genes

MDPI

Article

Population Structure and Diversity in European Honey Bees (*Apis mellifera* L.)—An Empirical Comparison of Pool and Individual Whole-Genome Sequencing

Chao Chen [1,2,*,†], Melanie Parejo [3,4,*,†], Jamal Momeni [5], Jorge Langa [3], Rasmus O. Nielsen [5], Wei Shi [1,2], SMARTBEES WP3 DIVERSITY CONTRIBUTORS [‡], Rikke Vingborg [5], Per Kryger [6,§], Maria Bouga [7,§], Andone Estonba [3,§] and Marina Meixner [8,§]

1 Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing 100093, China; shiweibri@126.com
2 Key Laboratory of Pollinating Insect Biology, Ministry of Agriculture and Rural Affairs, Beijing 100093, China
3 Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), 48940 Leioa, Spain; jorgeeliseo.langa@ehu.es (J.L.); andone.estonba@ehu.eus (A.E.)
4 Swiss Bee Research Center, Agroscope, 3003 Bern, Switzerland
5 Eurofins Genomics, 8200 Aarhus, Denmark; JamalMomeni@eurofins.dk (J.M.); bioinformatics@ron.dk (R.O.N.); RIVIN@vikinggenetics.com (R.V.)
6 Department of Agroecology, Aarhus University, 4200 Slagelse, Denmark; per.kryger@agro.au.dk
7 Lab of Agricultural Zoology and Entomology, Agricultural University of Athens, 11855 Athens, Greece; mbouga@aua.gr
8 LLH Bee Institute Kirchhain, 35274 Kirchhain, Germany; marina.meixner@llh.hessen.de
* Correspondence: chao_chen@outlook.com (C.C.); melanie.parejo@ehu.es (M.P.)
† Shared first co-authorship.
‡ Collaborators of the SMARTBEES WP3 DIVERSITY are listed in the Appendix A.
§ These authors contributed equally to this work.

Abstract: Background: Whole-genome sequencing has become routine for population genetic studies. Sequencing of individuals provides maximal data but is rather expensive and fewer samples can be studied. In contrast, sequencing a pool of samples (pool-seq) can provide sufficient data, while presenting less of an economic challenge. Few studies have compared the two approaches to infer population genetic structure and diversity in real datasets. Here, we apply individual sequencing (ind-seq) and pool-seq to the study of Western honey bees (*Apis mellifera*). Methods: We collected honey bee workers that belonged to 14 populations, including 13 subspecies, totaling 1347 colonies, who were individually (139 individuals) and pool-sequenced (14 pools). We compared allele frequencies, genetic diversity estimates, and population structure as inferred by the two approaches. Results: Pool-seq and ind-seq revealed near identical population structure and genetic diversities, albeit at different costs. While pool-seq provides genome-wide polymorphism data at considerably lower costs, ind-seq can provide additional information, including the identification of population substructures, hybridization, or individual outliers. Conclusions: If costs are not the limiting factor, we recommend using ind-seq, as population genetic structure can be inferred similarly well, with the advantage gained from individual genetic information. Not least, it also significantly reduces the effort required for the collection of numerous samples and their further processing in the laboratory.

Keywords: *Apis mellifera*; population structure; diversity; whole-genome sequencing; pool-sequencing

check for updates

Citation: Chen, C.; Parejo, M.; Momeni, J.; Langa, J.; Nielsen, R.O.; Shi, W.; SMARTBEES WP3 DIVERSITY CONTRIBUTORS; Vingborg, R.; Kryger, P.; Bouga, M.; et al. Population Structure and Diversity in European Honey Bees (*Apis mellifera* L.)—An Empirical Comparison of Pool and Individual Whole-Genome Sequencing. *Genes* **2022**, *13*, 182. https://doi.org/10.3390/genes13020182

Academic Editor: Erich Bornberg-Bauer

Received: 15 December 2021
Accepted: 30 December 2021
Published: 21 January 2022

1. Introduction

Studying population genetic structure and diversity is the basis of our understanding of biodiversity and the conservation of species [1]. Due to the recent rapid developments in sequencing technology, it is now possible to gain insights into the genomic structure of populations with unprecedented power and accuracy [2]. To make the best use of limited

resources, different sampling and sequencing approaches to study population structure are being adopted, which can be summarized as (i) covering whole genomes, but sampling few individuals (e.g., Ref. [3]), (ii) sampling many individuals, but covering limited parts of the genome (e.g., restriction-site associated DNA (RAD) sequencing, exome capture, or genotype-by-sequencing (GBS)) [4,5] or (iii) whole-genome sequencing and pooling of many individuals (pool-seq) (e.g., Ref. [6]). While techniques relying on the reduced representation of the genome are mostly being used in non-model organisms (extensively reviewed, e.g., in [7,8]), the pooled sequence method has been advocated as an alternative, cost-effective approach identifying genome-wide patterns of genetic variation from large populations [9,10].

Allele frequencies are one of the key parameters to study population genetic structure and to estimate genetic distances between populations [1], and the power of many genetic analyses increases with the accuracy of the allele frequency estimates derived from population samples [10]. When sequencing individually a limited number of samples per population, the estimate of the allele frequencies and the sampling variance stems directly from the selection of individuals [11]. By pool-sequencing, on the contrary, numerous samples from a given population can be analyzed, notably reducing the sampling error. Thus, pool-seq has been statistically shown to produce more accurate estimates of population allele frequencies at a lower cost than sequencing of individuals [9,11–13].

The cost-effectiveness of pool-seq becomes obvious when considering the cost of individual handling and library preparation for sequencing: as one pool represents a single sample, only one library needs to be prepared [6,10]. As sequencing costs continue to decrease, library preparation is becoming an increasing factor to consider within the research budget.

Pool-seq is especially suitable for applications that require large sample sizes and the analysis of multiple samples/populations; however, for an optimal design of pools, previous knowledge of population structure is beneficial or even necessary. On the other hand, the disadvantage of pool-sequencing is the loss of genetic information at the individual level. Therefore, this technique may not be suitable for certain applications [10,11], for instance, in cases where population boundaries are not clear or gradual or after recent contact with populations. In such cases, individual whole-genome sequencing is better suited, but may not be feasible for several hundreds to thousands of individuals [14].

Most of the advantages and disadvantages of the pool and individual sequencing approaches have been discussed theoretically or by using simulation studies [6,10,11,15]. Thus, it is not entirely clear when, in practice, pool-seq or ind-seq would be the method of choice for a given population genetic study, and actual case studies are needed to reach an empirical consensus on the circumstances in which one approach is better suited over the other.

To directly compare the pool-seq and ind-seq approaches and evaluate their ability to infer population structure and diversity, we apply both methods to the same populations of European honey bees (*A. mellifera* L.). Both approaches, whole-genome ind-seq and pool-seq, have been previously applied in this species, albeit not with a comparative purpose (e.g., [16,17]). Europe, with its numerous autochthonous honey bee subspecies belonging to four different evolutionary lineages [18–23], holds a large fraction of the *A. mellifera* genetic heritage. Substantial geographic variation in European honey bee populations has been investigated and described in several previous studies [24–29]. Nonetheless, while the genetic variation of honey bees in some parts of Europe has been subject to detailed studies and a considerable level of knowledge has been accumulated ([29–35] amongst others), some other areas, especially in the eastern part of the continent, have received comparatively little scientific interest [36,37]. Consequently, the analysis and description of honey bee subspecific variation in Europe cannot be regarded as complete, and the taxonomic status of some populations is not yet fully resolved. Thus, comparatively little information is available regarding a global analysis of the genetic variability of *A. mellifera* in Europe. Moreover, because of modern apicultural management and activities, such as

queen trade or migratory beekeeping, the distribution and genetic diversity of European honey bees in many places no longer correspond to their natural state [32,38].

In this study, we empirically assess the consistency of two sequencing approaches, pool-seq and ind-seq, in estimating allele frequency, genetic diversity and documenting population structure in European *A. mellifera* honey bees. To this end, we sampled a total of 1347 worker bees from 14 populations, covering a large range of *A. mellifera* subspecies and diversity in Europe, and with a special focus on areas where studies had been scarce, especially towards the eastern range limit of the species. We sequenced the whole-genome of pools from about 90 workers (pool-seq), as well as ten individuals per population (ind-seq). Data obtained by either pool-seq and ind-seq were then compared to their ability to detect the population structure and diversity of European honey bees. Finally, we discuss the advantages and disadvantages of both approaches in practice and recommend future studies.

2. Materials and Methods

2.1. Sampling

A sampling strategy was devised to collect representative samples of worker bees from the entire range of *A. mellifera* in Europe (Figure 1, Table 1). Areas previously poorly studied, especially towards the eastern range limit of the species, were given special consideration. In total, we sampled 14 populations, each one represented by 80–100 worker bees from unrelated colonies (one worker bee per apiary). Following the *A. mellifera* subspecies nomenclature by Engel et al. [39], with some deviations (as in Momeni et al. [17]), the sampled populations in this study belong to four evolutionary lineages and 13 different subspecies: Lineage M: *A. m. mellifera* Linnaeus 1758, and *A. m. iberiensis* Engel 1999; Lineage A: *A. m. ruttneri* Sheppard et al. 1997 [20]; Lineage O: *A. m. anatoliaca* Maa 1953, *A. m. caucasia* Pollmann 1889, *A. m. remipes* Gerstaecker 1862, and *A. m. cypria* Pollman 1879; Lineage C: *A. m. cecropia* Kiesenwetter 1860, *A. m. carnica* Pollman 1879, *A. m. macedonica* Ruttner 1988, *A. m. adami* Ruttner 1975, *A. m. carpatica* Foti 1965 [40], and *A. m. rodopica* Petrov 1991 [41]. Based on mitochondrial DNA, *A. m. caucasia* has also been assigned to the C lineage [42]. In this study, we refer to the 14 populations based on the subspecies and country abbreviations, as presented in Table 1.

