scientific reports

OPEN

Admixture and breed traceability in European indigenous pig breeds and wild boar using genome-wide SNP data

Christos Dadousis[1]✉, Maria Muñoz[2], Cristina Óvilo[2], Maria Chiara Fabbri[1], José Pedro Araújo[3], Samuele Bovo[4], Marjeta Čandek Potokar[5], Rui Charneca[6], Alessandro Crovetti[1], Maurizio Gallo[7], Juan María García-Casco[2], Danijel Karolyi[8], Goran Kušec[9], José Manuel Martins[6], Marie-José Mercat[10], Carolina Pugliese[1], Raquel Quintanilla[11], Čedomir Radović[12], Violeta Razmaite[13], Anisa Ribani[4], Juliet Riquet[14], Radomir Savić[15], Giuseppina Schiavo[4], Martin Škrlep[5], Silvia Tinarelli[7], Graziano Usai[16], Christoph Zimmer[17], Luca Fontanesi[4] & Riccardo Bozzi[1]

Preserving diversity of indigenous pig (*Sus scrofa*) breeds is a key factor to (i) sustain the pork chain (both at local and global scales) including the production of high-quality branded products, (ii) enrich the animal biobanking and (iii) progress conservation policies. Single nucleotide polymorphism (SNP) chips offer the opportunity for whole-genome comparisons among individuals and breeds. Animals from twenty European local pigs breeds, reared in nine countries (Croatia: Black Slavonian, Turopolje; France: Basque, Gascon; Germany: Schwabisch-Hällisches Schwein; Italy: Apulo Calabrese, Casertana, Cinta Senese, Mora Romagnola, Nero Siciliano, Sarda; Lithuania: Indigenous Wattle, White Old Type; Portugal: Alentejana, Bísara; Serbia: Moravka, Swallow-Bellied Mangalitsa; Slovenia: Krškopolje pig; Spain: Iberian, Majorcan Black), and three commercial breeds (Duroc, Landrace and Large White) were sampled and genotyped with the GeneSeek Genomic Profiler (GGP) 70 K HD porcine genotyping chip. A dataset of 51 Wild Boars from nine countries was also added, summing up to 1186 pigs (~ 49 pigs/breed). The aim was to: (i) investigate individual admixture ancestries and (ii) assess breed traceability via discriminant analysis on principal components (DAPC). Albeit the mosaic of shared ancestries found for Nero Siciliano, Sarda and Moravka, admixture analysis indicated independent evolvement

[1]Dipartimento Di Scienze e Tecnologie Agrarie, Alimentari, Ambientali e Forestali, Università Di Firenze, 50144 Firenze, Italy. [2]Departamento Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Crta. de la Coruña, km. 7, 5, 28040, Madrid, Spain. [3]Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Viana Do Castelo, Escola Superior Agrária, Refóios do Lima, 4990-706 Ponte de Lima, Portugal. [4]Department of Agricultural and Food Sciences, Division of Animal Sciences, University of Bologna, Viale Fanin 46, 40127 Bologna, Italy. [5]Kmetijski inštitut Slovenije, Hacquetova 17, 1000 Ljubljana, Slovenia. [6]MED-Mediterranean Institute for Agriculture, Environment and Development & Escola de Ciências E Tecnologia, Universidade de Évora, Pólo da Mitra, Ap. 94, 7006-554 Évora, Portugal. [7]Associazione Nazionale Allevatori Suini (ANAS), Via Nizza 53, 00198 Rome, Italy. [8]Department of Animal Science, Faculty of Agriculture, University of Zagreb, Svetošimunska c. 25, 10000 Zagreb, Croatia. [9]Faculty of Agrobiotechnical Sciences Osijek, Josip Juraj Strossmayer University of Osijek, Vladimira Preloga 1, 31000 Osijek, Croatia. [10]IFIP Institut du Porc, La Motte au Vicomte, BP 35104, 35651 Le Rheu Cedex, France. [11]Programa de Genética y Mejora Animal, Institute for Research and Technology in Food and Agriculture (IRTA), Torre Marimon, 08140 Caldes de Montbui, Barcelona, Spain. [12]Department of Pig Breeding and Genetics, Institute for Animal Husbandry, 11080 Belgrade-Zemun, Serbia. [13]Animal Science Institute, Lithuanian University of Health Sciences, Baisogala, Lithuania. [14]Génétique Physiologie Et Systèmes d'Elevage (GenPhySE), Université de Toulouse, INRA, Chemin de Borde-Rouge 24, Auzeville Tolosane, 31326 Castanet Tolosan, France. [15]Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11080 Belgrade-Zemun, Serbia. [16]AGRIS SARDEGNA, Loc. Bonassai, 07100 Sassari, Italy. [17]Bäuerliche Erzeugergemeinschaft Schwäbisch Hall, Schwäbisch Hall, Germany. ✉email: christos.dadousis@unipr.it

for the rest of the breeds. High prediction accuracy of DAPC mark SNP data as a reliable solution for the traceability of breed-specific pig products.

The process of domestication of pigs and the spread of the species around the world has been the subject of some studies in the recent past[1,2], demonstrating that pig domestication involved multiple pig populations including wild boars[3,4]. The domestication aspects have often been investigated by the study of mitochondrial DNA, while genetic diversity was initially studied using simple sequence repeat (SSR) and amplified fragment length polymorphism (AFLP) in intensively selected breeds[5,6] but also in indigenous populations of limited diffusion[7–10]. The development of single nucleotide polymorphisms (SNPs) panels with SNPs distributed across the entire genome provided new opportunities to investigate and decipher the complex relationship between indigenous pig breeds[11,12]. This is a topic that has to be taken on in order to enhance the safeguard of local pig populations. Considering the region of Europe and Caucasus, the Food and Agriculture Organization (FAO) identified 48 already extinct pig breeds, representing ~ 20% of the global pig breeds. Among the existing breeds in the region, 14 breeds are classified at critical risk of extinction, 5 are in a critical-maintained status, 24 are endangered, 11 are defined as endangered-maintained and 6 in a vulnerable situation (http://www.fao.org/dad-is/risk-status-of-animal-genetic-resources/en/). This means that more than 25% of the local European pig population is in a worrisome demographic status. The improvement of breeding and conservation programs for these indigenous breeds is becoming extremely important for multiple reasons. Firstly, it is well known that indigenous breeds are well-adapted to their local environment and are a unique genetic pool that might be essential, not only as pig biobank[13], but also for the sustainability of the global pork chain. In addition, local pig farming is strongly related to niche products of high quality, which contribute to the local economy development and sustainability[14]. No less important is the increasing demand for organic and high welfare animal-based food products[15], which has led consumers to prefer local breed products that are considered more nutritious, tasty, healthy and safe[16] and because animals are usually reared freely and outdoors[17].

It is important to note that the European pork production amounts on 21—22 thousand tonnes of meat per year (https://ec.europa.eu/eurostat/databrowser/view/tag00042/), heavily based on the use of cosmopolitan pig breeds. Moreover, Germany, Spain, France, Poland, and the Netherlands are the largest consumers in Europe. In this context, a powerful system to ensure pig breed traceability is required, that will enable products from pure local breeds to be clearly differentiated from their cosmopolitan counterparts and controlling fraud. Currently, the administrative traceability is not infallible, and the possibility of errors and frauds exists. The use of genetic markers could overcome these limits[18]. Microsatellites and SNP have been mainly used for traceability purposes, with the latter nowadays prevailing over the former, presenting many advantages such as easier laboratory handling, low mutation rate, and better suitability for standardization[19]. Several SNP based studies, often using a runs of homozygosity approach, aimed to detect candidate genes which allowed the identification of a specific breed[20] and/or focused on genomic regions which discriminated populations from each other[21]. A pairwise fixation index (F_{ST}) distances method was used to differentiate indigenous from commercial pig populations[11,22,23], and to determine breeds belonging to different production systems[24]. Moreover, SNP detection from genome wide sequencing was used to develop a SNP chip for discriminating between purebred or crossbred Iberian origin of live pigs, meat and dry-cured pig products[25]. Other methods applied to distinguish breeds from each other are the investigation of the proportion of ancestry shared among the breeds[26,27] and the clustering of genetically related individuals by discriminant analysis[28]. This latter method has been applied to trace sheep, using sets of SNPs able to separate breeds belonging to different geographic areas[29] and for assigning animals to their true population[30]. A similar approach has been applied in cattle, where Dimauro et al.[31] argued that the canonical discriminant analysis was able to efficiently distinguish the three breeds studied (Holstein, Brown Swiss, and Simmental). Moreover, various other methods exist in human[32] and animal studies[33–35] to identify a small set of ancestry informative SNPs, derived from genotyping or sequence data, that are helpful for population identification and breed traceability.

