Kappes, S. M., Lien, S., Matukumalli, L. K., McEwan, J. C., Nazareth, L. V., Schnabel, R. D., Weinstock, G. M., Wheeler, D. A., Ajmone-Marsan, P., Boettcher, P. J., … Dodds, K. G. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, *324*(5926), 528–532. https://doi.org/10.1126/science.1167936

Adam, B. (1986). *Bee-keeping at Buckfast Abbey (New edition)*. Northern Bee Books.

Cao, L.-F., Zheng, H.-Q., Pirk, C. W. W., Hu, F.-L., & Xu, Z.-W. (2016). High royal jelly-producing honeybees (*Apis mellifera ligustica*) (Hymenoptera: Apidae) in China. *Journal of Economic Entomology*, *109*(2), 510–514. https://doi.org/10.1093/jee/tow013

Carpenter, M. H., & Harpur, B. A. (2021). Genetic past, present, and future of the honey bee (*Apis mellifera*) in The United States of America. *Apidologie*, *52*(1), 63–79. https://doi.org/10.1007/s13592-020-00836-4

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, *16*(11), 627–640. https://doi.org/10.1038/nrg3933

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., Li, R., & Shi, W. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan* n. ssp. *Molecular Biology and Evolution*, *33*(5), 1337–1348. https://doi.org/10.1093/molbev/msw017

Chen, C., Parejo, M., Momeni, J., Langa, J., Nielsen, R. O., Shi, W., Smartbees WP3 Diversity Contributors, Vingborg, R., Kryger, P., Bouga, M., Estonba, A., & Meixner, M. (2022). Population structure and diversity in European honey bees (*Apis mellifera* L.)—An empirical comparison of pool and individual whole-genome sequencing. *Genes*, *13*(2), 182. https://doi.org/10.3390/genes13020182

Christmas, M. J., Wallberg, A., Bunikis, I., Olsson, A., Wallerman, O., & Webster, M. T. (2018). Chromosomal inversions associated with environmental adaptation in honeybees. *Molecular Ecology*, *8*(6), 1358–1374. https://doi.org/10.1111/mec.14944

Cornuet, J. M., Daoudi, A., & Chevalet, C. (1986). Genetic pollution and number of matings in a black honey bee (*Apis mellifera mellifera*) population. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *73*(2), 223–227. https://doi.org/10.1007/BF00289278

Cornuet, J. M., Fresnaye, J., Blanc, J., & Paris, R. (1979). Production de miel chez des hybrides interraciaux d'abeilles (A*pis mellifica* L.) lors de générations successives de rétrocroisement sur la race locale. *Apidologie*, *10*(1), 3–15. https://doi.org/10.1051/apido:19790101

Cornuet, J. M., Fresnaye, J., Lavie, P., Blanc, J., Hanout, S., & Mary-Lafargue, C. (1978). Étude biométrique de deux populations d'abeilles cévenoles. *Apidologie*, *9*(1), 41–55. https://doi.org/10.1051/apido:19780104

Cornuet, J.-M., Albisetti, J., Mallet, N., & Fresnaye, J. (1982). Étude biométrique d'une population d'abeilles landaises. *Apidologie*, *13*(1), 3–13. https://doi.org/10.1051/apido:19820101

De la Rúa, P., Jaffé, R., Dall'Olio, R., Muñoz, I., & Serrano, J. (2009). Biodiversity, conservation and current threats to European honeybees. *Apidologie*, *40*(3), 263–284. https://doi.org/10.1051/apido/2009027

Delaneau, O., Marchini, J., & Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*(2), 179–181. https://doi.org/10.1038/nmeth.1785

DeLory, T., Funderburk, K., Miller, K., Zuluaga-Smith, W., McPherson, S., Pirk, C. W., Costa, C., Weinstein-Teixeira, É., Dahle, B., & Rueppell, O. (2020). Local variation in recombination rates of the honey bee

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(*Apis mellifera*) genome among samples from six disparate populations. *Insectes Sociaux*, *67*(1), 127–138. https://doi.org/10.1007/s00040-019-00736-6

Dogantzis, K. A., Tiwari, T., Conflitti, I. M., Dey, A., Patch, H. M., Muli, E. M., Garnery, L., Whitfield, C. W., Stolle, E., Alqarni, A. S., Allsopp, M. H., & Zayed, A. (2021). Thrice out of Asia and the adaptive radiation of the western honey bee. *Science Advances*, *7*(49), eabj2151. https://doi.org/10.1126/sciadv.abj2151