2.2. DNA Extraction and Sequencing

For this study, we pooled about 90 individuals per population following Schlötterer et al. [10], who recommended a sampling size of 40 to 100 individuals. To assemble one pool for sequencing, the heads without eyes of up to 100 workers were ground together, and the DNA was extracted following standard methods [45]. 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 (Illumina Inc., San Diego, CA, USA)) one lane per pool.

DNA from the thorax of ten workers (each one from a different colony), except for *A. m. adami* with only nine workers available, from each of the 14 pools, were extracted individually using the CTAB method [46]. Sequencing libraries were generated using NEB Next® Ultra DNA Library Prep Kit for Illumina® (New England Biolabs Inc., Ipswich, MA, USA) following the manufacturer's recommendations. Libraries were sequenced on the Illumina X Ten platform (Illumina Inc., San Diego, CA, USA). The very same samples sequenced individually have also been included in the respective pools.

For the two approaches, two different types of DNA extractions were used from different parts of the bodies (head, thorax), followed by different library preparation protocols, and finally, two different sequencing platforms were used. This strategy was chosen in order to compare the two methods in very realistic settings, that is, we used the same initial populations, then applied the two approaches independently, including the sequencing at different facilities, and bioinformatics analyses by different researchers, such

that finding good agreement between the two approaches would indicate robustness to the method used.

Table 1. Sample sizes and origin of the 14 populations used in this study.

Lineage	Population	Subspecies	Country	Pool Sequencing Samples (N)	Individual Sequencing Samples (N)	Origin of Samples/ References
M	ibe_esp_eus	*A. m. iberiensis*	Spain	100	10	Miguel et al., 2007 [35]
	mel_irl	*A. m. mellifera*	Ireland	100	10	Hassett et al., 2018 [43]
	mel_rus	*A. m. mellifera*	Russia (Ural)	100	10	This study, Momeni et al., 2021 [17]
C	car_aut_hun	*A. m. carnica*	Austria & Hungary	100	10	This study, Momeni et al., 2021 [17]
	rod_bgr	*A. m. rodopica*	Bulgaria	95	10	This study, Momeni et al., 2021 [17]
	carp_rou_mda	*A. m. carpatica*	Romania & Moldova	90	10	This study, Momeni et al., 2021 [17]
	mac_mkd_grc	*A. m. macedonica*	North Macedonia & N-Greece	86	10	This study, Uzunov et al., 2014 [27]
	cec_grc	*A. m. cecropia*	Greece	93	10	This study, Momeni et al., 2021 [17]
	ada_grc	*A. m. adami*	Greece (Crete)	88	9	This study, Momeni et al., 2021 [17]
O	cyp_cyp	*A. m. cypria*	Cyprus	100	10	This study, Momeni et al., 2021 [17]
	ana_tur	*A. m. anatoliaca*	Turkey	100	10	This study, Francis et al., 2014 [44]
	rem_arm	*A. m. remipes*	Armenia	90	10	This study, Momeni et al., 2021 [17]
	cau_tur_geo	*A. m. caucasia*	NE-Turkey & Georgia	105	10	This study, Momeni et al., 2021 [17]
A	rut_mlt	*A. m. ruttneri*	Malta	100	10	This study, Momeni et al., 2021 [17]
TOTAL				1347	139	

2.3. Sequencing Data Processing

Bioinformatics processing of the generated pool sequence data was performed using best practices following Schlötterer et al. [10]. Illumina adaptors and low-quality bases were removed using Trimmomatic v0.32 [47], and read quality was checked with FastQC (http://broadinstitute.github.io/fastqc, accessed on 3 November 2019). High-quality sequences were mapped against the honey bee reference genome Amel4.5 [48] using bwa-mem 0.7.10 [49]. SAMtools v0.1.19 [50] and Picard-tools v1.124 (http://broadinstitute.github.io/picard/, accessed on 3 November 2019) were used to convert between SAM and BAM formats, remove duplicate reads, sort the BAM files, remove reads with low-quality mapping (MAPQ < 20), and keep only properly mapped pairs. Subsequently, the data were processed following the steps of the PoPoolation package [51]. The mapping files were split by chromosome to speed up the analyses, converted to mpileup, and indels were removed. Finally, using a minimum count of 3, the data for the different pools were subsampled to uniform coverage (50×) to allow comparability between under-sequenced and over-sequenced pools. The entire strategy for pooled sequence analysis described here was successfully applied in previous studies [17,52], and is available as an automatized Snakemake pipeline [53,54] at https://github.com/jlanga/smsk_popoolation (accessed on 3 November 2019).

For individual sequencing data, reads were filtered by fastp [55] to exclude those with excessive low-quality bases. Clean reads were mapped to the Amel4.5 reference genome [48] using the bwa-mem aligned [56]. Variants were called using SpeedSeq pipeline [57] and default settings. We removed indels, SNPs within ten bp of indels, and SNPs within the repeat regions. The remaining SNPs were further filtered, and only high-quality ones meeting the following criteria were kept: (1) biallelic; (2) quality score > 30; (3) missing genotype <10%.

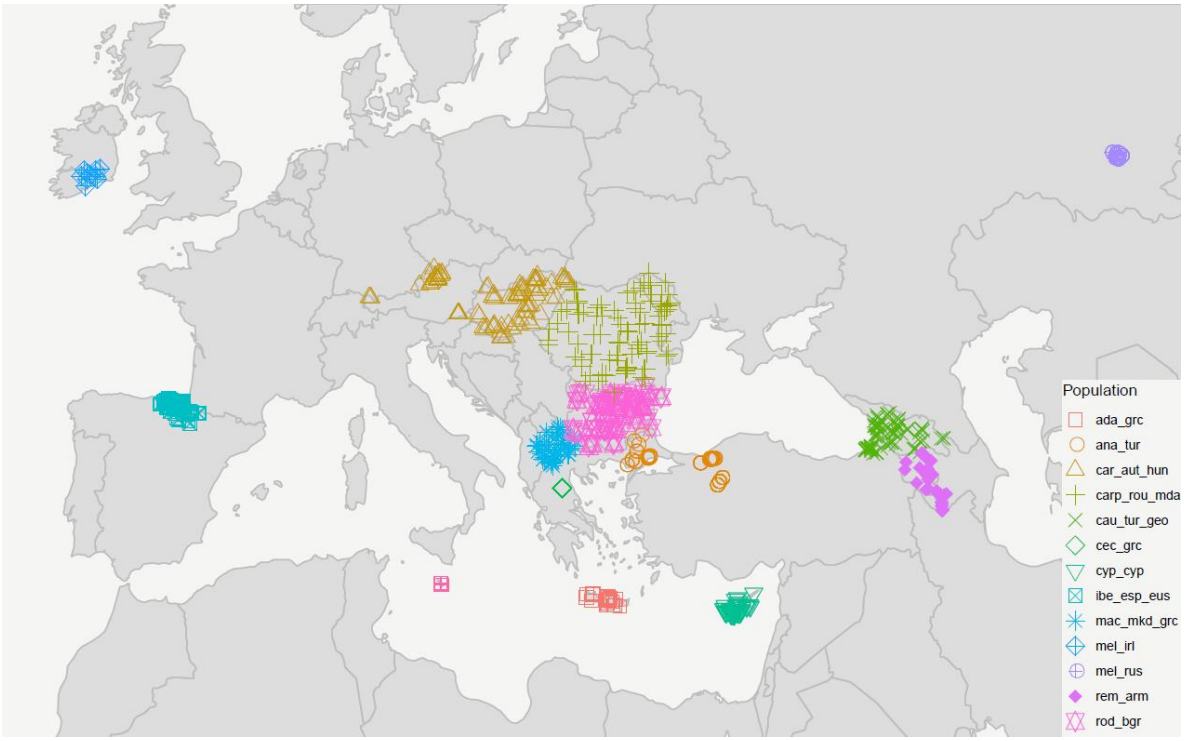

Figure 1. Sampling locations in Europe and adjacent regions plotted in R using ggplot2. M-lineage: ibe_esp_eus = *A. m. iberiensis* from Spain; mel_irl = *A. m. mellifera* from Ireland; mel_rus = *A. m. mellifera* from Russia; C-lineage: car_aut_hun = *A. m. carnica* from Austria and Hungary; rod_bgr = *A. m. rodopica* from Bulgaria; carp_rou_mda = *A. m. carpatica* from Romania and Moldova; mac_mkd_grc = *A. m. macedonica* from North Macedonia a Northern Greece; cec_grc = *A. m. cecropia* from Greece; ada_grc = *A. m. adami* from Crete, Greece; O-lineage: cyp_cyp = *A. m. cypria* from Cyprus; ana_tur = *A. m. anatoliaca* from Turkey; rem_arm = *A. m. remipes* from Armenia; cau_tur_geo = *A. m. caucasia* from North-East Turkey and Georgia; and A-lineage: rut_mlt = *A. m. ruttneri* from Malta. Locations of mel_irl samples are exemplary, as exact coordinates are unavailable.

2.4. Allele Frequency Correlation

To compare the genetic variability identified based on the two different sequencing approaches, the allele frequencies of the commonly called SNPs in each population were calculated using PLINK1.9 [58] for ind-seq and a custom-built Python script for pool-seq. For each population, the allele frequencies estimated by both approaches were correlated with Pearson's correlation coefficient [59] using the R [60] package ggpubr [61] and plotted using ggplot2 [62].

