


METHODOLOGY ARTICLE

Open Access



# Authoritative subspecies diagnosis tool for European honey bees based on ancestry informative SNPs

Jamal Momeni<sup>1\*†</sup>, Melanie Parejo<sup>2,3†</sup>, Rasmus O. Nielsen<sup>1</sup>, Jorge Langa<sup>2</sup>, Iratxe Montes<sup>2</sup>, Laetitia Papoutsis<sup>4</sup>, Leila Farajzadeh<sup>5</sup>, Christian Bendixen<sup>5^</sup>, Eliza Căuia<sup>6</sup>, Jean-Daniel Charrière<sup>3</sup>, Mary F. Coffey<sup>7</sup>, Cecilia Costa<sup>8</sup>, Raffaele Dall'Olio<sup>9</sup>, Pilar De la Rúa<sup>10</sup>, M. Maja Drazic<sup>11</sup>, Janja Filipi<sup>12</sup>, Thomas Galea<sup>13</sup>, Miroljub Golubovski<sup>14</sup>, Ales Gregorc<sup>15</sup>, Karina Grigoryan<sup>16</sup>, Fani Hatjina<sup>17</sup>, Rustem Ilyasov<sup>18,19</sup>, Evgeniya Ivanova<sup>20</sup>, Irakli Janashia<sup>21</sup>, Irfan Kandemir<sup>22</sup>, Aikaterini Karatasou<sup>23</sup>, Meral Kekecoglu<sup>24</sup>, Nikola Kezic<sup>25</sup>, Enikő Sz. Matray<sup>26</sup>, David Mifsud<sup>27</sup>, Rudolf Moosbeckhofer<sup>28</sup>, Alexei G. Nikolenko<sup>19</sup>, Alexandros Papachristoforou<sup>29</sup>, Plamen Petrov<sup>30</sup>, M. Alice Pinto<sup>31</sup>, Aleksandr V. Poskryakov<sup>19</sup>, Aglyam Y. Sharipov<sup>32</sup>, Adrian Siceanu<sup>6</sup>, M. Ihsan Soysal<sup>33</sup>, Aleksandar Uzunov<sup>34,35</sup>, Marion Zammit-Mangion<sup>36</sup>, Rikke Vingborg<sup>1†</sup>, Maria Bouga<sup>4†</sup>, Per Kryger<sup>37†</sup>, Marina D. Meixner<sup>34†</sup> and Andone Estonba<sup>2\*†</sup> 

## Abstract

**Background:** With numerous endemic subspecies representing four of its five evolutionary lineages, Europe holds a large fraction of *Apis mellifera* genetic diversity. This diversity and the natural distribution range have been altered by anthropogenic factors. The conservation of this natural heritage relies on the availability of accurate tools for subspecies diagnosis. Based on pool-sequence data from 2145 worker bees representing 22 populations sampled across Europe, we employed two highly discriminative approaches (PCA and  $F_{ST}$ ) to select the most informative SNPs for ancestry inference.

**Results:** Using a supervised machine learning (ML) approach and a set of 3896 genotyped individuals, we could show that the 4094 selected single nucleotide polymorphisms (SNPs) provide an accurate prediction of ancestry inference in European honey bees. The best ML model was Linear Support Vector Classifier (Linear SVC) which correctly assigned most individuals to one of the 14 subspecies or different genetic origins with a mean accuracy of  $96.2\% \pm 0.8$  SD. A total of 3.8% of test individuals were misclassified, most probably due to limited differentiation between the subspecies caused by close geographical proximity, or human interference of genetic integrity of reference subspecies, or a combination thereof.

(Continued on next page)

\* Correspondence: [JamalMomeni@eurofins.dk](mailto:JamalMomeni@eurofins.dk); [andone.estonba@ehu.eus](mailto:andone.estonba@ehu.eus)

<sup>†</sup>Jamal Momeni and Melanie Parejo are shared first author.

<sup>^</sup>Christian Bendixen is deceased.

<sup>†</sup>Rikke Vingborg, Maria Bouga, Per Kryger, Marina D. Meixner and Andone Estonba contributed equally to this work.

<sup>1</sup>Eurofins Genomics Europe Genotyping A/S (EFEG), (Former GenoScan A/S), Aarhus, Denmark

<sup>2</sup>Laboratory Genetics, University of the Basque Country (UPV/EHU), Leioa, Bilbao, Spain

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

**Conclusions:** The diagnostic tool presented here will contribute to a sustainable conservation and support breeding activities in order to preserve the genetic heritage of European honey bees.

**Keywords:** *Apis mellifera*, European subspecies, Conservation, Machine learning, Prediction, Biodiversity

## Background

Honey bees (*Apis mellifera* L.) are the most important managed pollinators and currently under threat due to a multitude of pressures worldwide [1, 2]. The species shows considerable variation across its natural range and is comprised of at least 30 described subspecies belonging to different evolutionary lineages [3–6]. Europe holds a large fraction of this honey bee diversity with numerous endemic subspecies representing four evolutionary lineages, namely the African lineage (A), Central and Eastern European lineage (C), Western and Northern European lineage (M), and Near East and Central Asian lineage (O) [7, 8]. However, this diversity and the natural distribution range of European honey bees have been influenced by anthropogenic factors to an extent that several locally adapted populations are at risk due to introgression and crossbreeding [9–11]. Large-scale queen breeding, commercial trade and long distance migratory beekeeping may reduce genetic diversity and can lead to genetic homogenization of admixed populations [9, 12] and potential subsequent loss of local adaptations. In fact, it has been demonstrated that locally adapted honey bees have higher survivability [13] from which follows that the conservation of the underlying genotypic variation must be a priority for the long-term sustainability of populations [14]. To conserve the honey bees' natural heritage and thereby its adaptive potential to future global change, there is a need to promote the sustainable breeding of certified local subspecies.

Numerous conservation efforts for native honey bees have been initiated across Europe [9, 10, 15, 16]. The success of such conservation efforts including genetic improvement programs [17, 18] depends on mating within the population of interest, which is complicated by the honey bees' mating system where virgin queens mate freely with multiple drones from surrounding colonies [19, 20]. Beyond the use of isolated mating apiaries or artificial insemination, successful mating control measures can include different management techniques of queens and drones [21] and regular monitoring of genetic origin and parentage. In some countries and regions in Europe, queen importations are restricted to the native honey bee subspecies [22, 23] or ecotypes [24, 25]. In such instances, when trading queens or colonies across national borders, queen origin needs to be verified. Additionally, authentication of the genetic origin of bee products in terms of a certifiable native bee label,

could help beekeepers to better market their hive products [26]. Thus, to implement effective border control, increase economic value of bee products and to support informed conservation and breeding management decisions across Europe, there is a demand for diagnostic genetic test to reliably infer the subspecies of origin.

With the advances of high-throughput sequencing and genotyping technology in the last decade, reference genomes, whole-genome sequence data, and thousands of individual genotypes are now available for many species. Within these oftentimes massive data sets, it is possible to mine for highly informative single nucleotide polymorphisms (SNPs) that can then be exploited to genotype a larger number of individuals [27, 28]. Such genotyping panels based on a selected set of informative SNPs have been developed for numerous species, including humans, and can be used to infer introgression, genetic ancestry, population structure, genetic stock identification, and food forensics [29–31].

