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

Short title: Haploid drone sequence for population and genome analysis

David Wragg[1,2*], Sonia E. Eynard[1], Benjamin Basso[3,4], Kamila Canale-Tabet[1], Emmanuelle Labarthe[1], Olivier Bouchez[5], Kaspar Bienefeld[6], Małgorzata Bieńkowska[7], Cecilia Costa[8], Aleš Gregorc[9], Per Kryger[10], Melanie Parejo[11,12], M. Alice Pinto[13], Jean-Pierre Bidanel[14], Bertrand Servin[1], Yves Le Conte[4], Alain Vignal[1]

[1] GenPhySE, Université de Toulouse, INRAE, INPT, INP-ENVT, 31326 Castanet Tolosan, France

[2] Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, UK

[3] Institut de l'abeille (ITSAP), UMT PrADE, 8914 Avignon, France

[4] INRAE, UR 406 Abeilles et Environment, UMT PrADE, 84914 Avignon, France

[5]GeT-PlaGe, Genotoul, INRAE, Castanet Tolosan, France

[6]Bee Research Institute, F.-Engels-Straße 32, 16540 Hohen Neuendorf, Germany

[7]National Research Institute of Horticulture, Apiculture Division, 24–100 Puławy, Poland

[8]CREA Research Centre for Agriculture and Environment, via di Saliceto 80, Bologna, Italy

[9]University of Maribor, Faculty of Agriculture and Life Sciences, Pivola, Slovenia.

[10]Department of Agroecology, Science and Technology, Aarhus University, Slagelse, Denmark

[11]Agroscope, Swiss Bee Research Centre, Bern, Switzerland

[12]Applied Genomics and Bioinformatics, Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, Leioa, Spain

[13]Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Bragança, Portugal

[14]GABI, INRAE, AgroParisTech, Université Paris-Saclay, 78352 Jouy-en-Josas, France

David Wragg (david.wragg@roslin.ed.ac.uk)

Sonia E. Eynard (sonia.eynard@inrae.fr)

Benjamin Basso (benjamin.basso@inrae.fr)

Kamila Canale-Tabet (kamila.tabet@inrae.fr)

Emmanuelle Labarthe (emmanuelle.labarthe@inrae.fr)

Olivier Bouchez (olivier.bouchez@inrae.fr)

Kaspar Bienefeld (kaspar.bienefeld@hu-berlin.de)

Małgorzata Bieńkowska (malgorzata.bienkowska@inhort.pl)

Cecilia Costa (cecilia.costa@crea.gov.it)

Aleš Gregorc (ales.gregorc@um.si)

Per Kryger (per.kryger@agro.au.dk)

Melanie Parejo (melanieparejo@gmail.com)

M. Alice Pinto (apinto@ipb.pt)

Jean-Pierre Bidanel (jean-pierre.bidanel@inrae.fr)

Bertrand Servin (bertrand.servin@inrae.fr)

Yves Le Conte (yves.le-conte@inrae.fr)

Alain Vignal (alain.vignal@inrae.fr)

Keywords : genome, population, honey bee

Abstract

Honey bee subspecies originate from specific geographic areas in Africa, Europe and the Middle East. The interest of beekeepers in specific phenotypes has led them to import subspecies to regions outside of their original range. The resulting admixture complicates population genetics analyses and populations stratification can be a major problem for association studies. As a typical example, the case of the French population is studied here. We sequenced 870 haploid drones for SNP detection and identified nine genetic backgrounds in 629 samples. Five correspond to subspecies, two to isolated populations and two to human-mediated population management. We also highlight several large haplotype blocks, some of which coinciding with the position of centromeres. The largest is 3.6 Mb long on chromosome 11, representing 1.6 % of the genome and has two major haplotypes, corresponding to the two dominant genetic background identified.

Introduction

The honey bee *Apis mellifera* comprises more than 30 subspecies, each of which defined according to morphological, behavioural, physiological and ecological characteristics suited to their local habitat [1–4]. 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) [1]. 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* [4]. 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 [5]. 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 [6,7]. However, during the 20th century much interest arose amongst apiarists in developing hybrids between the endemic *A. m.*

mellifera and other subspecies including *A. m. ligustica*, *A. m. carnica* and the Caucasian *A. m. caucasia* from Georgia [1,8,9]. Apiarists found the hybrids to perform better with regards to the production of honey and royal jelly than the native *A. m. mellifera*, spurring further interest in these subspecies which were also reported to be more docile and easier to manage [1]. *A. m. ligustica* is a very popular subspecies worldwide 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 [10]. *A. m. ligustica* queens are also frequently exported worldwide, and most of the honey bees imported during the last centuries into the New World were also of Italian origin [10,11]. Apiculture involving *A. m. carnica*, is also very popular among apiarists [12]. *A. m. carnica* became increasingly popular for further selection throughout central and western Europe [13,14] on account of their calm temperament and higher honey yield compared to *A. m. mellifera* [1], to the point where *A. carnica* almost replaced entirely *A.m. mellifera* in Germany [15]. *A. m. caucasia* is a subspecies that was also imported to France, to generate *A. m. ligustica* x *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 [16]. Following the extensive imports of queens from "exotic" subspecies, the genetic makeup of honey bee populations in France was unclear. The genetic pollution of local populations soon emerged with clear phenotypic consequences such as changes in the colour of the cuticle [17]. 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 [18–22].

As a result of the different breeding practices, the necessity for a study targeted towards *A. m. mellifera* conservatories and French bee breeders specialized in rearing and selling queens arose and in this context the genomic diversity project "SeqApiPop" was started. Within this project, samples from

97 French conservatories, from individual French breeders and breeder organisations were analysed,
98 including Buckfast samples. Traditionally, such wide diversity studies have been performed using a
99 small number of molecular markers such as microsatellites [23] or limited sets of single-nucleotide
100 polymorphisms (SNPs) [24–27], enabling population stratification, introgression and admixture levels
101 to be characterized. However, to understand complex population admixture events, as has happened
102 for the managed honey bee populations in France and elsewhere, or to identify signatures of natural
103 [25,28–30] or artificial [31,32] selection in the genome, a much higher density of markers is required.
104 As no high-density SNP chip was available for honey bee at the onset of the project, and as the honey
105 bee genome is very small compared to most animal genomes, being only 226.5 Mb long [33], we
106 employed a whole-genome sequencing approach [28,34]. Although the sequencing of honey bee
107 workers has proved successful for detecting selection signatures or admixture events [28,34–36],
108 analysing haploid drones allows to sequence at a lower depth and with greater accuracy in variant
109 detection [20,29,31]. An additional advantage of sequencing haploids is that the alleles are phased,
110 which is invaluable for studies investigating genome dynamics such as recombination hotspots and
111 haplotype structure. Although some insights into recombination patterns in the honey bee have been
112 made through the analysis of drones from individual colonies [37,38] and Linkage Disequilibrium
113 (LD)-based approaches [39,40], a deep understanding of the recombination landscape, essential for
114 fine-scale genetic analyses, requires hundreds of phased genomes. Such 'HapMap' projects have been
115 conducted in humans and cattle, initially using SNP arrays [41,42] and more recently by whole-
116 genome sequencing as in the "1000 genome" projects [43,44].
117 Therefore, as a first step towards a deep understanding of French and Western European managed
118 honey bee populations and of their genome dynamics, we undertook the sequencing of a large dataset
119 of haploid drones. This set is composed of samples from French conservatories or commercial
120 breeders in addition to samples from several European countries each representing potentially pure *A.*
121 *m. ligustica*, *A. m. carnica, A. m. mellifera* and *A. m. caucasia* populations typically imported by

122 French breeders. Finally, *A. m. iberiensis*, the Iberian subspecies only separated from the native
123 French *A. m. mellifera* by the natural barrier of the Pyrenees was also studied. In total, 870 samples
124 were sequenced for SNP detection and 629 were used for a detailed genetic analysis of present-day
125 honey bee populations in France.
126
127 **Methods**
128 *Sampling and sequencing*
129 For the population genomics analyses, one individual drone per colony was sampled before
130 emergence, from colonies throughout France, Spain, Germany, Switzerland, Italy, the UK, Slovenia,
131 Poland, Denmark, China and from a French beekeeper having imported queens from Georgia,
132 amounting to a total of 642 samples (Supplementary figure 1). To improve the robustness of the
133 primary SNP detection and filtering steps, additional samples of similar genetic backgrounds from two
134 other projects in progress were used. These included 198 samples from two other projects and 30
135 colony duplicates from the present population genomics project. Thus, although 642 colonies were
136 included for population genomics analyses, in total 870 samples were used for SNP detection
137 (supplementary table 1).
138 DNA was extracted from the thorax of adult bees or from pupae as described in Wragg *et al*. (2016).
139 Briefly, drones were sampled at either the pupae/nymph or larval stage and stored in absolute ethanol
140 at -20°C. DNA was extracted from the thorax or from diced whole larvae. Tissue fragments were first
141 incubated 3 hours at 56° in 1 mL of a solution containing 4 M urea, 10 mM Tris-HCl pH 8, 300 mM
142 NaCl, 1% SDS, 10 mM EDTA and 0.25 mg proteinase K, after which 0.25 mg proteinase K was
143 added for an incubation over-night at 37°C. Four hundred µL of a saturated NaCl solution was added
144 to the incubation, which was then gently mixed and centrifuged for 30 minutes at 15000 g. The
145 supernatant was treated for 5 minutes at room temperature with RNAse (Qiagen) and then centrifuged
146 again, after which the DNA in the supernatant was precipitated with absolute ethanol and re-suspended

147 in 100 µL TE 10/0.1. Pair-end sequencing was performed on Illumina™ HiSeq 2000, 2500 and 3000
148 sequencing machines with 20 samples per lane, or on a NovaSeq machine with 96 samples per lane,
149 following the manufacturer's protocols for library reparations.
150
151 *Mapping and genotype calling*
152 Sequencing reads were mapped to the reference genome Amel_HAv3.1 [33] using BWA-MEM
153 (v0.7.15) [45], and duplicates marked with Picard (v2.18.2;) (http://broadinstitute.github.io/picard/).
154 Libraries that were sequenced in multiple runs were merged with Samtools (v1.8) merge [46] prior to
155 marking duplicates. Local realignment and base quality score recalibration (BQSR) were performed
156 using GATK (v4.1.2.) [47], using single-nucleotide polymorphisms (SNPs) called with GATK
157 HaplotypeCaller as covariates for BQSR. Each drone was processed with the pipeline independently,
158 and genotyped independently with HaplotypeCaller. Although the drones sequenced are haploid,
159 variant calling was performed using a diploid model to allow the detection and removal of SNPs that
160 called heterozygous genotypes in > 1% of samples, which might have arisen for example as a result of
161 short-tandem repeats repeats and could highlight Copy Number Variants on the genome. Individual
162 gVCF files were combined with CombineGVCFs, and then jointly genotyped with GenotypeGVCFs,
163 resulting in a single VCF file for the 870 samples containing 14.990.574 raw variants detected,
164 10.601.454 potential SNPs were retained once Indels were removed with GATK SelectVariants.
165 Sequencing depth was estimated using Mosdepth [48]. Further details are given in supplementary file
166 SeqApiPop_1_MappingCalling.pdf.
167
168 *Quality filters on SNPs*
169 The first run of filters concerns technical issues related to the sequencing and alignment steps and
170 were therefore used for the total dataset of 870 samples, to benefit from its large size for SNP
171 detection and validation (supplementary figure 2). These were (i) strand biases and mapping quality

172 metrics (SOR ≥ 3; FS ≤ 60 and MQ ≥ 40), (ii) genotyping quality metrics (QUAL > 200 and QD < 20)
173 and (iii) individual SNP genotyping metrics (heterozygote calls < 1%; missing genotypes < 5%, allele
174 number < 4 and genotypes having individual GQ < 10 < 20%). Distribution and ECDF plots of values
175 for all the filters used on the dataset were used to select thresholds and are shown in supplementary
176 file SeqApiPop_2_VcfCleanup.pdf.
177
178 *Haplotype blocks detection, LD pruning, PCA, Admixture, Treemix, RFMix*
179 Haplotype blocks were detected in Plink (v1.9) [49] using the blocks function, "--blocks no-pheno-req
180 no-small-max-span", with the parameter "--blocks-max-kb 5000". LD pruning was performed in Plink
181 with the indep-pairwise function. Principal component analyses were performed in Plink and the
182 contribution of individual SNPs to the principal components were estimated using smartpca from the
183 eigensoft package v7.2.1. Further details are given in supplementary file
184 SeqApiPop_3_LDfilterAndPCAs.pdf. Admixture analysis was performed with the program Admixture
185 v 1.3.0 [50], with values of K ranging from 2 to 16. Fifty runs were performed each time using a
186 unique random seed. The Pong software [51] was used for aligning runs with different K values and
187 for grouping results from runs into clustering modes, setting the similarity threshold to -s = 0.98.
188 Further details are given in supplementary file SeqApiPop_4_Admixture.pdf. Population migration
189 analysis was performed with TreeMix [52], with the option for grouping SNPs set to -k=500, testing
190 between 0 and 9 migrations and performing 100 runs per migration with a unique random seed. The
191 optimum number of migrations was estimated with the R package OptM (Fitak, R. R.:
192 https://github.com/cran/OptM) using the Evanno method provided [53]. Tree summaries for the 100
193 runs per migration tested were performed with DendroPy [54] and drawn with FigTree v1.4.4
194 (http://tree.bio.ed.ac.uk/software/figtree/). Further details are given in supplementary file
195 SeqApiPop_5_TreeMix.pdf. Local ancestry inference and positioning of haplotype switches were
196 performed with RFMix v2.03-r0 [55]. Three main genetic backgrounds were considered for this

197 analysis, corresponding to the three major groups highlighted in the PCA analysis.
198 Reference samples were selected as having > 95 % pure background. Before performing the RFMix
199 analysis, the vcf files were phased with shapeit.v2.904 [56]. RFMIX was run using genetic maps
200 generated from the data of Liu et al. [37]. Briefly, reads from the project SRP043350 were retrieved
201 from Short Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra), aligned to the reference genome
202 for SNP detection and recombinants were detected with the custom script find_crossing_overs.py to
203 produce a genetic map. Further details on the RFMix analysis are given in supplementary file
204 SeqApiPop_6_RFMix.pdf.
205
206 **Results**
207 *Sequencing and genotyping*
208 Sequencing of the honey bee drones for the SeqApiPop diversity project began in 2014 on Illumina
209 HiSeq instruments and some of the first samples had such low coverage that a second run (or even
210 three in the case of OUE8) sequencing was performed. For these samples, the resulting BAM files
211 were merged prior to variant calling. Only four samples of the diversity project were sequenced on
212 Novaseq instruments, for which higher sequencing depths were achieved. Therefore, to improve the
213 robustness of the SNP detection pipeline, we decided to add drone genome sequences from other
214 ongoing subsequent projects using the same genetic types, that were produced with Novaseq
215 instruments. Samples sequenced with the HiSeq and NovaSeq instruments have sequencing depths of
216 12.5 ± 6.1 and 33.5 ± 10.2 respectively (Supplementary Table ST1, Supplementary figures 3 and 4).
217 Genotyping the whole dataset of 870 drones with the GATK pipeline allowed the detection of
218 10,601,454 raw SNPs (supplementary figure 2). Results of the subsequent filtering steps are shown in
219 the Venn diagrams in supplementary figures 5, 6 and 7. A total of 7,023,976 high-quality SNPs
220 remained after filtering. The 198 samples from the other projects and 30 within-colony duplicate
221 samples from the present diversity project were removed from the dataset for downstream analyses.