2.5. Genetic Diversity between and within Populations

Genetic diversity between populations was inferred using F_{ST} distances [63]. For pools, we used PoPoolation2 [51] to calculate pairwise F_{ST} in overlapping window sizes of 20 kb and 10 kb step-size. For ind-seq data, we first used ANGSD [64] to estimate genotype likelihoods from the mapped reads, with loci from repeated regions, low mapping quality (minMapQ < 30), or low base quality (minQ < 20) removed. In addition, we only kept loci covered in at least 100 individuals. Based on the probabilities, we then used the realSFS function to calculate F_{ST} between pairs of populations, using a sliding windows approach (window size 20 kb, step size 10 kb).

Expected heterozygosity (He), as a measure of genetic diversity within populations, was calculated for each population with a custom R script using the above-estimated allele frequencies by pool-seq and ind-seq and the standard formula (2pq) [63].

Genetic diversity and F_{ST} results were plotted in R [60] using the ggplot2 package [62]. Pearson's correlation coefficients [59] were calculated between heterozygosity and pairwise F_{ST} estimates as calculated by the pool-seq and ind-seq approaches, respectively, using the ggpubr package in R [61].

2.6. Population Structure

To infer population structure with the pool-seq data, a principal component analysis was performed using the allele frequencies for each population. For ind-seq data, we first used ANGSD [64] to estimate genotype likelihoods as described above, but with an additional filter to keep loci with a minor allele frequency of no less than 0.05. Based on the posterior genotype probability, PCAngsd [65] was used to calculate the covariance matrix, and ngsDist [66] was used to calculate pairwise genetic distances.

For ind-seq data, population structure was further investigated using a model-based approach. NGSadmix [67] was used to estimate admixture proportions from K = 2 to K = 20, with ten runs for each K. Optimum numbers of K clusters were determined using DeltaK on CLUMPAK [68,69].

Principal components (PCs) and individual model-based ancestries were plotted in R [60] with the ggplot2 package [62].

3. Results

3.1. Sequence Data and Variants

For the pool sequencing approach, we obtained 4,372,477,988 raw reads in total, resulting in an overall genome depth of coverage >1800×. Mean coverage for each pool ranged from 70.3× (cec_grc) to 264.5× (rod_bgr) (Table S1). For the individual sequencing approach, a total of 4,060,632,652 raw reads were generated, resulting in an overall genome depth of coverage ~2600×, with a mean coverage of 17.7× (Table S2).

3.2. Allele Frequency Correlation

A total of 1.6 M and 3.7 M SNPs were called for pool- and ind-seq, respectively. Variants called by both approaches were extracted, leaving 607 K SNPs to calculate the correlations between the allele frequencies generated with either method in each population. The allele frequencies generated with the two methods were very highly correlated with each other, but high variation was observed for single variants (Figure 2; R = 0.92, $p > 0.001$).

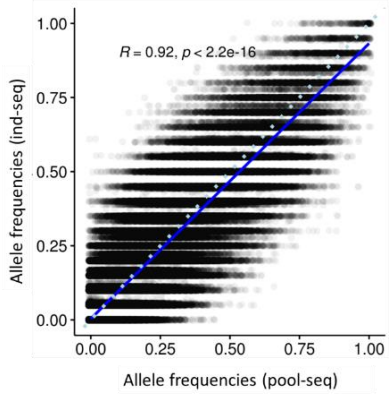

Figure 2. Allele frequency correlation between pool- and ind-seq in each population.

3.3. Genetic Diversity between and within Populations

The analysis of genetic diversity between populations revealed two hierarchical levels of differentiation that were consistently observed based on both pooled populations and individual sequence data: High divergence between populations of different evolutionary lineages (average $F_{ST\ Pools}$ = 0.41 ± 0.11 SD; $F_{ST\ Ind}$ = 0.51 ± 0.12 SD; Table S3) and low divergence between populations of the same lineage (average $F_{ST\ Pools}$ = 0.06 ± 0.03 SD;

$F_{ST\ Ind}$ = 0.09 ± 0.06 SD) (Figure 3A, Tables S4 and S5). While F_{ST} estimates inferred from ind-seq data ($F_{ST\ Ind}$ = 0.40 ± 0.22 SD; Table S5) were higher in all pairwise comparisons than the ones calculated using pool-seq data ($F_{ST\ Pools}$ = 0.27 ± 0.17 SD; Table S4), the correlation between both approaches is extremely high (Figure 3D, R = 0.98, $p < 0.001$), and its visualization as a distance heatmap revealed near identical results (Figure 3B).

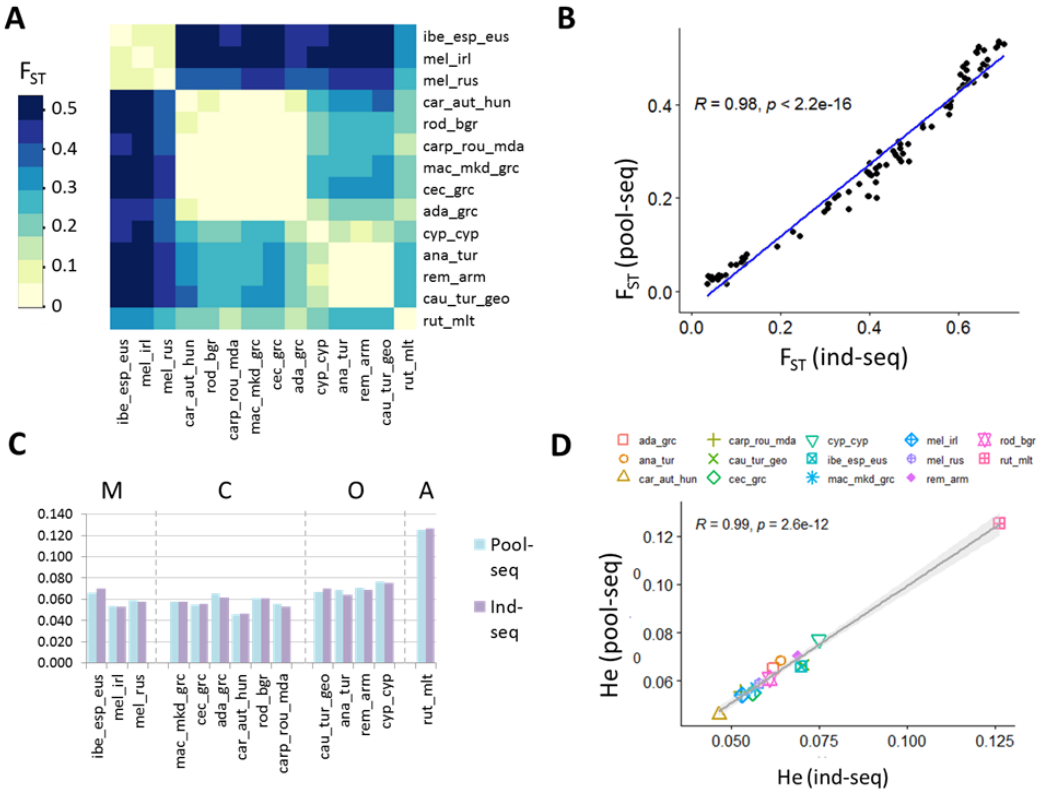

Figure 3. Genetic variation between and within populations. (**A**) Heatmap based on whole-genome pairwise average FST between each population pair calculated from pool-seq data. (**B**) Correlation of pairwise FST estimates between pool-seq and ind-seq data. (**C**) Expected heterozygosity for each population as estimated by pool and ind-seq. (**D**) Correlation between heterozygosities as estimated by pool-seq and ind-seq data. M-lineage: ibe_esp_eus = *A. m. iberiensis* from Spain; mel_irl = *A. m. mellifera* from Ireland; mel_rus = *A. m. mellifera* from Russia; C-lineage: car_aut_hun = *A. m. carnica* from Austria and Hungary; rod_bgr = *A. m. rodopica* from Bulgaria; carp_rou_mda = *A. m. carpatica* from Romania and Moldova; mac_mkd_grc = *A. m. macedonica* from North Macedonia a Northern Greece; cec_grc = *A. m. cecropia* from Greece; ada_grc = *A. m. adami* from Crete, Greece; O-lineage: cyp_cyp = *A. m. cypria* from Cyprus; ana_tur = *A. m. anatoliaca* from Turkey; rem_arm = *A. m. remipes* from Armenia; cau_tur_geo = *A. m. caucasia* from North-East Turkey and Georgia; and A-lineage: rut_mlt = *A. m. ruttneri* from Malta.

Similarly to the allele frequencies and genetic distances, expected heterozygosities were very highly correlated between pool- and ind-seq (Figure 3D; R = 0.98, p = 0.037), and revealed nearly identical results of diversity within populations as estimated by both approaches (Figure 3C): The highest genetic diversity by far was identified in the rut_mlt population which belongs to the African evolutionary lineage. Followed by O lineage populations which have a significantly higher mean diversity (He_{Pools} = 0.071, He_{Ind} = 0.070) than the mean C lineage (He_{Pools} = 0.057, He_{Ind} = 0.056, p = 0.002) and mean M lineage (He_{Pools} = 0.059, He_{Ind} = 0.060, p = 0.031) diversity. The lowest diversity overall, as estimated by pool-seq and ind-seq, was found in the car_aut_hun pool from Austria and Hungary belonging to the C-lineage.