To the best of our knowledge no similar studies have been performed in pig breeds. In this work, we describe a comprehensive approach of using principal component analysis (PCA), admixture and discriminant analysis of principal components (DAPC) to evaluate pig breeds (indigenous and commercial) and wild boars traceability via the whole set of SNPs revealed by the GGP Porcine HD Array. The last two methods allow to predict the breed of origin (DAPC), and the proportions of ancestry per pig (admixture analysis).

Results

Population stratification and ancestry.

Analyses were based on 1,186 pigs and 40,364 SNP (Table 1). A PCA analysis was applied on the matrix of 1,186 pig genotypes. The scatterplots of the first two and all of the first five PCs in pairwise combinations are shown in Fig. 1a,b. Mora Romagnola and Duroc were clearly distinguished from the rest of the breeds (bottom-right quarter, Fig. 1a). Moreover, PC1 placed closely the Turopolje, Alentejana, Iberian, Swallow-Bellied Mangalitsa, Majorcan Black and Basque (left part, Fig. 1a). Lithuanian White Old Type and Large White were also separated in the opposite direction of PC1 (top-right quarter, Fig. 1a) and were closely positioned. In close proximity to those two was the Landrace breed. Considering PC1 and PC2, pigs belonging to the rest of the breeds were largely overlapped showing considerable within breed variation. Despite this, Gascon was almost clearly differentiated and this differentiation was more profound in PC5 (Fig. 1b). Considering further axes, Basque and Apulo Calabrese were also distinguished (PC3 and PC5, respectively), while Turopolje was further separated. It should be noted however, that the eigenvalues where low, with the first 2 eigenvalues accounting cumulatively ~ 9.3% of the original variability, while the first 697 eigenvalues captured ~ 90% (Fig. 1c).

Breed name	Country of origin	N. pre-QC	N. post-QC
Indigenous			
Alentejana	Portugal	48	48
Apulo Calabrese	Italy	53	53
Basque	France	39	39
Bísara	Portugal	49	49
Black Slavonian (Crna Slavonska)	Croatia	49	49
Casertana	Italy	55	53
Cinta Senese	Italy	54	54
Gascon	France	48	48
Iberian	Spain	48	48
Krškopolje pig	Slovenia	52	52
Lithuanian Indigenous Wattle	Lithuania	48	48
Lithuanian White Old Type	Lithuania	48	48
Majorcan Black	Spain	48	48
Mora Romagnola	Italy	48	48
Moravka	Serbia	50	50
Nero Siciliano	Italy	50	48
Sarda	Italy	49	48
Schwäbisch-Hällisches Schwein (Swabian Hall pig)	Germany	51	49
Swallow-Bellied Mangalitsa	Serbia	50	50
Turopolje	Croatia	50	50
Commercial			
Duroc	Italy, Spain	53	53
Landrace	Italy, Spain	52	52
Large White	Italy, Spain	52	50
Wild			
Wild Boar	Finland, Greece, Hungary, Italy, Spain, Poland, Russia, The Netherlands, Tunisia	160	51

Table 1. Breed name, type, country of origin and number of pigs analysed before (pre-) and after (post-) quality control (QC) per breed.

Complementary to PCA, an admixture analysis was carried out, to estimate the proportion of ancestries per pig (Fig. S1). After cross-validation (CV), the model with 24 distinct groups was kept for further analysis (Fig. 2a). Results could be summarized in four main points (Fig. 2b, Fig. S1, S2 and Table S2): (i) in general, Alentejana, Basque, Gascon, Iberian and Mora Romagnola as indigenous breeds, and Duroc as commercial breed, showed the lowest levels of introgression, (ii) Casertana, Swallow-Bellied Mangalitsa and Turopolje consisted mainly of two group ancestries, (iii) the Italian breeds Nero Siciliano and Sarda showed a mosaic of different ancestries and (iv) Wild Boar ancestry contribution was mainly found in the Alentejana, Black Slavonian, Iberian, Nero Siciliano, Sarda and Swallow-Bellied Mangalitsa breeds.

Discriminant analysis.

Scenario 1 (semi-supervised learning). The overall successful assignment of pigs in breed of origin of the DAPC, averaged over the ten replicates, was 0.98 [0.967, 0.996] (Table 2). The number of PCs kept for DAPC ranged from 100 to 250 (52.4 and 69.4% of the original variance captured from the PCs, respectively). However, the number of PCs selected only marginally influenced the assignment success. The assignment success varied among breeds, with Black Slavonian, Cinta Senese, Krškopolje pig, Lithuanian White Old Type, Moravka, Nero Siciliano and Turopolje having < 100%, and the remaining breeds showing 100% accuracy (Fig. 3). The lowest value was observed for Black Slavonian (86%) with some pigs assigned to either as Cinta Senese or Turopolje (6 and 8%, respectively).

In general, a positive effect of the sample size on the correct assignment of the DAPC model was found (Fig. 4). Although the mean model accuracy was slightly influenced by sample size, implying the robustness of the DAPC analysis, increasing sample size produced higher mean accuracies and reduced variance.

Scenario 2 (un-supervised learning). In the second scenario, VAL sets consisted of separate breeds and the evaluated breed was entirely excluded from the TRN set, hence the pigs were assigned to the rest of the 23 breeds. Results (Fig. 5) could be summarized in the following points: (i) some breeds were 100% assigned to only one breed (Alentejana, Apulo Calabrese, Basque, Bísara, Casertana, Gascon, Iberian, Krškopolje pig, Nero Siciliano and Turopolje), (ii) Cinta Senese, Duroc, Landrace, Large White, Majorcan Black, Mora Romagnola, Moravka, Sarda, Schwabisch-Hällisches Schwein, Swallow-Bellied Mangalitsa and Wild Boar were assigned to two breeds, (iii) Black Slavonian, Lithuanian Indigenous Wattle and Lithuanian White Old Type were assigned to three