Different approaches have been used to select informative SNPs from larger genotyping panels or sequence data (reviewed in [32, 33]). The most common and popular method for selection is population differentiation as estimated by  $F_{ST}$ , which is based on allele frequency differences between populations expressing the variation among populations relative to the total population [34, 35]. Principal Component Analysis (PCA) has also been employed to identify informative SNPs, since it reduces feature dimensionality while only losing little information and is particularly advantageous with complex population structures [28, 36]. Given a set of informative SNP markers, supervised classification and so-called assignment tests are employed whereby an individual is assigned to predefined classes (i.e., subspecies or populations of origin). Classical applications of assignment testing in population genetics first used supervised parametric likelihood-based approaches [37, 38]. Recently, new methods, together referred to as supervised machine learning (ML), have emerged in computational population genomics [39]. The general approach for any supervised ML classifiers is to split the data into a reference (training) set to 'learn' a function that can discriminate between the given data classes [40]. This function is then used to predict the probability of an 'unknown sample' (test) of belonging to any given class (e.g. subspecies). The accuracy of the classification, expressed as the proportion of test individuals correctly classified

to their population of origin, is influenced by the properties of the training data set (i.e., number of samples, genetic diversity, levels of population differentiation, degree of overlap in data distribution and quality of reference samples) [41]. ML classifiers aim to optimize the predictive accuracy of an algorithm rather than performing parameter estimation of a probabilistic model, and they have the potential to be agnostic to the assessment of the given dataset, i.e. without assumptions of the processes leading to differentiation, including the evolutionary history [39].

For honey bees, different SNP panels have been designed, for instance to identify and estimate C-lineage introgression in M-lineage subspecies *A. m. iberiensis* and *A. m. mellifera* [15, 42–46]. The latter subspecies is native to northern and western Europe and once occupied a large fraction of the European territory, but is now threatened and even has been completely replaced in much of its range [10, 47, 48]. Moreover, SNP panels have also been developed to infer the level of Africanization and ancestry in honey bees of the New World and Australia [46, 49, 50]. However, for most *A. mellifera* subspecies, whose populations have been genetically examined to a lesser extent or not at all, molecular knowledge at this level of detail is still lacking. These subspecies and locally adapted populations or ecotypes appear more vulnerable due to the extant multiple threats to honey bees.

The SmartBees project was initiated with the purpose of developing new tools to describe and conserve honey bee diversity in Europe. We have designed a molecular tool consisting of highly informative SNP markers suitable for assigning honey bee individuals to their subspecies of origin, based on a comprehensive sampling of European honey bee diversity. Based on pool-sequence data from 1995 worker bees representing 22 populations, four evolutionary lineages and 14 subspecies, we selected 4400 informative SNPs employing two powerful and commonly used approaches ( $F_{ST}$  and PCA). Of these, 4165 SNPs, for which probes could be designed and which passed the BeadChip decoding quality metric, were genotyped in 3903 individual bees using the Illumina Infinium platform. Final quality control filtering left 4094 reliable SNPs to build a statistical model using machine learning (ML) algorithms for assignment of European honey bees to 14 different genetic origins. The best model was the Linear Support Vector Classifier (Linear SVC) which could correctly assign 96.2% of the tested samples to their genetic origin. Thus, the here presented method accurately identifies European subspecies, which is crucial to support management strategies in sustainable honey bee breeding and conservation programs.

## Results

### Samples and pool-sequencing

A total of 22 populations representing the four European evolutionary lineages and 14 subspecies have been sampled from their native ranges throughout Europe and adjacent regions (Tables 1 and S1). Each selected population included up to 100 worker bees from unrelated colonies, totaling 2145 samples, which represents the most comprehensive sampling effort for the study of European honey bees to date. The samples from each population were homogenized, pooled and their DNA extracted. Sequencing on an Illumina HiSeq 2500, produced 1.6 billion paired-end fragments (3.2 billion individual reads) with an average read length of 125 bp, and a total genome depth of coverage of 2800x. Sequencing and variant statistics can be found in Table S2.

### Selected SNPs

While main evolutionary lineages were easily differentiated with only few SNPs (Figure S1A), it was more challenging to differentiate closely related subspecies with a reduced number of genetic markers. Given the complex, hierarchical population structure of European honey bees, we employed two powerful and commonly used approaches, PCA (Figure S1) and  $F_{ST}$ , to identify the most discriminant markers to differentiate subspecies of European honey bees (see details in [Methods](#) and [supplementary materials and methods](#)). Based on the variants inferred from the pool-sequence data, we selected 4400 informative SNPs, of these, a total of 4165 SNPs passed the decoding quality metric for genotyping using the Illumina Infinium custom-designed BeadChip, indicating that 99% of the originally submitted probes were suitable for genotyping. The SNPs are distributed across all of the 16 honey bee chromosomes as well as in unplaced contigs (Table S3), with an average distance between SNPs of 64 kb. SNP information and genomic position of the 4165 SNPs selected to differentiate European honey bee subspecies are presented in Additional file 1.

### Sample genotyping and visualization

Of the 4165 SNPs, 4094 were successfully genotyped in 3896 individual bees using Illumina Infinium BeadChip technology (Table 1). With only 71 SNPs never producing any data, the genotyping success rate (SNP validation) rate was 98%. The average call rate per individual was 0.87, varying among samples of every subspecies from 0.84 in *A. m. cypria* to 0.89 in *A. m. adami* (Table S4). More than one-third of the samples have a call rate exceeding 0.9.

The genotype data of the individuals from the pool sequencing is visualized in a t-SNE plot [51] that reduces high-dimensional data to a two-dimensional map where

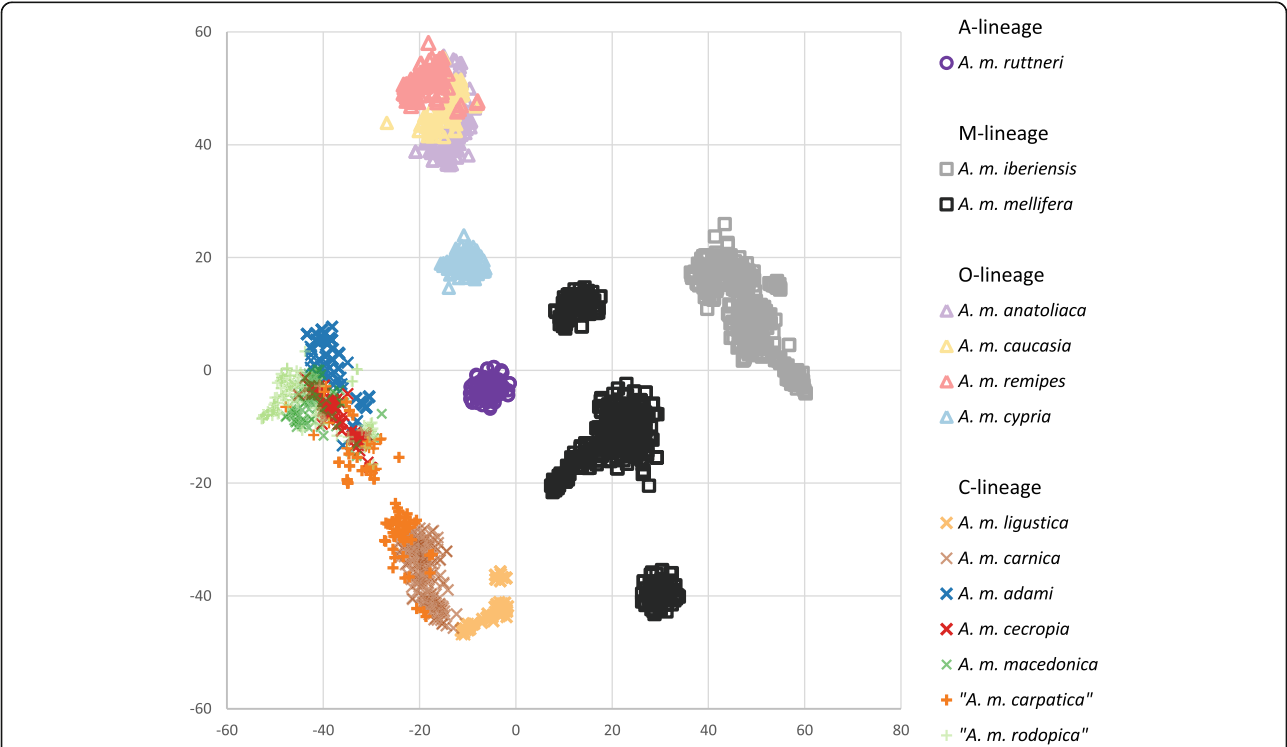
**Table 1** Samples individually genotyped for subspecies classification ( $N_{TOT} = 3896$ ) consisting of individual samples from the pool sequencing (in bold,  $N = 1998$ , excluding 62 outliers) and new independent samples ( $N = 1908$ ). Samples were collected from their native range and labelled based on previous studies, morphometric analysis or local knowledge (see [Methods](#) sections and Table S1). 70% of pool sequencing samples ( $N = 1391$ ) were used as training data for building the model, while the remaining 30% ( $N = 597$ ) together with the independent samples ( $N_{Total} = 2505$ ) were considered as out-of-sample data for subsequent validation

