PLOS | ONE

RESEARCH ARTICLE

Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (*Apis mellifera mellifera*)

Irene Muñoz[1], Dora Henriques[1], J. Spencer Johnston[2], Julio Chávez-Galarza[1], Per Kryger[3], M. Alice Pinto[1]*

1 Mountain Research Centre (CIMO), Polytechnic Institute of Bragança, Campus de Sta. Apolónia, Apartado 1172, 5301–855, Bragança, Portugal, **2** Department of Entomology, Texas A&M University, College Station, Texas, 77843–2475, United States of America, **3** Aarhus University, Department of Agroecology, Forsøgsvej 1, 4200, Slagelse, Denmark

* apinto@ipb.pt

Citation: Muñoz I, Henriques D, Johnston JS, Chávez-Galarza J, Kryger P, Pinto MA (2015) Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (*Apis mellifera mellifera*). PLoS ONE 10(4): e0124365. doi:10.1371/journal.pone.0124365

Academic Editor: Wolfgang Blenau, University of Cologne, GERMANY

Received: January 9, 2015

Accepted: March 10, 2015

Published: April 13, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Muñoz was supported by Fundación Séneca (Murcia, Spain) through the Post-doctoral fellowship 19149/PD/13-N whereas Dora Henriques and Julio Chávez-Galarza were supported by Fundação para a Ciência e Tecnologia (Portugal) through the PhD scholarships SFRH/BD/84195/2012 and SFRH/BD/68682/2010, respectively. This research was funded by Fundação para a Ciência e Tecnologia and COMPETE/QREN/EU through the projects PTDC/BIA-BEC/099640/2008 and

Abstract

Beekeeping activities, especially queen trading, have shaped the distribution of honey bee (*Apis mellifera*) subspecies in Europe, and have resulted in extensive introductions of two eastern European C-lineage subspecies (*A. m. ligustica* and *A. m. carnica*) into the native range of the M-lineage *A. m. mellifera* subspecies in Western Europe. As a consequence, replacement and gene flow between native and commercial populations have occurred at varying levels across western European populations. Genetic identification and introgression analysis using molecular markers is an important tool for management and conservation of honey bee subspecies. Previous studies have monitored introgression by using microsatellite, PCR-RFLP markers and most recently, high density assays using single nucleotide polymorphism (SNP) markers. While the latter are almost prohibitively expensive, the information gained to date can be exploited to create a reduced panel containing the most ancestry-informative markers (AIMs) for those purposes with very little loss of information. The objective of this study was to design reduced panels of AIMs to verify the origin of *A. m. mellifera* individuals and to provide accurate estimates of the level of C-lineage introgression into their genome. The discriminant power of the SNPs using a variety of metrics and approaches including the Weir & Cockerham's F_{ST}, an F_{ST}-based outlier test, Delta, informativeness (I_n), and PCA was evaluated. This study shows that reduced AIMs panels assign individuals to the correct origin and calculates the admixture level with a high degree of accuracy. These panels provide an essential tool in Europe for genetic stock identification and estimation of admixture levels which can assist management strategies and monitor honey bee conservation programs.

BiodivERsA/0002/2014. The open access publishing fees for this article have been covered by the Texas A&M University Online Access to Knowledge (OAK) Fund, supported by the University Libraries and the Office of the Vice President for Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The role of introgression and admixture in conservation is a dilemma: While natural admixture may be an important evolutionary force in speciation and maintenance of genetic diversity [1–2], admixture induced by human activities may contribute, either directly or indirectly, to the extinction of many taxa [3]. Introduction of species, subspecies and habitat modifications has caused increased rates of admixture with native flora and fauna and introgression that can generate extinction and irretrievable loss of combinations of genotypes throughout the entire genome [4].

The honey bee, *Apis mellifera* L., represents a valuable model to study human-mediated change. Beekeeping has been practiced in Europe for many centuries [5], which has led to loss of native genetic diversity through three major mechanisms: (i) replacement of native populations by human-selected more docile and productive colonies, (ii) spread of honey bee pests and parasites, such as the mite *Varroa destructor* and the microsporidian *Nosema ceranae*, that have contributed to worldwide population declines [6–7], and (iii) recurrent introductions of commercial colonies (reviewed by De la Rúa et al. [8]).

The genetic diversity harbored in native honey bee subspecies is amongst the most important legacies that we can leave to future generations of beekeepers and farmers [9–10]. Native honey bee subspecies are important reservoirs of local adaptations; their extinction means the loss of unique combinations of traits shaped by natural selection over extended periods of time. These combinations can be important for a more sustainable beekeeping, as shown by a recent pan-European experiment [11].

In Europe, honey bees show considerable differences in morphological, behavioural and biological characters across their range as a result of historical patterns of isolation and adaptation to environmental conditions [8]. Those differences are materialized in 10 extant European subspecies, among the 30 subspecies currently recognized worldwide [12–16], representing thereby a substantial component of the total honey bee diversity. These 10 European subspecies have been grouped by morphological and molecular tools [12, 17–21] into two evolutionary lineages: the M-lineage, in Western Europe, and the C-lineage, in Eastern Europe.

Subspecies-specific genetic footprints can still be identified in Europe [22–28], in spite of centuries of beekeeping [5], although introgression and admixture events have also been detected in eastern [28–30] and western [9, 26, 31–32] European populations. The M-lineage *A. m. mellifera* (dark honey bee) has been recognized as the most threatened, with most of the threat due to introgression from the C-lineage [9, 31–32]. In addition to the documented intentional replacement of *A. m. mellifera* by *A. m. carnica* in Germany [33–34], the increasing trade of commercial breeds (mainly C-lineage *A. m. carnica*, *A. m. ligustica* and the hybrid buckfast) is threatening the genetic integrity of the native *A. m. mellifera* as many beekeepers prefer using commercial as opposed to native honey bees.

Increasing awareness that native honey bee diversity represents a valuable asset for sustainable beekeeping is fuelling local breeding and conservation efforts across Europe. One of the earliest, and until recently the single conservation program enacted by law, is that implemented by the Danish Beekeepers Association and the Læsø Beekeepers Association on behalf of the Danish Government in 1993 and the European Union in 1998 [35] to create a reserve and protect the *A. m. mellifera*. Following approval by the Scottish government of an order to protect the *A. m. mellifera* on the islands of Colonsay and Oronsay [The Bee Keeping (Colonsay and Oronsay) Order 2013], a second European reserve was recently created in the United Kingdom. Other *A. m. mellifera* conservation efforts, although not enacted by law, are underway in France, Holland, Norway, Switzerland, Ireland, and Belgium, among others (see the website "http://www.sicamm.org" run by the International Association for the Protection of the

European Dark bee). The success or failure of all these efforts will be tightly linked to efforts that monitor the integrity of these protected populations.

Assessing introgression is an important activity in honey bee breeding programs, especially when conservation of native subspecies is a major concern. This activity requires molecular tools that are reliable, inexpensive and preferably automated. Previous studies have monitored introgression between the endemic *A. m. mellifera* and introduced C-lineage subspecies using microsatellite and PCR-RFLP markers [31–32, 36]. However, with the publication of the honey bee genome [37], development of single-nucleotide polymorphism (SNP) markers [20, 38], and next generation sequencing becoming fast and affordable, particularly for a small genome as that of the honey bee (236 Mb), increasingly powerful tools are available to measure genomic ancestry and admixture levels occurring in both native and introduced honey bee populations [21, 39–40]. However, the genomic approach is not always cost-effective and low quality and/or degraded DNA can be a handicap to using genomic re-sequencing. Alternatively, ancestry can be estimated using a subset of highly informative SNPs ranging in number from a few dozens to several hundreds. The selected SNPs, commonly known as Ancestry-Informative Markers (AIMs), are those that exhibit large allele frequency differences between populations. AIMs can be used for inferring geographic origin of individuals [41–43], detecting illegal trade and translocation of animals [44], food authentication [45], for estimating overall admixture proportions efficiently and inexpensively [43, 46], among others. It is possible, using a panel of AIMs distributed throughout the genome, to estimate the relative ancestral proportions in admixed individuals, and infer the time since the admixture process [47–48].