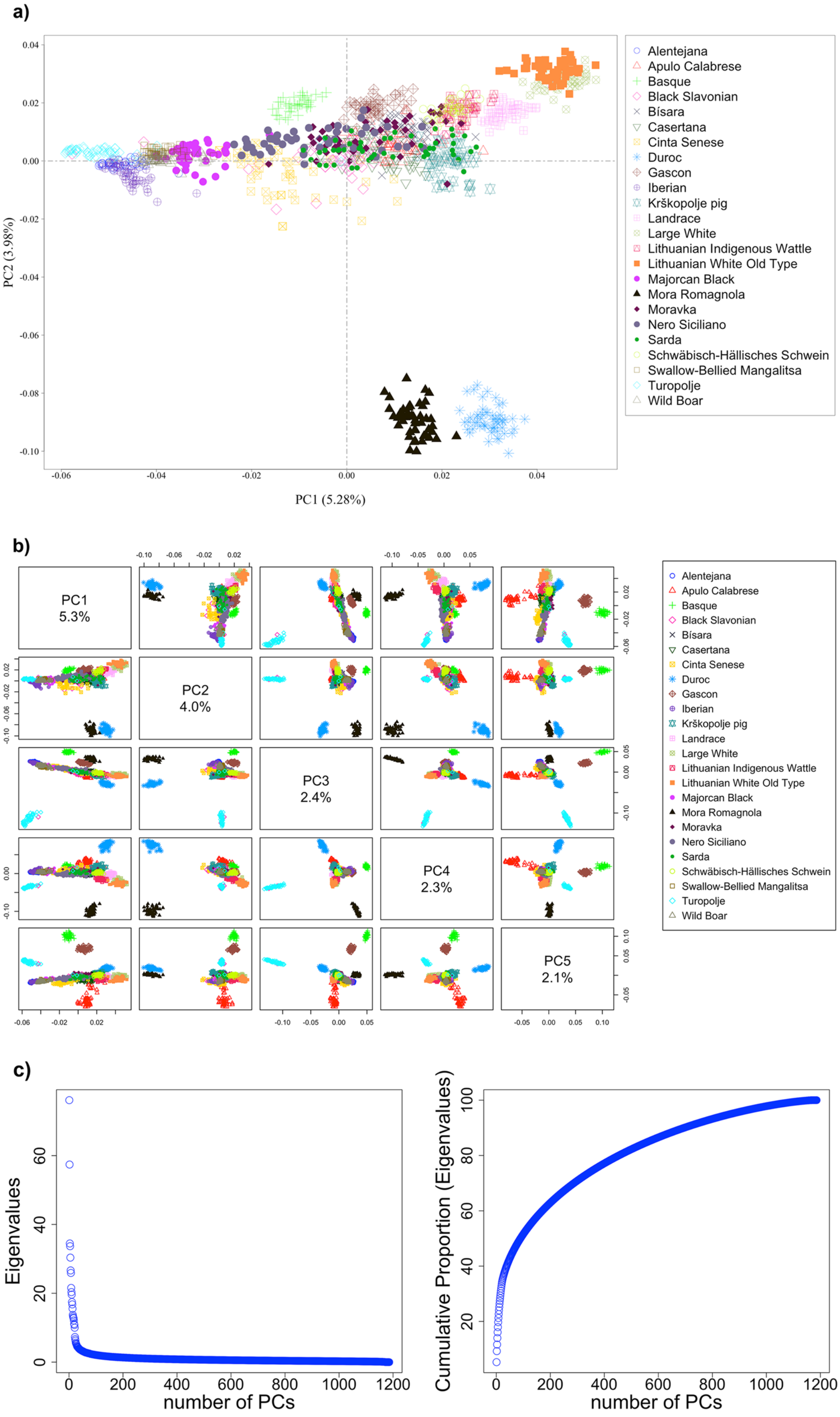


Figure 1. Results of the principal component analysis using the genotypes of 1,186 pigs: (**a**) Scatterplot of the first two principal components (PCs), (**b**) pairwise scatterplots of the first five PCs and (**c**) variance and cumulative variance explained by the PCs.

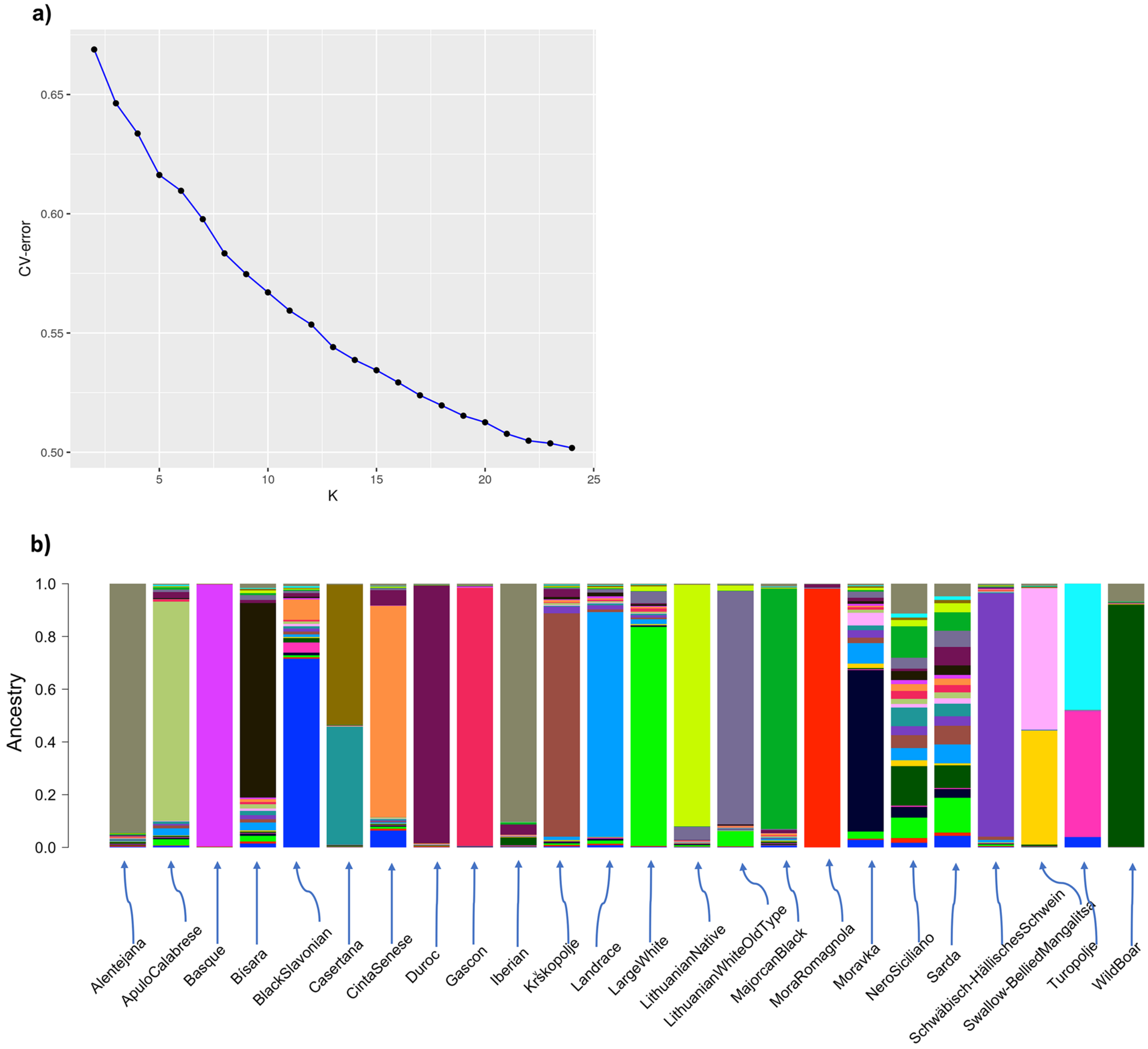
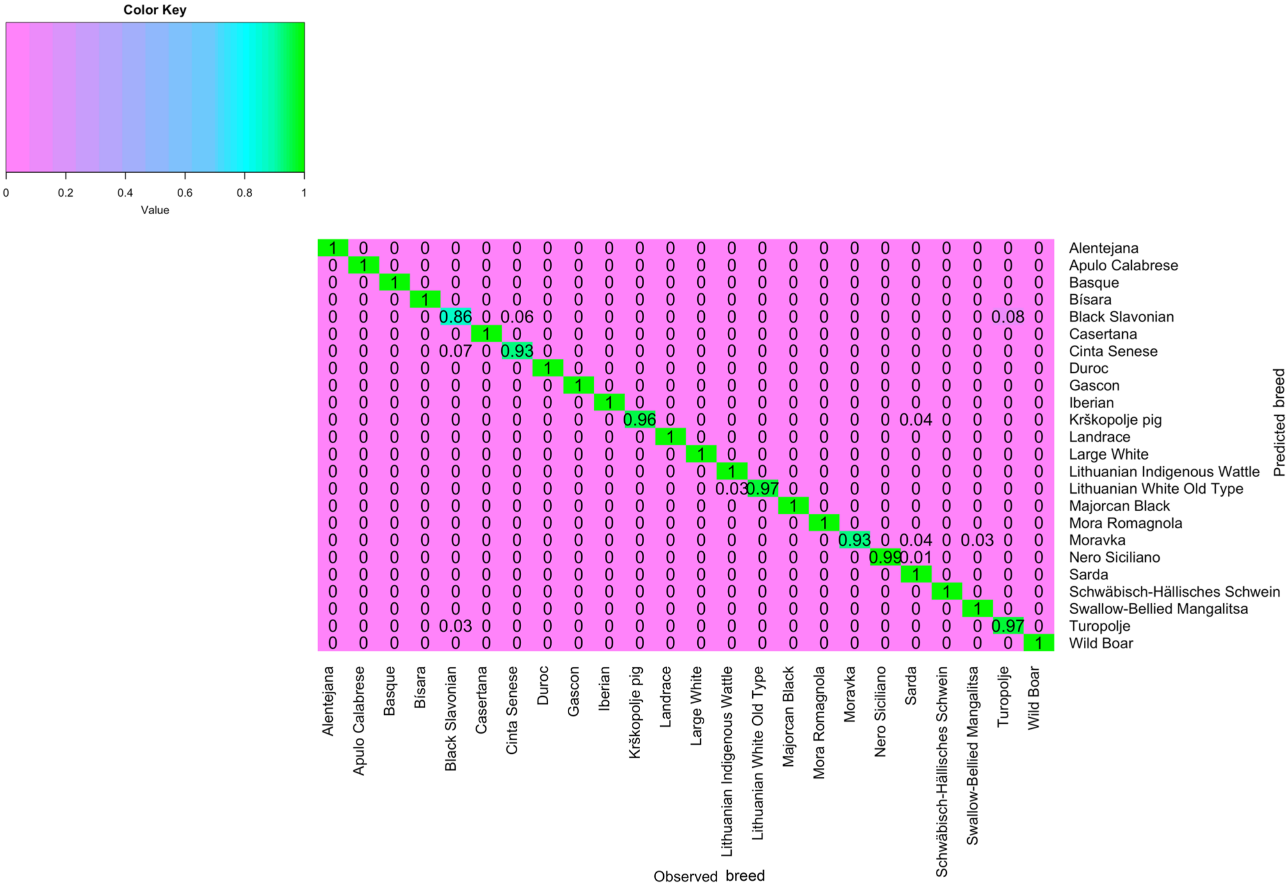