Evolutionary lineage	Subspecies	Sampling country	Pool name / Sampling group	N	$N_{TOT}$
A	<i>A. m. rutnieri</i>	Malta	<b>rut_mlt</b>	91	187
			MLT	96	
C	<i>A. m. adami</i>	Crete, Greece	<b>ada_grc</b>	82	82
	<i>A. m. carnica</i>	Austria & Hungary	<b>car_aut_hun</b>	93	825
		Croatia & Slovenia	<b>car_svn_hrv</b>	95	
		Croatia	HRV	94	
		Denmark	DNK	89	
		France	FRA	8	
		Germany	GER	282	
		Poland	POL	40	
		Serbia	SRB	49	
		Slovenia	SVN	75	
		Romania & Moldova	<b>carp_rou_mda</b>	86	
	<i>A. m. cecropia</i>	France	FRA	4	4
		Greece	<b>cec_grc</b>	93	140
	<i>A. m. ligustica</i>	Italy	GRC	47	
			<b>lig_ita</b>	84	143
			ITA	59	
	<i>A. m. macedonica</i>	N. Macedonia & N-Greece	<b>mac_mkd_grc</b>	86	429
		Greece	GRC	49	
		N. Macedonia	MKD	96	
		Germany	GER	198	
	<i>"A. m. rodopica"</i>	Bulgaria	<b>rod_bgr</b>	84	84
M	<i>A. m. iberiensis</i>	Spain & Portugal	<b>ibe_esp_west_prt</b>	94	460
		Spain	<b>ibe_esp_eus</b>	96	
			<b>ibe_esp_north</b>	91	
			<b>ibe_esp_south</b>	64	
			ESP	115	
	<i>A. m. mellifera</i>	Belgium	BEL	96	1066
		Denmark	<b>mel_dnk</b>	96	
			DNK	97	
			FIN	15	
		France	FRA	49	
		Ireland	<b>mel_irl</b>	96	
		Isle of Man	<b>mel_imn</b>	92	
		Norway	NOR	12	
		Poland	POL	33	
		Russia	<b>mel_rus</b>	96	
		Scotland	SCT	280	
		Sweden	SWE	8	
		Switzerland	<b>mel_che</b>	96	

**Table 1** Samples individually genotyped for subspecies classification ( $N_{TOT} = 3896$ ) consisting of individual samples from the pool sequencing (in bold,  $N = 1998$ , excluding 62 outliers) and new independent samples ( $N = 1908$ ). Samples were collected from their native range and labelled based on previous studies, morphometric analysis or local knowledge (see [Methods](#) sections and Table S1). 70% of pool sequencing samples ( $N = 1391$ ) were used as training data for building the model, while the remaining 30% ( $N = 597$ ) together with the independent samples ( $N_{Total} = 2505$ ) were considered as out-of-sample data for subsequent validation (Continued)

Evolutionary lineage	Subspecies	Sampling country	Pool name / Sampling group	N	$N_{TOT}$
O	<i>A. m. anatoliaca</i>	Turkey	<b>ana_tur</b>	94	94
	<i>A. m. remipes</i>	Armenia	<b>rem_arm</b>	90	90
	<i>A. m. caucasia</i>	Poland	<b>cau_tur_geo</b>	96	113
		Denmark	DNK	4	
		NE-Turkey & Georgia	POL	13	
	<i>A. m. cypria</i>	Cyprus	<b>cyp_cyp</b>	93	93
Total					3896

each individual is represented by a point (Fig. 1). The genotyped samples were grouped in several separated clusters according to their evolutionary lineage or subspecies of origin (Fig. 1). Within each lineage, most of the individuals from the same geographic origin were closely grouped together and generally well separated from neighboring groups. The only A-lineage subspecies in our study, *A. m. ruttneri*, was placed in the center intermediate to the other clusters. In the O-lineage, *A. m. cypria* bees were well separated from *A. m. anatoliaca*, *A. m. caucasia* and *A. m. remipes*, which appear less well differentiated. The two subspecies of the M-lineage were well differentiated, with *A. m. mellifera* populations grouped in three subclusters separating the distant (Burzyan region, Russia, top *A. m. mellifera* cluster in Fig. 1) or isolated (Læsø island, Denmark, bottom *A. m. mellifera*) sampling regions. C-lineage samples grouped into three subclusters: (i) *A. m. ligustica*, (ii) *A. m.*



**Fig. 1** Visualization using a t-SNE manifold plot of the 1988 honey bee samples from the pool sequencing individually genotyped for 4094 SNPs. Samples have been color-coded according to the subspecies reference populations corresponding to the 14 classes used for subsequent supervised machine learning classification



*carnica* bees including part of the “*A. m. carpatica*” samples and (iii) a heterogeneous subcluster of *A. m. macedonica*, *A. m. cecropia*, *A. m. adami*, “*A. m. rodopica*” and the rest of “*A. m. carpatica*” bees. A t-SNE plot with sample labels according to their pool of origin is presented in Figure S2.

### Sample classification using machine learning

We employed machine learning (ML) methods to build a model for the classification and assignment of European honey bees to its subspecies of origin. Out of the tested ML algorithms, the best performing model was the Linear SVC (Table S5). The model calculates the prediction probability for a sample to belong to any of the 14 reference populations. Each test sample was classified into the subspecies which showed the highest prediction probability ranging from as low as 0.29 to 1.0 with a median of 0.98 (Figure S3).

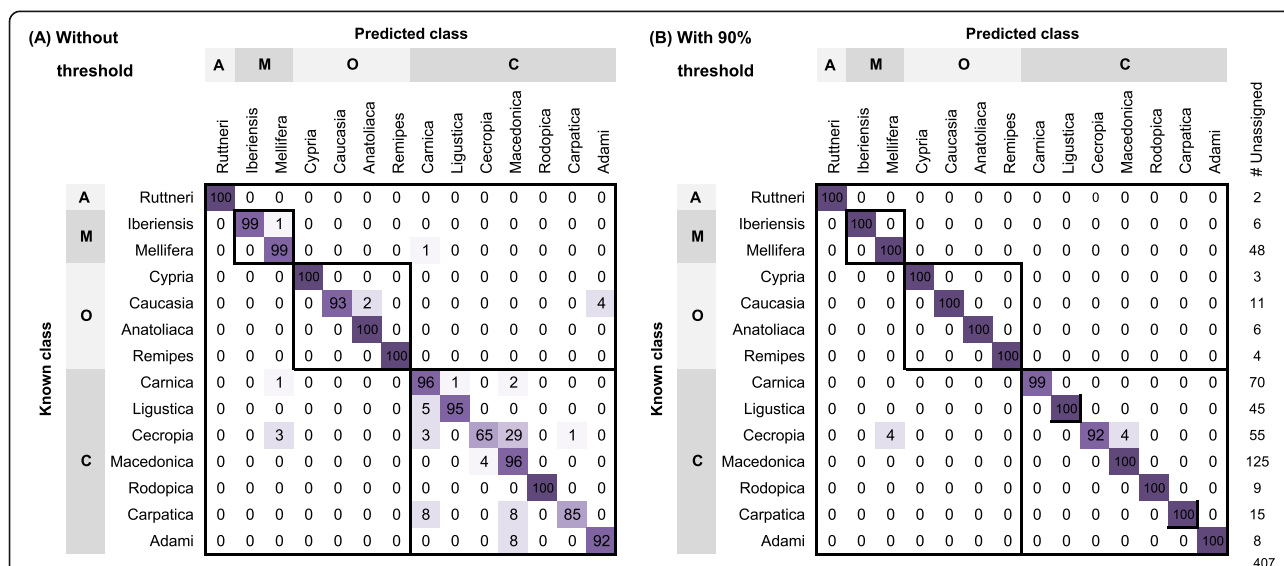
A confusion matrix was used to summarize, describe and visualize the performance of the Linear SVC classification model on a set of test data (out-of-sample data,  $N = 2505$ ) for which the true values (subspecies) were known. For the lineages, the model is capable of predicting all samples with 100% accuracy (Figure S4). For the subspecies, the confusion matrix revealed that for most of them the model accurately predicted the ancestry of the test samples ( $N = 2505$ ), with only a few exceptions (Fig. 2a). The accuracy ranged from 65 to 100%, indicating that some subspecies are easier to distinguish than others. In total 96.2% of test samples were correctly predicted, while 95 individuals (3.8%) were misclassified,