The ability of an AIMs panel to measure ancestry is generally evaluated empirically, by examining its performance on a given set of samples for which ancestry is known [49]. In this paper, we employed five analytical methods to select different combinations of SNPs to form five nested panels of 48-, 96-, 144-, 192- and 384-AIMs optimized to estimate admixture proportions of C-lineage (*A. m. ligustica* and *A. m. carnica*) into the M-lineage *A. m. mellifera*. This was done in two successive stages. In the first stage, we evaluated the performance of the five selection methods [Weir & Cockerham's F_{ST}, an F_{ST}-based outlier test, Delta, informativeness (I_n), and PCA] on a training dataset, in an effort to select AIMs and to rank them by decreasing level of informativeness. In the second stage, we tested the power of the reduced five designed panels and validated their performance on holdout and simulated sets, by comparing the admixture estimates produced by the panels with those produced by an initial dataset of 1183 SNPs.

Material and Methods

Samples, DNA Extraction and SNP Genotyping

A total of 113 honey bee haploid males were collected in 2010 and 2011 across the native range of *A. m. mellifera*, *A. m. ligustica* and *A. m. carnica* in Europe (see the sampling map in Pinto et al. [9]). The samples of *A. m. mellifera* (N = 77) were collected from apiaries located in England (N = 8), France (N = 15), Belgium (N = 3), Denmark (N = 10), Holland (N = 15), Switzerland (N = 6), Scotland (N = 10), and Norway (N = 10) from protected and unprotected populations [9]. Colonies of protected populations have been identified by morphological (B. Dahle, pers. comm.) and molecular tools (mtDNA $tRNA^{leu}$-cox2 and microsatellites; [31, 32, 50–51]) as the best representatives of *A. m. mellifera* and have therefore been integrated into conservation programs. To prevent C-lineage introgression and assure pure breeding, these colonies have been maintained in islands or in isolated mating stations. Despite careful management to protect the threatened *A. m. mellifera* from C-lineage introgression, a recent SNP survey detected variable, although generally low, levels of introgression in these protected

populations (see Pinto et al. [9] for details). A reference collection of 36 samples representing C-lineage diversity was obtained from the natural range of *A. m. carnica* in Serbia (N = 8) and Croatia (N = 11) and from the natural range *A. m. ligustica* in Italy (N = 17). The owners of all the sampled apiaries gave permission to collect honey bee individuals from the hives. In each location, samples were taken from the inner part of hives, placed into absolute ethanol and stored at −20°C until molecular analysis.

Using a phenol/chloroform isoamyl alcohol (25:24:1) protocol [52], total DNA was extracted from the thorax of the 113 individuals, each representing a single colony. A total of 1536 SNP loci were genotyped for those individuals using Illumina's BeadArray Technology and the Illumina GoldenGate Assay with a custom Oligo Pool Assay (Illumina, San Diego, CA, USA) following manufacturer's protocols. The Oligo Pool consisted of the 1536 SNPs, which included the 768 most informative SNPs of Whitfield et al. [20] and 768 newly developed SNPs employed by Chávez-Galarza et al [38]. The 1536 SNP array was used previously to study diversity and introgression levels in populations of *A. m. mellifera* sampled across Western Europe [9] and to detect signatures of selection in the Iberian honey bee genome [38]. Genotype calling was performed using Illumina's GenomeStudio Data Analysis software. Of the initial 1536 SNPs, 353 did not meet the quality criteria for analysis and were therefore excluded from the dataset. The SNP filtering was as follows: 124 exhibited poorly separated intensity clusters or low signal intensity when visualized in the GenomeStudio software; 167 were monomorphic (defined by a cut-off criterion of >0.98 for the most common allele, as in Chávez-Galarza et al. [38]) across all populations; 54 did not map in the honey bee genome assembly Amel_4.0; and 8 hit two different genomic positions (the first with 100% identity and the second with 96–98%) in the honey bee genome assembly Amel_4.0 during the mapping process using the 100 bp flanking sequence. Allele frequencies were calculated for each of the remaining 1183 bi-allelic SNPs (S1 Table) in each population using the program Plink [53].

Selection of AIMs

Five different methods were employed on the initial 1183 SNP dataset for estimating marker information content. The first method, which has been one of the most popular for selecting informative loci, was the pairwise F_{ST} of Weir & Cockerham [54] as calculated at each locus using Genepop software [55]. The second method was the F_{ST}–based outlier test developed by Foll & Gaggiotti [56], which employs a Bayesian likelihood approach to detect loci deviating from neutral expectations (outliers). This outlier test was implemented in Bayescan 2.01 [56] using 20 pilot runs of 5 000 iterations (sample size of 5 000 and thinning interval of 10) and an additional burn-in of 50 000 iterations. The third method was based on the estimate of allele-frequency differential (Delta), which is one of the most straightforward ways to evaluate the information content of a SNP. For a bi-allelic marker, like a SNP, the Delta value is estimated as $|pA_i—pA_j|$, where pAi and pA_j are the frequencies of allele A in the i^{th} and j^{th} populations, respectively. When more than two populations were analyzed, the Delta value for each SNP locus was estimated as the mean across all pair-wise comparisons. The fourth method was the informativeness for assignment (I_n, natural logarithm of the number of populations) proposed by Rosenberg et al. [41]. I_n provides the amount of information gained about population assignment from observation of a single randomly chosen allele at a locus. This method assumes a uniform prior across K potential source populations for the origin of the allele. For a given set of populations, the minimum value of I_n (0) occurs when all alleles have equal frequencies in all populations whereas the maximum value (1) occurs when alleles are not shared among populations. I_n was calculated using the software Infocalc available at http://www.stanford.edu/group/rosenberglab/infocalc.html. Finally, the fifth selection method was principal component

analysis (PCA), which was performed using the PAST software [57]. The first eight principal components were used to calculate the information content of each SNP following the approach of Paschou et al. [58]. The loadings for each SNP were squared and summed over the eight most significant principal components to produce an estimate of informativeness.

SNPs were ranked and panels of SNPs tested using reference populations and the Anderson's Simple Training and Holdout method to reduce the potential for upward bias, which is introduced when loci are ranked and assessed using the same individuals [59]. To that end, a total of 34 pure (*sensu* Soland-Reckeweg et al. [32]) individuals of *A. m. mellifera*, previously identified in Pinto et al. [9], and all reference individuals (17 *A. m. ligustica* and 19 *A. m. carnica*) were used for SNP ranking (training set = 70) and the remaining 43 individuals of *A. m. mellifera* were reserved for panel testing (holdout set = 113). To minimize the effect of clusters of populations on the selection of the AIMs [41, 60–61], the five selection methods were tested using four training datasets. The first dataset consisted of 70 individuals: 34 pure *A. m. mellifera* and 36 C-lineage individuals, with no distinction between the *A. m. carnica* and *A. m. ligustica* subspecies (dataset I). The second dataset consisted of 51 individuals: 34 pure *A. m. mellifera* and 17 *A. m. ligustica* (dataset II). The third dataset consisted of 53 individuals: 34 pure *A. m. mellifera* and 19 *A. m. carnica* (dataset III). Finally, the fourth dataset consisted of 70 individuals: 34 pure *A. m. mellifera*, 17 *A. m. ligustica* and 19 of *A. m. carnica* (dataset IV).

Ranking of SNPs

The five selection methods were implemented on the four training datasets producing a total of 20 information content values for each of the 1183 SNPs. These values were ranked and analyzed individually and then were averaged in two steps to obtain a single global value per SNP. In the first step the information content values were averaged across the four training datasets for each of the five selection methods. In the second step the information content values produced by each selection method were converted into a 0–1 scale and then averaged to obtain a global score for each of the 1183 SNPs. After standardizing the values produced by the five selection methods, the global ranking was obtained for the 1183 SNPs using the global score. Given that linked loci yield redundant information, having therefore similar resolving power, markers were excluded if they were within a predefined genetic distance (<1 cM) of higher ranking selected SNPs. The genetic distance of the remaining SNPs ranged from 1.01 to 24.25 cM with a mean of 4.64 cM. Prior to obtaining the global score for each SNP, pairwise associations between information content values produced by the five methods and between the four training datasets were calculated using the Spearman`s rank correlation coefficient, in order to compare the five selection methods and examine the effect of clusters of populations.