Estoup, A., Garnery, L., Solignac, M., & Cornuet, J. M. (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: Hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics*, *140*(2), 679–695.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Eynard, S. E., Vignal, A., Basso, B., Le Conte, Y., Decourtye, A., Genestout, L., Labarthe, E., Mondet, F., Tabet, K., & Servin, B. (2021). From group to individual—Genotyping by pool sequencing eusocial colonies/ [Preprint]. *Genomics*. https://doi.org/10.1101/2021.11.08.467442

Fontana, P., Costa, C., Prisco, G. D., Ruzzier, E., Annoscia, D., Battisti, A., Caoduro, G., Carpana, E., Contessi, A., Dal, A., Dall, R., Cristofaro, A. D., Felicioli, A., Floris, I., Gardi, T., Lodesani, M., Malagnini, V., Manias, L., Manino, A., … Segrè, A. (2018). Appeal for biodiversity protection of native honey bee subspecies of *Apis mellifera* in Italy (San Michele all'Adige declaration). *Bulletin of Insectology*, *71*(2), 257–271.

Franck, P., Garnery, L., Celebrano, G., Solignac, M., & Cornuet, J. M. (2000). Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and sicily (*A. m. sicula*). *Molecular Ecology*, *9*(7), 907–921.

Fresnaye, J., Lavie, P., & Boesiger, E. (1974). La variabilité de la production du miel chez l'abeille de race noire (*Apis mellifica* L.) et chez quelques hybrides interraciaux. *Apidologie*, *5*(1), 1–20.

Gershoni, M., Shirak, A., Raz, R., & Seroussi, E. (2022). Comparing BeadChip and WGS genotyping: Non-technical failed calling is attributable to additional variation within the probe target sequence. *Genes*, *13*(3), 485. https://doi.org/10.3390/genes13030485

Gilly, A., Southam, L., Suveges, D., Kuchenbaecker, K., Moore, R., Melloni, G. E. M., Hatzikotoulas, K., Farmaki, A.-E., Ritchie, G., Schwartzentruber, J., Danecek, P., Kilian, B., Pollard, M. O., Ge, X., Tsafantakis, E., Dedoussis, G., & Zeggini, E. (2019). Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics*, *35*(15), 2555–2561. https://doi.org/10.1093/bioinformatics/bty1032

Gregorc, A., & Lokar, V. (2010). Selection criteria in an apiary of Carnolian honey bee (*Apis mellifera carnica*) colonies for queen rearing. *Journal of Central European Agriculture*, *11*(4), 401–408.

Gregorc, Ă., Lokar, V., & Škerl, M. I. S. (2008). Testing of the isolation of the Rog-Ponikve mating station for Carniolan (*Apis mellifera carnica*) honey bee queens. *Journal of Apicultural Research*, *47*(2), 137–140. https://doi.org/10.1080/00218839.2008.11101440

Han, F., Wallberg, A., & Webster, M. T. (2012). From where did the Western honeybee (*Apis mellifera*) originate? *Ecology and Evolution*, *2*(8), 1949–1957. https://doi.org/10.1002/ece3.312

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M. D., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(7), 2614–2619. https://doi.org/10.1073/pnas.1315506111

Hassett, J., Browne, K. A., McCormack, G. P., Moore, E., Society, N. I. H. B., Soland, G., & Geary, M. (2018). A significant pure population of the dark European honey bee (*Apis mellifera mellifera*) remains in Ireland. *Journal of Apicultural Research*, *57*(3), 337–350. https://doi.org/10.1080/00218839.2018.1433949

Henriques, D., Browne, K. A., Barnett, M. W., Parejo, M., Kryger, P., Freeman, T. C., Muñoz, I., Garnery, L., Highet, F., Jonhston, J. S., McCormack, G. P., & Pinto, M. A. (2018a). High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: An accurate and cost-effective SNP-based tool. *Scientific Reports*, *8*(1), 8552. https://doi.org/10.1038/s41598-018-26932-1

Henriques, D., Wallberg, A., Chávez-Galarza, J., Johnston, J. S., Webster, M. T., & Pinto, M. A. (2018b). Whole genome SNP-associated signatures of local adaptation in honeybees of the Iberian Peninsula. *Scientific Reports*, *8*(1), 11145. https://doi.org/10.1038/s41598-018-29469-5

Hoshiba, H. (1984). Karyotype and banding analyses on haploid males of the honey bee (*Apis mellifera*). *Proceedings of the Japan Academy, Series B*, *60*(5), 122–124. https://doi.org/10.2183/pjab.60.122