222 Although a filter on genotyping rate ≥ 95% was applied in the primary filtering steps, the final filter on
223 heterozygote calls was set to keep SNPs with up to 1% of heterozygote samples, and these remaining
224 heterozygous genotypes were set to missing (supplementary figure 2). After this, a final filter on
225 missing data in samples was applied and 15 samples were removed due to the fraction of missing
226 genotypes exceeding 10 %. The ultimate diversity dataset comprised 629 drones (supplementary table
227 1) and 7,012,891 SNPs, and was used for all subsequent analyses unless stated otherwise.
228
229 *Contribution of SNPs to the variance in PCAs: detection of large haplotype blocks*
230 Principal component analysis was performed on the 629 samples and 7 million SNPs, retaining 20
231 eigenvectors, resulting in clear differentiation of three groups of samples. The first principal
232 component, representing 47.4 % of the total variance over the 20 dimensions estimated, broadly
233 differentiates M lineage bees, *A. m. mellifera* and *A. m. iberiensis*, from the *A. m. ligustica*, *A. m.*
234 *carnica* and *A. m. caucasia* bees. The second principal component, representing 13.8 % of the
235 variance, separates the O lineage *A. m. caucasia* bees from the C lineage *A. m. ligustica* and *A. m.*
236 *carnica* bees (supplementary figure 8). The remaining principal components each represent 5.2 % or
237 less of the variance (ean = 2.1 %). When looking at the individual contributions of SNPs to the
238 variance, we can see that only a very small proportion of the ~7 million markers contribute
239 significantly to PC1 (red lines on supplementary figure 9) and that this proportion is even much
240 smaller for PCs 2 and 3. Two reasons for such a limited contribution to the variance of the majority of
241 markers is the low informativity of markers of low minor allele frequency and the redundancy of
242 markers that are in strong linkage disequilibrium (LD). Therefore, to thin the dataset, we tested the
243 effect of several minor allele frequency (MAF) filters and chose the most pertinent one for subsequent
244 testing of various LD pruning values. The effects of these filters were estimated by inspecting the
245 contributions of the SNPs to the principal components. The MAF filters tested showed clearly that
246 datasets containing only SNPs with MAF > 0.01 or MAF > 0.05 are sufficient to allow a higher

247 proportion of markers contributing to the PCs, with a notable increase of SNPs contributing to PC2
248 and PC3 (supplementary figure 9). To avoid losing too many potential population-specific markers
249 present at low frequency in the data, we chose to use the lowest MAF threshold tested, leaving a
250 dataset of 3,285,296 SNPs having MAF > 0.01 for subsequent analyses. On inspecting the
251 contributions of individual SNPs to principal components along the genome, a striking feature we
252 observe is that for several large chromosomal regions, five of which being larger than 1 Mb, most
253 SNPs make a significant contribution, with the observed values being amongst the strongest observed
254 genome-wide (Supplementary Figures 10 and 11). Such observations suggest the existence of large
255 haplotype blocks driving differentiation along principal components, in particular principal component
256 1. To explore this further we compared these genomic regions to the haplotype blocks detected with
257 Plink (supplementary table 2) revealing significant overlap by visual inspection (Supplementary
258 Figure 12). The largest of these blocks spans 3.6 Mb on chromosome 11, which is close to 1.6 % of
259 the honey bee genome size, and four others on chromosomes 4, 7 and 9 are larger than 1 Mb (figure 1,
260 supplementary figures 10, 11, 12, supplementary table 2).
261
262 *LD filtering*
263 Population structure and admixture analysis rely on the assumption that markers along the genome are
264 independent Indeed, markers in strong LD such as those in haplotype blocks, have drastic effects by
265 shaping the genetic structure. Therefore, we sought to investigate the impact of LD pruning on
266 population structure inference. The number of SNPs used in a window for LD pruning was determined
267 such that most windows would correspond to a physical size of 100 kb. To achieve this, we used the
268 mode of the distribution of the number of SNPs in 100 kb bins, which is 1749 for the dataset of
269 3,285,296 SNPs with MAF > 0.01 (supplementary figures 13 and 14). LD pruning was thus performed
270 with a window size of 1749 SNPs and 175 bp (10 %) overlap and various values were tested, spanning
271 between $0.1 \leq$ LD $r^2 \leq 0.9$. PCA following these various thresholds show that with LD $r^2 < 0.3$ the

272 global structure of the dataset is lost, with only one population (*A. m. iberiensis*) contributing strongly
273 to the variance (supplementary figure 15), whereas with LD $r^2 > 0.3$, the contributions to the variance
274 in PC1 is not so widely distributed (supplementary figure 16). The effect of LD pruning on the
275 haplotype blocks is drastic, with the few SNPs retained having a distribution of their contributions to
276 the variance in PC1 and PC2 similar to that of the rest of the genome (supplementary figure 17). After
277 pruning for LD $r^2 < 0.3$, 601,945 SNPs were left in the dataset which were subsequently used in the
278 analysis of population structure.
279
280 *Analysis of population structure*
281 The PCA revealed distinct population structure within the data. For instance, some populations from
282 French breeder's organisations, such as the Royal Jelly breeder organisation (GPGR: Groupement des
283 Producteurs de Gelée Royale), or the Corsican breeder's organization (AOC Corse) appear quite
284 homogenous (figure 2), the first clustering close to the *A. m. ligustica* and *A. m. carnica* reference
285 populations and the latter appearing as a group between the C lineage *A. m. ligustica* and *A. m. carnica*
286 on one side and the M lineage *A. m. mellifera* and *A. m. iberiensis* on the other. Other populations
287 from French breeders appear much less homogenous, with individuals scattered across the whole
288 graph, see for examples Tarn 2 on figure 2, suggesting various degrees of admixture between the three
289 principal genetic groups (supplementary figure 18).
290 To further investigate the genetic structure and the effects of human-mediated breeding, we performed
291 admixture analyses. Our dataset consists of reference samples from thirteen origins, including two
292 islands, in addition to samples from several commercial breeders and conservatories. The genetic
293 makeup is therefore expected to be complex and the first task was to estimate the optimal number of
294 genetic backgrounds (K). We performed 50 independent runs with the Admixture software for each
295 value of $2 \leq K \leq 16$ on the LD-pruned dataset, totaling 750 independent analyses. Cross-validation
296 (CV) error estimates of the results computed by the software are shown in figure 3A. Results suggest

297 that the most likely number of genetic backgrounds is 8 or 9, with K = 8 having runs with the lowest
298 CV values overall, and K = 9 having the lowest median CV value over its 50 runs. The resulting Q
299 matrices were jointly analyzed using Pong [51], where for each value of K runs are grouped together
300 by similarity into modes and the mode containing the largest number of similar runs is defined as the
301 major mode. As Pong failed to find disjoint modes with the default similarity threshold of 0.97, we
302 increased the stringency of this value to 0.98 for our analyses. Naturally, for low values of K, such as
303 2 or 3, most of the Q matrices are very similar and the major modes contain most runs, if not all.
304 Typically, for K = 2, all 50 runs are in a single mode and for K = 3, the major mode contains 49 out of
305 all 50 runs and reflects the three main groups from the PCA analysis. Amongst the values of K having
306 the lowest CV values (figure 3A), K = 9 stands out as having a major mode containing 33 runs out of
307 50. While K = 8 had the lowest overall mean CV value, its major mode contains only 12 runs,
308 indicating K = 9 to be the better model (figure 3B and supplementary table 3). Interestingly, the
309 pattern observed when considering only K = 3 genetic backgrounds, recapitulates the general pattern
310 observed in the PCAs, in which the reference populations separate into three groups. These groups
311 reflect the main evolutionary lineages present in the dataset, being the M lineage (*A. m. mellifera* and
312 *A. m. iberiensis*), C lineage (*A. m. ligustica*, *A. m. carnica*), and O lineage (*A. m. caucasia*). For K = 2,
313 these *A. m. caucasia* bees are considered as having the same genetic background as the *A. m. ligustica*
314 and *A. m. carnica* samples, also reflecting the results from the PCA (figure 2, supplementary figure
315 19). Our results support the assumption that *A. m. caucasia* bees are assigned to the O lineage by
316 morphometry 20/09/2021 16:51:00 and to the C lineage by mtDNA [57]. Some admixture can be
317 observed for a small proportion of the reference samples. For instance, the reference samples from the
318 Savoy conservatory appear to have a small proportion of genetic background from *A. m. ligustica*
319 and/or *A. m. carnica,* which is consistent with the PCA results (figure 2). Likewise, the *A. m. carnica*
320 samples from Poland have a small proportion of genetic background from *A. m. caucasia*. Finally, the
321 *A. m. carnica* from Switzerland show some proportion of *A. m. mellifera* genetic background.

322 When examining the admixture pattern representing the 33 runs at K = 9 genetic backgrounds, the
323 three main groups are now separated. The M lineage group from the K = 3 backgrounds is now
324 composed of four genetic backgrounds: *A. m. iberiensis* is now separated from *A. m. mellifera* and the
325 *A. m. mellifera* bees are separated in three groups from mainland France, and the two islands of
326 Ouessant and Colonsay. The other three subspecies *A. m. ligustica*, *A. m. carnica* and *A. m. caucasia*
327 each have their own genetic background. An eighth background corresponds to the samples from the
328 bees selected for the production of royal jelly and a ninth appears in the two populations that were
329 noted as Buckfast bees. Although it is a major background in these two populations, a majority of
330 samples have also a large proportion of *A. m. carnica* and, to a lesser extent, of *A. m. ligustica*
331 backgrounds. This ninth background can also be found in other breeder's populations principally in
332 Hérault and Tarn1 (figure 3C). Apart from the royal jelly population, all honey bees from breeders
333 show high levels of admixture. Moreover, there is a great variability in the genetic origins and
334 proportions of backgrounds, even for samples coming from a same location (figure 4). The exception
335 is the population from Corsica, for which all samples show proportions close to 75% - 25% of *A. m.*
336 *mellifera* and *A. m. ligustica* backgrounds respectively.
337
338 *Migrations between populations*
339 Due to the commercial interest expressed by beekeepers for the Buckfast bees and the peculiar genetic
340 composition observed in the Corsican population, we performed a population migration analysis with
341 TreeMix [52]. All samples having more than 80% ancestry from one of the 9 backgrounds detected in
342 the Admixture analysis were selected from one of the K = 9 major mode Q-matrices (supplementary
343 table 4), and the list supplemented with the 43 Corsican samples, making our data set composed of ten
344 representative groups for the European populations.
345 Estimations on the number of migrations (m) between the populations in the dataset, based on the
346 Evanno method [53], return a mode of m = 1, strongly suggesting a single migration, and a relatively

347 high Δm value for m = 2 supports the existence of a second migration. The Δm values for 3 or more
348 migrations are close to zero, suggesting that more than 2 migrations between populations in the dataset
349 is unlikely (figure 5A). For m = 1 the 100 TreeMix runs indicated a migration from *A. m. ligustica* to
350 the Corsican population. For m = 2 the 100 TreeMix runs show the two migrations as being from *A. m.*
351 *ligustica* to the Corsica population, and from *A. m. caucasia* to the Buckfast bees (figure 5B).
352 Summaries of the resulting trees with DendroPy [54] are shown in figure 5C, indicating that when the
353 two migrations are taken into account, the Corsican samples are grouped with the *A. m. mellifera* M
354 lineage bees, and the Buckfast bees group with the *A. m. ligustica* and *A. m. carnica* C lineage bees.
355
356 *Haplotype conservation in the admixed populations*
357 To investigate further the haplotype blocks detected, we performed a local ancestry inference on our
358 dataset with RFMix. Reference samples were selected as bees having > 95 % ancestry for a given
359 background following the Admixture analysis at K = 3 (figure 3C), resulting in 131 samples for group
360 1, 148 for group 2 and 17 for group 3, while the remaining 333 samples formed the query dataset. To
361 perform the local ancestry inference, we constructed a genetic map from cross overs identified in the
362 sequence data of 43 males from 3 colonies [37] aligned to the HAv3.1 reference genome. Results
363 indicate that few historical recombination events have occurred in the large haplotype blocks since the
364 admixture between the subspecies. The most notable example is that of the 3.6 Mb haplotype block
365 between positions 3.7 and 7.3 Mb on chromosome 11, in which almost all 333 samples from the query
366 dataset show one continuous stretch for one of the three backgrounds. Typically, only one of the 43
367 samples from Corsica presents two different ancestral haplotypes within this interval, with a switch
368 from a group 1 to a group 2 haplotype at position ~4.5 Mb on chromosome 11, within the 3.6 Mb
369 haplotype block, whereas numerous switches can be observed on the rest of the chromosome (figure
370 6A). When counting the haplotype switches detected in all 333 query samples, only 28 are situated
371 within the 3.6 Mb haplotype block on chromosome 11, whereas other regions of the chromosome can

372 have more than 50 switches per 100 kb (figure 6B and see supplementary figure 20 for the other
373 chromosomes). Interestingly, *LOC724287,* which is the largest gene described in the Gnomon
374 annotation set for the genome assembly HAv3.1, is found in this block at position 11:5,292,072-
375 6,161,805. This gene is 869,734 bp long and encodes protein rhomboid transcript variant X2, its large
376 size being largely due to intron 4, which is 596,047 bp long. However, on investigating a possible
377 relationship between haplotype block and gene sizes in the honey bee genome no obvious association
378 could be found (data not shown).
379
380 **Discussion**
381 *SNP detection in a haploid dataset*
382 Our complete dataset of haploid drones is composed of 870 samples sequenced using Illumina's HiSeq
383 and NovaSeq technologies. Results clearly show that although a few of the early sequences produced
384 on the HiSeq are of lower depth, only 15 samples were eliminated due to the fraction of missing
385 genotypes exceeding 10%. By contrast, the fraction of missing genotypes over the ~7 million SNPs
386 detected was considerably lower in samples sequenced on the NovaSeq sequencing platform. Having
387 sequenced haploids, the removal of heterozygote SNPs in individual samples is recommended to
388 reduce the likelihood of "pseudo SNPs", as we have shown previously that heterozygote SNPs tend to
389 cluster together [31] and co-locate with repetitive elements (data not shown). This set of ~7 million
390 markers can now be used as a basis for the realization of high-density SNP chips, allowing selections
391 of markers according to optimized spacing and to defined MAF values in the main subspecies of
392 interest. Indeed, an important technical issue in SNP chip design, is that very high SNP densities, such
393 as found in the honey bee, can potentially cause allele dropout when genotyping, due to interference in
394 the probe designs. Deep knowledge of SNP and indel positions will help select candidates flanked by
395 monomorphic sequences. Conversely, for lower density chips, the spacing of markers can be
396 optimized by taking the haplotype structure into account, thus avoiding redundancy while maintaining

397 the highest possible level of genetic information. Another advantage of sequencing haploid samples, is
398 that the whole dataset represents phased chromosomes. Notably, the present dataset will be invaluable
399 for genotype imputation in future studies using lower density genotyping, such as DNA chips or low-
400 pass sequencing [58–60].
401
402 *Population structure in managed honey bees*
403 The deep understanding of European honey bee populations and of their recent admixture via imports
404 of genetic stocks by breeders is not a simple task. The analyses of admixture events in complex
405 population structures can be sensitive to a number of parameters and sometimes yields misleading
406 results, especially if one or several populations went through a recent bottleneck [61]. PCA on all ~7
407 million markers indicate that our dataset is structured into three main genetic types (supplementary
408 figure 8). The first principal component, representing 47.4 % of the variance, separates two major
409 groups corresponding respectively to subspecies from north-western (M lineage) and south-eastern
410 Europe (C lineage). These two groups are represented by several populations, including the Savoy and
411 Porquerolles conservatories from South-East France on one side, and bees that are not so far
412 geographically from Italy or Slovenia on the other. This large genetic distance despite relatively close
413 geographic proximity of the populations supports the hypothesis of the colonization of Europe by
414 honey bees via distinct western and eastern routes [1,24,62,63], and the separation between subspecies
415 due to the Alps forming a natural barrier preventing genetic exchange [5]. Along the second principal
416 component, representing 13.8 % of the variance, the population originating from *A. m. caucasia*
417 separates from the south-eastern European populations (supplementary figure 8). Prior to investigating
418 admixture we pruned SNPs in LD taking care to maximize the removal of redundancy while
419 maintaining the general structure of the data (figure 2 and supplementary figures 15, 16, 17).
420 We explored a range of K number of genetic backgrounds, running multiple iterations of each, to
421 determine the most likely admixture pattern (figure 3). Our results indicate that this approach is