Panel Testing

Five panels of 48-, 96-, 144-, 192- and 384-SNPs (sets defined by multiplex sizes of commercial assays) were designed from the top-ranked SNPs. These nested panels were tested against a holdout set and a simulated set to obtain the admixture proportions estimated by each SNP panel. The holdout set (113 individuals) consisted of 34 pure individuals plus 43 reserved individuals of *A. m. mellifera* and the reference *A. m. ligustica* (17 individuals) and *A. m. carnica* (19 individuals), as described above. The simulated set (1000 individuals) was generated with the program ONCOR [62] using the function "simulate a single mixture". Ten populations, each with 100 simulated genotypes, were simulated using different levels of introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%).

Two approaches were used to validate the five reduced AIMs panels. First, a PCA was performed with SNPs in each AIMs panel on the holdout set using the software PAST to generate

two-dimensional PCA and to visualize the stability of population assignment produced by the panels. Second, ancestry and admixture was analyzed. Admixture proportions were estimated with SNPs in each AIMs panel for the holdout and simulated sets using a model-based maximum likelihood estimation of individual ancestries implemented in the software Admixture v1.23 [63]. Coancestry spanning 1–6 populations (K = 1–6, using the default termination criterion that stops the runs when the log-likelihood increases by less than $\varepsilon < 0.0001$ between iterations) was explored for each AIMs panel and the optimal K was identified with the inferred number of populations producing the lowest cross-validation error (CV) during the clustering analysis.

The performance of each reduced panel was examined using different approaches. First, the pairwise differences between admixture proportions inferred from the initial 1183 SNP dataset and the five panels were tested using a Mann-Whitney test. Second, the precision of each panel was tested against the initial 1183 SNP dataset by calculating linear regression coefficients (r^2) and the standard deviations of the differences between admixture proportions. Finally, the accuracy of the reduced panels was estimated via percentage of absolute error of admixture estimates obtained with the five panels in relation to the initial 1183 SNP dataset.

Results

Identification and Ranking of AIMs

The majority of the 1183 SNPs assessed in this study using five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) contain high levels of information content (Fig 1, S2 Table), facilitating the design of reduced panels for genetic identification and introgression analysis in the dark honey bee, *A*. *m*. *mellifera*. The distribution of frequency histograms and percentiles of genetic information content of the 1183 SNPs estimated by each selection method and training dataset are shown in Fig 1. The 50th percentile ranges of the four training datasets were 0.6974–0.7712, 0.5459–0.6362, 0.5532–0.7601, 0.3345–0.3583 and 0.0038–0.0040 for the Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, respectively, indicating a high level of information content for most SNPs and a similar pattern among the four training datasets (Fig 1).

The level of similarity (Spearman's rank correlation, r_s) between the different estimates of genetic information content produced by the five selection methods across the four training datasets is shown in Table 1. The highest correlation values were observed for Weir & Cockerham's F_{ST}, Delta and I_n ($0.7648 \leq r_s \leq 0.9985$, $P<0.001$) whereas a moderate correlation was detected between the F_{ST}-based outlier test and Weir & Cockerham's F_{ST}, Delta and I_n ($0.2864 \leq r_s \leq 0.6592$, $P<0.001$). The lowest correlations were observed between PCA and the other four methods ($-0.2228 \leq r_s \leq 0.1025$, $0.000 \leq P \leq 0.9412$). Regarding the four training datasets (Fig 2), high correlation values were observed across selection methods ($0.7557 \leq r_s \leq 0.9727$, $P<0.001$).

Using an information content cutoff value ≥ 0.25, which indicates very great genetic differentiation [64], a total of 627 AIMs were identified by the methods of Weir & Cockerham's F_{ST}, Delta, I_n, and F_{ST}-based outlier test. Of these, the top-ranked 384 AIMs were selected using the five methods and the four training datasets. The extent of overlap of the 384 AIMs across the five selection methods and the four training datasets is shown in Fig 2. Overlap between any two methods and across datasets ranged between 382 (Weir & Cockerham's F_{ST} and Delta for dataset I; Fig 2A) and 134 (Delta and PCA for dataset III; Fig 2C). The number of AIMs that were simultaneously selected by the five methods was lower, ranging from 82 (dataset I; Fig 2A) to 97 (dataset IV; Fig 2D). A substantially higher amount of overlap (273 AIMs; Fig 2E), supported by high correlation values ($r_s \geq 0.7557$, $P<0.001$; Fig 2F), was observed across the

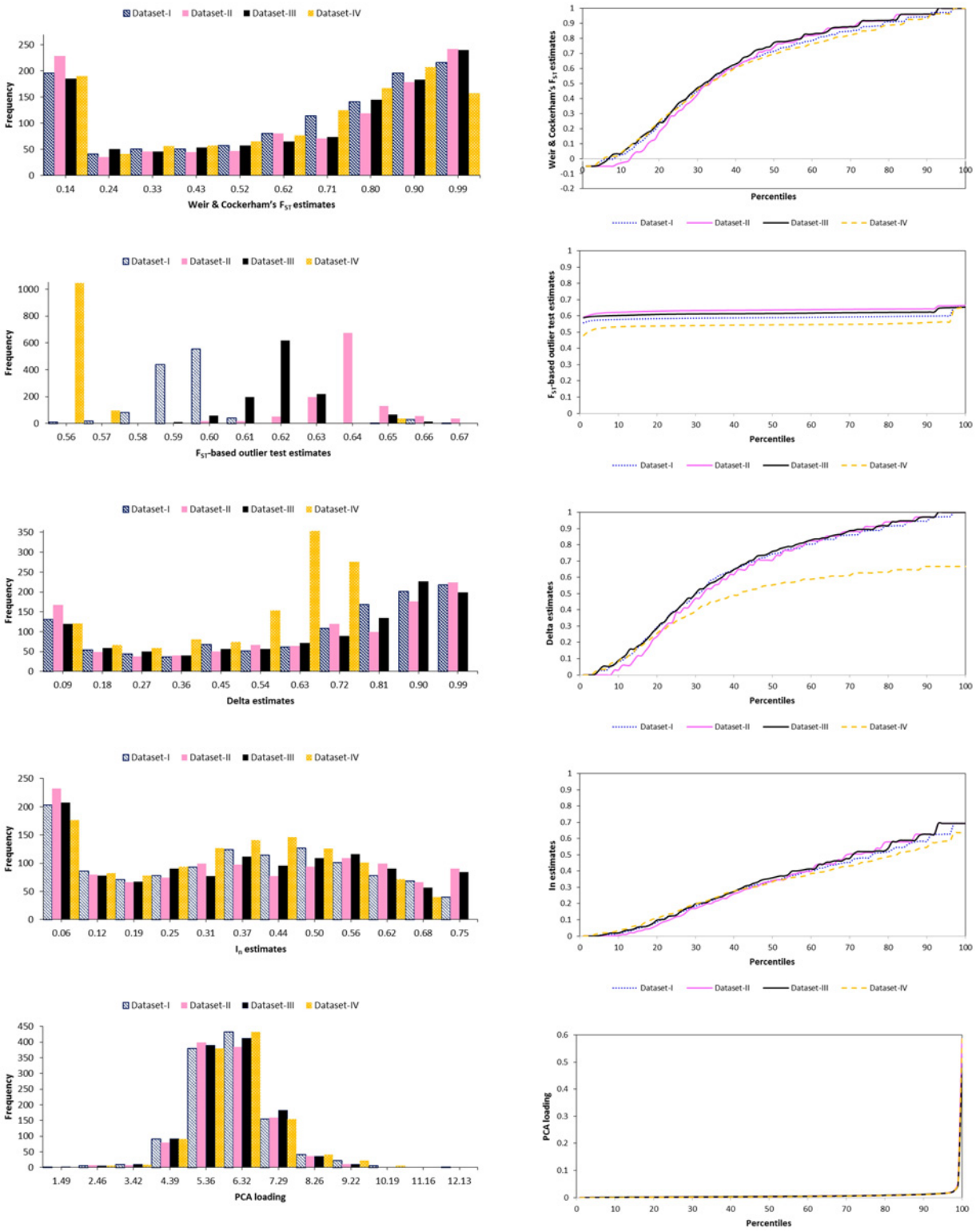


Fig 1. Frequency histograms and percentiles of the estimates of genetic information contained in the initial 1183 SNP dataset. Information content produced by the five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) is shown for the four training datasets (I, II, III and IV).