3.4. Population Structure

Overall, the population structure inferred by principal component analysis (PCA) based on data generated with both sequencing approaches was nearly identical, showing a clear separation of the populations into the four main lineages (Figure 4): The first component (PC1) separates the M-lineage from the O and C-lineage, that in turn are separated by the second component (PC2) (Figure 4). In these PCA plots, rut_mlt samples are placed close to the center, and they are clearly distinguished from the rest of the samples on PC3 (Figure 4). PCA by ind-seq further identifies two outliers (one individual each of the ibe_esp_eus and rem_arm population) that are placed distantly to their group members.

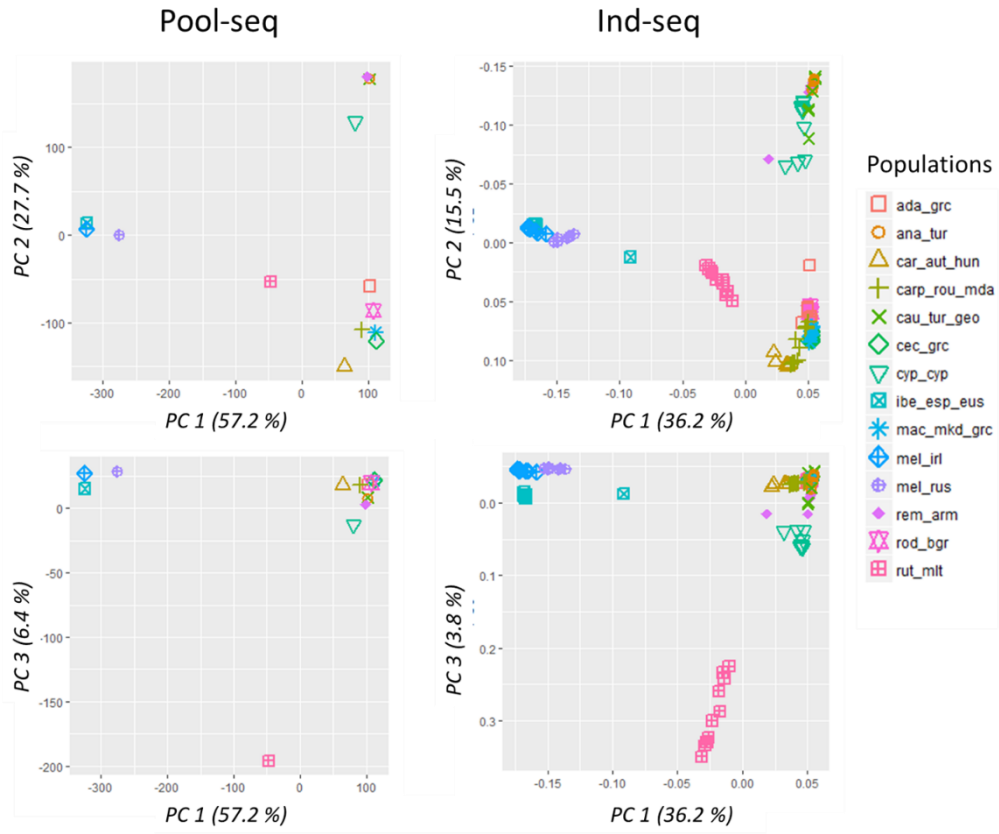

Figure 4. Principal component analysis (PCA) of pool-seq data (**left two panes**) and ind-seq data (**right panes**). Upper panels show the first and second principal components explaining most of the variance, while the lower panels display the first and third components, which only account for 6.4% and 3.8% of the total variation, respectively. M-lineage: ibe_esp_eus = *A. m. iberiensis* from Spain; mel_irl = *A. m. mellifera* from Ireland; mel_rus = *A. m. mellifera* from Russia; C-lineage: car_aut_hun = *A. m. carnica* from Austria and Hungary; rod_bgr = *A. m. rodopica* from Bulgaria; carp_rou_mda = *A. m. carpatica* from Romania and Moldova; mac_mkd_grc = *A. m. macedonica* from North Macedonia a Northern Greece; cec_grc = *A. m. cecropia* from Greece; ada_grc = *A. m. adami* from Crete, Greece; O-lineage: cyp_cyp = *A. m. cypria* from Cyprus; ana_tur = *A. m. anatoliaca* from Turkey; rem_arm = *A. m. remipes* from Armenia; cau_tur_geo = *A. m. caucasia* from North-East Turkey and Georgia; and A-lineage: rut_mlt = *A. m. ruttneri* from Malta.

Model-based ancestry was further investigated in individual samples. The optimal number of clusters as inferred by Evanno's DeltaK was K = 3 (Figure S1) that separated the individuals into three major clusters coinciding with the lineages M, O, and C, and leaving rut_mlt individuals with an intermediate mixed genetic background (Figure 5, top panel). The second-best K = 6 (Figure S2) separates rut_mlt into its own cluster and, within the O-lineage, differentiates cyp_cyp from the other three subspecies. Also, within this lineage, a substructure within cau_tur_geo becomes visible, where about half the samples

display two different ancestries. At K = 6, we further observe a differentiation within the C-lineage, between car_aut_hun from the northern part of the distribution and all other populations (mac_mkd_grc, cec_grc, ada_grc, rod_bgr) in the southern part of the lineage range; however, the individuals of carp_rou_mda display mixed ancestry with varying degrees of both genetic backgrounds. Individuals with a mixed genetic background were also identified in a few other populations, for instance, one each of ibe_esp_eus and rem_arm has already seen in the PCA, but also several rut_mlt and cyp_cyp individuals.

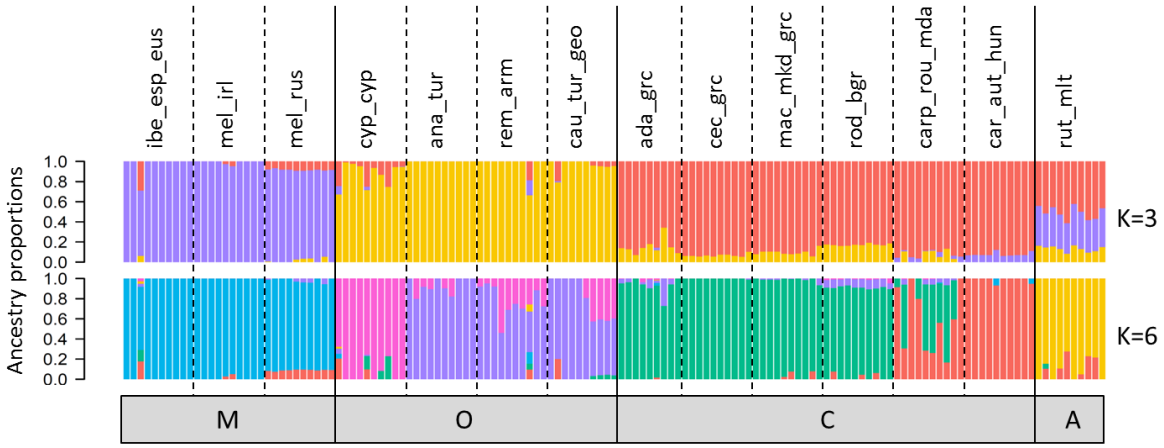

Figure 5. Model-based ancestry as calculated with NGSAdmix for best (K = 3) and second-best (K = 6) number of K ancestral populations. Each individual is represented by a vertical bar and colored according to the proportion of the genome that was derived from one of K clusters. Samples are ordered according to evolutionary lineage and sampling population. M-lineage: ibe_esp_eus = *A. m. iberiensis* from Spain; mel_irl = *A. m. mellifera* from Ireland; mel_rus = *A. m. mellifera* from Russia; C-lineage: car_aut_hun = *A. m. carnica* from Austria and Hungary; rod_bgr = *A. m. rodopica* from Bulgaria; carp_rou_mda = *A. m. carpatica* from Romania and Moldova; mac_mkd_grc = *A. m. macedonica* from North Macedonia a Northern Greece; cec_grc = *A. m. cecropia* from Greece; ada_grc = *A. m. adami* from Crete, Greece; O-lineage: cyp_cyp = *A. m. cypria* from Cyprus; ana_tur = *A. m. anatoliaca* from Turkey; rem_arm = *A. m. remipes* from Armenia; cau_tur_geo = *A. m. caucasia* from North-East Turkey and Georgia; and A-lineage: rut_mlt = *A. m. ruttneri* from Malta.

4. Discussion

Whole-genome sequencing of individuals provides genetic data at the highest resolution. However, this rather expensive approach can only be applied to a limited sample size, while pool-seq grants data for a much larger sample set. In this study, we empirically evaluated these approaches to infer population structure and diversity in a real dataset of European *A. mellifera* populations. Comparing the two different sequencing approaches (pool-seq and ind-seq), we found that both revealed a population genetic structure and genetic diversities of European honey bees that were nearly identical. Moreover, the results are in good concordance with previous findings (e.g., [16,18,19,70]), although most previous studies were based on different populations and smaller datasets. As either method comes with specific advantages, cost-effectiveness in the case of pool-seq and depth of information in the case of ind-seq, a cost-efficient strategy could be to combine both approaches.

4.1. The Limited Ability of Pool-Seq to Identify Low-Frequency Variants

The pool-seq approach also identified only a fraction of the SNPs found by ind-seq. However, this difference is expected, considering that the total sequencing depth of individuals (2600× for 139 individuals) was higher than the one for pools (1800× for 14 pools of 100 individuals each). In addition, by subsampling pool sequence data to uniform coverage and requiring a strict filter of minimum read count 3 to call a variant, we could only detect SNPs with a minor allele frequency of >0.06. This was already pointed

out by Cutler and Jensen [11], who showed that low-frequency variants are lost in pool-seq experiments when appropriate call filters are set that consider sequencing error rates. Moreover, variants may remain undetected if equal molar concentrations represent not all individuals in the pool. While we cannot rule out this possibility in our experiments, any variance because of pooling is expected to be small, as about 90 individuals were used in each pool. In any case, the inference of population structure is not influenced by the exact number of variants. On the contrary, rare and low-frequency variants are typically filtered for such analyses (e.g., Refs. [6,71,72]). Also, the importance rather lies in the accurate estimates of common variants [10]. However, for other applications such as the association of rare variants to disease or specific phenotypes [73,74], this point might be more critical and needs to be considered in the sampling and sequencing strategy, as advised elsewhere [10,11].