422 necessary to ensure the results from each K model are stable prior to interpretation. We observe from
423 our Admixture analyses that CV outliers within a K model are common. For instance, at K = 8, the
424 mode with the lowest CV is only represented by 8 out of 50 Admixture runs, whereas the major mode
425 has 12 runs. On examining the admixture patterns from these two modes, the major mode suggests the
426 *A. m. mellifera* bees from conservatories on mainland France to be hybrids between bees from
427 Ouessant and Spain, with roughly 50% of each genetic background moreover on the same mode, the
428 *A. m. iberiensis* background represents also 50% of the M lineage background in the bees from
429 Corsica (supplementary figure 19). This is extremely unlikely given the geography of Western Europe
430 and our knowledge of the history of the bees of Ouessant. Indeed, Ouessant is a very small island
431 (15.6 km^2) off the coast of western Brittany, isolated from the rest of the French honey bee population
432 since its installation in 1987 and the prohibition of imports since 1991 mostly for sanitary reasons.
433 Contrarywise the mode with the 8 runs and lowest CV presents a better separation of *A. m. mellifera*
434 and *A. m. iberiensis,* which is also found in the major mode at K = 9 backgrounds. A smaller level of
435 admixture can still be found between *A. m. mellifera* and *A. m. iberiensis,* that is quite likely due to the
436 shared ancestry between these two subspecies.
437 The major mode at K = 9 is represented by 33 out of 50 runs and returned the lowest mean CV value.
438 This mode identifies mainland France *A. m. mellifera* samples as having a distinct genetic background
439 and suggests that honey bees from Ouessant may have been re-introduced in the mainland
440 conservatories. This mode also identifies a distinct genetic background in French and Swiss Buckfast
441 bees. Buckfast bees were developed by Brother Adam, and are described on page 14 of "Beekeeping
442 at the Buckfast Abbey" as a cross performed around 1915 between "the leather-coloured Italian bee
443 and the old native English variety" [16]. Brother Adam also notes that the Italian bees that were
444 imported in later years were distinct from the ones used in the development of the Buckfast strain. Our
445 analysis of migrations between populations with TreeMix suggests that the Buckfast in our dataset
446 were subject to introgression with genetic material from *A. m. caucasia* (figure 5B), although the

447 timing of this potential admixture event could not be determined. When the two migrations of *A. m.*
448 *ligustica* into Corsica and *A. m. caucasia* into the Buckfast are considered, which is a likely scenario
449 suggested by the Evanno analysis, the latter is close to *A. m. carnica*, as seen in figures 5B and 5C.
450 Interestingly, a whole genome sequence study of Italian honey bees, also suggest that the Buckfast
451 bees are closer to *A. m. carnica*, than to *A. m. ligustica* [64] and no proximity of the Buckfast bees
452 with M lineage bees were found neither in this study nor in ours, despite the cross at the origin of the
453 Buckfast including an old native variety. Further investigations including more Buckfast samples and
454 additional honey bee subspecies will be needed to fully elucidate this question. The *A. m. carnica*
455 samples from Slovenia, Germany, France, Switzerland and Poland all share the same genetic
456 background, reflecting their identical origin, probably recent imports.
457 The population of bees from Corsica has the distinct characteristic of being homogenous in
458 composition, despite being admixed, with all samples showing mean proportions of 75% and 25% of
459 *A. m. mellifera* and *A. m. ligustica* backgrounds, respectively (figures 2 and 3). The introgression of
460 Italian bees is confirmed by the TreeMix migration analysis and when this is accounted for, the
461 Corsican samples group with *A. m. mellifera* bees from mainland France instead of being situated
462 between the two main genetic subgroups of western and eastern European bees (figures 2B and 2C).
463 This result likely reflects the fact that Italian bees may have been imported on the Island until the
464 1980's, following which the import of foreign genetic material was prohibited. As beekeepers
465 generally prefer the *A. m. ligustica* Italian bees over *A. m. mellifera*, it is very likely that the latter is
466 the original population, as also suggested by Ruttner [1]. Although the hypothesis of the separation of
467 the two subspecies on the mainland by the Alps seems appropriate [5], the situation of the
468 Mediterranean islands in the region is not so clear. Based on physical geography alone, Corsica being
469 at a closer distance to Italy than to France, the chances would have been greater to have originally M
470 lineage rather than C lineage Italian bees. Moreover, Corsica was under the control of Pisa, then fell to
471 Genoa in 1284 and was only purchased by France in 1768. Further studies including samples from

472 Sardinia would certainly help defining the Mediterranean boundaries between the M lineage and C
473 lineage honey bees and confirm observations based on morphology [1].
474 Apart from the subspecies references and the royal jelly populations, the honey bees provided by
475 breeders are largely admixed, exhibiting high variability in background proportions - even for samples
476 sourced from the same region. A typical example is that of the Tarn1 and Tarn2 populations, revealing
477 that two breeders situated very close to one another (less than 100 km), have very different genetic
478 management strategies. Tarn1 samples are mainly composed of Buckfast and *A. m. carnica*
479 backgrounds, whereas in Tarn2 a large proportion of *A. m. mellifera* background is also present and
480 the population is far less homogenous (figures 2, 3 and 4). This shows the great heterogeneity of the
481 managed populations found in France and a question that needs further investigation is the influence
482 of the mating strategies used by the breeders, such as artificial insemination, mating stations, with
483 drone producing hives to saturate the environment with the desired genetic strains, or open mating.
484 These strategies influence variable levels of control on the genetic makeup of a breeder's stock. The
485 higher proportion of *A.m. mellifera* background in the Tarn2 population can either be deliberate or due
486 to a lower level of control over the mating of the queens, with a proportion of queens mating to *A. m.*
487 *mellifera* drones from the environment.
488 The Royal Jelly population is the inverse: beekeepers from all over France exchange their genetic
489 stock within a selection programme and controlled mating. As a result, a specific background with
490 individuals presenting very little admixture, is found for this population at very distant locations. Most
491 of the worldwide production of Royal Jelly comes from China, where high Royal Jelly-producing
492 lineages of honey bees were developed from an imported *A. m. ligustica* lineage [65]. Interestingly, in
493 our dataset, only three *A. m. ligustica* and all of the bees from China have some Royal Jelly genetic
494 background.
495
496 *Large haplotype blocks in the honey bee genome, specific to the M and C lineages*

497 When investigating the contribution of SNPs to variance in the PCA, we noted several large genomic
498 regions, up to 3.5 Mb long, in which almost all markers contributed very strongly to the first principal
499 component, separating bees from north-western (M lineage) and south-eastern Europe (C lineage).
500 These regions were noted to coincide with haplotype blocks detected with Plink. To investigate the
501 matter further, we performed local ancestry inference in the admixed samples with RFMix, using
502 samples exhibiting 95% ancestry for the three main genetic backgrounds as references. A low
503 recombination rate is confirmed by the observation of only very few switches between the three main
504 genetic backgrounds within these haplotype blocks. Interestingly, some of our regions, including the
505 largest one detected on chromosome 11, coincide with regions of low recombination rate detected in
506 other studies. These include a LD map produced with 30 diploid sequences from African worker bees
507 [39], ancestry inference in an admixed population [35], low resolution genetic maps produced by Rad
508 or ddRAD sequencing, with microsatellite or SNP markers, ddRAD sequencing [66,67] or higher
509 resolution genetic maps produced by whole genome sequencing of European [37] and African
510 subspecies [38].
511 Most of these regions coincide with the position of the centromeres such as described in the reference
512 genome assembly, which is primarily based on the combination of the location of *Ava*I repeats, that
513 were previously assigned to centromeres by cytogenetic analysis, and of a low GC content [33,68].
514 However, the *Ava*I repeats only represent a very small fraction of the centromeric regions described,
515 with the largest one only covering 14 kb [33], whereas the estimation of the extent of the centromeres,
516 based on a GC content lower than the genome average is much larger although imprecise and supposes
517 a similar organization as for the AT-rich alpha-satellite repeats in vertebrates, such as human [69].
518 Whereas in some cases the boundaries of our regions of low recombination rate coincide with the
519 actual positioning of the centromere on the genome assembly [33], such as in chromosomes 5 or 8, in
520 other instances, such as in chromosome 12, the region defined is much narrower. Due to the
521 difficulties in interpreting banding patterns in honey bee chromosomes, the position of the centromeres

522 is not well settled. Some evidence based on G- and C- banding suggests there are four metacentric and
523 12 submetacentric or subtelocentric chromosomes [70], whereas other evidence based on the
524 fluorescent *in situ* hybridization of a centromere probe suggests there are two metacentric, four
525 submetacentric, two subtelocentric and eight telocentric chromosomes [71]. Our evidence suggests at
526 least six chromosomes that could be telocentric or acrocentric: chromosomes 3, 5, 6, 9, 14 and 15.
527 Some of the haplotype blocks/regions of low recombination can seem very large, such as representing
528 up to 21 % in the case of chromosome 11 (figure 6). This may seem a lot, but recent findings in a
529 complete sequencing of the human genome give a similar proportion for chromosomes 9, in which 40
530 Mb of satellite arrays represent 20 % of the chromosome [72]. One important difference, however, is
531 that the block on honey bee chromosome 11 contains some genes, except in the central region,
532 whereas the satellite array described on human chromosome 9 does not. This suggests that the exact
533 position of the centromere in honey bee chromosome 11 and other chromosomes still has to be
534 refined. The specific case of the acrocentric chromosomes in terms of gene content (supplementary
535 figure 20) seems to compare better to the situation described in human, as the sequencing of the p-arm
536 of the five human acrocentric chromosomes has allowed the discovery of novel genes within the
537 satellite repeat-containing regions [69].
538 Some haplotype blocks may have another origin than centromeric DNA. For instance, some may have
539 maintained genetic divergence by limiting recombination via the presence of structural variants such
540 as inversions. Indeed, two of the blocks described here, between positions 4.0 - 5.1 Mb and 5.8 - 6.9
541 Mb on chromosome 7, seem to coincide at least partially with two regions of haplotype divergence
542 possibly due to inversions, detected between positions 3.9 – 4.3 and 6.3 – 7.3 Mb on the same
543 chromosome, in a highland versus lowland study of East African bees [36]. The slight differences in
544 coordinates found between the two studies can be due to the fact that a different version of the HAv3
545 Hav3.0 had been used. If confirmed, this finding suggests that haplotype blocks differing between M
546 lineage and C lineage bees such as found here, can coincide with blocks found in other subspecies in

547 Africa. Another study identifying the thelytoky locus (*Th*) in the South African Cape honey bee *Apis*
548 *mellifera capensis* showed it was in a non-recombining region over 100 kb long on chromosome 1,
549 although long-read mapping failed to detect any inversion [73].
550 Given the current hypotheses on the colonization of Europe by honey bees via distinct western and
551 eastern routes [1,24,62,63], it is not surprising that the haplotype blocks described here, whether or not
552 representing centromeric regions, tend to separate mainly the M and C lineage bees. Further analyses
553 will be necessary to define the centromeric regions more precisely and study their implication,
554 together with the other haplotype blocks, in the sub species structure of the honey bee populations.
555

556 **Conclusion**

557 The sequencing of close to 900 haploid honey bee drones, was shown here to be an invaluable
558 approach for variant detection and for understanding the fine genetic makeup of a complex population
559 having gone through multiple events of admixture. In addition, the extent of regions of extremely low
560 recombination rate could be defined with a higher precision than previously. The dataset generated
561 here is a solid base for future research involving other honey bee populations and for any analyses
562 requiring data phasing or imputation.
563