Ilyasov, R. A., Lee, M., Takahashi, J., Kwon, H. W., & Nikolenko, A. G. (2020). A revision of subspecies structure of western honey bee *Apis mellifera*. *Saudi Journal of Biological Sciences*, *27*(12), 3615–3621. https://doi.org/10.1016/j.sjbs.2020.08.001

Jones, J. C., Du, Z. G., Bernstein, R., Meyer, M., Hoppe, A., Schilling, E., Ableitner, M., Juling, K., Dick, R., Strauss, A. S., & Bienefeld, K. (2020). Tool for genomic selection and breeding to evolutionary adaptation: Development of a 100K single nucleotide polymorphism array for the honey bee. *Ecology and Evolution*, *10*(13), 6246–6256. https://doi.org/10.1002/ece3.6357

Jones, J. C., Wallberg, A., Christmas, M. J., Kapheim, K. M., & Webster, M. T. (2019). Extreme differences in recombination rate between the genomes of a solitary and a social bee. *Molecular Biology and Evolution*, *36*(10), 2277–2291. https://doi.org/10.1093/molbev/msz130

Kawakami, T., Wallberg, A., Olsson, A., Wintermantel, D., de Miranda, J. R., Allsopp, M., Rundlöf, M., & Webster, M. T. (2019). Substantial heritable variation in recombination rate on multiple scales in honeybees and bumblebees. *Genetics*, *212*(4), 1101–1119. https://doi.org/10.1534/genetics.119.302008

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258. https://doi.org/10.1038/s41467-018-05257-7

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J. H., Mazur, C. A., Berisa, T., & Pickrell, J. K. (2021). Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Research*, *31*(4), 529–537. https://doi.org/10.1101/gr.266486.120

Liu, H., Zhang, X., Huang, J., Chen, J.-Q., Tian, D., Hurst, L. D., & Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biology*, *16*, 15. https://doi.org/10.1186/s13059-014-0566-0

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., &

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meixner, M. D., Pinto, M. A., Bouga, M., Kryger, P., Ivanova, E., & Fuchs, S. (2013). Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. *Journal of Apicultural Research*, *52*(4), 1–28. https://doi.org/10.3896/IBRA.1.52.4.05

Minozzi, G., Lazzari, B., De Iorio, M. G., Costa, C., Carpana, E., Crepaldi, P., Rizzi, R., Facchini, E., Gandini, G., Stella, A., & Pagnacco, G. (2021). Whole-genome sequence analysis of Italian honeybees (*Apis mellifera*). *Animals*, *11*(5), 1311. https://doi.org/10.3390/ani11051311

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Parejo, M., Henriques, D., Pinto, M. A., Soland-Reckeweg, G., & Neuditschko, M. (2018). Empirical comparison of microsatellite and SNP markers to estimate introgression in Apis mellifera mellifera. *Journal of Apicultural Research*, *57*(4), 504–506. https://doi.org/10.1080/00218839.2018.1494894

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using whole-genome sequence information to foster conservation efforts for the European dark honey bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution*, *4*, 140. https://doi.org/10.3389/fevo.2016.00140

Parejo, M., Wragg, D., Henriques, D., Charrière, J.-D., & Estonba, A. (2020). Digging into the genomic past of Swiss honey bees by whole-genome sequencing museum specimens. *Genome Biology and Evolution*, *12*(12), 2535–2551. https://doi.org/10.1093/gbe/evaa188

Parejo, M., Wragg, D., Henriques, D., Vignal, A., & Neuditschko, M. (2017). Genome-wide scans between two honeybee populations reveal putative signatures of human-mediated selection. *Animal Genetics*, *48*(6), 704–707. https://doi.org/10.1111/age.12599

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Pinto, M. A., Henriques, D., Chávez-Galarza, J., Kryger, P., Garnery, L., van der Zee, R., Dahle, B., Soland-Reckeweg, G., de la Rúa, P., Dall' Olio, R., Carreck, N. L., & Johnston, J. S. (2014). Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: A genome-wide assessment using SNPs and mtDNA sequence data. *Journal of Apicultural Research*, *53*(2), 269–278. https://doi.org/10.3896/IBRA.1.53.2.08

Privé, F., Luu, K., Vilhjálmsson, B. J., & Blum, M. G. B. (2020). Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Molecular Biology and Evolution*, *37*(7), 2153–2154. https://doi.org/10.1093/molbev/msaa053

Puškadija, Z., Kovačić, M., Raguž, N., Lukić, B., Prešern, J., & Tofilski, A. (2021). Morphological diversity of Carniolan honey bee (*Apis mellifera carnica*) in Croatia and Slovenia. *Journal of Apicultural Research*, *60*(2), 326–336. https://doi.org/10.1080/00218839.2020.1843847

Rinderer, T. E. (2013). *Bee genetics and breeding*. Academic press.