4.2. Sampling and the Importance of Previous Knowledge for Pool-Seq

The total sample size of individuals that are pooled is a crucial parameter that influences the accuracy of the allele frequency estimations [75,76]. By performing a large-scale and comprehensive sampling and including ~90 individuals per pool, we ensured that the allele frequencies obtained by pool-seq could be regarded as representative of the populations of the subspecies and regions under study. In contrast, allele frequency estimation based on ind-seq relies on the few samples chosen per population, ten in our case. Therefore, pooling is advantageous if it can be based on prior knowledge of the population structure. It is thus important to include additional data, for instance from morphology or previous genetic studies, when deciding which individuals should be included in a pool. Although we were able to base the sampling in some of our populations on previous genetic studies (see references in [17]), in some populations admixed individuals were identified (e.g., rem_arm, cau_tur_geo, carp_rou_mda, rut_mlt). Here, pool-based estimates could potentially yield biased information, as the presence of hidden substructures will obscure the estimated mean allele frequencies. Moreover, any possibly present Wahlund effect [77] caused by hidden substructures would not be detectable either, since the observed heterozygosity is not accessible via pool-seq. In other words, pooling limits us to the overall view by assuming genetic homogeneity between the individuals that constitute the population sample, so any potentially present heterogeneity within the population cannot be detected. In consequence, it is recommendable to use caution in choosing the individuals for a sequencing pool and, if in doubt about genetic homogeneity, to exclude individuals or to set up separate pools.

4.3. Near Identical Inference of Population Structure and Diversity by Pool-Seq and Ind-Seq

A remarkable finding in our study is that, despite the huge difference in sampling coverage between pool-seq and ind-seq, allele frequencies obtained with the two methods correlate extremely well with each other (R = 0.92, Figure 2), indicating that with as few as 10 individuals, good estimates of average population structure can be achieved. This is further evidenced by the correlation of F_{ST} values (Figure 3B) which reveals near identical results between the two approaches. Equally highly correlated were the genetic diversities within populations as estimated by expected heterozygosity (Figure 3D). Similar to our results, Dorant et al. [6] evaluated genotyping-by-sequencing, pool-seq, and RAD capture approaches and identified very congruent results by the three tested methods in identifying weak population structure of a Homarus americanus population (correlation coefficients $R > 0.9$). Also, in natural populations of *Arabidopsis halleri*, a non-model plant species, highly correlated allele frequencies ($R > 0.98$) were identified between pool-seq and ind-seq. Thus, although only few studies empirically evaluated the accuracy of allele frequency estimates derived from pool-seq and ind-seq in natural populations, all arrive at very high correlations and demonstrate that either approach applies to population genomics studies.

4.4. Both Approaches Compare Well to Established Studies

Regarding the population structure and differentiation, it is quite remarkable that our results based on using whole-genome sequence data at a high depth and of a comprehensive and unprecedented sample set are comparable to the population genetic structure of European honey bees proposed by F. Ruttner in the 1980s based on morphometric analyses [18]. Namely, high genetic divergences are found between the four lineages, while moderate and minor differences appear between subspecies within lineages. This finding is consistent with other published literature throughout the years studying different populations and using different tools from classical morphometry [18,30,36], over geometric-morphometrics [37], to microsatellites [27,31,35] and SNP markers [16,17]. Hence, while advanced technologies allow us to sequence samples at the whole-genome level and thereby gain maximal data, if the aim is simply to only infer population structure, whole-genome sequencing can be an overshoot, in particular in conservation genomics applications [1,78]. In contrast, if the aim is to identify local adaptations in natural populations, whole-genome sequencing enables the identification of signatures of selection [79–82], a limited number of genetic markers cannot identify that. For this application, pool-seq has also been successfully applied before [52,83–85].

The highest genetic diversity by far was identified in the rut_mlt population as a representative of the African evolutionary lineage (Figure 3A), which is known to be the lineage with the highest genetic diversity as identified in previous studies [16,72,86]. The lowest diversity was found in the car_aut_hun population from Austria and Hungary belonging to the C-lineage (Figure 3A). Already in other studies *A. m. carnica* has been identified as the subspecies with the lowest genetic diversity [16,38]. A possible explanation of the lower diversity consistently identified in this subspecies could be genetic drift caused by selective breeding, as *A. m. carnica* is one of the most popular honey bee subspecies used for breeding [87]. Among the O lineage, the cyp_cyp population, despite originating from a small island, revealed the highest expected heterozygosity in this group, most likely due to hybridization with other subspecies that are imported to the island, and known to increase diversity [86,88]. Moreover, within the M lineage the highest genetic diversity was identified in the ibe_eus_esp population, in concordance with the Iberian Peninsula being described as a glacial refuge of M-lineage diversity [89,90], while the two *A. m. mellifera* populations reflect distant (mel_rus) or isolated (mel_irl) populations, which may thus potentially have lost diversity through genetic drift during recolonization. In particular, in the case of the *A. m. mellifera* island population of Ireland, where limited human inference through importation, is suspected [43,91].

4.5. The Cost-Benefit Ratio between Pool-Seq and Ind-Seq

Since similar results are obtained with both sequencing approaches and results are concordant with published literature, the issue of costs involved with either approach gains additional weight. In general, the cost of pool-seq is considered lower [9,12,13]. To enable a direct comparison, we recently (2021) enquired sequencing quotes similar to the magnitude of our study with pool-seq (14 pools, 14 extractions, 14 library preparations, and total target depth 1400×) and ind-seq (140 individuals, 140 extractions, 140 library preparations, total depth 1400×) from a European and from a Chinese company. For ind-seq the estimates were 19,689 € and 44,800 ¥, respectively, while for pool-seq they were 7003 € and 19,600 ¥, that is, 65% to 55% less expensive. This considerable price difference between the two approaches is certainly a relevant point of consideration for scientists with low to medium research budgets. It is to note that for organisms with relatively small genome size, such as the honey bee (236 Mb), the extraction and library preparation steps constitute a large portion of the costs in comparison with the cost of actual sequencing, and therefore, applying the pool-seq approach becomes more cost-efficient. In contrast, when studying species with large genomes, such as plants (e.g., Ref. [15]), the cost of the actual amount of sequence data (Gb) needed will account for a higher proportion of the total cost and increase for both methods, resulting in a smaller difference between them (Figure 6).

On the other hand, for reduced representation approaches such as RAD-seq, GBS, or exome capture, for which the actual sequenced part of the genome can be very small, pool-seq will become much more affordable, e.g., the cost ratio approaching 10, as the sequencing cost would be very low (Figure 6), highlighting the advantages of pool-seq for ecological studies of non-model organisms [6,92,93].

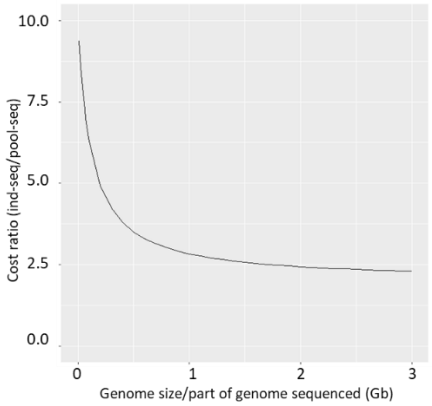

Figure 6. Cost ratio ind-seq/pool-seq. The x-axis represents the genome size or part of the genome sequenced (Gb) in case of reduced representation techniques. The y-axis is the ratio of the total cost of ind-seq/pool-seq (assuming 1400× total coverage for ind-seq and 700× total coverage for pool-seq).

On the other hand, besides the cost of sequencing itself, we need to consider the actual cost of collecting the samples required for either approach. In our case, the sampling of ~90 individuals per population, each one originating from a different colony and apiary and a total of 16 different countries, constituted a huge effort and generated extra costs. In comparison, the sampling of only ten individuals per population is less complex and labor-intensive.

4.6. Additional Insights from Ind-Seq

While pool-seq and ind-seq reveal similar overall results on population structure and genetic diversity, albeit, at different costs, ind-seq can provide additional information at the individual level. In this way, we could identify individuals with a mixed genetic background that is indicative of the existence of hybridization between subspecies and that would give rise to a greater resemblance between populations when analyzed by pooling. For instance, by natural hybridizations in the contact areas of two subspecies, as we find between *A. m. remipes* and *A. m. caucasia* on the borders between Turkey and Armenia, and between lineages as has been reported previously [29,89,94–96]. Or hybridizations due to contemporary human-mediated processes, as we see in the case of *A. m. ruttneri* on the island of Malta [97].

An additional advantage of individual whole-genome sequence data is that further analyses can be performed, such as the inference of evolutionary and demographic histories [72,98–100]. Although not the focus of this study, such analyses can give additional insights that may be particularly important for non-model species and in a conservation context [1,101].