564 **Figures**

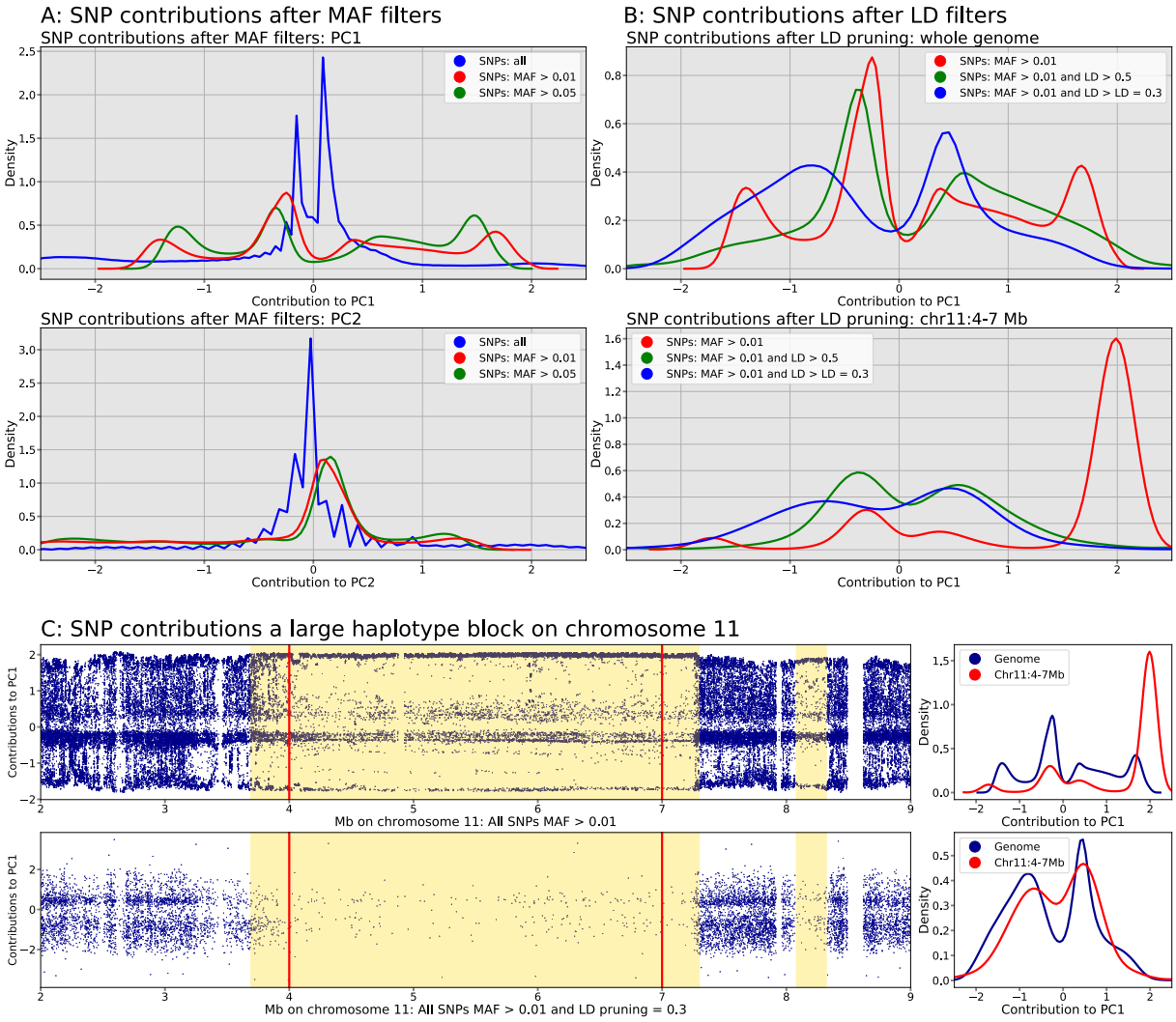
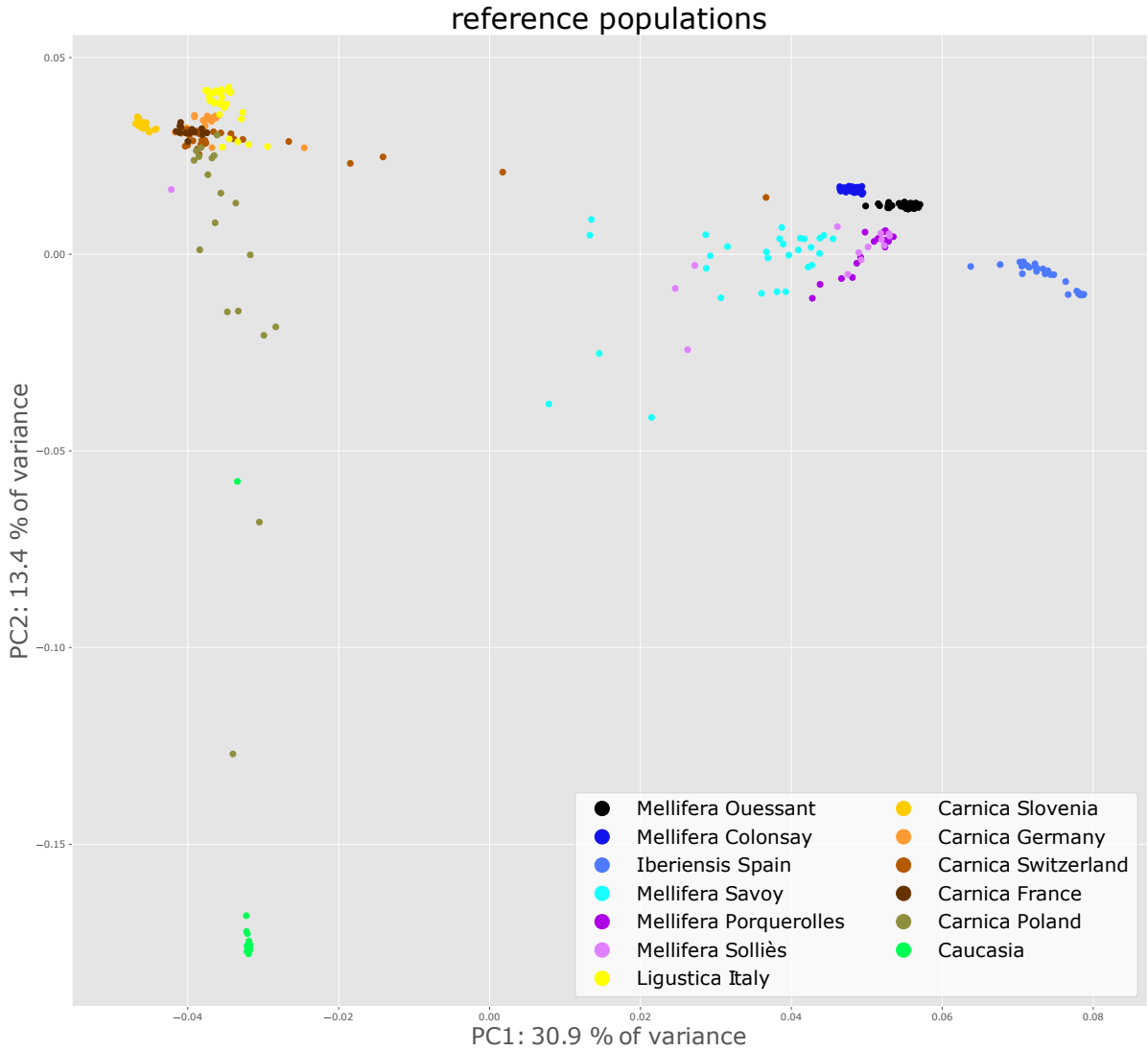
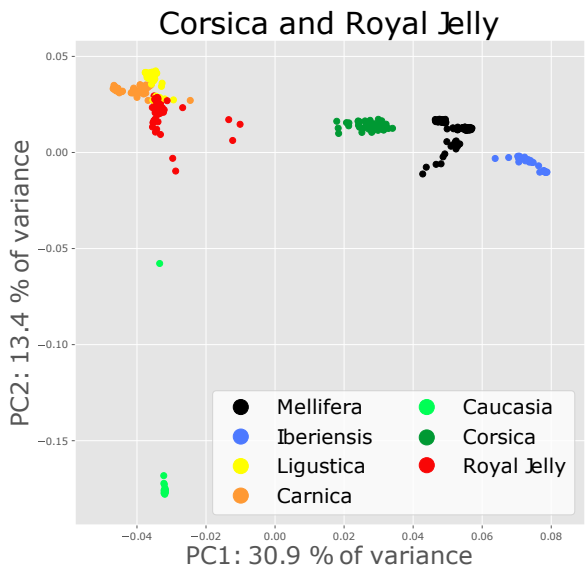
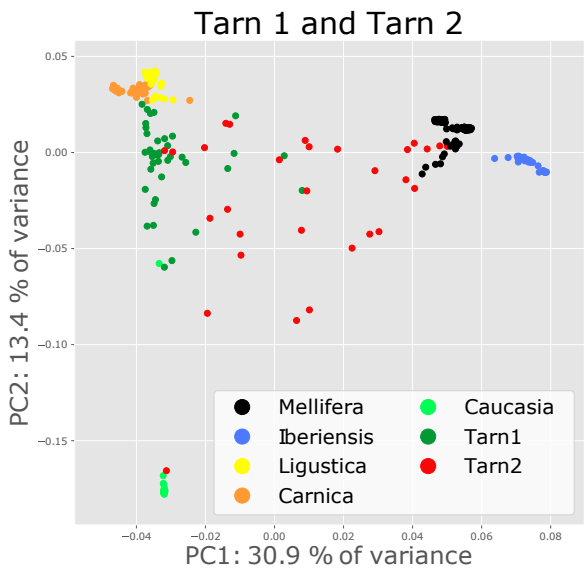


Figure 1: Contribution of SNPs to the principal components, MAF and LD filters and detection of large haplotype blocks. A: contribution to PC1 (top) and PC2 (bottom). When all 7 million SNPs are analysed simultaneously, the majority share a small contribution to PC1 and have no contribution to PC2 (blue). When retaining only markers with MAF > 0.01 or 0.05, (3,285,296 and 2,525418 SNPs; green and red lines, respectively), markers retained have a stronger contribution to PC1 and a higher proportion of markers contribute to PC2. B: LD pruning on the 3,285,296 SNPs with MAF > 0.01 (red line). Top: out of the 1,011,918 and 601,945 SNPs retained after pruning at LD = 0.5 (green line) or 0.3 (blue line), the distribution of the contribution of the markers is more even. Bottom: contribution of SNPs to PC1 in a 3 Mb region of chromosome 11. Almost all markers in this region show

575 contributions amongst the highest genome-wide (red line). The distribution of these contributions is
576 improved by LD pruning (green and blue lines). C: blue points show the contribution of individual
577 SNPs along a 6 Mb region of chromosome 11 containing two haplotype blocks of > 3 Mb and ~200 kb
578 (yellow backgrounds) before (top) and after (bottom) LD pruning. The LD pruning eliminates
579 successfully the markers in the haplotype blocks and the distribution of marker approaches that of the
580 rest of the genome, as shown in the corresponding density plots on the right.
581


reference populations
PC2: 13.4 % of variance
PC1: 30.9 % of variance
0.05
0.00
−0.05
−0.10
−0.15
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Mellifera Ouessant
Mellifera Colonsay
Iberiensis Spain
Mellifera Savoy
Mellifera Porquerolles
Mellifera Solliès
Ligustica Italy
Carnica Slovenia
Carnica Germany
Carnica Switzerland
Carnica France
Carnica Poland
Caucasia


Corsica and Royal Jelly
PC2: 13.4 % of variance
PC1: 30.9 % of variance
0.05
0.00
−0.05
−0.10
−0.15
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Mellifera
Iberiensis
Ligustica
Carnica
Caucasia
Corsica
Royal Jelly


Tarn 1 and Tarn 2
PC2: 13.4 % of variance
PC1: 30.9 % of variance
0.05
0.00
−0.05
−0.10
−0.15
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Mellifera
Iberiensis
Ligustica
Carnica
Caucasia
Tarn1
Tarn2

582

583 **Figure 2: PCA on the reference populations and on a sample of representative breeder**

584 **populations.** The 601,945 SNPs obtained after MAF filtering and LD pruning were used. Left:

585 reference populations only, with a colouring scheme according to their origin. Middle and left: only

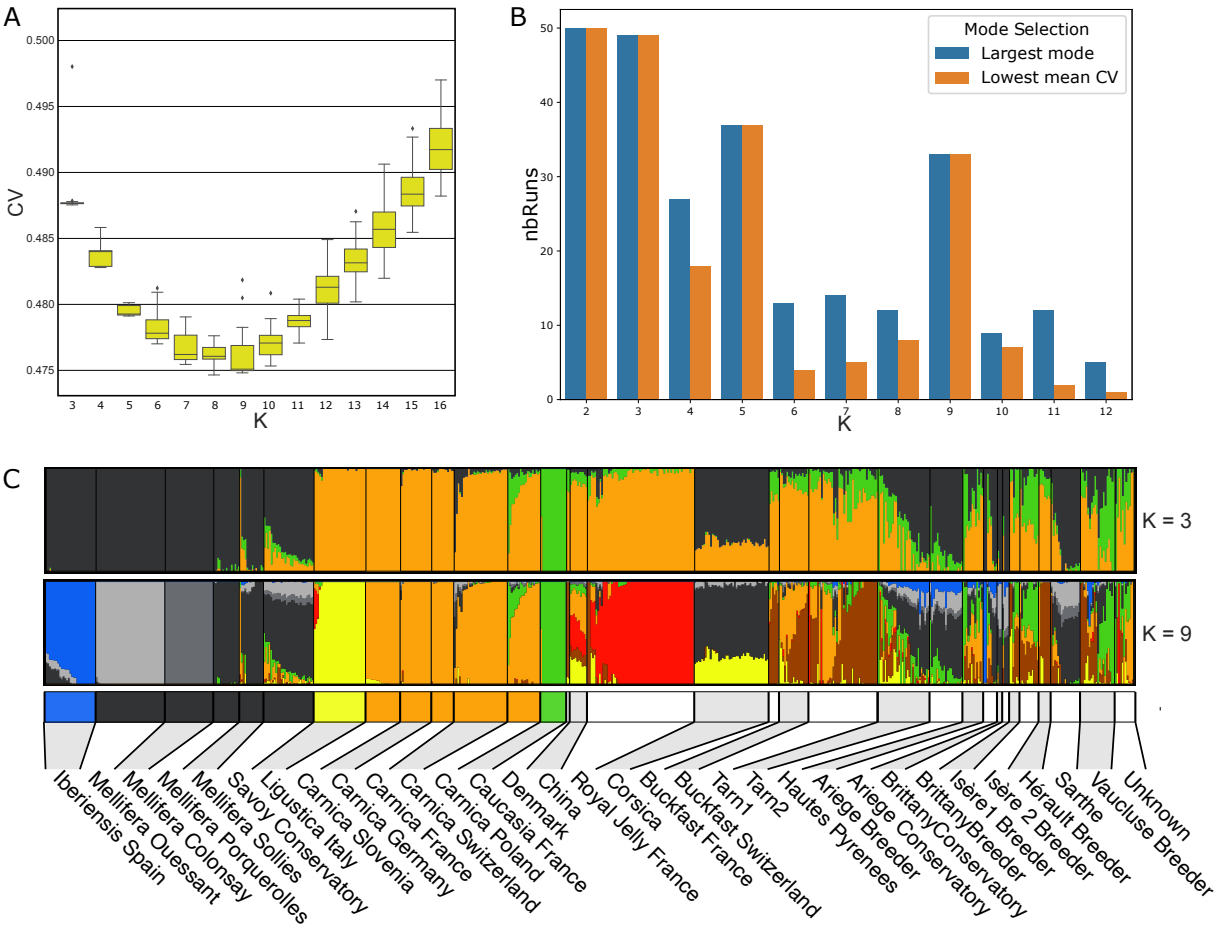
586 the reference populations with a high proportion of pure background individuals, as observed after

587 Admixture analysis, were kept and coloured according to the five subspecies. Some breeder

588 populations appear homogeneous, such as the honey bees selected for Royal Jelly or those from

589 Corsica. Others are heterogeneous, such as populations Tarn1 and Tarn2, from breeders.

590



591

592 **Figure 3: Admixture analysis.** A: estimation of Cross validation error for 50 runs of Admixture for 3

593 ≤ K ≤ 16. B: Major modes and modes with the lowest mean cross-validation (CV) error for Admixture

594 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 (see figure 12), K = 9 stands out as having a major mode containing 33 runs out of 50, 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, 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.

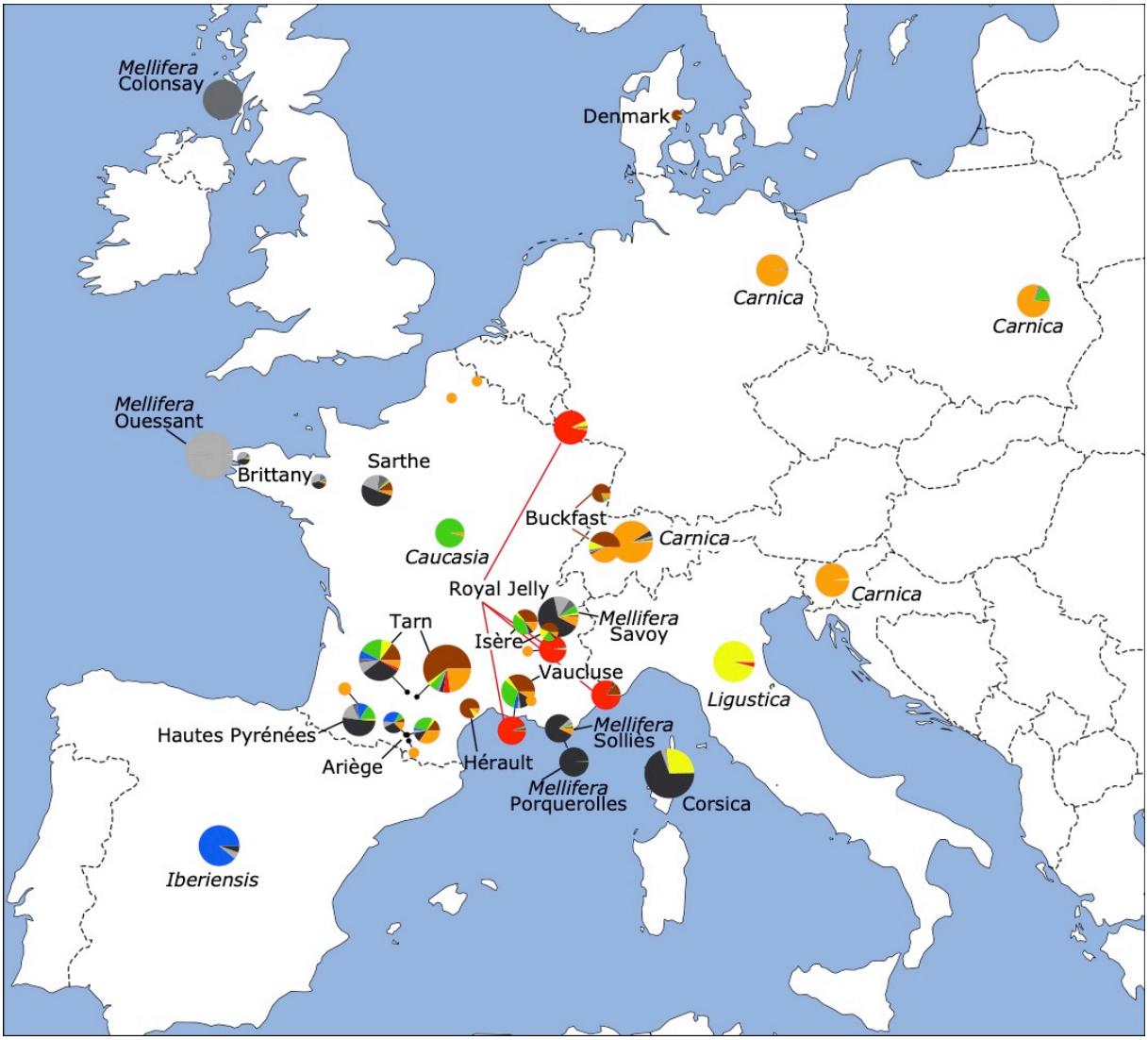


Figure 4: Admixture proportions and location of sample populations used in the diversity study. The size of the pie charts indicates the number of samples from a given location, with the number ranging from 2 samples (e.g. Denmark) to 43 samples (Corsica). Positions in France indicate the coordinates 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, so as the four breeders from the Royal Jelly breeder organisation (GPGR: Groupement des Producteurs de Gelée Royale) having provided samples.