Rosin, L. F., & Mellone, B. G. (2017). Centromeres drive a hard bargain. *Trends in Genetics*, *33*(2), 101–117. https://doi.org/10.1016/j.tig.2016.12.001

Ross, C. R., DeFelice, D. S., Hunt, G. J., Ihle, K. E., Amdam, G. V., & Rueppell, O. (2015). Genomic correlates of recombination rate and its variability across eight recombination maps in the western honey bee (*Apis mellifera* L.). *BMC Genomics*, *16*, 107. https://doi.org/10.1186/s12864-015-1281-2

Ruttner, F. (1988). *Biogeography and taxonomy of honeybees*. Springer-Verlag.

Storey, J. D., Bass, A. J., Dabney, A., Robinson, D., & Warnes, G. (2022). *qvalue: Q-value estimation for false discovery rate control (2.26.0) [Computer software]*. Bioconductor version: Release (3.14). https://doi.org/10.18129/B9.bioc.qvalue

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Techer, M. A., Clémencet, J., Turpin, P., Volbert, N., Reynaud, B., & Delatte, H. (2015). Genetic characterization of the honeybee (*Apis mellifera*) population of Rodrigues Island, based on microsatellite and mitochondrial DNA. *Apidologie*, *46*(4), 445–454. https://doi.org/10.1007/s13592-014-0335-9

Tihelka, E., Cai, C., Pisani, D., & Donoghue, P. C. J. (2020). Mitochondrial genomes illuminate the evolutionary history of the Western honey bee (*Apis mellifera*). *Scientific Reports*, *10*(1), 14515. https://doi.org/10.1038/s41598-020-71393-0

Wallberg, A., Bunikis, I., Pettersson, O. V., Mosbech, M.-B., Childers, A. K., Evans, J. D., Mikheyev, A. S., Robertson, H. M., Robinson, G. E., & Webster, M. T. (2019). A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*, *20*(1), 275. https://doi.org/10.1186/s12864-019-5642-0

Wallberg, A., Glémin, S., & Webster, M. T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genetics*, *11*(4), e1005189. https://doi.org/10.1371/journal.pgen.1005189

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z. L. P., Allsopp, M. H., Kandemir, I., De la Rúa, P., Pirk, C. W., & Webster, M. T. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature Genetics*, *46*(10), 1081–1088. https://doi.org/10.1038/ng.3077

Wasik, K., Berisa, T., Pickrell, J. K., Li, J. H., Fraser, D. J., King, K., & Cox, C. (2021). Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics*, *22*(1), 197. https://doi.org/10.1186/s12864-021-07508-2

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), 642–645. https://doi.org/10.1126/science.1132772

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J.-P., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Scientific Reports*, *6*(1), 27168. https://doi.org/10.1038/srep27168

Wragg, D., Techer, M. A., Canale-Tabet, K., Basso, B., Bidanel, J.-P., Labarthe, E., Bouchez, O., Le Conte, Y., Clémencet, J., Delatte, H., & Vignal, A. (2018). Autosomal and mitochondrial adaptation

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

following admixture: A case study on the honeybees of Reunion Island. *Genome Biology and Evolution*, *10*(1), 220–238. https://doi.org/10.1093/gbe/evx247

Zayed, A., & Whitfield, C. W. (2008). A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(9), 3421–3426. https://doi.org/10.1073/pnas.0800107105

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wragg, D., Eynard, S. E., Basso, B., Canale-Tabet, K., Labarthe, E., Bouchez, O., Bienefeld, K., Bieńkowska, M., Costa, C., Gregorc, A., Kryger, P., Parejo, M., Pinto, M. A., Bidanel, J-P, Servin, B., Le Conte, Y., & Vignal, A. (2022). Complex population structure and haplotype patterns in the Western European honey bee from sequencing a large panel of haploid drones. *Molecular Ecology Resources*, *22*, 3068–3086. https://doi.org/10.1111/1755-0998.13665

17550998, 2022, 8, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13665 by Cochrane Portugal, Wiley Online Library on [13/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License