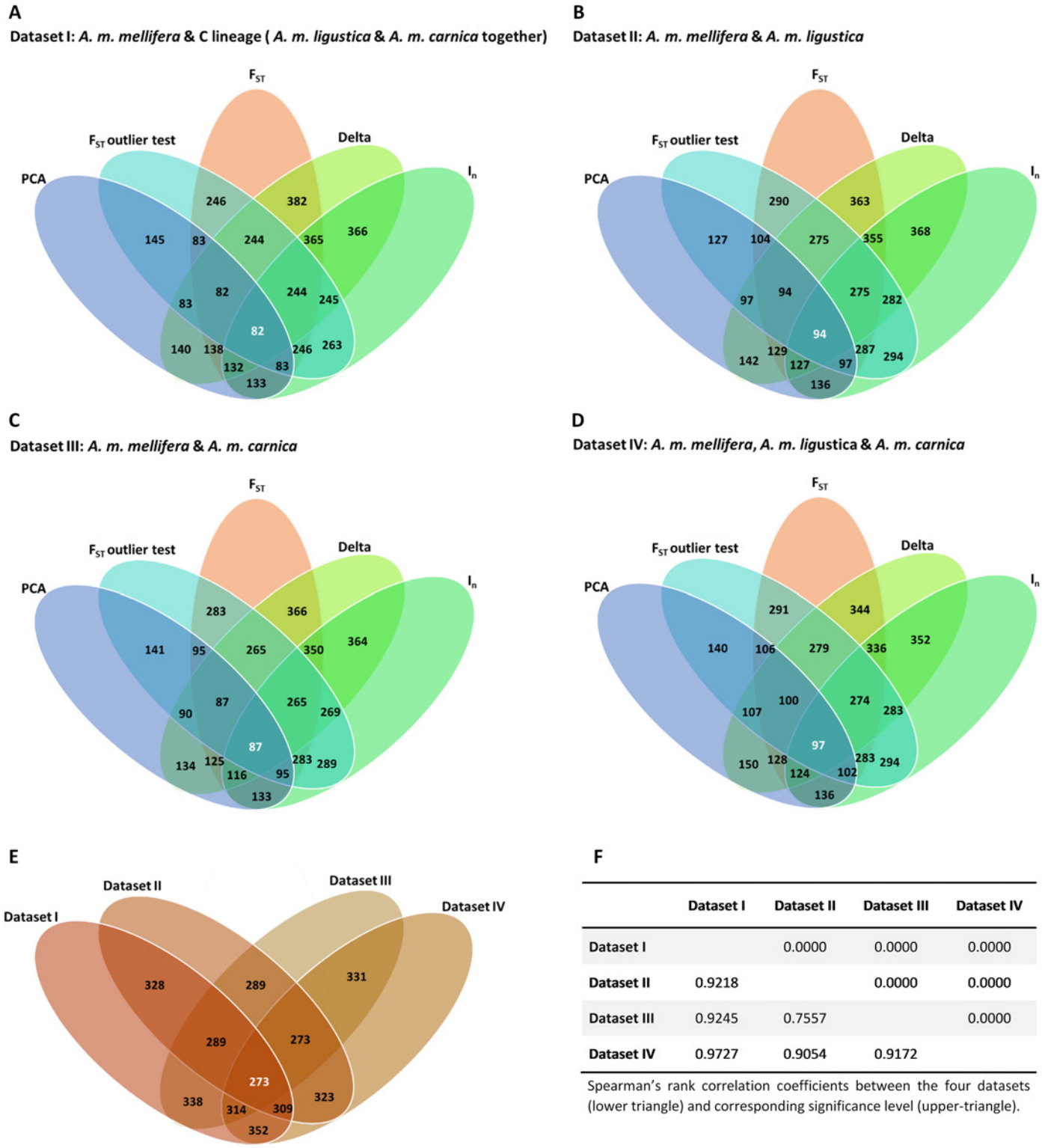
doi:10.1371/journal.pone.0124365.g001

Table 1. Comparison of selection methods and training datasets.

		Dataset I: *A. m. mellifera* & C lineage (*A. m. ligustica* & *A. m. carnica* together)					Dataset II: *A. m. mellifera* & *A. m. ligustica*					Dataset III: *A. m. mellifera* & *A. m. carnica*					Dataset IV: *A. m. mellifera A. m. ligustica* & *A. m. carnica*				
		F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test
Dataset I	F_{ST}		0.0000	0.0000	0.7134	0.0000	0.0000	0.0000	0.0000	0.0620	0.0000	0.0000	0.0000	0.0000	0.6130	0.0000	0.0000	0.0000	0.0000	0.7134	0.0000
	Delta	0.9985		0.0000	0.6139	0.0000	0.0000	0.0000	0.0000	0.0439	0.0000	0.0000	0.0000	0.0000	0.6772	0.0000	0.0000	0.0000	0.0000	0.6139	0.0000
	I_n	0.9977	0.9957		0.7284	0.0000	0.0000	0.0000	0.0000	0.0753	0.0000	0.0000	0.0000	0.0000	0.4980	0.0000	0.0000	0.0000	0.0000	0.7284	0.0000
	PCA	0.0107	0.0147	0.0101		0.0000	0.9412	0.8104	0.7135	0.0000	0.0000	0.0390	0.0241	0.0357	0.0000	0.0711	0.4901	0.1065	0.5267	0.0000	0.0001
	F_{ST} outlier test	0.5974	0.5778	0.6280	-0.1195		0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dataset II	F_{ST}	0.9364	0.9370	0.9392	-0.0021	0.5666		0.0000	0.0000	0.3782	0.0000	0.0000	0.0000	0.0000	0.8227	0.0000	0.0000	0.0000	0.0000	0.9412	0.0000
	Delta	0.9338	0.9357	0.9313	-0.0070	0.5205	0.9906		0.0000	0.3980	0.0000	0.0000	0.0000	0.0000	0.7760	0.0000	0.0000	0.0000	0.0000	0.8104	0.0000
	I_n	0.9329	0.9324	0.9342	-0.0107	0.5581	0.9923	0.9960		0.5187	0.0000	0.0000	0.0000	0.0000	0.8981	0.0000	0.0000	0.0000	0.0000	0.7135	0.0000
	PCA	0.0543	0.0586	0.0517	0.8545	-0.1074	0.0256	0.0246	0.0188		0.0000	0.0017	0.0004	0.0008	0.0000	0.1510	0.0986	0.0426	0.1329	0.0000	0.0000
	F_{ST} outlier test	0.4404	0.4235	0.4637	-0.2018	0.8128	0.4998	0.4732	0.5124	-0.2228		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dataset III	F_{ST}	0.9328	0.9317	0.9353	0.0600	0.5751	0.7965	0.7705	0.7719	0.0909	0.3301		0.0000	0.0000	0.2647	0.0000	0.0000	0.0000	0.0000	0.0390	0.0000
	Delta	0.9347	0.9353	0.9327	0.0656	0.5337	0.7825	0.7690	0.7663	0.1025	0.2864	0.9925		0.0000	0.4242	0.0000	0.0000	0.0000	0.0000	0.0241	0.0000
	I_n	0.9331	0.9312	0.9349	0.0611	0.5742	0.7829	0.7648	0.7656	0.0969	0.3150	0.9943	0.9960		0.3212	0.0000	0.0000	0.0000	0.0000	0.0357	0.0000
	PCA	-0.0147	-0.0121	-0.0197	0.6450	-0.1333	-0.0065	0.0083	0.0037	0.5534	-0.1435	-0.0325	-0.0233	-0.0289		0.0000	0.3690	0.3671	0.4471	0.0000	0.0000
	F_{ST} outlier test	0.5524	0.5361	0.5793	-0.0525	0.8862	0.4451	0.3910	0.4189	-0.0418	0.6063	0.6337	0.6006	0.6444	-0.1440		0.0000	0.0000	0.0000	0.0711	0.0000
Dataset IV	F_{ST}	0.9883	0.9862	0.9875	0.0201	0.6028	0.9301	0.9216	0.9222	0.0480	0.4567	0.9411	0.9387	0.9384	-0.0261	0.5771		0.0000	0.0000	0.4901	0.0000
	Delta	0.9459	0.9485	0.9474	0.0470	0.5487	0.9177	0.9022	0.9014	0.0590	0.4237	0.9241	0.9169	0.9152	-0.0262	0.5475	0.9683		0.0000	0.1065	0.0000
	I_n	0.9825	0.9798	0.9832	0.0184	0.6081	0.9313	0.9277	0.9308	0.0437	0.4733	0.9273	0.9279	0.9300	-0.0221	0.5798	0.9948	0.9710		0.5267	0.0000
	PCA	0.0107	0.0147	0.0101	1.0000	-0.1195	-0.0021	-0.0070	-0.0107	0.8545	-0.2018	0.0600	0.0656	0.0611	0.6450	-0.0525	0.0201	0.0470	0.0184		0.0001
	F_{ST} outlier test	0.6043	0.5867	0.6267	-0.1154	0.9119	0.5549	0.5292	0.5634	-0.1250	0.8250	0.5819	0.5589	0.5943	-0.1376	0.8758	0.6344	0.6104	0.6592	-0.1154	