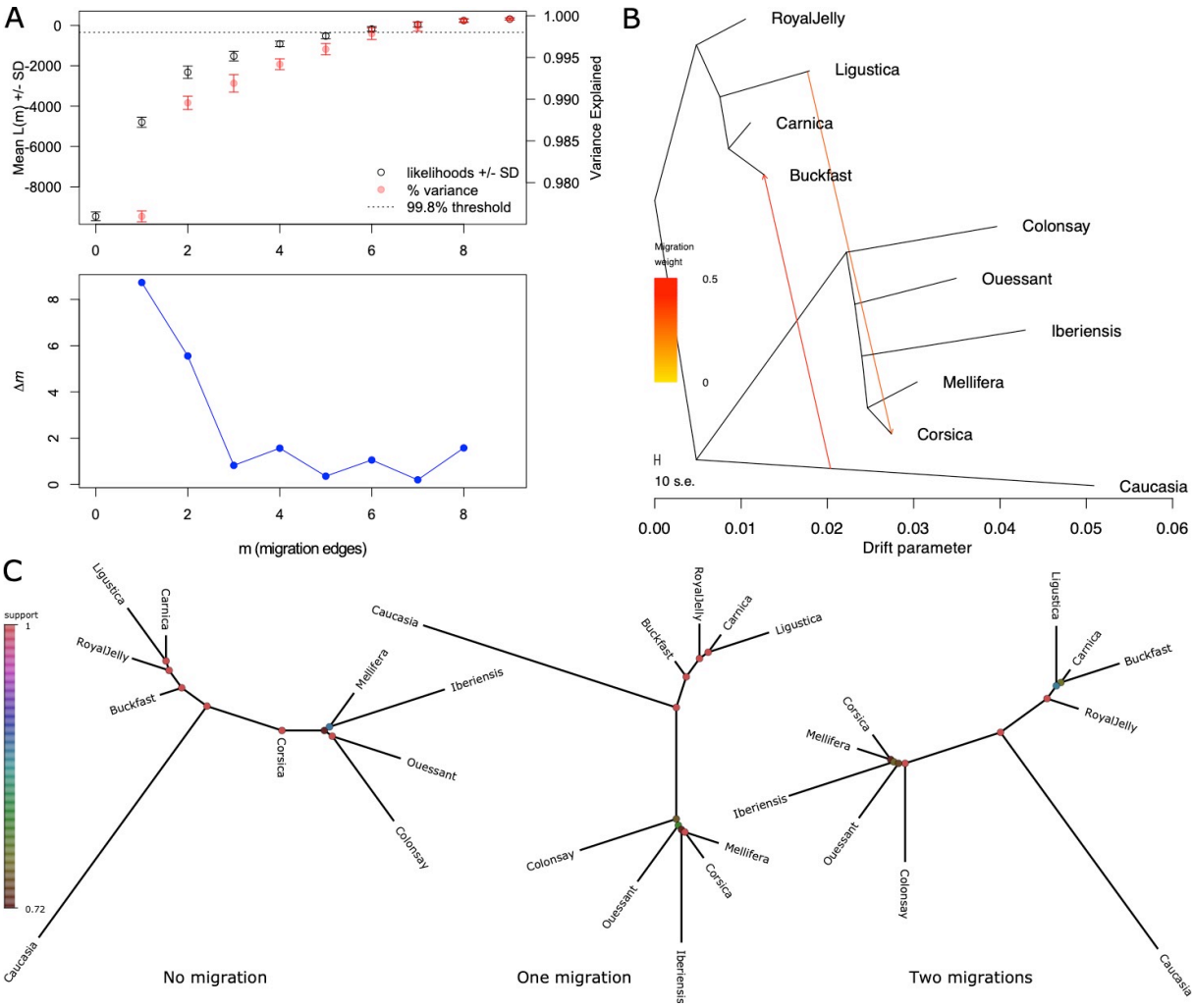


Figure 5: Analysis of migrations with Treemix. A: the OptM package was used to determine the optimal number of migrations between populations and backgrounds. The Δm values suggest one or two migrations. B: summaries of trees from TreeMix, estimated from 100 runs per migration with DendroPy. C: TreeMix graph selected amongst the 100 runs showing the two migrations identified.

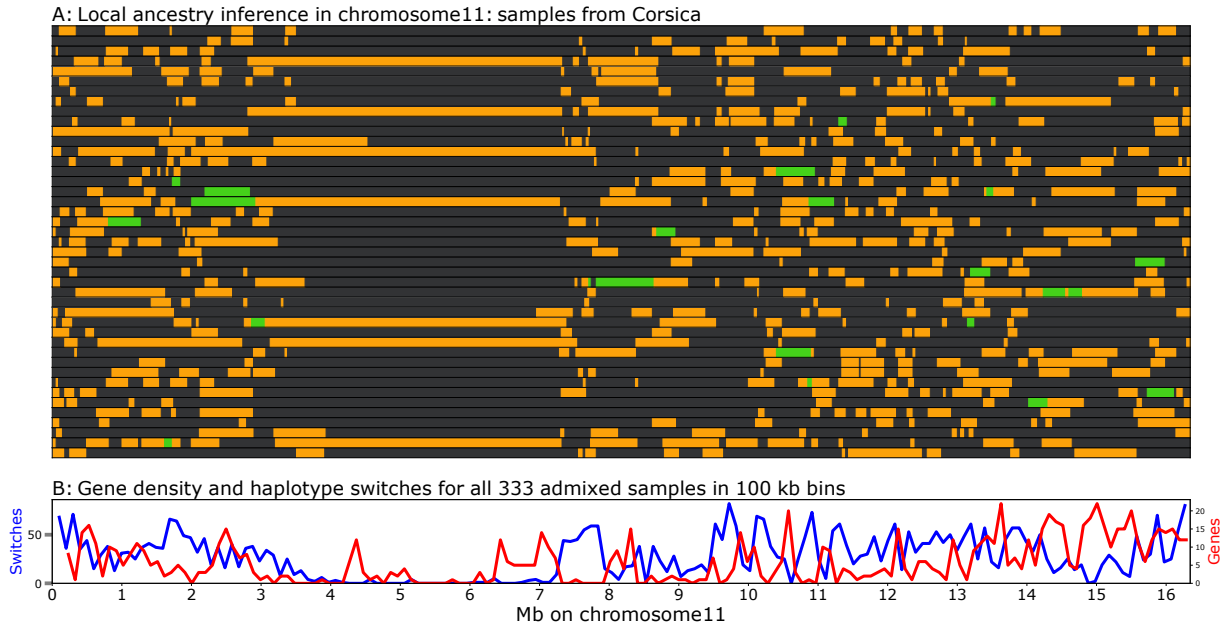


Figure 6: Local ancestry inference on chromosome 11 in the admixed samples from Corsica. A: each horizontal line represents the ancestry inference on one of the 43 individual samples from Corsica. Grey: *A. m. mellifera* and *A. m. iberiensis* backgrounds; yellow: *A. m. ligustica* and *A. m. carnica* 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.

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 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 to this study by providing samples from their colonies.

Data Accessibility

DNA Sequences 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 honet 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 supplementary table 1.

Authors' 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 co-wrote the manuscript.

References

1. Ruttner F. Biogeography and Taxonomy of Honeybees. Springer-Verlag; 1988.

2. Meixner MD, Pinto MA, Bouga M, Kryger P, Ivanova E, Fuchs S. Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. Journal of Apicultural Research. 2013;52:1–28.

3. Chen C, Liu Z, Pan Q, Chen X, Wang H, Guo H, et al. Genomic Analyses Reveal Demographic History and Temperate Adaptation of the Newly Discovered Honey Bee Subspecies *Apis mellifera sinisxinyuan* n. ssp. Mol Biol Evol. 2016;33:1337–48.

4. Ilyasov RA, Lee M, Takahashi J, Kwon HW, Nikolenko AG. A revision of subspecies structure of western honey bee Apis mellifera. Saudi Journal of Biological Sciences. 2020;27:3615–21.

5. Rinderer TE. Bee Genetics and Breeding. Academic press; 2013.

6. Cornuet JM, Fresnaye J, Lavie P, Blanc J, Hanout S, Mary-Lafargue C. Étude biométrique de deux populations d'abeilles cévenoles. Apidologie. 1978;9:41–55.

7. Cornuet J-M, Albisetti J, Mallet N, Fresnaye J. Étude biométrique d'une population d'abeilles landaises. Apidologie. 1982;13:3–13.

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

9. Cornuet JM, Fresnaye J, Blanc J, Paris R. Production de miel chez des hybrides interraciaux d'abeilles (apis mellifica l.) lors de générations successives de rétrocroisement sur la race locale. Apidologie. 1979;10:3–15.

10. Franck P, Garnery L, Celebrano G, Solignac M, Cornuet JM. Hybrid origins of honeybees from italy (Apis mellifera ligustica) and sicily (A. m. sicula). molecular ecology. 2000;9:907–21.

11. Carpenter MH, Harpur BA. Genetic past, present, and future of the honey bee (Apis mellifera) in the United States of America. Apidologie. 2021;52:63–79.

12. Puškadija Z, Kovačić M, Raguž N, Lukić B, Prešern J, Tofilski A. Morphological diversity of Carniolan honey bee (Apis mellifera carnica) in Croatia and Slovenia. Journal of Apicultural Research. Taylor & Francis; 2021;60:326–36.

13. Gregorc A, Lokar V. Selection criteria in an apiary of Carnolian honey bee (Apis mellifera carnica) colonies for queen rearing. Journal of Central European Agriculture. Faculty of Agriculture, University of Zagreb; 2010;11:401–8.

14. Gregorc Ă, Lokar V, Škerl MIS. Testing of the isolation of the Rog-Ponikve mating station for Carniolan (Apis mellifera carnica) honey bee queens. Journal of Apicultural Research. Taylor & Francis; 2008;47:137–40.

15. Moritz RFA. The limitations of biometric control on pure race breeding in Apis mellifera. Journal of Apicultural Research. 1991;30:54–9.

16. Brother Adam. Bee-Keeping at Buckfast Abbey. New edition. Hebden Bridge: Northern Bee Books; 1986.

17. Cornuet JM, Daoudi A, Chevalet C. Genetic pollution and number of matings in a black honey bee (Apis mellifera mellifera) population. Theor Appl Genet. 1986;73:223–7.

18. De la Rúa P, Jaffé R, Dall'Olio R, Muñoz I, Serrano J. Biodiversity, conservation and current threats to European honeybees. Apidologie. 2009;40:263–84.

19. Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, et al. 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. 2014;53:269–78.

20. Parejo M, Wragg D, Gauthier L, Vignal A, Neumann P, Neuditschko M. Using Whole-Genome Sequence Information to Foster Conservation Efforts for the European Dark Honey Bee, Apis mellifera mellifera. Frontiers in Ecology and Evolution. 2016;4:140.

21. Hassett J, Browne KA, McCormack GP, Moore E, Society NIHB, Soland G, et al. A significant

706 pure population of the dark European honey bee (*Apis mellifera mellifera*) remains in Ireland. Journal
707 of Apicultural Research. 2018;57:337–50.
708 22. Fontana P, Costa C, Prisco GD, Ruzzier E, Annoscia D, Battisti A, et al. Appeal for biodiversity
709 protection of native honey bee subspecies of Apis mellifera in Italy (San Michele all'Adige
710 declaration). Bulletin of Insectology. 2018;71:257–71.
711 23. Techer MA, Clémencet J, Turpin P, Volbert N, Reynaud B, Delatte H. Genetic characterization of
712 the honeybee (Apis mellifera) population of Rodrigues Island, based on microsatellite and
713 mitochondrial DNA. Apidologie. 2014;46:445–54.
714 24. Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS, et al. Thrice out
715 of Africa: ancient and recent expansions of the honey bee, Apis mellifera. Science. 2006;314:642–5.
716 25. Zayed A, Whitfield CW. A genome-wide signature of positive selection in ancient and recent
717 invasive expansions of the honey bee Apis mellifera. Proc Natl Acad Sci USA. 2008;105:3421–6.
718 26. Henriques D, Browne KA, Barnett MW, Parejo M, Kryger P, Freeman TC, et al. High sample
719 throughput genotyping for estimating C-lineage introgression in the dark honeybee: an accurate and
720 cost-effective SNP-based tool. Sci Rep. 2018;8:8552.
721 27. Parejo M, Henriques D, Pinto MA, Soland-Reckeweg G, Neuditschko M. Empirical comparison of
722 microsatellite and SNP markers to estimate introgression in Apis mellifera mellifera. Journal of
723 Apicultural Research. Taylor & Francis; 2018;57:504–6.
724 28. Harpur BA, Kent CF, Molodtsova D, Lebon JMD, Alqarni AS, Owayss AA, et al. Population
725 genomics of the honey bee reveals strong signatures of positive selection on worker traits. Proc Natl
726 Acad Sci USA. 2014;111:2614–9.
727 29. Henriques D, Wallberg A, Chávez-Galarza J, Johnston JS, Webster MT, Pinto MA. Whole
728 genome SNP-associated signatures of local adaptation in honeybees of the Iberian Peninsula. Sci Rep.
729 2018;8:11145.
730 30. Parejo M, Wragg D, Henriques D, Charrière J-D, Estonba A. Digging into the Genomic Past of

731 Swiss Honey Bees by Whole-Genome Sequencing Museum Specimens. Genome Biology and
732 Evolution. 2020;12:2535–51.
733 31. Wragg D, Marti-Marimon M, Basso B, Bidanel J-P, Labarthe E, Bouchez O, et al. Whole-genome
734 resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly.
735 Scientific Reports. Nature Publishing Group; 2016;6:27168.
736 32. Parejo M, Wragg D, Henriques D, Vignal A, Neuditschko M. Genome-wide scans between two
737 honeybee populations reveal putative signatures of human-mediated selection. Animal Genetics.
738 2017;48:704–7.
739 33. Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, et al. A hybrid de
740 novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC
741 Genomics. 2019;20:275.
742 34. Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, et al. A worldwide survey of
743 genome sequence variation provides insight into the evolutionary history of the honeybee Apis
744 mellifera. Nat Genet. 2014;46:1081–8.
745 35. Wragg D, Techer MA, Canale-Tabet K, Basso B, Bidanel J-P, Labarthe E, et al. Autosomal and
746 Mitochondrial Adaptation Following Admixture: A Case Study on the Honeybees of Reunion Island.
747 Genome Biol Evol. Oxford Academic; 2018;10:220–38.
748 36. Christmas MJ, Wallberg A, Bunikis I, Olsson A, Wallerman O, Webster MT. Chromosomal
749 inversions associated with environmental adaptation in honeybees. molecular ecology. 2018;215:403.
750 37. Liu H, Zhang X, Huang J, Chen J-Q, Tian D, Hurst LD, et al. Causes and consequences of
751 crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. Genome
752 Biol. 2015;16:15.
753 38. Kawakami T, Wallberg A, Olsson A, Wintermantel D, de Miranda JR, Allsopp M, et al.
754 Substantial Heritable Variation in Recombination Rate on Multiple Scales in Honeybees and
755 Bumblebees. Genetics. 2019;212:1101–19.