i.e., predicted by the model with a different subspecies than the labeled one (true values), for instance: four *A. m. ligustica* bees were predicted as *A. m. carnica*, two “*A. m. carpatica*” bees each as either *A. m. carnica* or *A. m. macedonica*, and 23 *A. m. cecropia* bees were predicted as *A. m. macedonica*.

The model predicts the probability that a given sample belongs to one of the 14 subspecies under study. On this basis, the test samples were assigned to a certain subspecies based on the highest prediction probability, even if the probability was low (see above). Therefore, with the purpose of increasing the certainty of classification we set a probability threshold, so to ensure that only samples very likely belonging to any of the 14 subspecies were assigned, while test samples with low prediction probabilities were considered unassigned. In Fig. 2b, we show an example of setting a probability threshold at 90%. By setting this threshold, we increased the proportion of truly assigned samples from 96.1 to 99.6%, while the misclassification rate fell from 3.9 to 0.4%. However, 407 of the test individuals remained “unassigned”, for instance, 22 out of the 23 *A. m. cecropia* bees predicted as *A. m. macedonica* were no longer considered misclassified but enter the unassigned category.

### Discussion

In this study, we performed a large-scale and comprehensive sampling following a standardized procedure, and aimed to capture as much of the honey bee genetic diversity in Europe as possible by deep-sequencing of pooled populations. Further, we applied two powerful



**Fig. 2** Confusion matrix for test samples (out-of-sample data,  $N = 2505$ ) showing the (rounded) percentages of truly assigned individuals (diagonal) and percentages of individuals assigned to a different subspecies (misclassified; upper and lower triangles). **a** Assignment based on the highest prediction probability classifies each of the test individuals to a subspecies, while **b** using a probability threshold of 90% some samples are considered “unassigned” and excluded from the confusion matrix

SNP selection methods [32, 33] to address diversity at different levels of differentiation (lineages, subspecies, populations). Subsequently, these ancestry informative markers were employed to build a model to classify samples of European honey bees into subspecies.

The considerable honey bee diversity poses a challenge when it comes to providing a discriminative tool applicable across Europe. The four European lineages were easily distinguished genetically with only 200 SNPs due to their ancient divergence [52], but difficulties arose at a lower hierarchical level of differentiation. Subspecies from the same evolutionary lineage diverged only recently [53] and are, thus, genetically very close. Moreover, there are some areas in Europe where *A. mellifera* subspecies variation has not yet been exhaustively described, while in others human-mediated introgression contributes to blurring the natural boundaries between subspecies [42, 48, 54]. National breeding programs can also disrupt the natural gene flow and may contribute to changing the genetic background of the original subspecies [11, 12, 55, 56]. In fact, in our study applying a stringent filtering option we only identified few unique SNPs that were exclusive to one population. Similarly, other population genomics studies have found a high degree of allele sharing across and within evolutionary lineages [7, 53]. In contrast, we found variation in the average call rate per individual between subspecies which may, in part, be explained by the presence of null alleles (alleles producing no signal), suggesting sequence variation or subspecies-specific deletions within the probe site. Probes that did not work for certain subspecies (i.e. missing data), in fact, contain valuable information and even enriched our model.

We employed a machine learning (ML) approach to build a model for subspecies classification. ML takes advantage of high dimensional input and provides an improvement of prediction accuracy in a model-free approach [39, 40]. In this way, subtle differences can be revealed which was particularly relevant in our study, due to the high number of closely related subspecies we wanted to discriminate. Our best performing model was Linear SVC, member of the family of Support Vector Machines (SVMs), which are known to generalize well because they are designed to maximize the margin between any two classes (subspecies) [57]. Typical biological applications of SVMs include protein function prediction, transcription initiation site prediction and gene expression data classification (reviewed in 57). In the field of population genetics, a thorough ML approach to select the best model is generally not yet commonly implemented, although specific models have been developed for ancestry inference [58, 59]. Here, we employ a comprehensive ML approach based on genotype data for honey bee subspecies diagnosis.

Despite the comprehensive sampling effort, the careful SNP selection and the application of the latest classification methods, some limits remain in the diagnostic system. For instance, within the C-lineage we have experienced problems in differentiating samples according to the alleged subspecies. Such misclassification of individuals can be explained by various factors coming together: (i) this lineage is of comparatively recent origin [53] and (ii) consists of multiple highly interrelated subspecies within close geographical proximity (see Figure S1D); (iii) the taxonomic status of some populations has not yet been fully resolved [60–62]; and (iv) the genetic background of some populations is being altered by introgression due to human interference [63]. Furthermore, labelling errors of the out-of-data samples could not be ruled out as an additional source of misclassification, especially if we refer to those samples for which the model predicted a different subspecies with high probability. Supervised ML relies on the qualities of the reference data for classification, thus, in the future, we aim to refine the training data to improve the model prediction accuracy and reduce the misclassification rate.

It is also important to note, that by setting a probability threshold for the assignment of any subspecies, the misclassification rate was reduced, for some subspecies considerably. While such a threshold increases the confidence in subspecies prediction, it also implied, however, that quite a few individuals were left “unassigned”. What threshold is used as a cut-off for subspecies classification depends on the specific circumstances and the application. For example, for the conservation of a small endangered population the threshold might be set lower in order to maintain genetic diversity, than for instance in a pure breeding line under selection for specific traits.

Overall, earlier methods based on morphometry, mtDNA variation, microsatellite loci, or even SNPs have been effective in differentiating between evolutionary lineages and, to some extent, between subspecies of the same lineage [22, 42, 45, 64–67]. Yet, our diagnostic tool is the most comprehensive tool to date to reliably classify European honey bees into subspecies in a single analysis. Moreover, the advantage of our approach is that it is a dynamic tool that can be updated to include more subspecies by genotyping new samples and adding their data to rebuild a classification model using ML with additional subspecies. Ongoing research indicates that this approach is applicable to *A. m. siciliana* from Sicily. Furthermore, individual bees from South Africa tested with our system were rejected as being of European origin (*i. e.*, low prediction probability to any of the subspecies). This dynamic tool, therefore, could easily incorporate new populations to be discriminated, and would even have the potential to be optimized to differentiate populations/ecotypes within subspecies, or to evaluate the degree of introgression.

## Conclusions

The main finding of the study is that our model can classify bees into each of the European subspecies with high accuracy. Consequently, as the bees included in this project were collected in a vast area ranging from Russia and Armenia in the East to Portugal in the West, and from Malta in the South to Scotland in the North, we conclude that much of the natural diversity of European honey bees can still be considered extant, in spite of human interference since more than 150 years. The in situ conservation of this genetic heritage is our duty [68], and we believe that the honey bee subspecies diagnostic tool presented will make a useful contribution. It is of value in an array of applications: for beekeepers who want to know the subspecies of their bees; for conservation managers in Europe, where subspecies diagnosis is essential to monitor the hybridization rate of colonies within conservatories; for veterinarians to control queen trade; for bee breeders to certify the subspecies origin of their queens; and for beekeepers to authenticate their bee products.