Spearman's rank correlation coefficients (lower triangle), and corresponding *P*-values (upper triangle), between information content estimates produced by the five selection methods (Weir & Cockerham's F_{ST}, Delta, informativeness (I_n), PCA, F_{ST}-based outlier test) using the four training datasets (I, II, III and IV).

doi:10.1371/journal.pone.0124365.t001



F

	Dataset I	Dataset II	Dataset III	Dataset IV
Dataset I		0.0000	0.0000	0.0000
Dataset II	0.9218		0.0000	0.0000
Dataset III	0.9245	0.7557		0.0000
Dataset IV	0.9727	0.9054	0.9172	

Spearman's rank correlation coefficients between the four datasets (lower triangle) and corresponding significance level (upper-triangle).

Fig 2. (A-E) Venn diagrams showing the extent of overlap of the top-ranked 384 AIMs. (A-D) Overlap among the five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) and the four training datasets (I, II, III and IV). (E) Overlap among the four training datasets, after averaging the information content obtained with the five selection methods, and (F) corresponding Spearman's rank correlation coefficients.

doi:10.1371/journal.pone.0124365.g002

four training datasets, suggesting that the different population groupings have a small effect on the AIMs ranking. The global ranking of the 384 AIMs was used to design reduced panels of 192-, 144-, 96-, and 48 that included SNPs with the highest respective global scores. The performance of these reduced panels was subsequently assessed using the holdout and simulated sets.

Validation of the AIMs Panels

The performance of the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) was first validated by using PCA to produce a visual summary of the observed genetic variation carried by the holdout set (Fig 3). The overall diversity pattern is characterized by the presence of two distinct

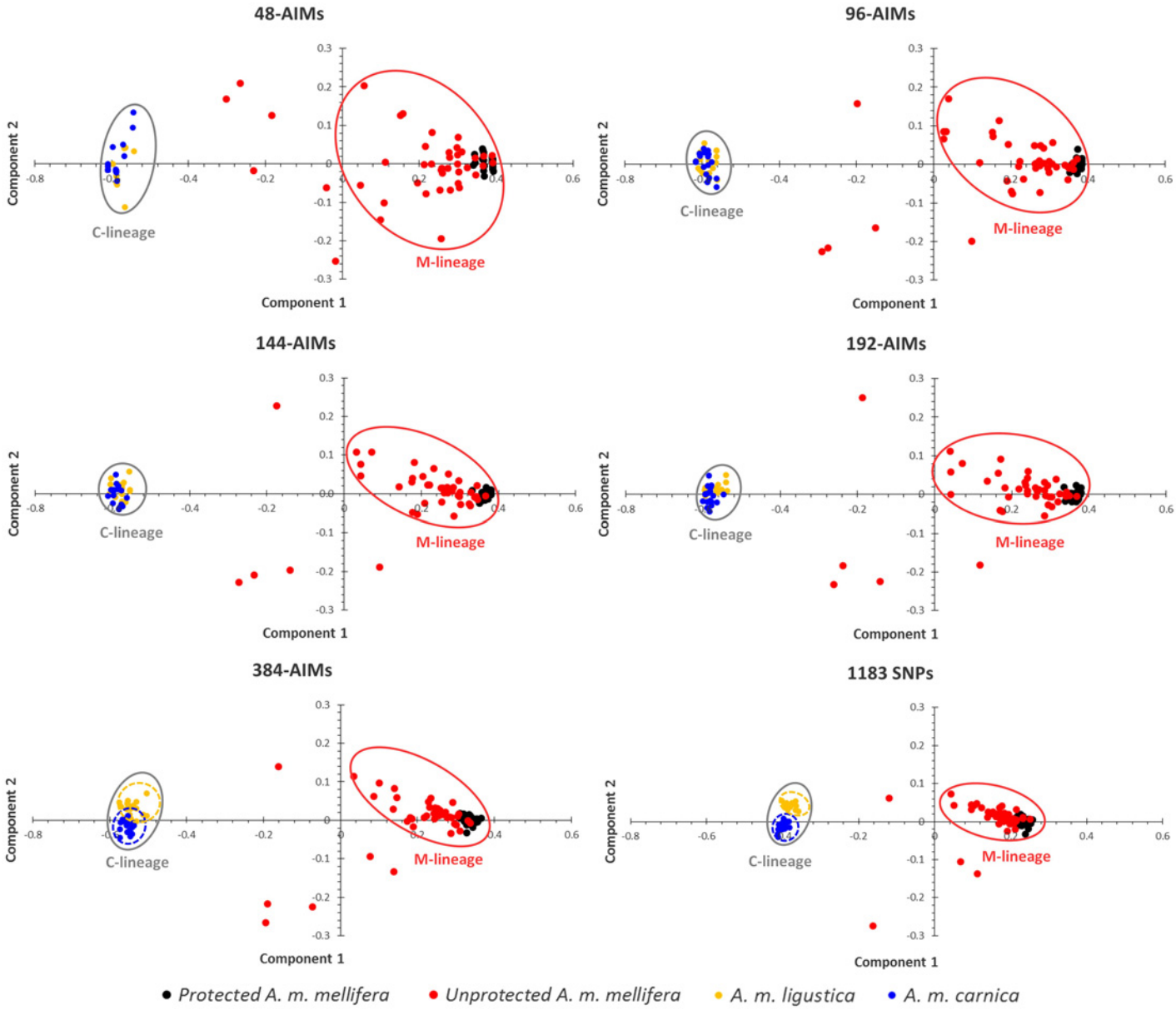


Fig 3. Principal components analysis. Plots obtained for the holdout set using the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset.

doi:10.1371/journal.pone.0124365.g003

clusters, which are coincidental with the M and C evolutionary lineages. This pattern was captured by every single AIMs panel, although a greater dispersion was observed for the smaller panels. Additionally, the panels with less than 192 AIMs were unable to distinguish the two C-lineage subspecies, *A.m. ligustica* and *A.m. carnica*, which were clearly identified by the initial 1183 SNPs and, to a lesser degree, by the 384-AIMs panel.

Ancestry and admixture analyses based on admixture estimates confirm the overall pattern captured by the PCA (S3 Table and S1 Fig). At the optimal K = 2 (inferred by the initial 1183 SNP dataset and the five AIMs panel), the two clusters corresponded to the C and M-lineages. However, C-lineage individuals formed a more homogeneous cluster than those of the M-lineage individuals. While membership proportions in the C-lineage cluster were greater than 95% for the five AIMs panels, the M-lineage cluster comprised 13 (384-AIMs and 1183 SNPs), 14 (48- and 192-AIMs) and 15 (96- and 144-AIMs) individuals with membership proportions lower than 85%, a pattern that was already evident in the PCA plots.

The introgression levels exhibited by individuals of the M-lineage cluster were significantly higher (Student's t-test, $P<0.001$) in unprotected (13.76–15.18%, with 1183 SNPs and 48 AIMs, respectively) than in protected individuals (0.08–0.52%, with 96 AIMs and 1183 SNPs, respectively) for any AIMs panel. The overall estimates of C-lineage introgression into *A. m. mellifera* varied with the panel (8.4, 7.9, 7.8, 7.9, 7.5 and 7.7% with 48-, 96-, 144-, 192-, 384-AIMs and 1183 SNPs, respectively), although the differences were not statistically significant (Mann-Whitney test, $0.8225 \leq P \leq 0.9983$; S4 Table).

In addition to the admixture analyses using the holdout set, the AIMs panels were further validated using a simulated set of 10 different levels of C-lineage introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%). As for the analyses with the holdout set, the simulated set produced two clusters corresponding to M and C lineages with no significant differences in admixture proportions between the different AIMs panels and the initial 1183 SNP dataset (Mann-Whitney test, $P \geq 0.2313$; S5 Table).