Figure 2. Results of admixture analysis: (**a**) fivefold cross-validation minimum error from K = 2–24; (**b**) summary per breed of admixture ancestries at K = 24.

Replicate	Assignment success, %	nPCs	VarPCs, %
rep1	0.983	100	0.524
rep2	0.988	200	0.646
rep3	0.975	100	0.525
rep4	0.979	200	0.649
rep5	0.992	100	0.526
rep6	0.988	100	0.526
rep7	0.996	250	0.695
rep8	0.967	200	0.649
rep9	0.992	200	0.646
rep10	0.975	250	0.694

Table 2. Summary results of the DAPC model on the complete dataset. nPCs = number of principal components selected for the DAPC model; VarPCs = percentage of original variance explained by the selected principal components. The total number of pigs was 1,186, the number of pigs in the TRN set was 944, and the number of pigs in the validation set was 242.



Alentejana	Apulo Calabrese	Basque	Bísara	Black Slavonian	Casertana	Cinta Senese	Duroc	Gascon	Iberian	Krškopolje pig	Landrace	Large White	Lithuanian Indigenous Wattle	Lithuanian White Old Type	Majorcan Black	Mora Romagnola	Moravka	Nero Siciliano	Sarda	Schwäbisch-Hällisches Schwein	Swallow-Bellied Mangalitsa	Turopolje	Wild Boar	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Alentejana
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Apulo Calabrese
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Basque
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Bísara
0	0	0	0	0.86	0	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.08	0	Black Slavonian
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Casertana
0	0	0	0	0.07	0	0.93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Cinta Senese
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Duroc
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Gascon
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Iberian
0	0	0	0	0	0	0	0	0	0	0.96	0	0	0	0	0	0	0	0	0.04	0	0	0	0	Krškopolje pig
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Landrace
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Large White
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Lithuanian Indigenous Wattle
0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0.97	0	0	0	0	0	0	0	0	0	Lithuanian White Old Type
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Majorcan Black
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Mora Romagnola
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	0	0.04	0	0.03	0	0	Moravka
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0.01	0	0	0	0	Nero Siciliano
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Sarda
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Schwäbisch-Hällisches Schwein
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	Swallow-Bellied Mangalitsa
0	0	0	0	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.97	0	Turopolje
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Wild Boar

Figure 3. Heatmap of the DAPC assignment in the semi-supervised scenario with percentage of correct assignment per breed (in a scale of 0–1). Heatmap was constructed using the R[36] package gplots[37] and the function *heatmap.2*.

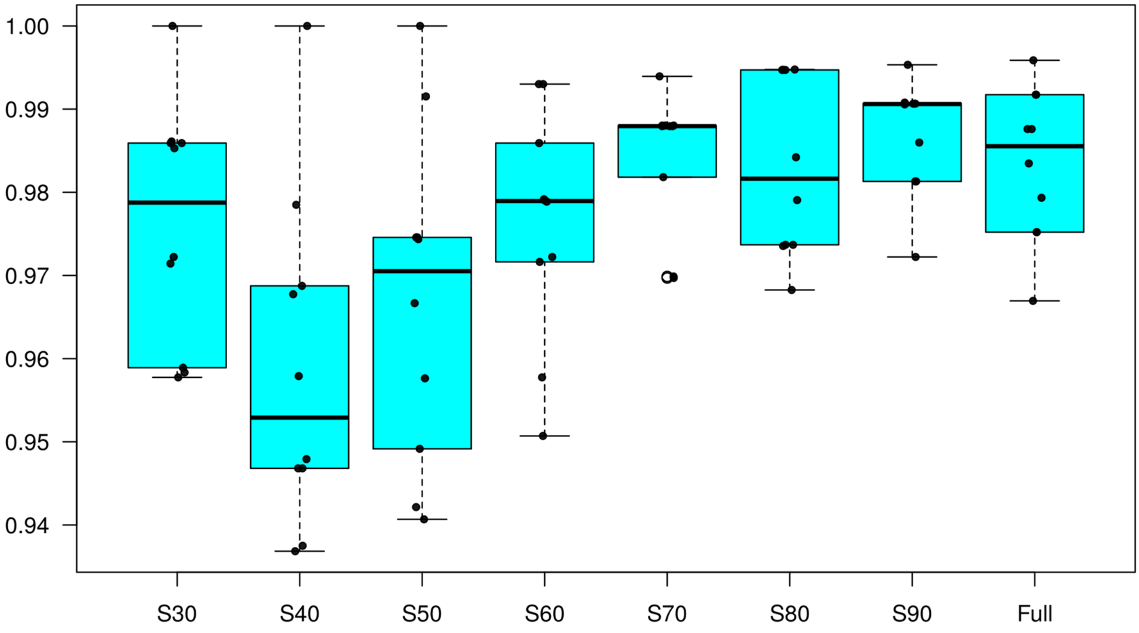


Figure 4. Boxplot of the overall successful assignment over different sampling (S) proportions of the data (30 to 100%) using DAPC. Median (black horizontal lines within the boxplots) over ten replicates (black dots).