## Methods

### Pool-sequencing samples

For this study, in total 22 populations have been sampled, all within their native range (Tables 1 and S1), and are referred to as different subspecies and genetic origins according to the classification of Ruttner [8] and the most recent revision of the genus *Apis* by Engel [62]: *A. m. iberiensis* Engel 1999, *A. m. mellifera* Linnaeus 1758, *A. m. carnica* Pollman 1879, *A. m. caucasia* Pollmann 1889, *A. m. ligustica* Spinola 1806, *A. m. macedonica* Ruttner 1988, *A. m. cecropia* Kiesenwetter 1860, *A. m. cypria* Pollman 1879, *A. m. adami* Ruttner 1975, *A. m. anatoliaca* Maa 1953, *A. m. remipes* Gerstaecker 1862; in addition we include *A. m. ruttneri* Sheppard et al. 1997 [69], “*A. m. carpatica*” Foti 1965 [60], and “*A. m. rodopica*” Petrov 1991 [61]. There exist some uncertainty and unresolved taxonomic status of some populations, and subspecies descriptions in literature have not always been performed according to the standards laid down in the International Code of Zoological Nomenclature (ICZN) [62]. Thus, different views are found in literature to what is to be considered a subspecies or ecotype. In this paper, we do not aim to resolve or justify any classification. Finally, we considered 14 subspecies/genetic origins (listed above) for our diagnostic tool, which were used as categories in the machine learning classification model.

Each selected population included up to about 100 (ranging from 86 to 100) worker bees from unrelated colonies that were used for subsequent pool-sequencing. Effort was undertaken to cover the entire distribution range of any subspecies, while taking into account

within-subspecies variability when appropriate. We focused on collecting representative samples for each subspecies by primarily sampling from beekeepers that were known not to import bees in order to minimize the risk of including hybrids. Moreover, we only chose one worker bee per apiary to avoid related individuals and to include as much diversity per population as possible. Also in order to secure the subspecies-origin of the collected samples, in some cases (where possible), a morphometric analysis was performed and/or we relied on already genotyped bees [55, 65, 66, 70–72]. Detailed information on sample origin and respective references are presented in Table S1.

### DNA extraction, library preparation, and pool-sequencing

The heads or thoraxes of up to 100 bees (Table S1) from each pool were homogenized, DNA was extracted from all samples by using a magnetic bead-based purification method (NucleoMag® Blood 100 µL, Macherey-Nagel, Germany). Subsequently, sequencing libraries of each pool-DNA were constructed with the TruSeq DNA PCR-Free library preparation kit and sequenced on an Illumina HiSeq 2500 platform. Bioinformatic processing, including trimming, mapping and variant calling of the generated pool sequence data, was performed using best practices and standard software (details in [supplementary material and methods](#)). The pipeline for the analysis of the pool sequence data is available at [https://github.com/jlanga/smsk\\_popoolation](https://github.com/jlanga/smsk_popoolation).

### Selection of ancestry informative markers

Several studies have selected a limited number of SNPs to differentiate between the main evolutionary lineages [15, 45, 46], however, for closely related subspecies more markers and a more refined selection approach are needed. Thus, we used two different approaches (PCA and  $F_{ST}$ ) [28, 34] to identify and select informative SNPs, in order to capture the most discriminant markers at different levels: (i) SNPs to differentiate the four main evolutionary lineages, (ii) SNPs to discriminate subspecies within evolutionary lineages, and (iii) SNPs to identify specific populations within subspecies (e.g. ecotypes).

First, we created a matrix with the minor allele frequencies for each SNP and sequenced pool, which was used to perform PCA to select SNPs that differentiate the main evolutionary lineages (Figure S1A). Second, PCA was performed separately on the subsets of pools from each lineage in order to select informative SNPs to discriminate subspecies within each lineage (Figure S1B–D). We used the FactoMineR R package [73] and custom-made R scripts to select at each hierarchical level the SNPs with the highest contributions to the significant PCs. Using this procedure, 300 PCA-informative SNPs were selected for discriminating the four



evolutionary lineages, 200 SNPs for the M-lineage, 600 SNPs for the O-lineage and 1100 SNPs for the most complex C-lineage (Figure S1D). Preliminary simulations using allele frequencies from the pool-sequencing revealed that this approach was highly effective in discriminating lineages and subspecies (Figure S1).

To select additional SNPs that can differentiate between pools, pairwise  $F_{ST}$  values [74] between all population were calculated for each SNP with two settings (loose and stringent options) using PoPoolation2 [75]. The loose setting option will return more SNPs with less certainty and lower quality, which in turn potentially reduces genotyping success. This drawback is counterbalanced, since the loose option increases the chance of identifying highly informative population-specific (unique) SNPs. For either setting option (loose and stringent), the pairwise  $F_{ST}$  values of each pool against all other pools were summed up for each SNP and then ranked according to the highest summed  $F_{ST}$  value. A fixed and unique SNP in one pool is expected to have a maximum sum of 21, which means this variant is only present in this specific population. A reasonable trade-off between unique and reliable SNPs was achieved by selecting the top 20 SNPs with the highest summed  $F_{ST}$  from the loose option and the top 80 SNPs from the stringent option for each pool. With 22 pools, a total of 2200 informative population-specific SNPs were selected using  $F_{ST}$ .

Overall, 4400 ancestry informative SNPs were selected based on PCA and  $F_{ST}$  (Table S3). These highly informative markers are not only important for the assignment of individuals to subspecies as presented in this study, but, because of their varied allele frequencies in different populations, they can be used, for instance, for classification of new subspecies and for further follow-up studies.

### Probe design

Probes for the 4400 selected SNPs were evaluated for genotyping on the Illumina Infinium platform using Illumina's DesignStudio® software which requires as input the flanking region of 50 bp up and downstream of each SNP. SNPs were discarded if no probe could be designed in the flanking region or if the probes had more than one hit when aligned to the honey bee reference genome. The final list of 4197 SNPs was submitted to Illumina for probe design and production. The SNPs are distributed across all of the 16 honey bee chromosomes as well as in unplaced contigs (Table S3; Additional file 1), with an average distance between SNPs of 64 kb.

### Validation samples and genotyping

A total of 3958 individual bees were genotyped for the selected SNPs, including 2050 same individual worker bees that were used for pool sequencing, as well as 1908

newly collected individuals (Table 1). These new additional samples were received from several different sources and of variable quality, including whole honey bees in ethanol, honey bees squeezed on FTA cards, tissue samples from flight muscle or purified DNA. These originated from SmartBees breeding apiaries [76] and from colonies examined for Varroa-sensitive hygienic behavior within the SmartBees project [77]. The samples were genotyped using the custom-made BeadChip array Infinium iSelect XT 96. The results were analyzed using Illumina's GenomeStudio® software, and the genotypes of each sample were exported for further analysis. For an initial visualization of the genotyping results, we created t-distributed stochastic neighbor embedding (t-SNE) manifold plots. This technique visualizes high-dimensional data by giving each data point a location in a two- or three-dimensional map [78]. Outliers and samples that were labeled as one subspecies, but were clearly grouped with another cluster, were removed, in total 62 samples, leaving  $N = 1988$  pool sequence reference samples. This was done with the objective to create a high-quality and representative reference data set for subspecies assignment.