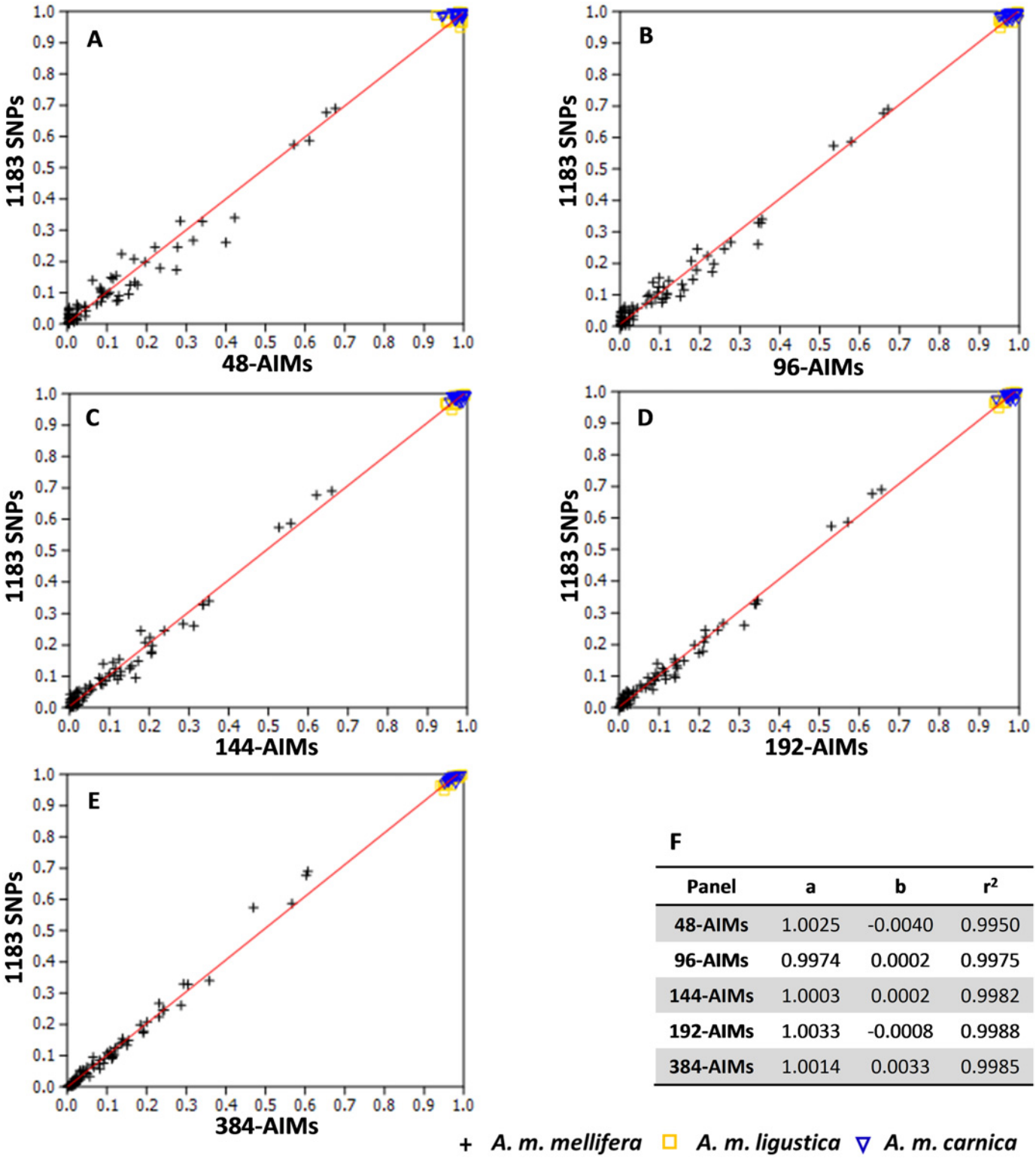
Assignment's precision and accuracy

The power of the reduced AIMs panels in identifying *A. m. mellifera* and estimating admixture proportions was evaluated on the holdout set. Estimates of C-lineage introgression into *A. m. mellifera* inferred from the five panels were greatly concordant with those inferred from the initial 1183 SNP dataset, as indicated by the high correlation values ($r \geq 0.997$; Fig 4). Despite the high correlations obtained for each comparison, the error rate in admixture estimates, which is very low for all the panels (0.0012–0.0042 with the simulated set and 0.4–1.3 with the holdout set), does increase as the size of the panel decreases (S2 Fig). Nevertheless, the reduced AIMs panels provide good precision in estimating admixture proportions.

As another assessment of the performance of the panels, the accuracy was calculated via absolute error. The success of assignment of the 113 individual genotypes of the holdout set to genetic origin and level of admixture inferred from the different AIMs panels is shown in Fig 5. The average percentage of correct assignment was high varying from 98.2, 98.8, 99.0, 99.2 to 99.4% for the 48-, 96-, 144-, 192- and 384-AIMs panels, respectively. The chosen AIMs panels accurately distinguish M/C admixture, therefore these results suggest that a small number of AIMs are sufficient to identify *A. m. mellifera* and estimate introgression from C-lineage colonies with great accuracy.

Discussion

The recognition that native honey bee genetic diversity is fundamental for sustainable beekeeping and for facing the challenges of a rapidly changing world (e.g. climate change, novel



F

Panel	a	b	r^2
48-AIMs	1.0025	-0.0040	0.9950
96-AIMs	0.9974	0.0002	0.9975
144-AIMs	1.0003	0.0002	0.9982
192-AIMs	1.0033	-0.0008	0.9988
384-AIMs	1.0014	0.0033	0.9985

Fig 4. Linear regression. (A-E) Plots between admixture proportions inferred from the initial 1183 SNP dataset and those inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) using individuals of the holdout set. (F) Parameters and coefficients for each AIMs panel.

doi:10.1371/journal.pone.0124365.g004

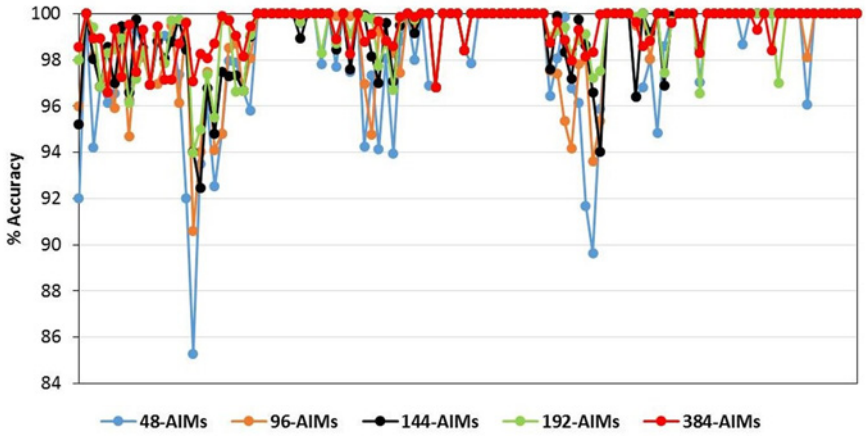


Fig 5. Assignment accuracy. Percentage obtained with the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) for each of the 113 individuals of the holdout set.

doi:10.1371/journal.pone.0124365.g005

diseases and parasites) is stimulating implementation of conservation programs across Europe in an attempt to recover and protect *A. m. mellifera*, which is the European honey bee subspecies with the widest natural range [12], and at the same time the most threatened by introgression [9, 31]. The need of a reliable, high-throughput, and cost-effective tool for identifying candidate *A. m. mellifera* colonies targeted for conservation, a crucial step when managing conservatoires, motivated the design of reduced AIMs panels containing the most informative SNPs to verify ancestry and introgression from C-lineage subspecies. In this study we developed, validated and tested the first reduced AIMs panels for honey bees. Our results provide strong confidence in a panel of 384 AIMs and show that even smaller subsets of 192-, 144-, 96- and 48-AIMs are able to identify ancestry and estimate introgression with great accuracy. These reduced panels promise to be a useful tool for routine identification of *A. m. mellifera* colonies maintained in the breeding populations of conservation programs.