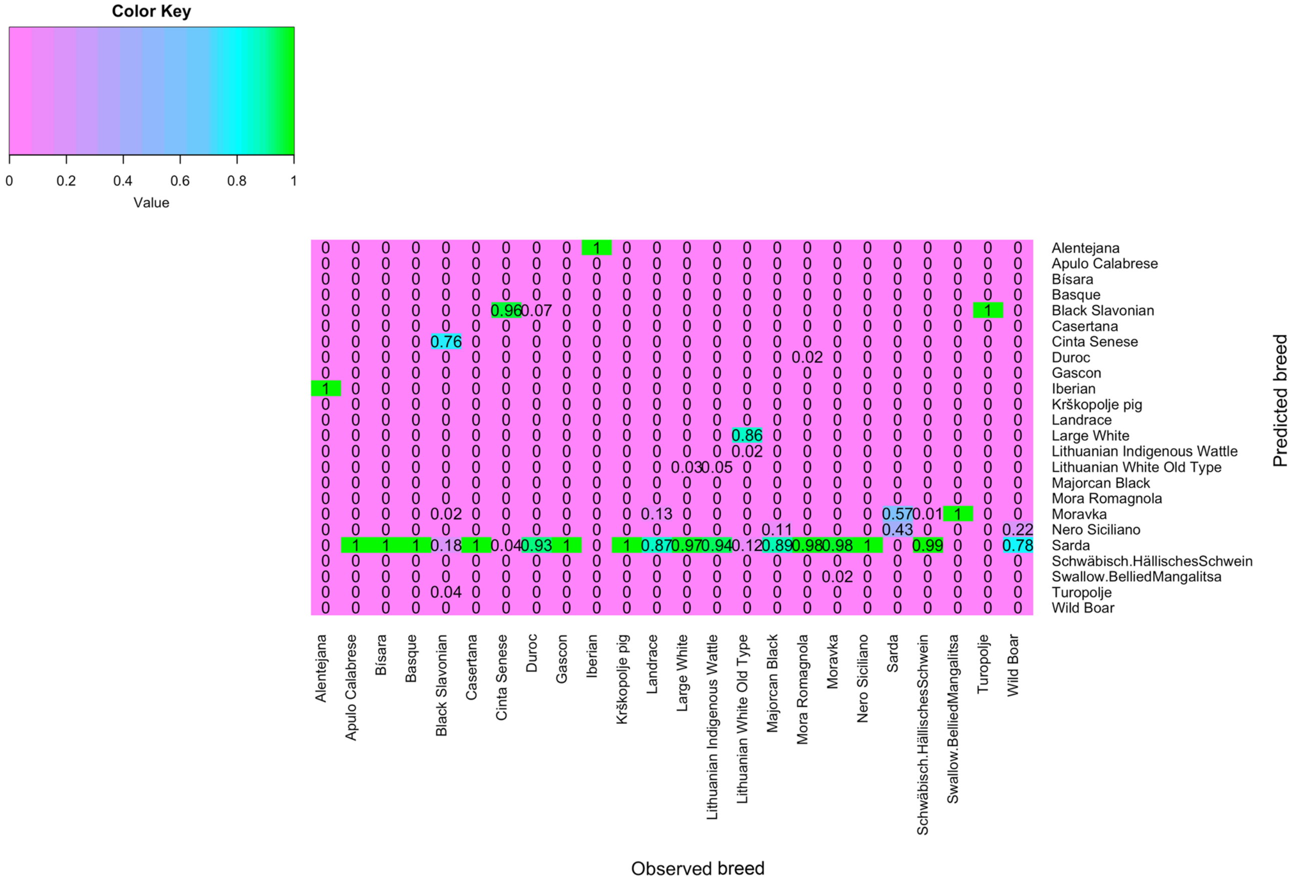


Figure 5. Heatmap of the DAPC assignment in the un-supervised scenario with percentage of external assignment per breed (in a scale of 0 to 1). Heatmap was constructed using the R[36] package gplots[37] and the function *heatmap.2*.

breeds, (iv) when the evaluated set of pigs was assigned to more than one breed, Sarda always appeared as one of the assigned breeds, so presenting mostly the highest assignment rate (except in the case of Black Slavonian, Lithuanian White Old Type and Swallow-Bellied Mangalitsa), (v) Alentejana was 100% assigned to Iberian and the other way around. That was the only case found of such a relationship between two breeds. For instance, Apulo-Calabrese, Basque, Bísara, Casertana, Krškopolje, and Nero Siciliano matched 100% to Sarda, but Sarda pigs were aligned only to Moravka and Nero Siciliano, (vi) Wild Boar was assigned mainly to Sarda and a small number to Nero Siciliano. The second most frequent breed to be assigned was Moravka with Black Slavonian, Landrace, Sarda, Schwabisch-Hällisches Schwein and Swallow-Bellied Mangalitsa being assigned to this breed.

These results were, in general, consistent and the sample size in the TRN set only marginally influenced the assignment of the breeds (Fig. 6). It is interesting that even with 30% of the dataset (~ 340 pigs), assignments were fairly consistent with results obtained utilizing the full dataset (~ 1,138 pigs). Sarda was in all subsets the breed mostly assigned. The percentage of classification of a specific breed to Sarda was either increased or decreased with an increasing sample size. For example, the proportion of the Black Slavonian classified as Sarda was medium (~ 40–50%) at a small sample size (30–60% of the data) and reduced to 10–20% with accumulated data, with the majority of the Black Slavonian pigs being assigned to Cinta Senese (~ 70–80%). Similarly, Lithuanian White Old Type had a ~ 40% assignment to Sarda and ~ 50% to Large White with ~ 340 pigs in the TRN set, and this ratio changed to 10–90% (Sarda and Large White, respectively) when all pigs from the remaining 23 breeds were considered in the TRN. In contrast, the percentage of Wild Boars assigned to Sarda was increased from 50 to 80% when increasing the sample size. The relationship between Alentejana – Iberian was not influenced in any scenario, resulting in 100% assignment of pigs of one breed to the other in all the cases.

Discussion

Nowadays, modern pig farming worldwide is mostly highly intensive, utilizing few commercial breeds undergoing intense selection. Nevertheless, successful applications of indigenous pig farming exist, perhaps with the most prominent example being the Iberian pig in Spain. Disease outbreaks, such as the African swine-fever, threaten global pig production. Indigenous pig breeds consist of a unique genetic pool that might be proved of a great importance in the future, not only for the sustainability of the global pork chain but also for human research as in the case of the pig biobank[13,39]. However, indigenous pig farming is greatly based on outdoor rearing, making it vulnerable not only to disease outbreaks but also to natural disasters.