### Sample classification using machine learning (ML) algorithms

In order to build a model to classify and predict the subspecies assignment of unknown samples of European honey bees, we employed ML methods using the scikit-learn python environment [79]. First, the 1988 genotyped individuals from the pools were shuffled, then 70% of them ( $N = 1391$ ) were used as training data. The remaining 30% ( $N = 597$ ), together with the additional newly collected individuals ( $N = 1908$ ) were considered as out-of-sample data ( $N_{\text{Total}} = 2505$ ) for subsequent validation (Table 1) [40]. Different supervised ML algorithms were tested, including RandomForest, LogisticRegression, SupportVector Machine (SVM), and Linear SupportVectorClassifier (SVC) (Table S5; detailed information on model selection in [supplementary materials and methods](#)). Briefly, the genotype data was converted to a matrix compatible with machine learning (one-hot encoding) [80]. Class information such as lineage and subspecies of each sample was added to the matrix, which was used to train the different machine learning models to predict the sample ancestry. Linear SVC was one of the best performing models according to average accuracy estimated using cross-validation and was finally selected (Table S5, Figure S5).

After training the Linear SVC model, it was used to classify out-of-sample data ( $N = 2505$ ). Samples were classified according to the highest prediction probability belonging to any of the subspecies. A confusion matrix [81] was created to summarize and visualize the

performance on out-of-sample data for which the true values are known. Each row of the matrix represents the true class, while each column represents the predicted class based on the highest probability for each subspecies. The resulting percentages compare a list of expected values with a list of predictions from the model.

In order for the model to be applied in practical conservation and breeding, we defined a threshold of 90% based on the observed distribution of the prediction probabilities (Figure S3), which are in accordance with values found in bee literature [43, 82]. If the prediction probability for any given sample is less than the threshold of 90%, it is considered “unassigned”, while if it exceeded the threshold it was assigned to the respective subspecies.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-07379-7>.

**Additional file 1:** Honey bee subspecies informative markers. SNP information and genomic position of the 4165 SNPs selected in this study to differentiate European honey bee subspecies.

### Additional file 2: Supplementary materials and methods.

Supplementary materials and methods describing in detail the datasets used, the laboratory methods, the bioinformatic pipeline, the SNP selection approach, and the sample classification using Machine Learning algorithms.

**Additional file 3: Figure S1.** PCA-Plots with the PCA-selected SNPs and 100 simulated individuals based on allele frequencies of the pools. **(A)** Using 300 SNPs, the evolutionary lineages M, C and O were well separated with the first two PCs, while lineage A can be differentiated with the third component (not shown). **(B)** Within the M-lineage, harboring only two European subspecies (*A. m. mellifera* and *A. m. iberiensis*), the first PC using 200 SNPs already contributes sufficient information. **(C)** In the O-lineage, four subspecies are represented that are separated with 600 SNPs using 4 PCs. **(D)** For lineage C that contains numerous subspecies (subspecies complex), 1100 SNPs were selected to obtain a better resolution.

**Additional file 4: Figure S2.** Visualization using a t-SNE manifold plot of the 1988 honey bee samples from the pool sequencing individually genotyped for 4094 SNPs. Samples have been color-coded according to the pool name which represents subspecies and country of origin as listed in Table 1.

**Additional file 5: Figure S3.** Histogram of the prediction probabilities for the out-of-sample data. The 90% assignment threshold includes 2098 out of the 2505 samples (=84%).

**Additional file 6: Figure S4.** Confusion matrix for out-of-sample data prediction of evolutionary lineage: African lineage (A), Central and Eastern European lineage (C), Western and Northern European lineage, and Near East and Central Asian lineage (O). Each row of the matrix represents the true class (lineage), while each column represents the predicted class based on the highest prediction probability. The resulting percentages compare a list of expected values with a list of predictions from the model.

**Additional file 7: Figure S5.** Learning curve for the best performing model, the Linear SVC, with average and standard deviation of 10 fold cross validation scores.

**Additional file 8: Table S1.** Additional sampling information, references and sample origin of pool sequencing samples.

**Additional file 9: Table S2.** Pool sequencing and variants statistics.

**Additional file 10: Table S3.** Distribution of selected informative SNPs across the honey bee genome.

**Additional file 11: Table S4.** Average call rate per sample for each subspecies.

**Additional file 12: Table S5.** Accuracy statistics for the different tested models estimated using 10-fold cross validation.

## Abbreviations

DNA: Deoxyribonucleic acid; FTA: Flinders Technology Associates; ICZN: International Code of Zoological Nomenclature; ML: Machine learning; mtDNA: Mitochondrial deoxyribonucleic acid; PCA: Principal component analysis; PCR: Polymerase chain reaction; SD: Standard deviation; SNP: Single-nucleotide polymorphism; SVC: Support Vector Classifier; SVM: Support Vector Maschine; t-SNE: t-Distributed stochastic neighbor embedding

## Acknowledgements

We are especially grateful to all beekeepers, bee breeders and other contributors who provided the valuable samples for this work; in representation we express our gratitude to Timothea Ioannou from the Department of Agriculture, Limassol, Cyprus. We also wish to give a special thanks to all technicians involved in the project, in particular to Mahesha Perera for her great contribution in the laboratory to process the many samples. Finally, we thank our SmartBees colleagues for valuable discussions and the COLOSS research association for providing a networking platform.

## Authors' contributions

AE, MDM, PK, MB, & RV designed the study. JM, RON, & JL analyzed the sequencing data and selected SNPs with input from IM & MP. RON and JM designed array probes and created the classification model. MP and JM wrote the manuscript with major contributions from AE, MDM, PK, & MB. CB was responsible for the sequencing of the samples. LF contributed to the sequencing and bioinformatic analysis. LP contributed to the population analysis. SMARTBEES WP3 DIVERSITY COLLABORATORS: EC, JDC, MFC, CC, RD, PDLR, MMD, JF, TG, MG, AG, KG, FH, RI, EI, IJ, IK, AK, MK, NK, ESM, DM, RM, AGN, AP, PP, MAP, AVP, AYS, AS, MIS, AU & MZM contributed to the design of the research by developing local sampling protocols, observing the need for unrelated bees, covering the whole range, and avoiding bees that may originate from imported populations. Whenever possible, they provided samples based on previous genetic analyses in order to ensure the suitability of the samples submitted (Table S1). All authors substantively reviewed the manuscript contributing with important comments that much improved the manuscript and approved the final version.

## Funding

The SmartBees project was funded by the European Commission under its FP7 KBBE programme (2013.1.3–02, SmartBees Grant Agreement number 613960) <https://ec.europa.eu/research/fp7>. MP was supported by a Basque Government grant (IT1233–19). The funders provided the financial support to the research, but had no role in the design of the study, analysis, interpretations of data and in writing the manuscript.

## Availability of data and materials

All sequence data from the pools analyzed during the current study have been submitted to the NCBI Short Read Archive (SRA) under the BioProject accession number PRJNA666033: <https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA666033>. The pipeline for the analysis of the pool sequence data is available at [https://github.com/jlangua/smsk\\_population](https://github.com/jlangua/smsk_population).

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

JM, RON and RV were GenoScan employees at the time the project was designed. GenoScan, now owned by Eurofins Genomics, was an SME project partner in the SmartBees project. At present, JM is a bioinformatician at Eurofins Genomics, who is the genotyping service provider of the SNP chip presented in this study.