5. Conclusions

Our case study on *A. mellifera* is a good example to evaluate two methods. The species is well studied, and we could compare our results with previous studies that used different tools and populations. While overall results of both approaches were very similar, our final verdict for the empirical comparison between pool-seq and ind-seq to investigate population structure in the honey bee is that ind-seq, albeit more expensive, could give us similar or even equal information, while additionally providing insights into individual-based admixture. Based on our experience, we would thus not necessarily recommend

to only applying the pool-seq approach in similar studies. Nevertheless, pool-seq can be a useful and cost-efficient option, for instance for variant discovery [17,102], and/or in combination with ind-seq for other less well-studied species.

Both approaches, pool-seq and ind-seq, enabled us to get a global vision of European honey bee diversity. The population structure of some of the studied populations was genetically examined for the first time. It will be interesting to analyze them more in-depth to shed light on complex patterns of diversity. For instance, within the C lineage, where multiple highly interrelated subspecies exist in close geographical proximity, we found some level of mixed genetic background in several populations. The comprehensive genomic dataset generated by this study will therefore be the basis for future studies to explore further the genetic variation within and among subspecies and to identify signatures of local adaptations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13020182/s1, Table S1: mapping statistics pool-seq data, Table S2: mapping statistics individual whole-genome seq data, Table S3: lineage differentiation (F_{ST}) based on pool and individual seq data, Table S4: differentiation between populations (F_{ST}) based on pool-seq data, Table S5: differentiation between populations (F_{ST}) based on individual seq data, Figure S1: delta k plot of Evanno's test based on NGSadmix analysis for K = 2–20, Figure S2: delta K plot of Evanno's test based on NGSadmix analysis for K = 3–20.

Author Contributions: Conceptualization, M.M., A.E., P.K., M.B., W.S. and R.V.; Formal analysis, C.C., J.L., M.P., J.M. and R.O.N.; writing—original draft preparation, M.P. and C.C.; writing—review and editing, A.E., M.M., P.K. and M.B.; resources, SMARTBEES WP3 DIVERSITY COLLABORATORS; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (Grant No. 31902219) and Modern Agro-industry Technology Research System (Grant No. CARDS-45-KXJ1). The SmartBees project was funded by the European Commission under its FP7 KBBE programme (2013.1.3-02, SmartBees Grant Agreement number 613960). MP and J.L were supported by the Applied Genomics and Bioinformatics research group (IT1233-19) funded by the Basque Government grant IT1233-19. Additionally, JL was funded by the grant PRE_2017_2_0169 from the Department of Education of the Basque Government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data from pool sequencing is available on NCBI's short read archive (SRA) under accession PRJNA666033. Individual sequencing data is deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0001986. The pipeline for the analysis of the pool sequence data is available at https://github.com/jlanga/smsk_popoolation (accessed on 3 December 2021).

Acknowledgments: We are especially grateful to all beekeepers, bee breeders and other contributors who provided the valuable samples for this work. We also wish to give a special thanks to all technicians involved in the project, particularly Mahesha Perera for her great contribution to 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.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Collaborators of the SMARTBEES WP3 DIVERSITY(in alphabetical order):

Eliza Căuia (Institutul de Cercetare Dezvoltare pentru Apicultura SA, Bucharest, Romania), Mary F. Coffey (University of Limerick, Limerick, Ireland), Thomas Galea (Breeds of Origin, Haz-Zebbug, Malta), Miroljub Golubovski (MacBee Association, Skopje, North Macedonia), Karina Grigoryan (Yerevan State University, Yerevan, Armenia), Fani Hatjina (Department of Apiculture, Agricultural Organization 'DEMETER', Thessaloniki, Greece), Rustem Ilyasov (Vavilov Institute of General Genetics of Russian Academy of Sciences,

Moscow, Russia), Timothea Ioannou (Department of Agriculture, Limassol, Cyprus), Dimosthenis Isaakidis (Beekeeping Center of Crete, Heraklion, Crete, Greece), Evgeniya Ivanova (University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria), Irakli Janashia (Agricultural University of Georgia, Tbilisi, Georgia), Irfan Kandemir (Ankara University, Ankara, Turkey), Aikaterini Karatasou (Federation of Greek Beekeepers' Associations, Larissa, Greece), Enikö Sz. Matray (Hungarian Bee Breeders Association, Budapest, Hungary), David Mifsud (Division of Rural Sciences and Food Systems, Institute of Earth Systems, University of Malta, Msida, Malta), Rudolf Moosbeckhofer (Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH, Wien, Austria), Alexei G. Nikolenko (Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia), Alexandros Papachristoforou (Cyprus University of Technology, Limassol, Cyprus), Plamen Petrov (Agricultural University of Plovdiv, Plovdiv, Bulgaria), Aleksandr V. Poskryakov (Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia), Aglyam Y. Sharipov (Shulgan-Tash Nature Reserve, Burzyansky District, Russia), Adrian Siceanu (Institutul de Cercetare Dezvoltare pentru Apicultura SA, Bucharest, Romania), Aleksandar Uzunov (Landesbetrieb Landwirtschaft Hessen, Bee Institute Kirchhain, Kirchhain, Germany; Faculty of Agricultural Sciences and Food, University Ss. Cyril and Methodius, Skopje, Republic of Macedonia), Marion Zammit-Mangion (Department of Physiology and Biochemistry, University of Malta, Msida, Malta).

References

1. Allendorf, F.W.; Luikart, G.; Aitken, S.N. *Conservation and the Genetics of Populations*, 2nd ed.; Wiley-Blackwell: Hoboken, NJ, USA, 2012.
2. Luikart, G.; Kardos, M.; Hand, B.K.; Rajora, O.; Aitken, S.; Hohenlohe, P.A. *Population Genomics: Advancing Understanding of Nature*; Springer: Cham, Germany, 2018.
3. Foote, A.D.; Vijay, N.; Avila-Arcos, M.C.; Baird, R.W.; Durban, J.W.; Fumagalli, M.; Gibbs, R.A.; Hanson, M.B.; Korneliussen, T.S.; Martin, M.D.; et al. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat. Commun.* **2016**, *7*, 11693. [CrossRef]
4. Baird, N.A.; Etter, P.D.; Atwood, T.S.; Currey, M.C.; Shiver, A.L.; Lewis, Z.A.; Selker, E.U.; Cresko, W.A.; Johnson, E.A. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* **2008**, *3*, e3376. [CrossRef]
5. Choi, M.; Scholl, U.I.; Ji, W.; Liu, T.; Tikhonova, I.R.; Zumbo, P.; Nayir, A.; Bakkaloglu, A.; Ozen, S.; Sanjad, S.; et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19096–19101. [CrossRef]
6. Dorant, Y.; Benestan, L.; Rougemont, Q.; Normandeau, E.; Boyle, B.; Rochette, R.; Bernatchez, L. Comparing Pool-seq, Rapture, and GBS genotyping for inferring weak population structure: The American lobster (*Homarus americanus*) as a case study. *Ecol. Evol.* **2019**, *9*, 6606–6623. [CrossRef]
7. Davey, J.W.; Hohenlohe, P.A.; Etter, P.D.; Boone, J.Q.; Catchen, J.M.; Blaxter, M.L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **2011**, *12*, 499–510. [CrossRef]
8. Andrews, K.R.; Good, J.M.; Miller, M.R.; Luikart, G.; Hohenlohe, P.A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **2016**, *17*, 81–92. [CrossRef]
9. Futschik, A.; Schloetterer, C. The Next Generation of Molecular Markers from Massively Parallel Sequencing of Pooled DNA Samples. *Genetics* **2010**, *186*, 207–218. [CrossRef]
10. Schloetterer, C.; Tobler, R.; Kofler, R.; Nolte, V. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **2014**, *15*, 749–763. [CrossRef]
11. Cutler, D.; Jensen, J. To Pool, or Not to Pool? *Genetics* **2010**, *186*, 41–43. [CrossRef]
12. Gautier, M.; Foucaud, J.; Gharbi, K.; Cezard, T.; Galan, M.; Loiseau, A.; Thomson, M.; Pudlo, P.; Kerdelhue, C.; Estoup, A. Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Mol. Ecol.* **2013**, *22*, 3766–3779. [CrossRef]
13. Saelao, P.; Simone-Finstrom, M.; Avalos, A.; Bilodeau, L.; Danka, R.; de Guzman, L.; Rinkevich, F.; Tokarz, P. Genome-wide patterns of differentiation within and among US commercial honey bee stocks. *BMC Genom.* **2020**, *21*, 704. [CrossRef]
14. Yunusbaev, U.B.; Kaskinova, M.D.; Ilyasov, R.A.; Gaifullina, L.R.; Saltykova, E.S.; Nikolenko, A.G. The Role of Whole-genome Research in the Study of Honey Bee Biology. *Russ. J. Genet.* **2019**, *55*, 778–787. [CrossRef]
15. Rellstab, C.; Zoller, S.; Tedder, A.; Gugerli, F.; Fischer, M.C. Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species. *PLoS ONE* **2013**, *8*, e80422. [CrossRef]