756 39. Wallberg A, Glémin S, Webster MT. Extreme Recombination Frequencies Shape Genome
757 Variation and Evolution in the Honeybee, Apis mellifera. PLoS Genet. 2015;11:e1005189.
758 40. Jones JC, Wallberg A, Christmas MJ, Kapheim KM, Webster MT. Extreme Differences in
759 Recombination Rate between the Genomes of a Solitary and a Social Bee. Molecular Biology and
760 Evolution. 2019;36:2277–91.
761 41. Bansal V, Bashir A, Bafna V. Evidence for large inversion polymorphisms in the human genome
762 from HapMap data. Genome Res. 2007;17:219–30.
763 42. Bovine HapMap Consortium, Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, et
764 al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science.
765 2009;324:528–32.
766 43. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated
767 map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
768 44. Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human
769 genomes. Nat Rev Genet. 2015;16:627–40.
770 45. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
771 arXiv:13033997 [q-bio] [Internet]. 2013 [cited 2020 Oct 19]; Available from:
772 http://arxiv.org/abs/1303.3997
773 46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
774 format and SAMtools. Bioinformatics. 2009;25:2078–9.
775 47. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
776 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
777 Genome Res. 2010;20:1297–303.
778 48. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes.
779 Bioinformatics. 2018;34:867–8.
780 49. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK:

781 rising to the challenge of larger and richer datasets. GigaSci. 2015;4:7.
782 50. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry
783 estimation. BMC Bioinformatics. 2011;12:246.
784 51. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of
785 latent clusters in population genetic data. Bioinformatics. 2016;32:2817–23.
786 52. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele
787 Frequency Data. PLOS Genetics. Public Library of Science; 2012;8:e1002967.
788 53. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software
789 structure: a simulation study. Molecular Ecology. 2005;14:2611–20.
790 54. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing.
791 Bioinformatics. 2010;26:1569–71.
792 55. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A Discriminative Modeling Approach
793 for Rapid and Robust Local-Ancestry Inference. The American Journal of Human Genetics.
794 2013;93:278–88.
795 56. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of
796 genomes. Nature Methods. Nature Publishing Group; 2012;9:179–81.
797 57. Ilyasov R, Nikolenko A, Tuktarov V, Goto K, Takahashi J-I, Kwon HW. Comparative analysis of
798 mitochondrial genomes of the honey bee subspecies A. m. caucasica and A. m. carpathica and
799 refinement of their evolutionary lineages. Journal of Apicultural Research. Taylor & Francis;
800 2019;58:567–79.
801 58. Gilly A, Southam L, Suveges D, Kuchenbaecker K, Moore R, Melloni GEM, et al. Very low-depth
802 whole-genome sequencing in complex trait association studies. Hancock J, editor. Bioinformatics.
803 2019;35:2555–61.
804 59. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and
805 decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res.

806 2021;31:529–37.
807 60. Wasik K, Berisa T, Pickrell JK, Li JH, Fraser DJ, King K, et al. Comparing low-pass sequencing
808 and genotyping for trait mapping in pharmacogenetics. BMC Genomics. 2021;22:197.
809 61. Lawson DJ, van Dorp L, Falush D. A tutorial on how not to over-interpret STRUCTURE and
810 ADMIXTURE bar plots. Nat Commun. 2018;9:3258.
811 62. Estoup A, Garnery L, Solignac M, Cornuet JM. Microsatellite variation in honey bee (Apis
812 mellifera L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise
813 mutation models. Genetics. 1995;140:679–95.
814 63. Han F, Wallberg A, Webster MT. From where did the Western honeybee (Apis mellifera)
815 originate? Ecol Evol. 2012;2:1949–57.
816 64. Minozzi G, Lazzari B, De Iorio MG, Costa C, Carpana E, Crepaldi P, et al. Whole-Genome
817 Sequence Analysis of Italian Honeybees (Apis mellifera). Animals. 2021;11:1311.
818 65. Cao L-F, Zheng H-Q, Pirk CWW, Hu F-L, Xu Z-W. High Royal Jelly-Producing Honeybees
819 (Apis mellifera ligustica) (Hymenoptera: Apidae) in China. J Econ Entomol. 2016;109:510–4.
820 66. Ross CR, DeFelice DS, Hunt GJ, Ihle KE, Amdam GV, Rueppell O. Genomic correlates of
821 recombination rate and its variability across eight recombination maps in the western honey bee (Apis
822 mellifera L.). BMC Genomics. 2015;16:107.
823 67. DeLory T, Funderburk K, Miller K, Zuluaga-Smith W, McPherson S, Pirk CW, et al. Local
824 variation in recombination rates of the honey bee (Apis mellifera) genome among samples from six
825 disparate populations. Insect Soc. 2020;67:127–38.
826 68. Beye M, Moritz RF. Characterization of honeybee (Apis mellifera L.) chromosomes using
827 repetitive DNA probes and fluorescence in situ hybridization. J Hered. 1995;86:145–50.
828 69. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, et al. Complete
829 genomic and epigenetic maps of human centromeres [Internet]. Genomics; 2021 Jul. Available from:
830 http://biorxiv.org/lookup/doi/10.1101/2021.07.12.452052

70. Hoshiba H. Karyotype and banding analyses on haploid males of the honey bee (apis mellifera). Proc Jpn Acad, Ser B. 1984;60:122–4.

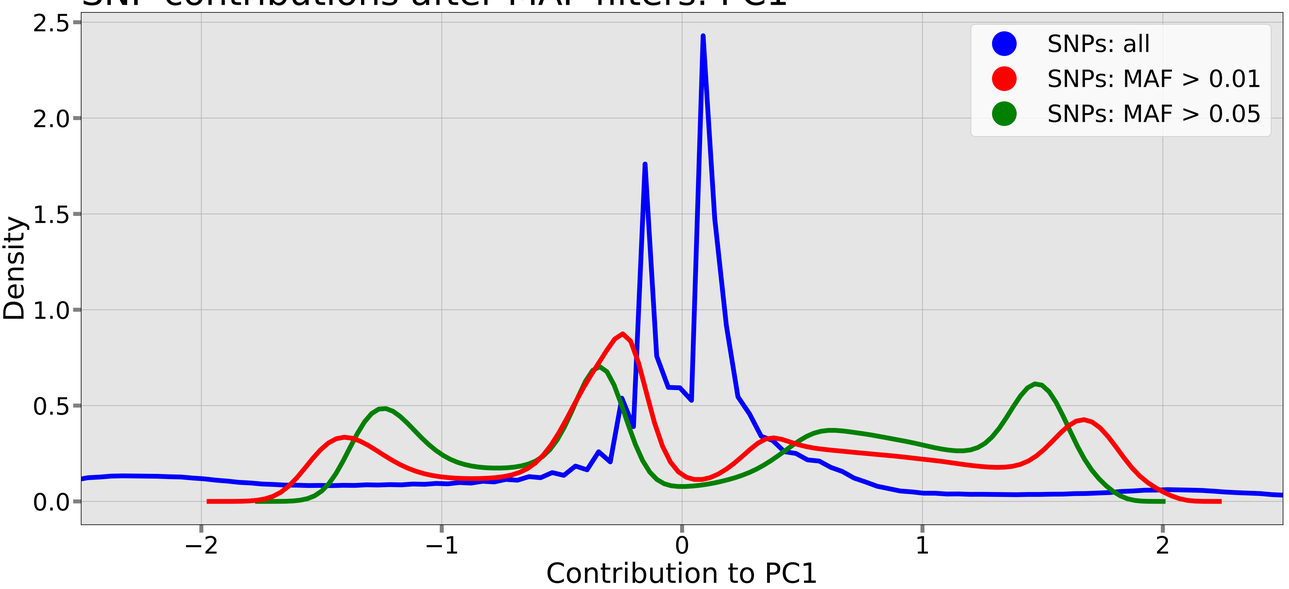
71. Beye M, Moritz RFA. A Centromere-Specific Probe for Fluorescence in-Situ Hybridization on Chromosomes of Apis-Mellifera L. Apidologie. 1994;25:322–6.

72. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome [Internet]. Genomics; 2021 May. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.05.26.445798

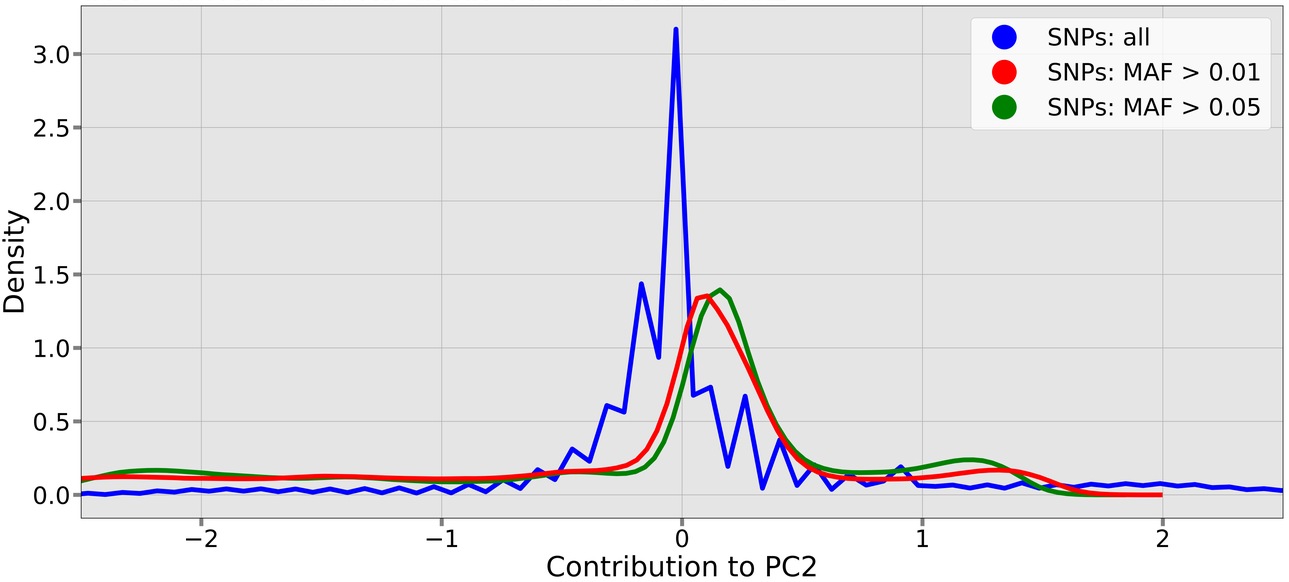
73. Aumer D, Stolle E, Allsopp M, Mumoki F, Pirk CWW, Moritz RFA. A Single SNP Turns a Social Honey Bee (*Apis mellifera*) Worker into a Selfish Parasite. True J, editor. Molecular Biology and Evolution. 2019;36:516–26.

A: SNP contributions after MAF filters

SNP contributions after MAF filters: PC1

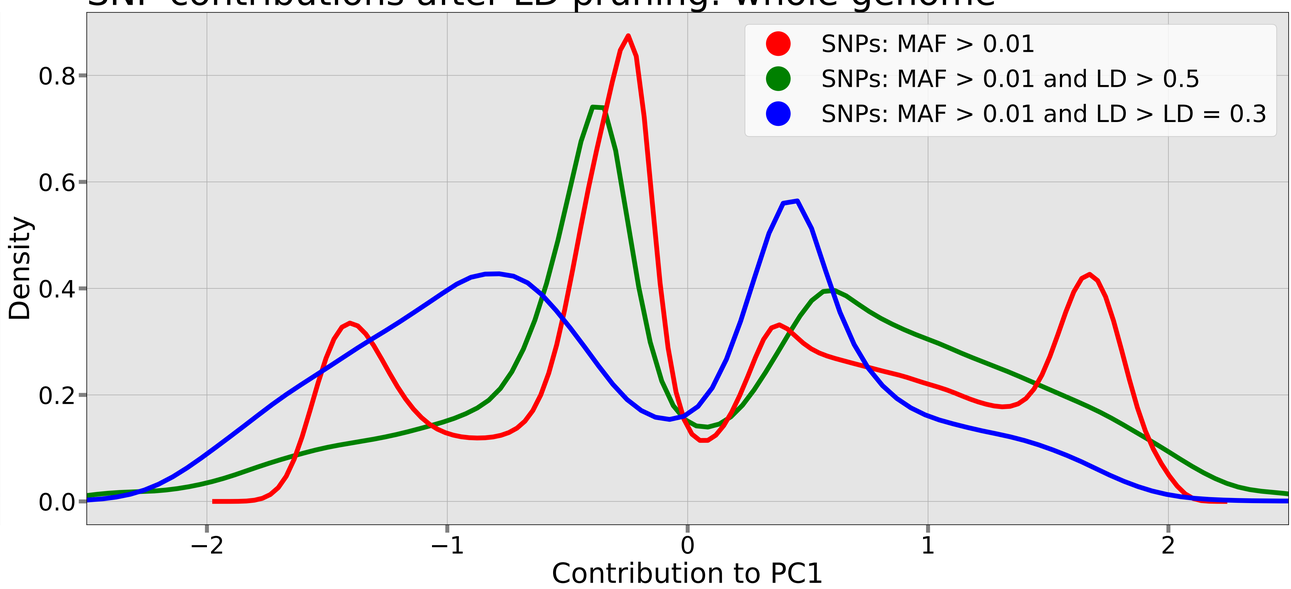


SNP contributions after MAF filters: PC2

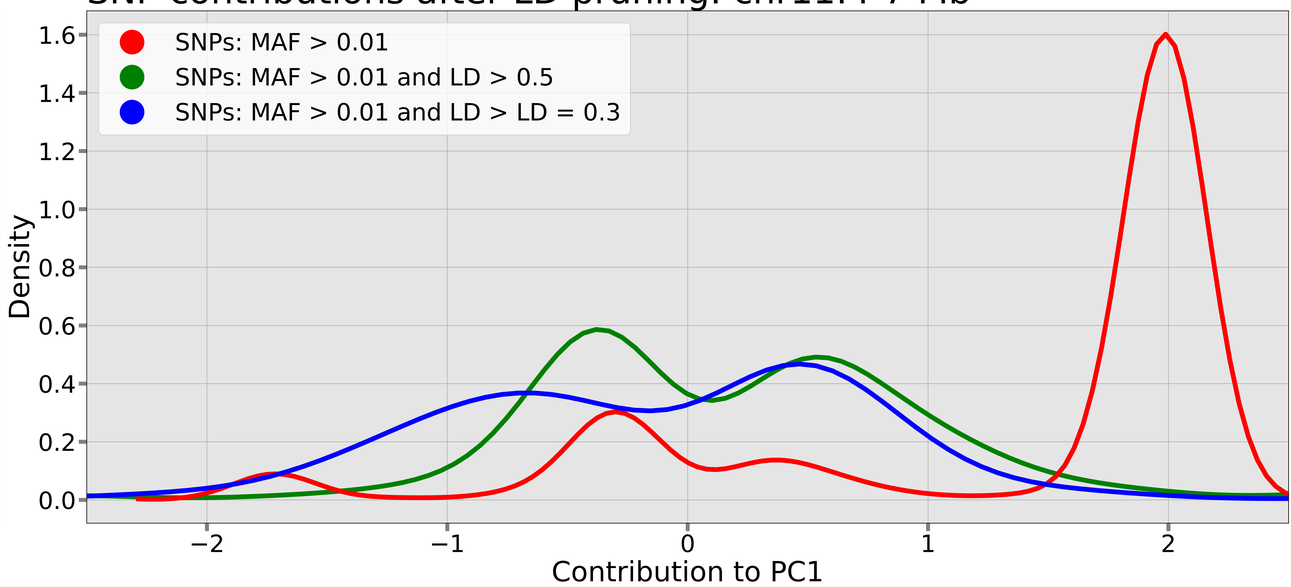


B: SNP contributions after LD filters

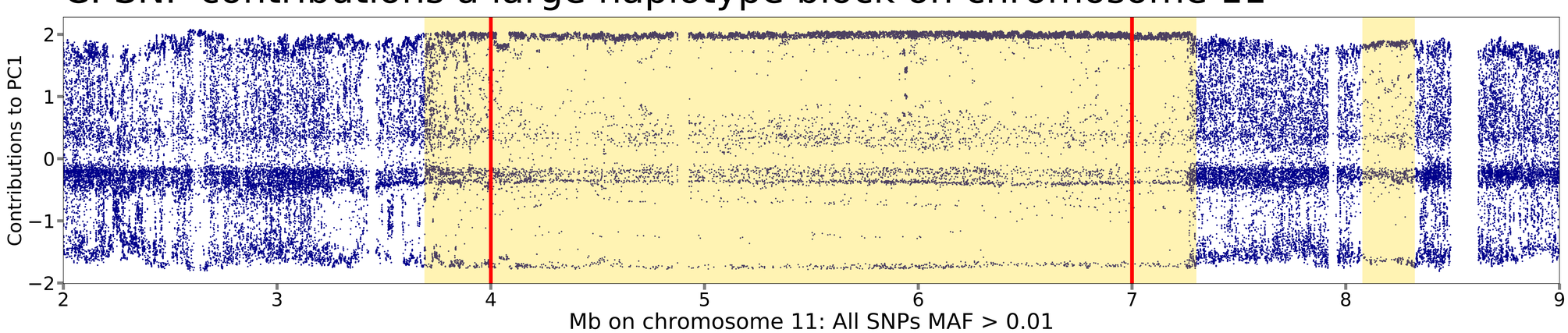
SNP contributions after LD pruning: whole genome