## Author details

<sup>1</sup>Eurofins Genomics Europe Genotyping A/S (EFEG), (Former GenoScan A/S), Aarhus, Denmark. <sup>2</sup>Laboratory Genetics, University of the Basque Country (UPV/EHU), Leioa, Bilbao, Spain. <sup>3</sup>Swiss Bee Research Center, Agroscope, Bern, Switzerland. <sup>4</sup>Laboratory of Agricultural Zoology and Entomology, Agricultural University of Athens, Athens, Greece. <sup>5</sup>Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. <sup>6</sup>Institutul de Cercetare Dezvoltare pentru Apicultura SA, Bucharest, Romania. <sup>7</sup>University of Limerick, Limerick, Ireland. <sup>8</sup>CREA Research Centre for Agriculture and Environment, Bologna, Italy. <sup>9</sup>BeeSources, Bologna, Italy. <sup>10</sup>Veterinary Faculty, University of Murcia, Murcia, Spain. <sup>11</sup>Croatian Ministry of Agriculture, Zagreb, Croatia. <sup>12</sup>Department of Ecology, Agronomy and Aquaculture, University of Zadar, Zadar, Croatia. <sup>13</sup>Breeds of Origin, Haz-Zebbug, Malta. <sup>14</sup>MacBee Association, Skopje, North Macedonia. <sup>15</sup>Faculty of Agriculture and Life Sciences, University of Maribor, Maribor, Slovenia. <sup>16</sup>Yerevan State University, Yerevan, Armenia. <sup>17</sup>Department of Apiculture, Agricultural Organization 'DEMETER', Thessaloniki, Greece. <sup>18</sup>Division of Life Sciences, Major of Biological Sciences, and Convergence Research Center for Insect Vectors, Incheon National University, Incheon, Korea. <sup>19</sup>Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia. <sup>20</sup>University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria. <sup>21</sup>Agricultural University of Georgia, Tbilisi, Georgia. <sup>22</sup>Ankara University, Ankara, Turkey. <sup>23</sup>Federation of Greek Beekeepers' Associations, Larissa, Greece. <sup>24</sup>Düzce University, Düzce, Turkey. <sup>25</sup>University of Zagreb, Zagreb, Croatia. <sup>26</sup>Hungarian Bee Breeders Association, Budapest, Hungary. <sup>27</sup>Division of Rural Sciences and Food Systems, Institute of Earth Systems, University of Malta, Msida, Malta. <sup>28</sup>Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH, Wien, Austria. <sup>29</sup>Cyprus University of Technology, Limassol, Cyprus. <sup>30</sup>Agricultural University of Plovdiv, Plovdiv, Bulgaria. <sup>31</sup>Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Bragança, Portugal. <sup>32</sup>Shulgan-Tash Nature Reserve, Burzyansky District, Russia. <sup>33</sup>Tekirdag University, Tekirdag, Turkey. <sup>34</sup>Landesbetrieb Landwirtschaft Hessen, Bee Institute Kirchhain, Kirchhain, Germany. <sup>35</sup>Faculty of Agricultural Sciences and Food, University Ss. Cyril and Methodius, Skopje, Republic of Macedonia. <sup>36</sup>Department of Physiology and Biochemistry, University of Malta, Msida, Malta. <sup>37</sup>Department of Agroecology, Aarhus University, Slagelse, Denmark.

Received: 29 May 2020 Accepted: 8 January 2021

Published online: 03 February 2021

## References

- Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE. Global pollinator declines: trends, impacts and drivers. *Trends Ecol Evol*. 2010;25:345–53.
- IPBES. Summary for policymakers of the assessment report of the intergovernmental science-policy platform on biodiversity and ecosystem services on pollinators, pollination and food production. Bonn: Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services; 2016.
- Dogantzis KA, Zayed A. Recent advances in population and quantitative genomics of honey bees. *Curr Opin Insect Sci*. 2019;31:93–8.
- 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 sinixinyuan* n. ssp. *Mol Biol Evol*. 2016;33:1337–48.
- Sheppard WS, Meixner MD. *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. *Apidologie*. 2003;34:367–75.
- Meixner MD, Leta MA, Koeniger N, Fuchs S. The honey bees of Ethiopia represent a new subspecies of *Apis mellifera*—*Apis mellifera simensis* n. ssp. *Apidologie*. 2011;42(3):425–37.
- Cridland JM, Tsutsui ND, Ramírez SR. The complex demographic history and evolutionary origin of the western honey bee, *Apis mellifera*. *Genome Biol Evol*. 2017;9(2):457–72.
- Ruttner F. Biogeography and taxonomy of honeybees. Berlin: Springer Verlag; 1988.
- 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.
- 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. *J Apic Res*. 2014;53(2):269–78.
- Bouga M, Harizan PC, Kilias G, Alahiotis S. Genetic divergence and phylogenetic relationships of honey bee *Apis mellifera* (Hymenoptera: Apidae) populations from Greece and Cyprus using PCR–RFLP analysis of three mtDNA segments. *Apidologie*. 2005;36(3):335–44.
- Dall'olio R, Marino A, Lodesani M, Moritz RF. Genetic characterization of Italian honeybees, *Apis mellifera ligustica*, based on microsatellite DNA polymorphisms. *Apidologie*. 2007;38(2):207–17.
- Büchler R, Costa C, Hatjina F, Andonov S, Meixner MD, Conte YL, et al. The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera* L. colonies in Europe. *J Apic Res*. 2014;53(2):205–14.
- Frankham R, Ballou JD, Briscoe DA. Introduction to conservation genetics. Cambridge: Cambridge University Press; 2002.
- 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*. *Front Ecol Evol*. 2016;4:140.
- Bertrand B, Alburaki M, Legout H, Moulin S, Mougé F, Garnery L. MtDNA COI–COL marker and drone congregation area: an efficient method to establish and monitor honeybee (*Apis mellifera* L.) conservation centres. *Mol Ecol Resour*. 2015;15:673–83.
- Büchler R, Uzunov A. Selecting for Varroa resistance in German honey bees. *Bee World*. 2016;93(2):49–52.
- Uzunov A, Brascamp EW, Büchler R. The basic concept of honey bee breeding programs. *Bee World*. 2017;94(3):84–7.
- Baudry E, Solignac M, Garnery L, Gries M, Cornuet J, Koeniger N. Relatedness among honeybees (*Apis mellifera*) of a drone congregation. *Proc R Soc Lond B Biol Sci*. 1998;265(1409):2009–14.
- Tarpy DR, Delaney DA, Seeley TD. Mating frequencies of honey bee queens (*Apis mellifera* L.) in a population of feral colonies in the northeastern United States. *PLoS One*. 2015;10(3):e0118734.
- Büchler R, Andonov S, Bienefeld K, Costa C, Hatjina F, Kezic N, et al. Standard methods for rearing and selection of *Apis mellifera* queens. *J Apic Res*. 2013;52(1):1–30.
- Bouga M, Alaux C, Bienkowska M, Büchler R, Carreck NL, Cauia E, et al. A review of methods for discrimination of honey bee populations as applied to European beekeeping. *J Apic Res*. 2011;50(1):51–84.
- Lodesani M, Costa C. Bee breeding and genetics in Europe. *Bee World*. 2003;84(2):69–85.
- Muñoz I, Pinto MA, De la Rúa P. Temporal changes in mitochondrial diversity highlights contrasting population events in Macaronesian honey bees. *Apidologie*. 2013;44(3):295–305.
- Miguel I, Garnery L, Iriondo M, Baylac M, Manzano C, Steve Sheppard W, et al. Origin, evolution and conservation of the honey bees from La Palma Island (Canary Islands): molecular and morphological data. *J Apic Res*. 2015;54(5):427–40.
- Soares S, Grazina L, Mafra I, Costa J, Pinto MA, Oliveira MBP, et al. Towards honey authentication: differentiation of *Apis mellifera* subspecies in European honeys based on mitochondrial DNA markers. *Food Chem*. 2019;283:294–301.
- Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet*. 2003;73:1402–22.
- Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintrón W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet*. 2007;3:1672–86.
- Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P, Drineas P. Tracing cattle breeds with principal components analysis ancestry informative SNPs. *PLoS One*. 2011;6(4):e18007.
- Heaton MP, Leymaster KA, Kalbfleisch TS, Kijas JW, Clarke SM, McEwan J, et al. SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS One*. 2014;9(4):e94851.
- Montes I, Conklin D, Albaina A, Creer S, Carvalho GR, Santos M, et al. SNP discovery in European anchovy (*Engraulis encrasicolus*, L.) by high-throughput transcriptome and genome sequencing. *PLoS One*. 2013;8(8):e70051.
- Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, et al. Evaluation of approaches for identifying population informative markers from high density SNP chips. *BMC Genet*. 2011;12(1):45.
- Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsma C, et al. Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*. 2011;12(1):622.

34. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F<sub>ST</sub>. *Nat Rev Genet*. 2009;10(9):639.
35. Montes I, Laconcha U, Iriando M, Manzano C, Arizabalaga H, Estonba A. Reduced single nucleotide polymorphism panels for assigning Atlantic albacore and Bay of Biscay anchovy individuals to their geographic origin: toward sustainable fishery management. *J Agric Food Chem*. 2017;65(21):4351–8.
36. Liu N, Zhao H. A non-parametric approach to population structure inference using multilocus genotypes. *Hum Genomics*. 2006;2:353–64.
37. Paetkau D, Calvert W, Stirling I, Strobeck C. Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol*. 1995;4(3):347–54.
38. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. GENECLAS S2: a software for genetic assignment and first-generation migrant detection. *J Hered*. 2004;95(6):536–9.
39. Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet*. 2018;34(4):301–12.
40. Tarca AL, Carey VJ, Chen XW, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS Comput Biol*. 2007;3(6):116.
41. Guinand B, Topchy A, Page KS, Burnham-Curtis MK, Punch WF, Scribner KT. Comparisons of likelihood and machine learning methods of individual classification. *J Hered*. 2002;93(4):260–9.
42. Muñoz I, Henriques D, Johnston JS, Chávez-Galarza J, Kryger P, Pinto MA. Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLoS One*. 2015;10:e0124365.
43. Parejo M, Henriques D, Pinto MA, Soland-Reckeweg G, Neuditschko M. Empirical comparison of microsatellite and SNP markers to estimate introgression in *Apis mellifera mellifera*. *J Apic Res*. 2018;57(4):504–6.
44. Garnery L, Franck P, Baudry E, Vautrin D, Cornuet JM, Solignac M. Genetic diversity of the west European honey bee (*Apis mellifera mellifera*) and A. M Iber II Microsatellite. *Loci Genet Sel Evol*. 1998;30:49–74.
45. Henriques D, Parejo M, Vignal A, Wragg D, Wallberg A, Webster MT, et al. Developing reduced SNP assays from whole-genome sequence data to estimate introgression in an organism with complex genetic patterns, the Iberian honeybee (*Apis mellifera iberiensis*). *Evol Appl*. 2018;11(8):1270–82.
46. Chapman NC, Harpur BA, Lim J, Rinderer TE, Allsopp MH, Zayed A, et al. A SNP test to identify Africanized honeybees via proportion of 'African' ancestry. *Mol Ecol Resour*. 2015;15(6):1346–55.
47. Rortais A, Arnold G, Alburaki M, Legout H, Garnery L. Review of the Dral COI-COI test for the conservation of the black honeybee (*Apis mellifera mellifera*). *Conserv Genet Resour*. 2011;3:383–91.
48. Jensen AB, Palmer KA, Boomsma JJ, Pedersen BV. Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in Northwest Europe. *Mol Ecol*. 2005;14:93–106.
49. Harpur BA, Chapman NC, Krimus L, Maciukiewicz P, Sandhu V, Sood K, et al. Assessing patterns of admixture and ancestry in Canadian honey bees. *Insect Soc*. 2015;62(4):479–89.
50. Chapman NC, Bourgeois AL, Beaman LD, Lim J, Harpur BA, Zayed A, et al. An abbreviated SNP panel for ancestry assignment of honeybees (*Apis mellifera*). *Apidologie*. 2017;48(6):776–83.
51. Platzer A. Visualization of SNPs with t-SNE. *PLoS One*. 2013;8(2):56883.
52. Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS, et al. Thrive out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. *Science*. 2006;314:642–5.
53. Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat Genet*. 2014;46(10):1081–8.
54. Franck P, Garnery L, Celebrano G, Solignac M, Cornuet JM. Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). *Mol Ecol*. 2000;9:907–21.
55. Uzunov A, Meixner MD, Kiprijanovska H, Andonov S, Gregorc A, Ivanova E, et al. Genetic structure of *Apis mellifera macedonica* in the Balkan peninsula based on microsatellite DNA polymorphism. *J Apic Res*. 2014;53:288–95.
56. Nedić N, Francis RM, Stanisavljević L, Pihler I, Kezić N, Bendixen C, et al. Detecting population admixture in honey bees of Serbia. *J Apic Res*. 2014; 53(2):303–13.
57. Yang ZR. Biological applications of support vector machines. *Brief Bioinform*. 2004;5(4):328–38.
58. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19:1655–64.
59. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–59.
60. Foti N, Lungu M, Pelimon C, Barac I, Copaitici M, Marza E. Researches on morphological characteristics and biological features of the bee population in Romania. In: *Proceedings of XXth Jubiliar International Congress of Beekeeping Apimondia*; 1965. p. 171–176.
61. Petrov P. Systematics of Bulgarian bees. *Pchelarstvo*. 1991;9:15–7.
62. Engel MS. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae: Apis). *J Hymenopt Res*. 1999;8:165–96.
63. Ivanova E, Bouga M, Staykova T, Mladenovic M, Rasic S, Charistos L, et al. The genetic variability of honey bees from the southern Balkan Peninsula, based on alloenzymic data. *J Apic Res*. 2012;51(4):329–35.
64. Garnery L, Cornuet JM, Solignac M. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Mol Ecol*. 1992;1:145–54.
65. Francis RM, Kryger P, Meixner M, Bouga M, Ivanova E, Andonov S, et al. The genetic origin of honey bee colonies used in the COLOSS genotype-environment interactions experiment: a comparison of methods. *J Apic Res*. 2014;53(2):188–204.
66. Meixner MD, Pinto MA, Bouga M, Kryger P, Ivanova E, Fuchs S. Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. *J Apic Res*. 2013;52:1–28.
67. Muñoz I, Henriques D, Jara L, Johnston JS, Chávez-Galarza J, De La Rúa P, et al. SNP s selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (*Apis mellifera mellifera*). *Mol Ecol Resour*. 2017;17(4):783–95.
68. Jenks DT. The convention on biological diversity—an efficient framework for the preservation of life on earth. *Nw J Intl Bus*. 1994;15:636.
69. Sheppard WS, Arias MC, Grech A, Meixner MD. *Apis mellifera ruttneri*, a new honey bee subspecies from Malta. *Apidologie*. 1997;28(5):287–93.
70. Chávez-Galarza J, Henriques D, Johnston JS, Azevedo JC, Patton JC, Muñoz I, et al. Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Mol Ecol*. 2013;22(23):5890–907.
71. Miguel I, Iriando M, Garnery L, Sheppard WS, Estonba A. Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie*. 2007;38(2):141–55.
72. Ilyasov RA, Poskryakov AV, Petukhov AV, Nikolenko AG. Molecular genetic analysis of five extant reserves of black honeybee *Apis mellifera mellifera* in the Urals and the Volga region. *Russ J Genet*. 2016;52(8):828–39.
73. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw*. 2008;25(1):1–18.
74. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984;38:1358–70.
75. Kofler R, Pandey RV, Schlötterer C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*. 2011;27(24):3435–6.
76. Büchler R, Uzunov A, Hoppe A, Bienefeld K. Field testing and selection on European honey bee populations (Smartbees project 2015–2018). In: *Abstract book of the 8th EurBee congress of Apidology*. Ghent: Ghent University; 2018.
77. Farajzadeh L, Wegener J, Momeni J, Nielsen O, Bienefeld K, Bendixen C. Whole-genome analysis of uncapping behaviour of individual honey bees towards Varroa destructor-parasitized brood. In: *Proceedings of the 46th International Apicultural Congress*. Montreal: Apimondia; 2019. p. 8–12.
78. Maaten LVD, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–605.
79. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
80. Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. *Bioinformatics*. 2018;34(15):2642–8.
81. Talbot J, Lee B, Kapoor A, Tan DS. EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM; 2009. p. 1283–1292.
82. Henriques D, Browne KA, Barnett MW, Parejo M, Kryger P, Freeman TC, et al. High sample throughput genotyping for estimating C-lineage introgression in the dark honeybee: an accurate and cost-effective SNP-based tool. *Sci Rep*. 2018;8:8552.

## Publisher's Note

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



## Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

[onlineservice@springernature.com](mailto:onlineservice@springernature.com)