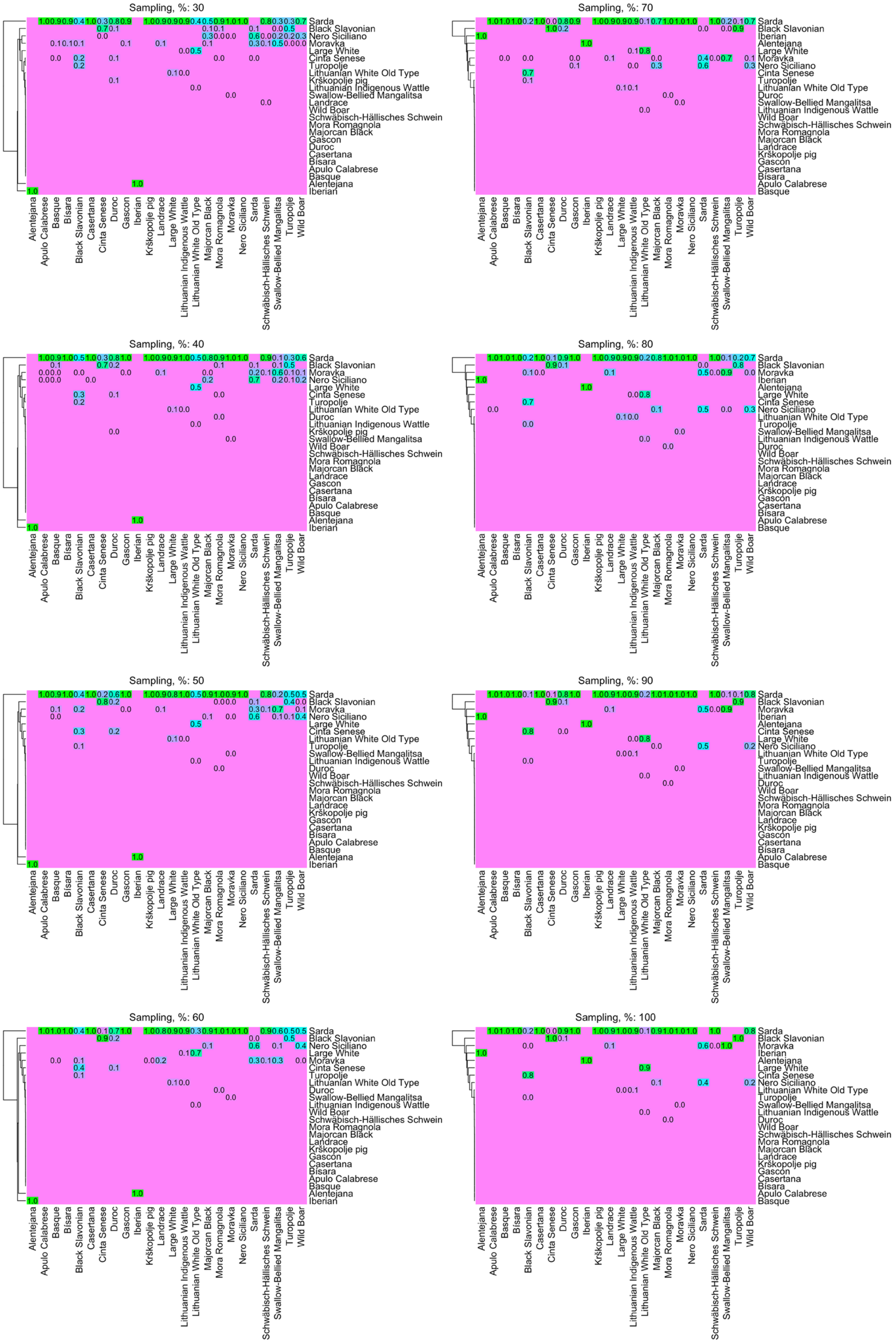


Figure 6. Heatmaps of the DAPC assignment in the un-supervised scenario, in increasing sample size, of percentage of external assignment per breed (in a scale of 0 to 1); x-axes show the observed and y-axes the predicted breed. Heatmaps were constructed using the R[36] package ComplexHeatmap[38].

Studying genetic diversity is essential for the characterization of indigenous animal populations and can be used for conservation policies and promotion of local breeds. To support local pig farming, the TREASURE project joined researchers from nine countries and twenty-four research institutes to collect data from twenty European indigenous breeds. Previous genomic analyses of the aforementioned breeds were focused on linkage disequilibrium analysis and selection signatures detection using genome-wide SNP markers[12], as well as genome sequencing data[40]. Studies on genetic diversity have also been performed, whether based on a candidate genes approach[41] or a runs of homozygosity method[42]. The present work complements these studies by further investigating the proportion of ancestry shared among these breeds, together with three of the most representative commercial breeds as well as a joined dataset of Wild Boar, originating from nine countries. To address the question of potential breed traceability via genomic data, we further investigated the ability to predict the breed of origin by SNP markers. Linear discriminant analysis is a widely used methodology, but it lacks efficiency with high dimensional data such as genomic data. To overcome this problem, the methodology of linear discriminant analysis on a reduced dimensionality space, consisting of few principal components derived from SNP, was used.

PCA and admixture results were generally in agreement with high within-breed variability observed for the Sarda, Nero Siciliano and the Moravka, while Duroc and Mora Romagnola were the breeds that diverged most from the rest. Furthermore, unique ancestries were detected with both approaches for the Alentejana, Iberian, Basque, Duroc, Gascon and Mora Romagnola. Regarding Mora Romagnola, PCA and DAPC analyses showed contradictory results compared to previous study using candidate genes approach[41]. To explain this, it can be hypothesized that in a population such as Mora Romagnola, characterized by a low number of individuals and high level of inbreeding, there may be different response when investigating loci under selective pressure compared to neutral loci.

Nevertheless, slight differences among the PCA and admixture were also observed. For instance, the PCA scatterplot of the first two axes (Fig. 1a) clustered Turopolje close to Alentejana and Iberian; however, admixture analysis showed that ancestries were shared with Black Slavonian, Cinta Senese and Sarda (Fig. 2b, Table S2).

Regarding the closeness of some local with the cosmopolitan breeds as revealed by PCA (i.e., Duroc with Mora Romagnola; Large White with Lithuanian White Old Type), the reason for this could be the sharing some parts of the genome linked to phenotypic characteristics and origin of Lithuanian White pigs; however, the amount of variability explained by the first PCs is largely limited with respect to the overall genetic variability possessed by populations in the entire dataset. Moreover, although in PCA based on the scatterplot of the first two PCs (Fig. 1a) Duroc and Mora Romagnola were closely placed, the two breeds had common ancestries close to zero (Fig. 2b, Table S2).

Admixture analysis revealed common ancestries shared between some indigenous and the commercial breeds. More precisely, Duroc shared ancestries mainly with Cinta Senese, Iberian and Sarda; Landrace with Bísara, Moravka, Nero Siciliano and Sarda; and Large White with Lithuanian White Old Type, Nero Siciliano, Sarda, and Lithuanian Indigenous Wattle. Regarding Wild Boars, our dataset consisted of a set of 51 samples from seven European countries, Tunisia, and Russia, to capture as much variability and to avoid country-specific bias. Indeed, a recent study investigating the history of the domesticated European pigs indicated an interbreeding between the local pig breeds and Wild Boars[43]. Previous analysis on the same local breeds reported a close relationship, based on neighbour-joining tree constructed with Nei's distances, between the Wild Boar and Alentejana and Iberian breeds[12]. In our analysis, introgression of Wild Boar was also found, besides the two aforementioned breeds, for the Italian breeds Nero Siciliano and Sarda. Common features between the PCA, admixture and the un-supervised DAPC were also observed, as explained below.

The un-supervised DAPC method could represent a real lab scenario for testing the "blind" or external to TRN set samples. In the un-supervised DAPC, many of the breeds, except Alentejana, Iberian, Black Slavonian, Cinta Senese, Lithuanian Wild Old Type and Turopolje, were mainly assigned as Sarda. This is not surprising, given the high admixture level of the Sarda breed. Black Slavonian was assigned to Cinta Senese in 76% of the cases, while Cinta Senese was predicted as Black Slavonian with 96% rate. Similarly, in the admixture analysis ~ 7.5% of the Black Slavonian was shared with Cinta Senese, while Turopolje was classified as Black Slavonian (100%). Interestingly, in the admixture analysis, Turopolje was assigned to two major ancestral groups sharing common ancestries mainly with Black Slavonian (Table S2). Regarding Lithuanian White Old Type, ancestries were mainly shared with Sarda (~ 6%), Lithuanian Indigenous Wattle (~ 5%) and Large White (~ 4.5%), so it would be expected to be predicted as Sarda. Nevertheless, the breed was assigned to a large extent to Large White (86%) followed by Sarda (~ 12%).

A second objective was to study traceability of pigs based on genome-wide SNP data. To resemble a practical application, the efficiency of the DAPC method was evaluated using an external validation. Furthermore, to assess the effect of sample size, the analyses were repeated several times with subsets of the dataset ranging from 30 to 90%. Although the correct assignment of the breeds was > 90% in all subsets, the variation of the correct assignment decreased with increased sample size, indicating a more robust model (Fig. 4). This level of correct reassignment of pigs is higher than the one reported by Muñoz et al.[41], where there were many breeds with percentages of correct reassignment < 80%. Moreover, the actual differences might be even higher, since in that analysis an external validation was not considered and the whole data were analysed simultaneously. The correct reassignment was further improved for the Moravka, Nero Siciliano and Sarda breeds that had the lowest values in the DAPC analysis by Muñoz et al.[41]. However, in that study only a limited number of 39 SNPs in candidate genes was used.

Using the complete dataset, the majority of the breeds were correctly assigned to its breed of origin, with the exceptions of Black Slavonian, Cinta Senese, Krškopolje, Lithuanian White Old Type, Moravka and Turopolje, with the lowest value (86%) being observed for Black Slavonian (Fig. 3). In the case of Black Slavonian, there were some cases where animals were classified either as Cinta Senese or Turopolje. This was consistent with the shared ancestries found among the breeds, even at a low degree (Table S2). The relation among these breeds was further highlighted with the un-supervised DAPC, in which Black Slavonian was assigned mainly as Cinta Senese, followed by Sarda and Turopolje.