16. Wallberg, A.; Han, F.; Wellhagen, G.; Dahle, B.; Kawata, M.; Haddad, N.; Paulino Simoes, Z.L.; Allsopp, M.H.; Kandemir, I.; De la Rua, P.; 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. [CrossRef]
17. Momeni, J.; Parejo, M.; Nielsen, R.O.; Langa, J.; Montes, I.; Papoutsis, L.; Farajzadeh, L.; Bendixen, C.; Cauia, E.; Charriere, J.-D.; et al. Authoritative subspecies diagnosis tool for European honey bees based on ancestry informative SNPs. *BMC Genom.* **2021**, *22*, 101. [CrossRef]
18. Ruttner, F. *Biogeography and Taxonomy of Honeybees*; Springer: Berlin/Heidelberg, Germany, 1988; pp. 163–257.
19. Garnery, L.; Cornuet, J.M.; Solignac, M. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Mol. Ecol.* **1992**, *1*, 145–154. [CrossRef]
20. Sheppard, W.S.; Arias, M.C.; Grech, A.; Meixner, M.D. *Apis mellifera ruttneri*, a new honey bee subspecies from Malta. *Apidologie* **1997**, *28*, 287–293. [CrossRef]
21. Kandemir, I.; Kence, M.; Sheppard, W.S.; Kence, A. Mitochondrial DNA variation in honey bee (*Apis mellifera* L.) populations from Turkey. *J. Apic. Res.* **2006**, *45*, 33–38. [CrossRef]
22. De la Rua, P.; Jaffe, R.; Dall'Olio, R.; Munoz, I.; Serrano, J. Biodiversity, conservation and current threats to European honeybees. *Apidologie* **2009**, *40*, 263–284. [CrossRef]
23. Ilyasov, R.A.; Lee, M.-L.; Takahashi, J.-I.; Kwon, H.W.; Nikolenko, A.G. A revision of subspecies structure of western honey bee *Apis mellifera*. *Saudi J. Biol. Sci.* **2020**, *27*, 3615–3621. [CrossRef]
24. Bouga, M.; Harizanis, P.C.; 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*, 335–344. [CrossRef]
25. Ivanova, E.; Bouga, M.; Staykova, T.; Mladenovic, M.; Rasic, S.; Charistos, L.; Hatjina, F.; Petrov, P. The genetic variability of honey bees from the Southern Balkan Peninsula, based on alloenzymic data. *J. Apic. Res.* **2012**, *51*, 329–335. [CrossRef]
26. Munoz, I.; Alice Pinto, M.; De la Rua, P. Temporal changes in mitochondrial diversity highlights contrasting population events in Macaronesian honey bees. *Apidologie* **2013**, *44*, 295–305. [CrossRef]
27. Uzunov, A.; Meixner, M.D.; Kiprijanovska, H.; Andonov, S.; Gregorc, A.; Ivanova, E.; Bouga, M.; Dobi, P.; Buechler, R.; Francis, R.; et al. Genetic structure of *Apis mellifera macedonica* in the Balkan Peninsula based on microsatellite DNA polymorphism. *J. Apic. Res.* **2014**, *53*, 288–295. [CrossRef]
28. Ilyasov, R.A.; Poskryakov, A.V.; Petukhov, A.V.; Nikolenko, A.G. Molecular genetic analysis of five extant reserves of black honeybee *Apis melifera melifera* in the Urals and the Volga region. *Russ. J. Genet.* **2016**, *52*, 828–839. [CrossRef]
29. Franck, P.; Garnery, L.; Celebrano, G.; Solignac, M.; Cornuet, J.M. Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A.* m. sicula). *Mol. Ecol.* **2000**, *9*, 907–921. [CrossRef]
30. Bouga, M.; Alaux, C.; Bienkowska, M.; Buechler, R.; Carreck, N.L.; Cauia, E.; Chlebo, R.; Dahle, B.; Dall'Olio, R.; De la Rua, P.; et al. A review of methods for discrimination of honey bee populations as applied to European beekeeping. *J. Apic. Res.* **2011**, *50*, 51–84. [CrossRef]
31. Canovas, F.; de la Rua, P.; Serrano, J.; Galian, J. Microsatellite variability reveals beekeeping influences on Iberian honeybee populations. *Apidologie* **2011**, *42*, 235–251. [CrossRef]
32. Alice Pinto, M.; Henriques, D.; Chavez-Galarza, J.; Kryger, P.; Garnery, L.; van der Zee, R.; Dahle, B.; Soland-Reckeweg, G.; de la Rua, P.; Dall'Olio, 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. [CrossRef]
33. Garnery, L.; Franck, P.; Baudry, E.; Vautrin, D.; Cornuet, J.-M.; Solignac, M. Genetic diversity of the west European honey bee (*Apis mellifera mellifera* and *A. m. iberica*). II. Microsatellite loci. *Genet. Sel. Evol.* **1998**, *30*, S49–S74. [CrossRef]
34. Garnery, L.; Franck, P.; Baudry, E.; Vautrin, D.; Cornuet, J.-M.; Solignac, M. Genetic diversity of the west European honey bee (*Apis mellifera mellifera* and *A. m. iberica*). I. Mitochondrial DNA. *Genet. Sel. Evol.* **1998**, *30*, S31–S47. [CrossRef]
35. Miguel, I.; Iriondo, M.; Garnery, L.; Sheppard, W.S.; 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. [CrossRef]
36. Meixner, M.D.; Worobik, M.; Wilde, J.; Fuchs, S.; Koeniger, N. *Apis mellifera mellifera* in eastern Europe-morphometric variation and determination of its range limits. *Apidologie* **2007**, *38*, 191–197. [CrossRef]
37. Kandemir, I.; Ozkan, A.; Fuchs, S. Reevaluation of honeybee (*Apis mellifera*) microtaxonomy: A geometric morphometric approach. *Apidologie* **2011**, *42*, 618–627. [CrossRef]
38. 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. [CrossRef]
39. Engel, M.S. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; Apis). *J. Hymenopt. Res.* **1999**, *8*, 165–196.
40. Foti, N. Researches on morphological characteristics and biological features of the bee population in Romania. In Proceedings of the XXth Jubiliar International Congress of Beekeeping Apimondia, Bucharest, Romania, 15–18 February 1965; pp. 171–176.
41. Petrov, P. *Systematics of Bulgarian Bees*; Pchelarstvo: Sofia, Bulgaria, 1991.
42. Ilyasov, R.; Nikolenko, A.; Tuktarov, V.; Goto, K.; Takahashi, J.-I.; Kwon, H.W. Comparative analysis of mitochondrial genomes of the honey bee subspecies *A. m. caucasica* and *A. m. carpathica* and refinement of their evolutionary lineages. *J. Apic. Res.* **2019**, *58*, 567–579. [CrossRef]

43. Hassett, J.; Browne, K.A.; McCormack, G.P.; Moore, E.; Soland, G.; Geary, M.; Native Irish Honey Bee, S. A significant pure population of the dark European honey bee (*Apis mellifera mellifera*) remains in Ireland. *J. Apic. Res.* **2018**, *57*, 337–350. [CrossRef]
44. Francis, R.M.; Kryger, P.; Meixner, M.; Bouga, M.; Ivanova, E.; Andonov, S.; Berg, S.; Bienkowska, M.; Büchler, R.; Charistos, L.; 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. [CrossRef]
45. Evans, J.D.; Schwarz, R.S.; Chen, Y.P.; Budge, G.; Cornman, R.S.; De la Rua, P.; de Miranda, J.R.; Foret, S.; Foster, L.; Gauthier, L.; et al. Standard methods for molecular research in *Apis mellifera*. *J. Apic. Res.* **2013**, *52*, 1–54. [CrossRef]
46. Doyle, J. Isolation of Plant DNA from Fresh Tissue. *Focus* **1990**, *12*, 13–15.
47. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
48. Elsik, C.G.; Worley, K.C.; Bennett, A.K.; Beye, M.; Camara, F.; Childers, C.P.; de Graaf, D.C.; Debyser, G.; Deng, J.; Devreese, B.; et al. Finding the missing honey bee genes: Lessons learned from a genome upgrade. *BMC Genom.* **2014**, *15*, 86. [CrossRef]
49. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
50. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data, P. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
51. Kofler, R.; Pandey, R.V.; Schloetterer, C. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **2011**, *27*, 3435–3436. [CrossRef]
52. Ruiz-Larranaga, O.; Langa, J.; Rendo, F.; Manzano, C.; Iriondo, M.; Estonba, A. Genomic selection signatures in sheep from the Western Pyrenees. *Genet. Sel. Evol.* **2018**, *50*, 9. [CrossRef]
53. Gruening, B.; Dale, R.; Sjoedin, A.; Chapman, B.A.; Rowe, J.; Tomkins-Tinch, C.H.; Valieris, R.; Koester, J.; Team, B. Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **2018**, *15*, 475–476. [CrossRef]
54. Koester, J.; Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **2012**, *28*, 2520–2522. [CrossRef]
55. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, 884–890. [CrossRef]
56. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
57. Chiang, C.; Layer, R.M.; Faust, G.G.; Lindberg, M.R.; Rose, D.B.; Garrison, E.P.; Marth, G.T.; Quinlan, A.R.; Hall, I.M. SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* **2015**, *12*, 966–968. [CrossRef]
58. Chang, C.C.; Chow, C.C.; Tellier, L.C.A.M.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, s13742-015. [CrossRef]
59. Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242.
60. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 3 February 2021).
61. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. R Package Version 0.4.0. Available online: https://CRAN.R-project.org/package=ggpubr (accessed on 3 February 2021).
62. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
63. Weir, B.S.; Cockerham, C.C. Estimating F-statistics for the analysis of population structure. *Evol. Int. J. Org. Evol.* **1984**, *38*, 1358–1370. [CrossRef]
64. Korneliussen, T.S.; Albrechtsen, A.; Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinform.* **2014**, *15*, 356. [CrossRef]
65. Meisner, J.; Albrechtsen, A. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics* **2018**, *210*, 719–731. [CrossRef]
66. Vieira, F.G.; Lassalle, F.; Korneliussen, T.S.; Fumagalli, M. Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biol. J. Linn. Soc.* **2016**, *117*, 139–149. [CrossRef]
67. Skotte, L.; Korneliussen, T.S.; Albrechtsen, A. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics* **2013**, *195*, 693–702. [CrossRef]
68. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [CrossRef]
69. Kopelman, N.M.; Mayzel, J.; Jakobsson, M.; Rosenberg, N.A.; Mayrose, I. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **2015**, *15*, 1179–1191. [CrossRef]
70. Franck, P.; Garnery, L.; Solignac, M.; Cornuet, J.M. Molecular confirmation of a fourth lineage in honeybees from the Near East. *Apidologie* **2000**, *31*, 167–180. [CrossRef]
71. Roesti, M.; Salzburger, W.; Berner, D. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* **2012**, *12*, 94. [CrossRef]
72. Cridland, J.M.; Tsutsui, N.D.; Ramirez, S.R. The Complex Demographic History and Evolutionary Origin of the Western Honey Bee, *Apis Mellifera*. *Genome Biol. Evol.* **2017**, *9*, 457–472. [CrossRef]
73. Fournier, T.; Abou Saada, O.; Hou, J.; Peter, J.; Caudal, E.; Schacherer, J. Extensive impact of low-frequency variants on the phenotypic landscape at population-scale. *Elife* **2019**, *8*, e49258. [CrossRef]
74. Momozawa, Y.; Mizukami, K. Unique roles of rare variants in the genetics of complex diseases in humans. *J. Hum. Genet.* **2021**, *66*, 11–23. [CrossRef]