The AIMs included in the five reduced panels were simultaneously selected by pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, in order to balance out the limitations of each individual method [41, 58, 65]. These selection methods have proved to be powerful, although with varying performances, in identifying population informative markers in a wide range of organisms [43, 58, 60–61, 65]. A great extent of overlap of top-ranked AIMs was obtained for the five selection methods, especially for pairwise Weir & Cockerham's F_{ST}, Delta, and I_n suggesting that they capture the same information. Nonetheless, the smaller panels (48-, 96-, 144-, 192-AIMs) did not necessarily include all AIMs simultaneously detected by the five methods as the global ranking depended on the average score. High pairwise correlation values were obtained for Weir & Cockerham's F_{ST}, Delta and I_n but not for PCA, as found by Wilkinson et al. [65]. PCA has been recommended for ranking markers because it has the advantage of generating an overall estimate for a single SNP locus whereas the other methods require estimate of an average from pairwise calculations when the number of populations is greater than two [58].

The five reduced panels tested with the holdout and simulated sets performed virtually as well as the initial 1183 SNP dataset, as revealed by the strong correlations obtained between admixture estimates and low associated error rates. The assignment power was high across the five panels with average values of correct assignment varying between 98.2 and 99.4%, although the accuracy decreased slightly with panel size. Nonetheless, even the 48-AIMs panel exhibited high accuracy levels, which is not surprising as it includes the AIMs with the greatest resolution

power. Studies on other organisms have also found good performances with panels of similar sizes [43, 45, 60, 65], detecting sharp drops in accuracy for a number of SNPs below 25 [45, 60].

Evaluation of different combinations of the focal *A. m. mellifera* and the two most common sources of foreign genes, *A. m. ligustica* and *A. m. carnica*, revealed a negligible effect of population groupings on the AIMs ranking. These results suggest that the designed panels are suited for identifying and assessing introgression of *A. m. ligustica*, *A. m. carnica* or both into *A. m. mellifera*. While these panels will possibly perform well in the presence of other C-lineage subspecies, more complex combinations that include sources of different evolutionary lineages will require further testing and, most likely, new panels developed from broader baseline datasets. Additionally, it should be noted, that these reduced panels are not suitable for standard population genetic analyses, including determining allelic diversity or measuring isolation by distance, genetic drift or bottleneck effect. The bias introduced through selection for markers that segregate among target populations would seriously compromise these calculations [66–67].

Ancestry identification of honey bee subspecies is undergoing steady development (reviewed by Meixner et al. [68]) from classical morphometry, analysis of allozymes, mitochondrial DNA, nuclear microsatellites, and now SNP tools. Because researchers must balance the cost of genotyping many samples versus many loci, herein we developed five nested reduced panels that include AIMs with the highest resolution power for discriminating subspecies of the divergent M and C evolutionary lineages. While the 384-AIMs panel is also capable of discriminating the C-lineage *A. m. ligustica* and *A. m. carnica*, for estimating C-lineage introgression into *A. m. mellifera* we recommend using the 96-AIMs panel because it is accurate; and high-throughput 96-plex genotyping assays can be outsourced at an affordable cost ($8 900 for 480 samples), representing a saving of 92.4% when compared with the 1536-plex assay ($116 800 for 480 samples).

In conclusion, the proposed AIMs panels can be actively used as a tool in conservation management of *A. m. mellifera* populations that suffer from hybridization and introgression with the most commonly introduced and beekeepers' preferred *A. m. ligustica* and *A. m. carnica* subspecies. This can be an important advance because the current European regulation on organic beekeeping states that "preference shall be given to the use of European breeds of *Apis mellifera* and their local ecotypes" and several conservation programs have been undertaken in Europe (reviewed by De la Rúa et al. [8]). The use of these panels will apply well to monitoring, management and conservation programs of *A. m. mellifera* in Western Europe, which usually require high-sample throughput, and will be a resource for the honey bee community to obtain accurate genetic information at reduced costs.

Supporting Information

S1 Fig. Ancestry estimates. Global estimates (y-axis), for the 113 individuals of the holdout set (x-axis), inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset using the model-based approach implemented in the ADMIXTURE software. Results are shown for the optimal $K = 2$, which distinguishes the M (red) and C (cyan) evolutionary lineages of *A. mellifera*.
(TIFF)

S2 Fig. Standard deviation (SD) of admixture proportions. Precision estimates obtained using the SD of the differences between admixture proportions inferred from the initial 1183 SNP dataset and the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) using the holdout (blue line) and simulated (orange line) sets.
(TIFF)

S1 Table. Input file containing the 1183 coded SNPs for the 113 honey bee samples.
(XLSX)

S2 Table. Information content values of the initial 1183 SNP dataset estimated by the five selection methods (Weir & Cockerham's F_{ST}, Delta, informativeness (I_n), PCA and the F_{ST}-based outlier test) and for the four training datasets (I to IV). The SNPs are ordered from high to low information content. The top 48, 96, 144, 192 and 384 SNPs were included in the five reduced panels. SNPs marked with an asterisk (*) were excluded from the reduced panels because they were within a genetic distance <1 cM of other informative SNPs.
(DOCX)

S3 Table. Admixture proportion estimates inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset for the holdout set. The holdout set consisted of 34 pure (training set) and 43 reserved individuals of *A. m. mellifera* and all reference individuals of *A. m. ligustica* (17) and *A. m. carnica* (19). * Samples marked with an asterisk (*) are of *A. m. mellifera* from protected populations (pure breeding for conservation purposes; see Pinto et al. 2014 [9] for details).
(DOCX)

S4 Table. *P*-values of Mann-Whitney pairwise several-sample-test. Values obtained from comparing individual admixture proportions estimated with the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset using the holdout set.
(DOCX)

S5 Table. *P*-values of Mann-Whitney pairwise several-sample-test. Values obtained from comparing admixture proportions inferred from the five AIMs panels and the 1183 initial SNP dataset using the simulated set. The simulated set was generated with the program ONCOR (Kalinowski et al. 2007) using the function "simulate a single mixture". Ten populations, each with 100 genotypes, were simulated using different levels of C-lineage introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%).
(DOCX)

Acknowledgments

We are deeply grateful to Andrew Abrahams, Bjørn Dahle, Gabriele Soland-Reckeweg, Gilles Fert, Lionel Garnery, Norman Carreck, Pilar de la Rúa, Raffaele Dall'Olio, and Romee Van der Zee for providing honey bee samples. DNA extractions and SNP genotyping were performed by Colette Abbey, with support from the TAMU Institute of Genomic Science and Society. An earlier version of the manuscript was improved by the constructive comments made by two anonymous reviewers.

Author Contributions

Conceived and designed the experiments: MAP IM DH. Analyzed the data: IM DH JC-G. Contributed reagents/materials/analysis tools: PK. Wrote the paper: MAP IM JSJ.

References

1. Dowling TE, Secor CL The role of hybridization and introgression in the diversification of animals. Annu Rev Ecol Evol Syst. 1997; 28: 593–619.
2. Nolte AW, Tautz D. Understanding the onset of hybrid speciation. Trends Genet. 2009; 26: 54–58.
3. Rhymer JM, Simberloff D. Extinction by hybridization and introgression. Annu Rev Ecol Evol Syst. 1996; 27: 83–109.