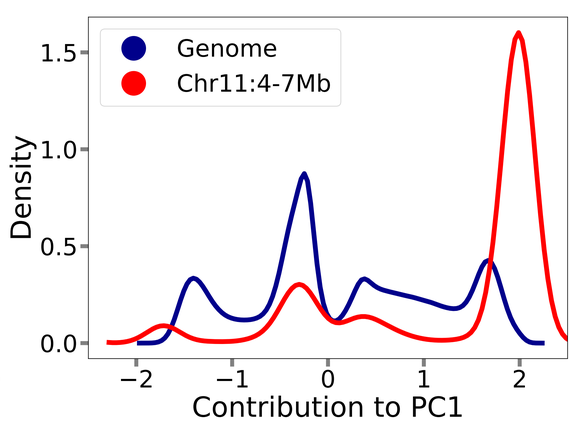


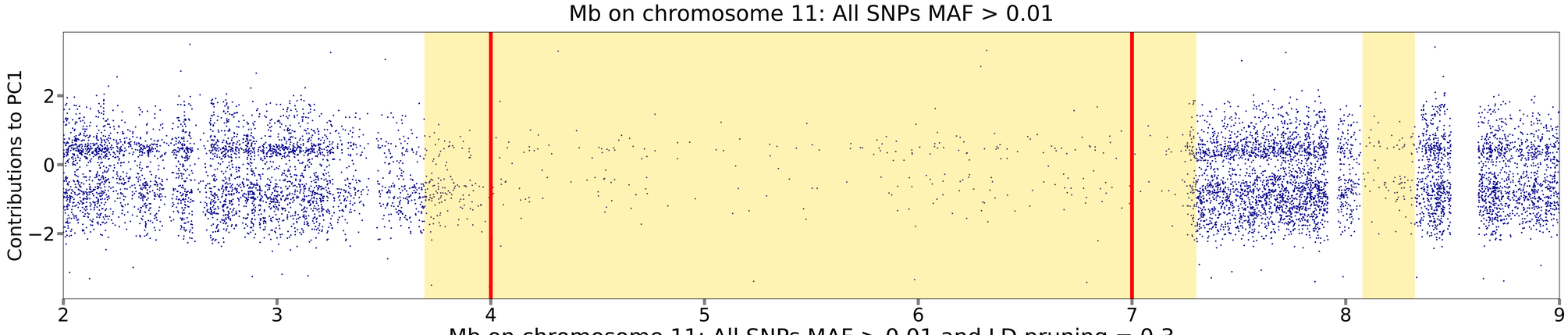
SNP contributions after LD pruning: chr11:4-7 Mb

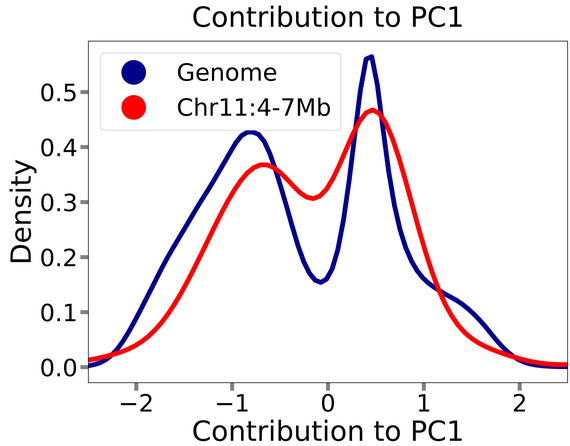


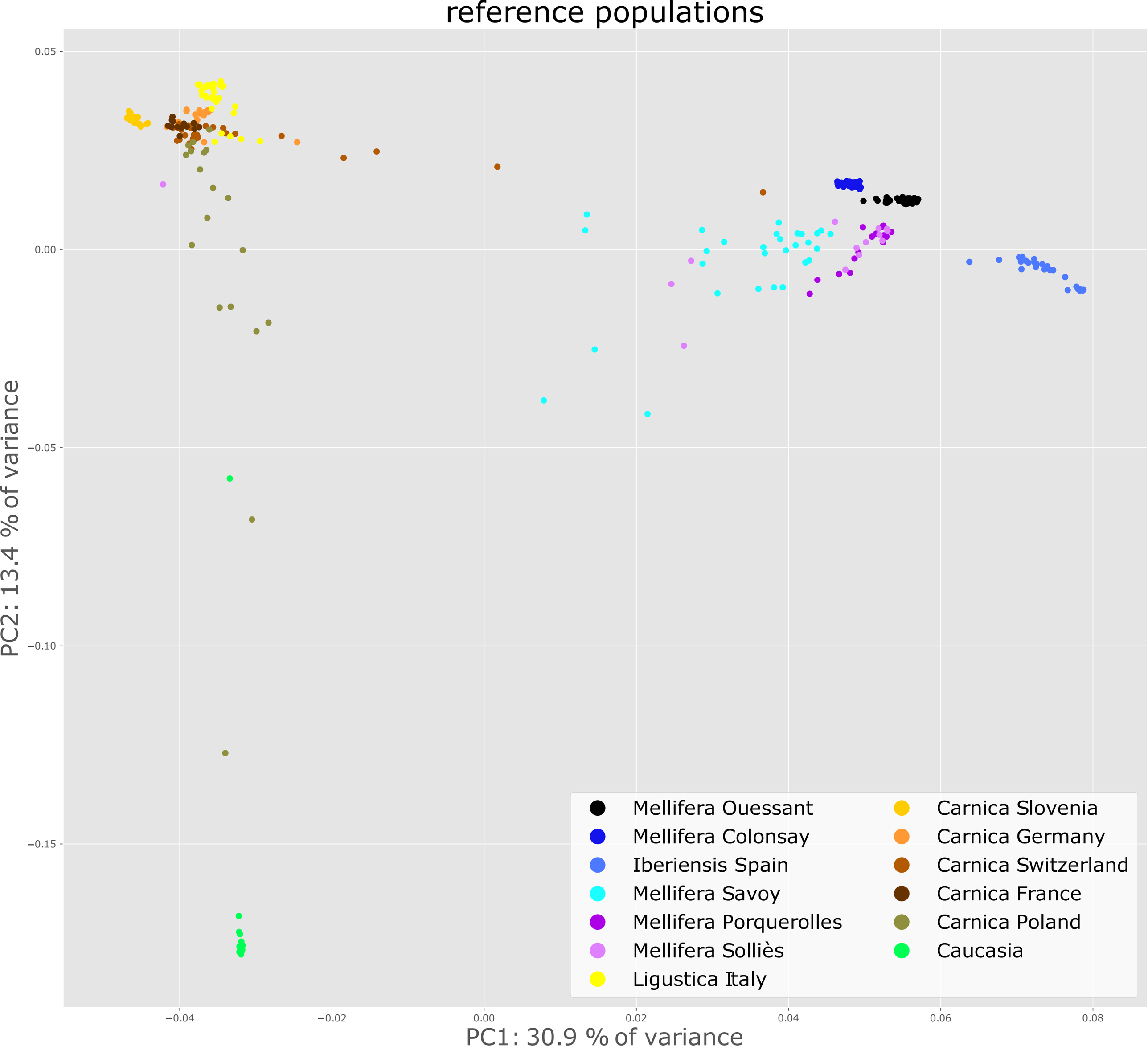
C: SNP contributions a large haplotype block on chromosome 11