75. Anderson, E.C.; Skaug, H.J.; Barshis, D.J. Next-generation sequencing for molecular ecology: A caveat regarding pooled samples. *Mol. Ecol.* **2014**, *23*, 502–512. [CrossRef]
76. Rode, N.O.; Holtz, Y.; Loridon, K.; Santoni, S.; Ronfort, J.; Gay, L. How to optimize the precision of allele and haplotype frequency estimates using pooled-sequencing data. *Mol. Ecol. Resour.* **2018**, *18*, 194–203. [CrossRef]
77. Garnier-Géré, P.; Chikhi, L. Population Subdivision, Hardy-Weinberg Equilibrium and the Wahlund Effect. *eLS* **2013**. [CrossRef]
78. Kristensen, T.N.; Pedersen, K.S.; Vermeulen, C.J.; Loeschcke, V. Research on inbreeding in the 'omic' era. *Trends Ecol. Evol.* **2010**, *25*, 44–52. [CrossRef]
79. Parejo, M.; Wragg, D.; Henriques, D.; Charriere, J.-D.; Estonba, A. Digging into the Genomic Past of Swiss Honey Bees by Whole-Genome Sequencing Museum Specimens. *Genome Biol. Evol.* **2020**, *12*, 2535–2551. [CrossRef]
80. Parejo, M.; Wragg, D.; Henriques, D.; Vignal, A.; Neuditschko, M. Genome-wide scans between two honeybee populations reveal putative signatures of human-mediated selection. *Anim. Genet.* **2017**, *48*, 704–707. [CrossRef]
81. Henriques, D.; Wallberg, A.; Chavez-Galarza, J.; Johnston, J.S.; Webster, M.T.; Alice Pinto, M. Whole genome SNP-associated signatures of local adaptation in honeybees of the Iberian Peninsula. *Sci. Rep.* **2018**, *8*, 11145. [CrossRef]
82. Wragg, D.; Marti-Marimon, M.; Basso, B.; Bidanel, J.-P.; Labarthe, E.; Bouchez, O.; Le Conte, Y.; Vignal, A. Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Sci. Rep.* **2016**, *6*, 27168. [CrossRef]
83. Bastide, H.; Betancourt, A.; Nolte, V.; Tobler, R.; Stoebe, P.; Futschik, A.; Schloetterer, C. A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in Drosophila melanogaster. *PLoS Genet.* **2013**, *9*, e1003534. [CrossRef]
84. Beissinger, T.M.; Hirsch, C.N.; Vaillancourt, B.; Deshpande, S.; Barry, K.; Buell, C.R.; Kaeppler, S.M.; Gianola, D.; de Leon, N. A Genome-Wide Scan for Evidence of Selection in a Maize Population Under Long-Term Artificial Selection for Ear Number. *Genetics* **2014**, *196*, 829–840. [CrossRef]
85. Lattorff, H.M.G.; Buchholz, J.; Fries, I.; Moritz, R.F.A. A selective sweep in a *Varroa destructor* resistant honeybee (*Apis mellifera*) population. *Infect. Genet. Evol.* **2015**, *31*, 169–176. [CrossRef]
86. Whitfield, C.W.; Behura, S.K.; Berlocher, S.H.; Clark, A.G.; Johnston, J.S.; Sheppard, W.S.; Smith, D.R.; Suarez, A.V.; Weaver, D.; Tsutsui, N.D. Thrice out of Africa: Ancient and recent expansions of the honey bee, *Apis mellifera*. *Science* **2006**, *314*, 642–645. [CrossRef]
87. Puskadija, Z.; Kovacic, M.; Raguz, N.; Lukic, B.; Presern, J.; Tofilski, A. Morphological diversity of Carniolan honey bee (*Apis mellifera carnica*) in Croatia and Slovenia. *J. Apic. Res.* **2020**, *60*, 326–336. [CrossRef]
88. Harpur, B.A.; Minaei, S.; Kent, C.F.; Zayed, A. Management increases genetic diversity of honey bees via admixture. *Mol. Ecol.* **2012**, *21*, 4414–4421. [CrossRef]
89. Chavez-Galarza, J.; Henriques, D.; Johnston, J.S.; Carneiro, M.; Rufino, J.; Patton, J.C.; Alice Pinto, M. Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: Maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Mol. Ecol.* **2015**, *24*, 2973–2992. [CrossRef]
90. Miguel, I.; Baylac, M.; Iriondo, M.; Manzano, C.; Garnery, L.; Estonba, A. Both geometric morphometric and microsatellite data consistently support the differentiation of the *Apis mellifera* M evolutionary branch. *Apidologie* **2011**, *42*, 150–161. [CrossRef]
91. Browne, K.A.; Hassett, J.; Geary, M.; Moore, E.; Henriques, D.; Soland-Reckeweg, G.; Ferrari, R.; Mac Loughlin, E.; O'Brien, E.; O'Driscoll, S.; et al. Investigation of free-living honey bee colonies in Ireland. *J. Apic. Res.* **2020**, *60*, 229–240. [CrossRef]
92. Nielsen, E.S.; Henriques, R.; Toonen, R.J.; Knapp, I.S.S.; Guo, B.; von der Heyden, S. Complex signatures of genomic variation of two non-model marine species in a homogeneous environment. *BMC Genom.* **2018**, *19*, 347. [CrossRef]
93. Kurland, S.; Wheat, C.W.; Mancera, M.d.l.P.C.; Kutschera, V.E.; Hill, J.; Andersson, A.; Rubin, C.-J.; Andersson, L.; Ryman, N.; Laikre, L. Exploring a Pool-seq-only approach for gaining population genomic insights in nonmodel species. *Ecol. Evol.* **2019**, *9*, 11448–11463. [CrossRef]
94. Canovas, F.; De la Rua, P.; Serrano, J.; Galian, J. Analysis of a contact area between two distinct evolutionary honeybee units: An ecological perspective. *J. Insect Conserv.* **2014**, *18*, 927–937. [CrossRef]
95. Nazzi, F. Morphometric analysis of honey bees from an area of racial hybridization in northeastern Italy. *Apidologie* **1992**, *23*, 89–96. [CrossRef]
96. Cornuet, J.-M.; Excoffier, L.; Franck, P.; Estoup, A. *Bayesian Inference under Complex Evolutionary Scenarios Using Microsatellite Markers: Multiple Divergence and Genetic Admixture Events in the Honey Bee, Apis mellifera. Genetic Diversity*; Mahoney, C.L., Springer, D.A., Eds.; Nova Science Publishers: New York, NY, USA, 2009.
97. Janczyk, A.; Meixner, M.D.; Tofilski, A. Morphometric identification of the endemic Maltese honey bee (*Apis mellifera ruttneri*). *J. Apic. Res.* **2021**, *60*, 157–164. [CrossRef]
98. Chen, C.; Liu, Z.; Pan, Q.; Chen, X.; Wang, H.; Guo, H.; Liu, S.; Lu, H.; Tian, S.; Li, R.; et al. Genomic Analyses Reveal Demographic History and Temperate Adaptation of the Newly Discovered Honey Bee Subspecies *Apis mellifera sinisxinyuan* n. ssp. *Mol. Biol. Evol.* **2016**, *33*, 1337–1348. [CrossRef]
99. Chen, C.; Wang, H.; Liu, Z.; Chen, X.; Tang, J.; Meng, F.; Shi, W. Population Genomics Provide Insights into the Evolution and Adaptation of the Eastern Honey Bee (*Apis cerana*). *Mol. Biol. Evol.* **2018**, *35*, 2260–2271. [CrossRef]
100. Steinrucken, M.; Kamm, J.; Spence, J.P.; Song, Y.S. Inference of complex population histories using whole-genome sequences from multiple populations. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 17115–17120. [CrossRef]

101. Steiner, C.C.; Putnam, A.S.; Hoeck, P.E.A.; Ryder, O.A. Conservation Genomics of Threatened Animal Species. *Annu. Rev. Anim. Biosci.* **2013**, *1*, 261–281. [CrossRef]
102. Fleming, D.S.; Koltes, J.E.; Fritz-Waters, E.R.; Rothschild, M.F.; Schmidt, C.J.; Ashwell, C.M.; Persia, M.E.; Reecy, J.M.; Lamont, S.J. Single nucleotide variant discovery of highly inbred Leghorn and Fayoumi chicken breeds using pooled whole genome resequencing data reveals insights into phenotype differences. *BMC Genom.* **2016**, *17*, 812. [CrossRef]