4. Allendorf FW, Luikart G. Conservation and the Genetics of Populations. 1st ed. Malden, Massachusetts: Blackwell Publishing; 2007.
5. Crane E. The World History of Beekeeping and Honey Hunting. 1st ed. New York: Routledge; 1999.
6. vanEngelsdorp D, Meixner MD. A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. J Invertebr Pathol. 2010; 103: 80–95.
7. Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE. Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol. 2010; 25: 345–353. doi: 10.1016/j.tree.2010.01.007 PMID: 20188434
8. De la Rúa P, Jaffé R, Dall'Olio R, Muñoz I, Serrano J. Biodiversity, conservation and current threats to European honeybees. Apidologie. 2009; 40: 263–284.
9. 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: 269–278.
10. Meixner MD, Costa C, Kryger P, Hatjina F, Bouga M, Ivanova E, et al. Conserving diversity and vitality for honey bee breeding. J Apic Res. 2010; 49: 85–92.
11. Büchler R, Costa C, Hatjina F, Andonov S, Meixner MD, Le Conte Y, 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: 205–214.
12. Ruttner F. Biogeography and Taxonomy of Honeybees. 1st ed. Berlin, Germany: Springer Verlag; 1988.
13. Hepburn HR, Radloff SE. (1998) Honey bees of Africa. Berlin, Germany: Springer. 370 p.
14. Engel MS. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; *Apis*). J Hymenopt Res. 1999; 8: 165–196.
15. Sheppard WS, Meixner MD. *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. Apidologie. 2003; 34: 367–375.
16. 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: 425–437.
17. Garnery L, Cornuet JM, Solignac M. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. Mol Ecol. 1992; 1: 145–154. PMID: 1364272
18. Garnery L, Solignac M, Celebrano G, Cornuet JM. A simple test using restricted PCR amplified mitochondrial DNA to study the genetic structure of *Apis mellifera* L. Experientia. 1993; 49: 1016–1021.
19. Arias MC, Sheppard WS. Molecular phylogenetics of honey bee subspecies (*Apis mellifera* L.) inferred from mitochondrial DNA sequence. Mol Phylogenet Evol. 1996; 5: 557–566. PMID: 8744768
20. Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS, et al. Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. Science. 2006; 314: 642–645. PMID: 17068261
21. 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: 1081–1088. doi: 10.1038/ng.3077 PMID: 25151355
22. 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. iberica*). I. Mitochondrial DNA. Genet Sel Evol. 1998a; 30: 31–47.
23. 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. iberica*). II. Microsatellite loci. Genet Sel Evol. 1998b; 30: 49–74.
24. Miguel I, Iriondo 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: 141–155.
25. Strange JP, Garnery L, Sheppard WS. Morphological and molecular characterization of the Landes honey bee (*Apis mellifera* L.) ecotype for genetic conservation. J Insect Conserv. 2008; 12: 527–537.
26. Oleksa A, Chybicki I, Tofilski A, Burczyk J. Nuclear and mitochondrial patterns of introgression into native dark bees (*Apis mellifera mellifera*) in Poland. J Apic Res. 2011; 50: 116–129.
27. Pinto MA, Muñoz I, Chávez-Galarza J, De la Rúa P. The Atlantic side of the Iberian Peninsula: a hotspot of novel African honey bee maternal diversity. Apidologie. 2012; 43: 663–673.
28. 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: 285–285.

29. Muñoz I, Dall'Olio R, Lodesani M, De la Rúa P. Population genetic structure of coastal Croatian honeybees (*Apis mellifera carnica*). Apidologie. 2009; 40: 617–626.

30. Nedić N, Francis RM, Stanisavljević L, Pihler I, Kezić N, Bendixen C, et al. Detecting population admixture in the honey bees of Serbia. J Apic Res. 2014; 53: 303–313.

31. 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. PMID: 15643954

32. Soland-Reckeweg G, Heckel G, Neumann P, Fluri P, Excoffier L. Gene flow in admixed populations and implications for the conservation of the Western honey bee, *Apis mellifera*. J Insect Conserv. 2009; 13: 317–328.

33. Dreher K. Gedanken zum Neuaufbau des Zuchtwesens. Die Hessische Biene. 1946; 81: 62–64.

34. Maul V, Hähnle A. Morphometric studies with pure bred stock of *Apis mellifera carnica* from Hessen. Apidologie. 1994; 25: 119–132.

35. Jensen AB, Pedersen BV. Honey bee Conservation: a case story from Læsø island, Denmark. In: Lodesani M, Costa C, editors. Beekeeping and conserving biodiversity of honey bee. Sustainable bee breeding. Theoretical and practical guide. Hebden Bridge: Northern Bee Books; 2005. pp. 142–164.

36. Rortais A, Arnold G, Alburaki M, Legout H, Garnery L. Review of the DraI COI—COII test for the conservation of the black honeybee (*Apis mellifera mellifera*). Conserv Genet Res. 2011; 3: 383–391.

37. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, et al. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 2006; 443: 931–949. PMID: 17073008

38. 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: 5890–5907. doi: 10.1111/mec.12537 PMID: 24118235

39. Harpur BA, Kent CF, Molodtsova D, Lebon JM, Alqarni AS, Owayssc AA, et al. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. P Natl Acad Sci USA. 2014; 111: 2614–2619. doi: 10.1073/pnas.1315506111 PMID: 24488971

40. Harpur BA, Minaei S, Kent CF, Zayed A. Admixture increases diversity in managed honey bees: reply to De la Rúa et al., 2013. Mol Ecol. 2013; 22: 3211–3215. doi: 10.1111/mec.12332 PMID: 24433573

41. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 2003; 73: 1402–1422. PMID: 14631557

42. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat. 2009; 30: 69–78. doi: 10.1002/humu.20822 PMID: 18683858

43. Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, et al. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet. 2012; 8: e1002554. doi: 10.1371/journal.pgen.1002554 PMID: 22412386

44. Frantz AC, Pourtois JT, Heuertz M, Schley L, Flamand MC, Krier A, et al. Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. Mol Ecol. 2006; 15: 3191–3203. PMID: 16968264

45. Wilkinson S, Archibald AL, Haley CS, Megens H-J, Crooijmans RPMA, Groenen MAM, et al. Development of a genetic tool for product regulation in the diverse British pig breed market. BMC Genomics. 2012; 13:580. doi: 10.1186/1471-2164-13-580 PMID: 23150935

46. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, et al. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet. 1998; 63: 1839–1851. PMID: 9837836

47. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164: 1567–1587. PMID: 12930761

48. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. Am J Hum Genet. 2004; 74: 965–978. PMID: 15088268

49. Pardo-Seco J, Martinón-Torres F, Salas A. Evaluating the accuracy of AIM panels at quantifying genome ancestry. BMC Genomics. 2014; 15: 543. doi: 10.1186/1471-2164-15-543 PMID: 24981136

50. 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: 188–204.

51. Bertrand B, Alburaki M, Legout H, Moulin S, Mougel F, Garnery L. MtDNA COI-COII marker and drone congregation area: An efficient method to establish and monitor honeybee (*Apis mellifera* L.) conservation centres. Mol Ecol Res. 2014.

52. Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. 2nd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1989.

53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007; 81: 559–575. PMID: 17701901

54. Weir RJ, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984; 38: 1358–1370.

55. Raymond M, Rousset F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Heredity. 1995; 86: 248–249.

56. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008; 180: 977–993. doi: 10.1534/genetics.108.092221 PMID: 18780740

57. Hammer Ø, Harper DAT, Ryan PD. PAST: paleontological statistics software package for education and data analysis. Palaeontol Electron. 2001; 4(1): art. 4.

58. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007; 3: 1672–1686. PMID: 17892327

59. Anderson EC. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. Mol Ecol Res. 2010; 10: 701–710. doi: 10.1111/j.1755-0998.2010.02846.x PMID: 21565075

60. Storer CG, Pascal CE, Roberts SB, Templin WD, Seeb LW, Seeb JE. Rank and order: Evaluating the performance of SNPs for individual assignment in a non-model organism. PLoS ONE. 2012; 7: e49018. doi: 10.1371/journal.pone.0049018 PMID: 23185290

61. Ozerov M, Vasemägi A, Wennevik V, Diaz-Fernandez R, Kent M, Gilber J, et al. Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. PLoS ONE. 2013; 8: e82434. doi: 10.1371/journal.pone.0082434 PMID: 24358184

62. Kalinowski ST, Manlove KR, Taper ML. ONCOR A computer program for Genetic Stock Identification; 2007. Available: Department of Ecology, Montana State University, Bozeman MT 59717. Accessed: http://www.montana.edu/kalinowski.

63. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19: 1655–1664. doi: 10.1101/gr.094052.109 PMID: 19648217

64. Wright S. Evolution and the genetics of population, variability within and among natural populations. Chicago: University Chicago Press. 1978.

65. 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 Genetics. 2011; 12: 45. doi: 10.1186/1471-2156-12-45 PMID: 21569514

66. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15(11): 1496–1502. PMID: 16251459

67. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010; 27(11): 2534–254768. doi: 10.1093/molbev/msq148 PMID: 20558595

68. Meixner MD, Pinto MA, Bouga M, Kryger P, Ivanova E, Fuchs S. Standard methods for characterizing subspecies and ecotypes of *Apis mellifera*. J Apic Res. 2013; 52(4): 1–27.