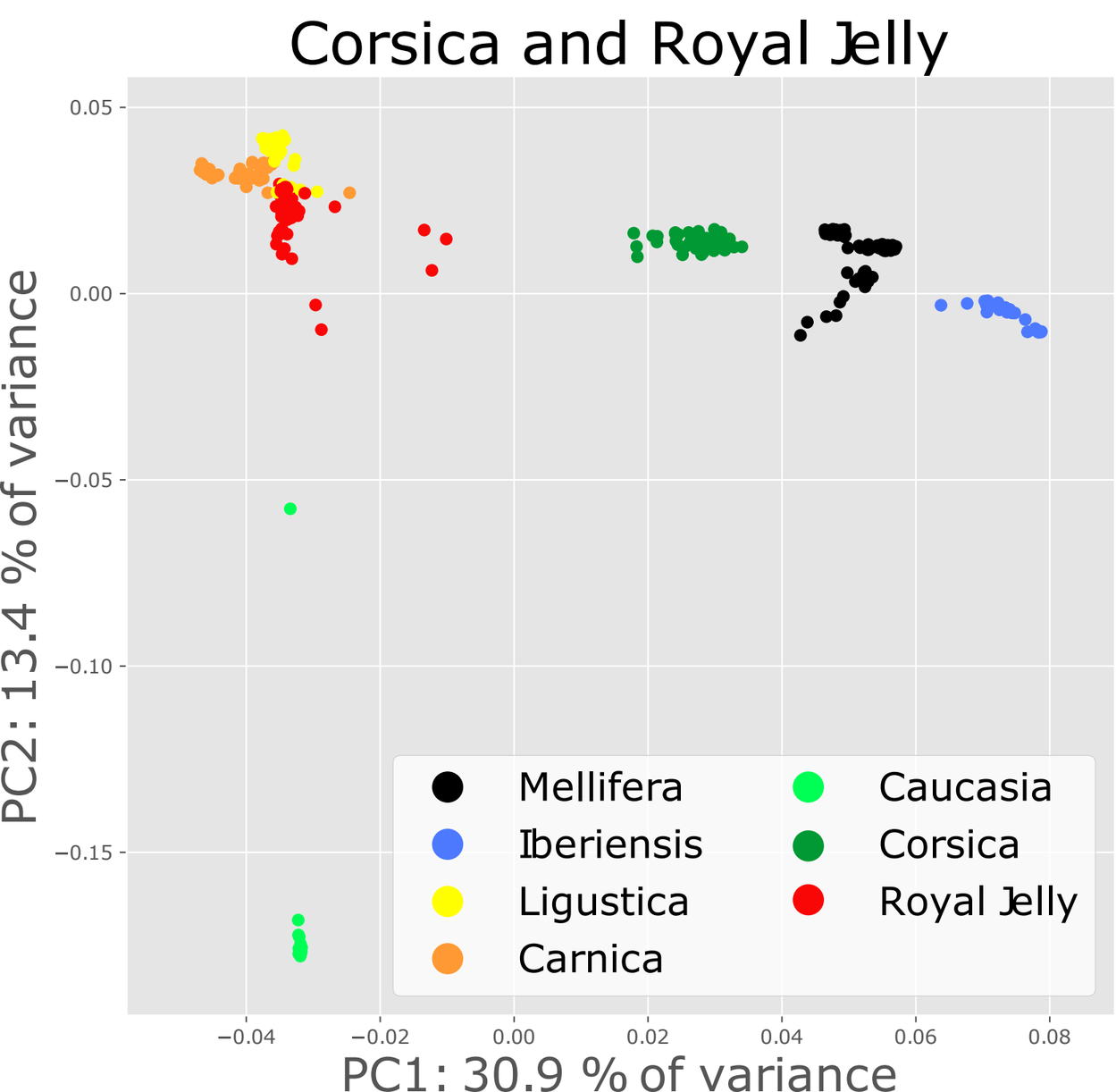


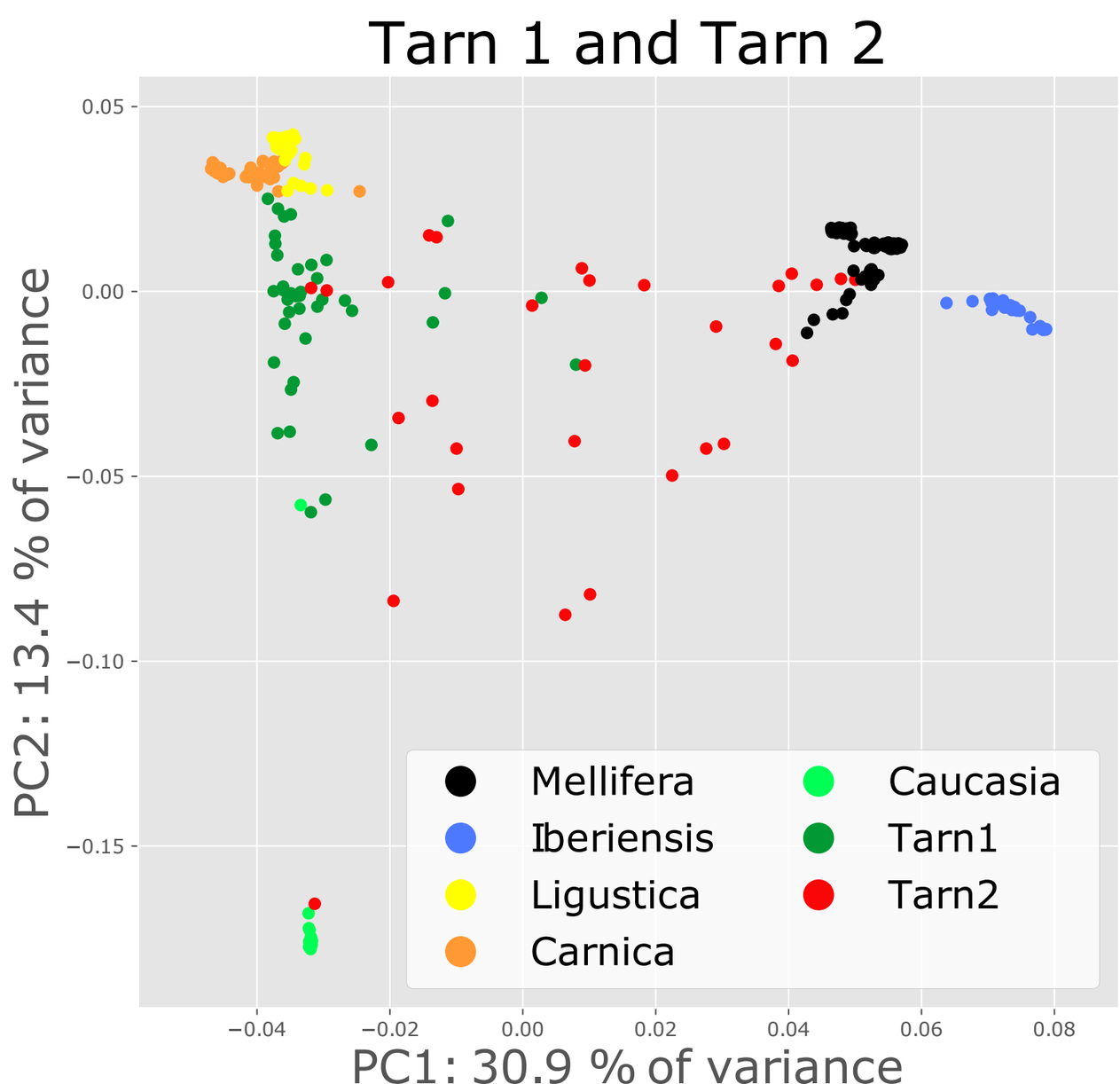


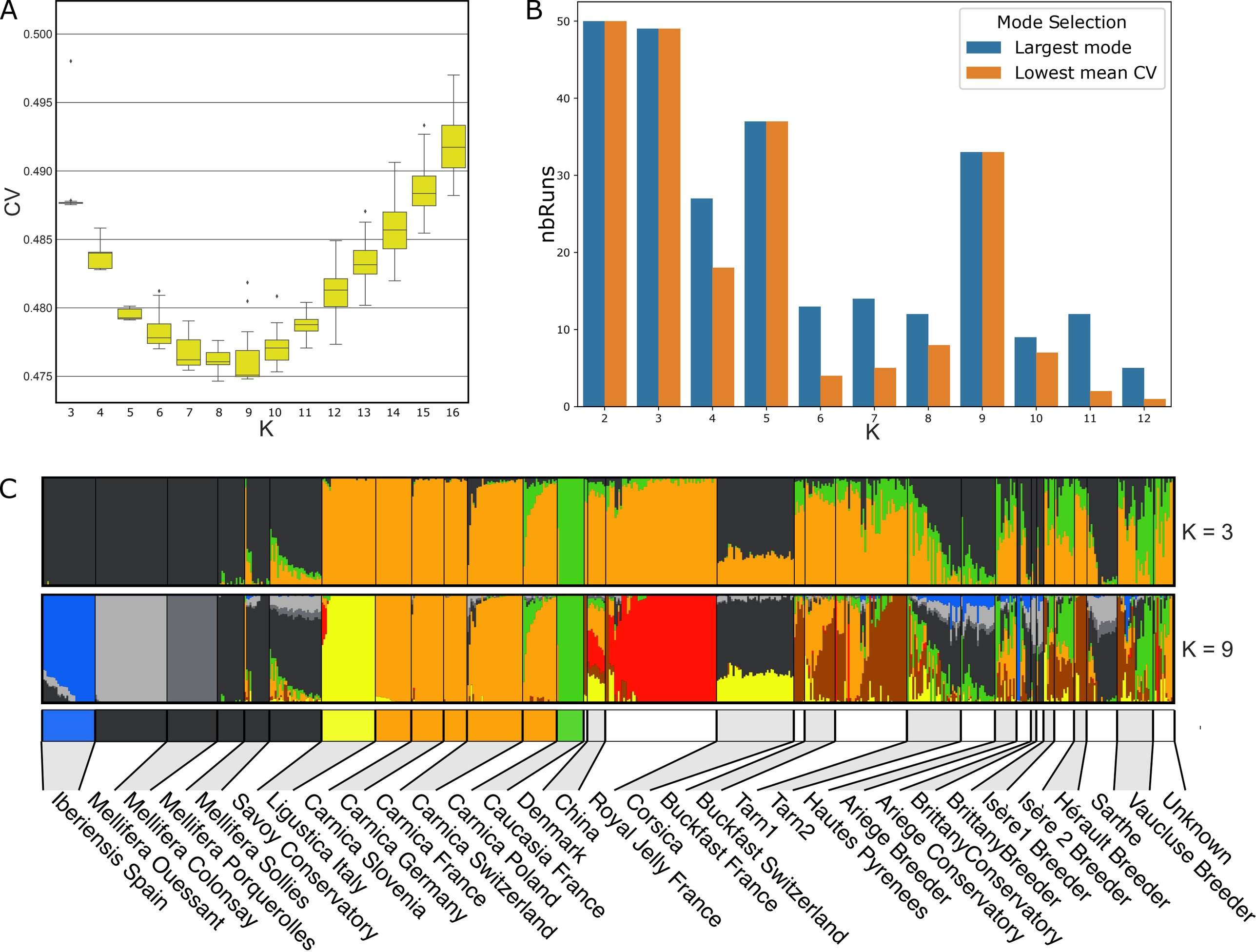


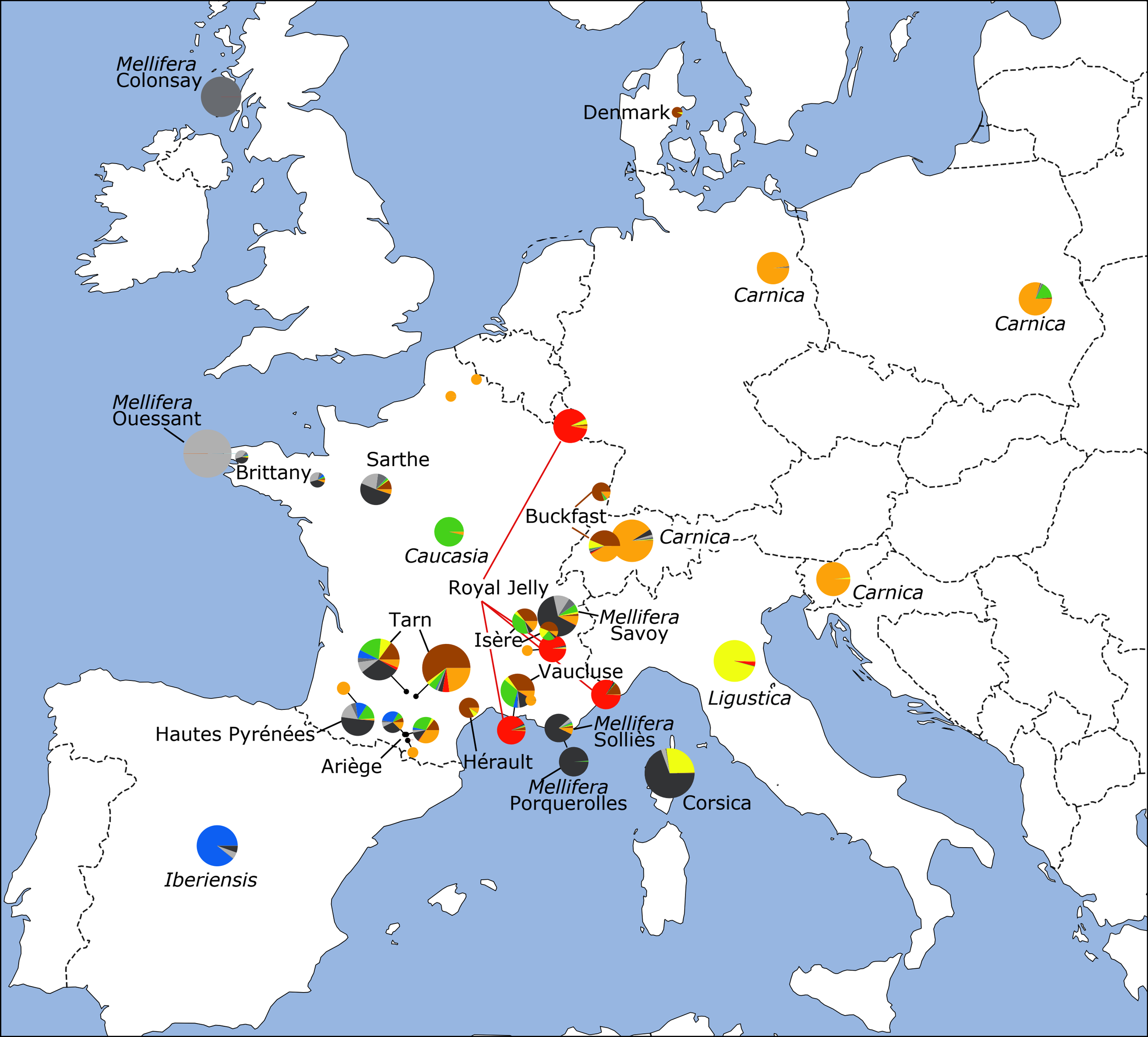


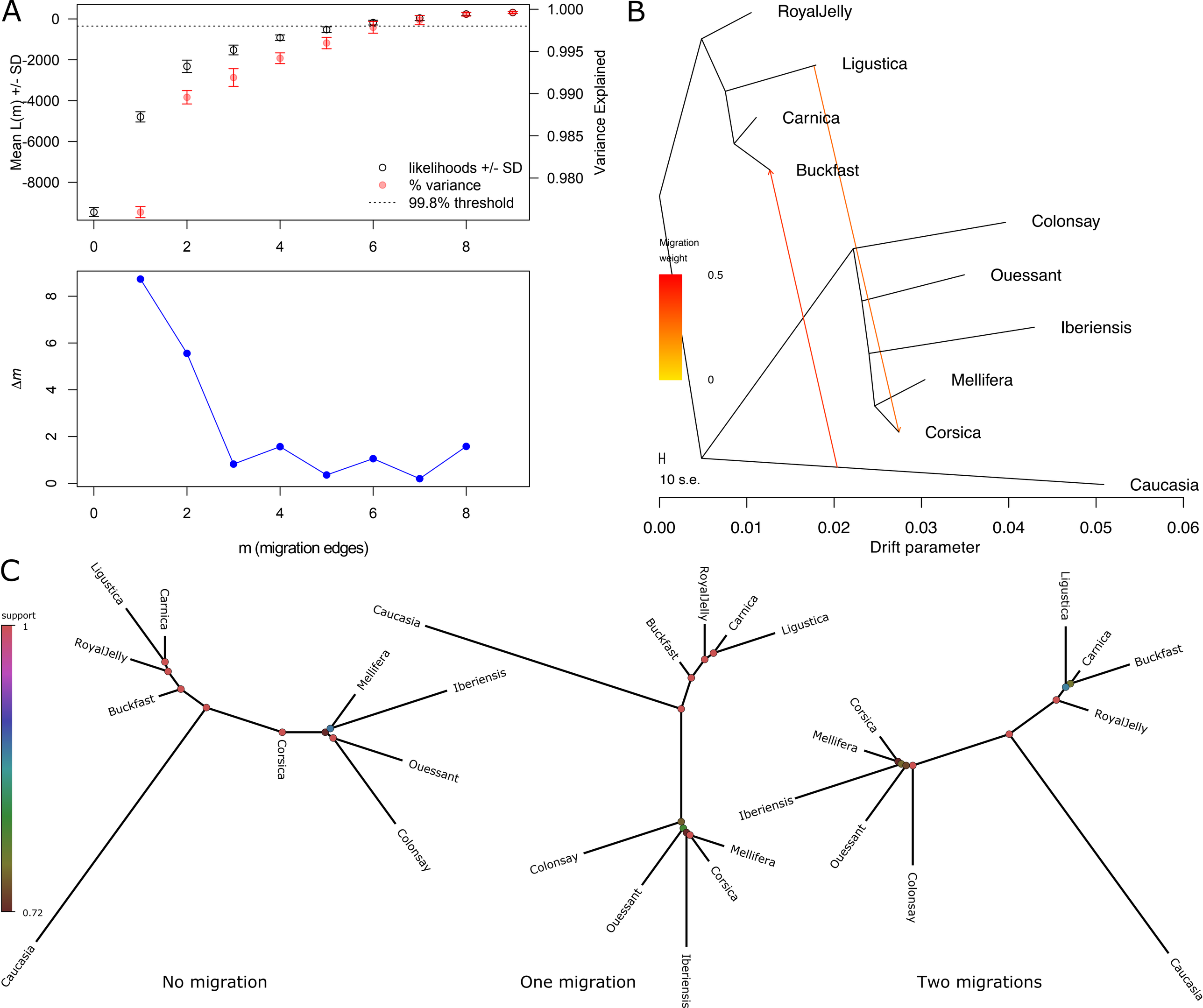

reference populations
PC2: 13.4 % of variance
PC1: 30.9 % of variance
0.05
0.00
−0.05
−0.10
−0.15
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Mellifera Ouessant
Mellifera Colonsay
Iberiensis Spain
Mellifera Savoy
Mellifera Porquerolles
Mellifera Solliès
Ligustica Italy
Carnica Slovenia
Carnica Germany
Carnica Switzerland
Carnica France
Carnica Poland
Caucasia

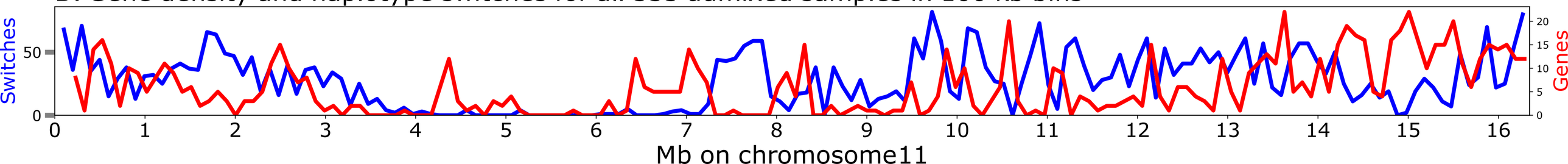

Corsica and Royal Jelly
PC2: 13.4 % of variance
PC1: 30.9 % of variance
0.05
0.00
−0.05
−0.10
−0.15
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Mellifera
Iberiensis
Ligustica
Carnica
Caucasia
Corsica
Royal Jelly


Tarn 1 and Tarn 2
PC2: 13.4 % of variance
PC1: 30.9 % of variance
0.05
0.00
−0.05
−0.10
−0.15
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Mellifera
Iberiensis
Ligustica
Carnica
Caucasia
Tarn1
Tarn2


A
CV
0.500
0.495
0.490
0.485
0.480
0.475
3 4 5 6 7 8 9 10 11 12 13 14 15 16
K
B
nbRuns
50
40
30
20
10
0
2 3 4 5 6 7 8 9 10 11 12
K
Mode Selection
Largest mode
Lowest mean CV
C
K = 3
K = 9
Iberiensis Spain
Mellifera Ouessant
Mellifera Colonsay
Mellifera Porquerolles
Mellifera Sollies
Savoy Conservatory
Ligustica Italy
Carnica Slovenia
Carnica Germany
Carnica France
Carnica Switzerland
Carnica Poland
Caucasia France
Denmark
China
Royal Jelly France
Corsica
Buckfast France
Buckfast Switzerland
Tarn1
Tarn2
Hautes Pyrenees
Ariege Breeder
Ariege Conservatory
BrittanyConservatory
BrittanyBreeder
Isère1 Breeder
Isère 2 Breeder
Hérault Breeder
Sarthe
Vaucluse Breeder
Unknown


Mellifera
Colonsay
Denmark
Carnica
Carnica
Mellifera
Ouessant
Brittany
Sarthe
Buckfast
Carnica
Caucasia
Carnica
Royal Jelly
Mellifera
Savoy
Tarn
Isère
Vaucluse
Ligustica
Hautes Pyrénées
Mellifera
Solliès
Ariège
Hérault
Mellifera
Porquerolles
Corsica
Iberiensis


A
Mean L(m) +/- SD
Variance Explained
likelihoods +/- SD
% variance
99.8% threshold
Δm
m (migration edges)
B
RoyalJelly
Ligustica
Carnica
Buckfast
Colonsay
Ouessant
Iberiensis
Mellifera
Corsica
Caucasia
Migration weight
0.5
0
H
10 s.e.
Drift parameter
C
support
1
0.72
No migration
One migration
Two migrations

A: Local ancestry inference in chromosome11: samples from Corsica

B: Gene density and haplotype switches for all 333 admixed samples in 100 kb bins