It should be noted that discrepancies between our results and previous genomic analyses on the same set of breeds were to some extent expected. There are two main reasons for this: (i) we considered three cosmopolitan breeds and a more diverse Wild Boar panel compared to Muñoz et al.[12] and (ii) a whole-genome analysis was conducted compared to the candidate gene approach and the 39 SNP of Muñoz et al.[41].

Conclusion

We report a whole genome SNP analysis on admixed ancestries and classification of 20 European indigenous pig breeds, together with three commercial breeds and Wild Boars. Our results confirm previous analysis on the genomic diversity of the local breeds. Classification results using the 70 K HD porcine SNP chip were reliable and robust, hence DAPC could be considered as a potential tool for local pig breed traceability in the future. Our results indicate that robustness of the model could further benefit with bigger sample sizes. Nevertheless, cost of genotyping might be a limiting factor for a wide scale application. To overcome this limitation, a search for the minimum set of SNPs, that could achieve similar results obtained with the medium density SNP chip, could be proposed. Indeed, it would be useful to genotype a high proportion of the individuals belonging to the breeds with the highest risk of extinction or in any case with a greater risk of introgression from other populations. The cost of the set of SNPs is therefore fundamental given that for many of the breeds considered in this study there is a limited budget for genotyping. Our results suggest that integration of statistical methodologies to investigate genomic variability within and between breeds should be considered. We hope our findings to contribute and enhance the indigenous pig farming.

Methods

Animals and genomic data.

Our initial pig genomic data (n = 1,195) were obtained from three sources: (i) 20 European indigenous breeds (n = 987) reared in 9 countries (Croatia: Black Slavonian, Turopolje; France: Basque, Gascon; Germany: Schwabisch-Hällisches Schwein; Italy: Apulo Calabrese, Casertana, Cinta Senese, Mora Romagnola, Nero Siciliano, Sarda; Lithuania: Indigenous Wattle, White Old Type; Portugal: Alentejana, Bísara; Serbia: Moravka, Swallow-Bellied Mangalitsa; Slovenia: Krškopolje pig; Spain: Iberian, Majorcan Black), and retrieved from the European funded project TREASURE (https://treasure.kis.si/). Blood samples were collected from each institution by specialized professionals, following standard guidelines. No interventions with animals were applied that would require ethical protocols (according to Directive 2010/63/EU-2010) (more details on sampling method can be found in Muñoz et al.[12]), (ii) three commercial breeds including Duroc (n = 53), Landrace (n = 52) and Large White (n = 52) and (iii) a sample of Wild Boars (n = 51) from Finland, Hungary, Italy, Spain, Poland, Russia, The Netherlands, Tunisia, and Greece was carefully selected from the Dryad Digital Repository: DOI: 10.5061/dryad.30tk6 (https://doi.org/10.5061/dryad.30tk6)[44]. Further details on the selection of the Wild Boars are provided in the Supplementary Information. In addition, a small Spanish Wild Boar sample (n = 7) was also added[12]. All pigs from the indigenous and the three commercial breeds were genotyped with the GeneSeek Genomic Profiler (GGP) 70 K HD porcine genotyping chip containing 68,516 SNPs. The Wild Boars were genotyped with the Illumina 60 K SNP data[45]. The merged data contained 42,464 autosomal SNP. Samples with more than 10% and SNPs with more than 5% of missing values were excluded. The final data consisted of 1,186 pigs and 40,364 SNP (Table 1).

Population stratification and ancestry.

Admixture and PCA were used to investigate the data structure in terms of distinct populations. The two approaches, are complementary to each other. More precisely, PCA produces orthogonal projections of the original data, variance driven (from the highest to the lowest), focusing on how different populations are structured (between and within). In contrast, an admixture analysis provides the proportions from each of the source populations in each sample, i.e., how the individual samples are related to the source populations (ancestries). The PCA was performed in R software[36], using the *prcomp* function, while the proportion of mixed ancestry was assessed using the *ADMIXTURE 1.22* software[46,47]. The number of ancestries (K) to be retained in admixture (K = 2–24) was evaluated via a fivefold cross-validation (CV) and the model with minimum CV error was selected for further analysis. Results were also summarized per breed for an easier representation.

Discriminant analysis.

DAPC[48] was applied to assess breed traceability, as implemented in the R package *adegenet*[36,49,50]. DAPC replaces the original SNP data with a small set of principal components (PCs) and then applies a linear discriminant analysis on the selected PCs. In this way, DAPC maximizes the differences among groups while overlooking at the variability within groups. The number of PCs to be used in the discriminant analysis is determined via CV and the targeting function can be either the lowest root mean squared error or the highest mean success. To select the best option both methods were evaluated: In brief, data were randomly sampled in sets starting from 30% and augmenting by 10% up to the complete dataset, one repetition each, having all the breeds represented (stratified sampling), and the overall model assignment accuracy was recorded (Table S1). For each set, a tenfold CV was applied, and repeated 30 times, to select the optimum number of PCs for the discriminant analysis. On average, minimum prediction error slightly outperformed the highest mean success, and this was the option kept in subsequent analysis. It should be noted that according to Jombart[49] this is also the recommended option.

The objective of DAPC was to represent real case scenarios, i.e., to identify an external individual membership to a group (external validation). In such a case, the discriminant function is developed in a training set (TRN) and then applied on genotypes of an external validation set (VAL). The function *predict.dapc* was used for this analysis. Two different approaches were applied:

- Scenario 1 (semi-supervised learning). Data were randomly (without replacement) split at 80–20% for the TRN-VAL set, and the split was repeated 10 times. Random sampling was conditioned such that all the breeds were present in both TRN and VAL sets (stratified sampling).
- Scenario 2 (un-supervised learning). Each breed was analysed separately and consisted of the VAL set. In this scenario, no pigs of the VAL set were present in the TRN set, hence pigs had to be classified in one of the other 23 breeds. The TRN set consisted of pigs from the rest of the 23 remaining breeds, randomly selected (without replacement). This procedure was repeated 10 times. Scenario 2 can be seen as a method to assess similarity among breeds.

In both scenarios, the design of the DAPC analysis included: (i) tenfold CV for the selection of the optimum number of the PCs, (ii) the maximum number of PCs tested was set to 300 and (iii) minimum prediction error as the target function for model selection. Results were summarized over the 10 repetitions. Moreover, to assess the effect of the sample size and the robustness of the model, the complete dataset was split in sets of 10% increase (from 30 up to 100%). The terms (semi/un)-supervised should not be confused with the terminology in machine learning. These terms were used to distinguish between the two scenarios of DAPC, and although they are analogous to same terms used in the statistical field of machine learning they are not identical.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials. The raw genetic datasets generated during the current study are available from the corresponding author on reasonable request. The external Wild Boars sample can be found in Dryad Digital Repository: DOI: 10.5061/dryad.30tk6 (https://doi.org/10.5061/dryad.30tk6).

Received: 15 October 2021; Accepted: 11 April 2022
Published online: 05 May 2022

References

1. Giuffra, E. *et al.* The origin of the domestic pig: Independent domestication and subsequent introgression. *Genetics* **154**, 1785–1791 (2000).
2. Larson, G. *et al.* Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**, 1618–1621 (2005).
3. Larson, G. *et al.* Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc. Natl. Acad. Sci. USA* **104**, 15276–15281 (2007).
4. Ottoni, C. *et al.* Pig domestication and human-mediated dispersal in western eurasia revealed through ancient DNA and geometric morphometrics. *Mol. Biol. Evol.* **30**, 824–832 (2013).
5. Laval, G. *et al.* Genetic diversity of eleven European pig breeds. *Genet. Sel. Evol. GSE* **32**, 187–203 (2000).
6. Boitard, S. *et al.* Genetic variability, structure and assignment of Spanish and French pig populations based on a large sampling. *Anim. Genet.* **41**, 608–618 (2010).
7. Ollivier, L. European pig genetic diversity: A minireview. *Animal* **3**, 915–924 (2009).
8. Foulley, J. L. *et al.* Genetic diversity analysis using lowly polymorphic dominant markers: the example of AFLP in pigs. *J. Hered.* **97**, 244–252 (2006).
9. SanCristobal, M. *et al.* Genetic diversity in European pigs utilizing amplified fragment length polymorphism markers. *Anim. Genet.* **37**, 232–238 (2006).
10. SanCristobal, M. *et al.* Genetic diversity within and between European pig breeds using microsatellite markers. *Anim. Genet.* **37**, 189–198 (2006).
11. Lukić, B. *et al.* Conservation genomic analysis of the croatian indigenous black slavonian and turopolje pig breeds. *Front. Genet.* **11**, 261 (2020).
12. Muñoz, M. *et al.* Genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds assessed with a high density SNP chip. *Sci. Rep.* **9**, 1–14 (2019).
13. Abbott, A. An inside look at the first pig biobank. *Nat. News* **519**, 397 (2015).
14. Picardy, J. A., Pietrosemoli, S., Griffin, T. S. & Peters, C. J. Niche pork: Comparing pig performance and understanding producer benefits, barriers and labeling interest. *Renew. Agric. Food Syst.* **34**, 7–19 (2019).
15. Alonso, M. E., González-Montaña, J. R. & Lomillos, J. M. Consumers' concerns and perceptions of farm animal welfare. *Animals* **10**, 385 (2020).
16. Edwards, S. A. Product quality attributes associated with outdoor pig production. *Livest. Prod. Sci.* **94**, 5–14 (2005).
17. García-Gudiño, J. *et al.* Understanding consumers' perceptions towards Iberian pig production and animal welfare. *Meat Sci.* **172**, 108317 (2021).
18. Goffaux, F., China, B., Dams, L., Clinquart, A. & Daube, G. Development of a genetic traceability test in pig based on single nucleotide polymorphism detection. *Forens. Sci. Int.* **151**, 239–247 (2005).
19. Fries, R. & Durstewitz, G. Digital DNA signatures for animal tagging. *Nat. Biotechnol.* **19**, 508–508 (2001).
20. Schachler, K., Distl, O. & Metzger, J. Tracing selection signatures in the pig genome gives evidence for selective pressures on a unique curly hair phenotype in Mangalitza. *Sci. Rep.* **10**, 22142 (2020).
21. Schiavo, G. *et al.* Runs of homozygosity islands in Italian cosmopolitan and autochthonous pig breeds identify selection signatures in the porcine genome. *Livest. Sci.* **240**, 104219 (2020).
22. Gurgul, A. *et al.* A genome-wide detection of selection signatures in conserved and commercial pig breeds maintained in Poland. *BMC Genet.* **19**, 1 (2018).
23. Qin, M., Li, C., Li, Z., Chen, W. & Zeng, Y. Genetic diversities and differentially selected regions between shandong indigenous pig breeds and western pig breeds. *Front. Genet.* **10**, 1 (2020).
24. Hlongwane, N. L., Hadebe, K., Soma, P., Dzomba, E. F. & Muchadeyi, F. C. Genome wide assessment of genetic variation and population distinctiveness of the pig family in South Africa. *Front. Genet.* **11**, 1 (2020).

25. Muñoz, M. *et al.* Development of a 64 SNV panel for breed authentication in Iberian pigs and their derived meat products. *Meat Sci.* **167**, 108152 (2020).
26. Mujibi, F. D. *et al.* Genetic diversity, breed composition and admixture of Kenyan domestic pigs. *PLOS ONE* **13**, e0190080 (2018).
27. Wang, J. *et al.* Genome-wide analysis reveals human-mediated introgression from western pigs to indigenous Chinese Breeds. *Genes* **11**, 275 (2020).
28. Deperi, S. I. *et al.* Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. *PLOS ONE* **13**, e0194398 (2018).
29. Dimauro, C. *et al.* Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. *Small Rumin. Res.* **128**, 27–33 (2015).
30. Moradi, M. H., Khaltabadi-Farahani, A. H., Khodaei-Motlagh, M., Kazemi-Bonchenari, M. & McEwan, J. Genome-wide selection of discriminant SNP markers for breed assignment in indigenous sheep breeds. *Ann. Anim. Sci.* **1**, 1 (2020).
31. Dimauro, C. *et al.* Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. *Anim. Genet.* **44**, 377–382 (2013).
32. Paschou, P. *et al.* PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. *PLOS Genet.* **3**, e160 (2007).
33. Lewis, J. *et al.* Tracing Cattle Breeds with Principal Components Analysis Ancestry Informative SNPs. *PLOS ONE* **6**, e18007 (2011).
34. Wilkinson, S. *et al.* Evaluation of approaches for identifying population informative markers from high density SNP Chips. *BMC Genet.* **12**, 45 (2011).
35. Ding, L. *et al.* Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genom.* **12**, 622 (2011).
36. R Core Team. *R:* A language and environment for statistical computing, Vienna, Austria. (2021).
37. Warnes, G. R. *et al.* gplots: Various R Programming Tools for Plotting Data. 3.1.1. https://CRAN.R-project.org/package=gplots. (2020).
38. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinforma. Oxf. Engl.* **32**, 2847–2849 (2016).
39. Blutke, A. *et al.* The Munich MIDY Pig Biobank—A unique resource for studying organ crosstalk in diabetes. *Mol. Metab.* **6**, 931–940 (2017).
40. Bovo, S. *et al.* Whole-genome sequencing of European autochthonous and commercial pig breeds allows the detection of signatures of selection for adaptation of genetic resources to different breeding and production systems. *Genet. Sel. Evol.* **52**, 33 (2020).
41. Muñoz, M. *et al.* Diversity across major and candidate genes in European local pig breeds. *PLOS ONE* **13**, e0207475 (2018).
42. Schiavo, G. *et al.* Runs of homozygosity provide a genome landscape picture of inbreeding and genetic history of European autochthonous and commercial pig breeds. *Anim. Genet.* **52**, 155–170 (2021).
43. Frantz, L. A. F. *et al.* Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proc. Natl. Acad. Sci.* **116**, 17231–17238 (2019).
44. Yang, B. *et al.* Genome-wide SNP data unveils the globalization of domesticated pigs. *Genet. Sel. Evol.* **49**, 71 (2017).
45. Ramos, A. M. *et al.* Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLOS ONE* **4**, e6524 (2009).
46. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
47. Alexander, D., Shringarpure, S., Novembre, J. & Lange, K. *ADMIXTURE Version 1.3.0*.
48. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
49. Jombart, T. adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
50. Jombart, T. & Ahmed, I. adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634476 for the project with acronym TREASURE. The content of this article reflects only the authors' view and the European Union Agency is not responsible for any use that may be made of the information it contains. MŠ and MČP received core-financing by the Slovenian Research Agency (grant number P4-0133).

Author contributions:

R.B., C.D., and M.C.F.; conceived and designed the study, M.Č.P, L.F., and C.O.; funding acquisition, C.D.; formal analysis and writing – original draft preparation, C.D., R.B., M.C.F., A.C., C.O., M.M., D.K., M.Č.P, M.Š, R.C., J.P.A., V.R., M.J.M., R.S.; writing—review and editing. All authors read and approved the submitted version.

Funding

This study has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 634476 to CO (project acronym TREASURE). The content of this paper reflects only the author's view, and the European Union Agency is not responsible for any use that may be made of the information it contains. MČP and MŠ were co-financed by Slovenian Research Agency (grant P4-0133).

Competing interests

One of the authors of this manuscript is an academic editor for Scientific Reports. Dr. Luca Fontanesi (L.F.), Department of Agricultural and Food Sciences, Division of Animal Sciences, University of Bologna, Viale Fanin 46, 40127 Bologna, Italy. The other authors declare no competing interest. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10698-8.

Correspondence and requests for materials should be addressed to C.D.