Ancestry and diversity of American village pigs

DOCTORAL THESIS

By

WILLIAM ORLANDO BURGOS PAZ
M.Sc. Universitat Autònoma de Barcelona, España, 2011
M.Sc. Universidad de Antioquia, Colombia, 2010
B.Sc. Universidad de Nariño, Colombia, 2007

Supervisors

Dr. Miguel Pérez Enciso **Dr. Sebastián Ramos Onsins**

UNIVERSITAT AUTÒNOMA DE BARCELONA
Departament de Ciencia Animal I dels Aliments
Facultat de Veterinarìa

CENTRE DE RECERCA EN AGRIGENÒMICA – CRAG
CSIC-IRTA-UAB-UB

Bellaterra, July 2014

El **Dr. Miguel Pérez Enciso**, investigador ICREA del Departament de Ciència Animal i dels Aliments de la Universitat Autònoma de Barcelona (UAB)

i

el **Dr. Sebastián E. Ramos Onsins**, investigador del Centre de Recerca de Agrigenòmica (CRAG),

CERTIFIQUEN:

Que **William Orlando Burgos Paz** ha realitzat sota la seva direcció el treball de recerca:

"**Ancestry and diversity of American village pigs**"

per a obtenir el grau de Doctor per la Universitat Autònoma de Barcelona.
Que aquest treball s'ha dut a terme al Departament de Ciència Animal i dels Aliments de la Facultat de Veterinària de la Universitat Autònoma de Barcelona (UAB) i al Centre de Recerca en Agrigenòmica (CRAG).

Bellaterra, mayo 30 de 2014

Dr. Miguel Pérez Enciso **Dr. Sebastián E. Ramos Onsins**

-

SUMMARY

-

Advances in high throughput genetic technologies are revolutionizing the understanding of domestic animal genomes, including their history and how demography and selective processes have shaped the variation of individuals' genomes. Here we studied for the first time a large survey of village pig populations from America and estimate their relatedness with worldwide pig populations. In complement, we also analysed an ancient pig genome of 16th century to provide new evidence on important historical and genetic events like domestication and admixture.

Both history and population relationships of pigs are complex and have not been well described for some populations like those extant in the Americas. Here, using single nucleotide polymorphism (SNP) arrays, we explored the genetic diversity of American village pigs, their relationships with worldwide pig populations and the Iberian pig ancestry, considering independently chromosome X (Chapter 3) and autosomes (Chapter 4).

These studies showed the high differentiation between European and Asian pig populations, being particularly pronounced for the non-pseudoautosomal region of chromosome X. Despite the Iberian origin of pigs firstly introduced to America, a substantial reduction of this ancestry was observed in almost all American village pig populations. The actual ancestry observed in America is likely the result of admixing Iberian pigs in 15th century and recent introgression of commercial pig breeds like Duroc, Landrace, Large White or Hampshire. Additionally, some Asian ancestry also was observed probably due to introgression of commercial breeds carrying Asian haplotype, although direct admixture with Chinese breeds cannot be ruled out. Because the large diversity of environmental condition in the American continent, we compared the allele frequencies observed between populations to estimate signatures of selection in the genome, detecting some genes related with cardiovascular system and limbs conformation.

Ancient DNA provides valuable information about the historical events that have modelled the genome of modern individuals. In chapter 5 we performed the analysis of the partial genome of a pig that lived in 16th century at North eastern Spain together three new modern genomes from Iberian pig, Spanish wild boar and a Guatemalan Creole pig obtained by whole genome shotgun sequencing. Archaeological and genomic data suggested that ancient pig was domestic, closely related to extant Iberian pigs and to European wild boar with some genetic signals of admixture with wild boar. Surprisingly, the comparison of ancient pig and modern Iberian pig to American sample from Guatemala, showed that they are equally close to American Creole

pigs, and could support the hypothesis of reduction of Iberian origin in American village pigs driven by introgression of other breeds. Finally, among the highly differentiated genes we found those involved in coat colour and an increase the reproductive performance, both known functions associated with early domestication process.

One of the analytical strategies to describe the population relationships of pigs used in this thesis, and widely used in similar studies is the principal component analysis (PCA). This technique is very useful in presence of large amount of data (e.g. genotypes from SNP arrays) and highly computationally efficient. Nonetheless, PCA projections are sensitive to unequal sample size. In Chapter 6 we evaluated a correction in PCA that consider either sample size of evaluated populations or their F_{ST} estimates to correct bias in individual projections. Simulations suggest that the proposed method improves the two-dimensional projections of PCA data and, in some cases, entirely recovers population relationships patterns, even when sample size is as low as n=1. The weighted PCA can recover a more realistic structure than inferred with traditional PCA in well-structured populations.

-

RESUMEN

-

Los avances en las tecnologías de genotipado masivo están revolucionando la comprensión de los genomas de animales domésticos, incluyendo su historia y cómo los eventos demográficos y selectivos han dado forma a la variación de los genomas de los individuos. En este trabajo analizamos por primera vez una amplia muestra de poblaciones de cerdos criollos de América y se estimó su relación genética con las poblaciones de cerdos en todo el mundo. Adicionalmente, se analizó el genoma de un cerdo que vivió en el siglo XVI a fin de proporcionar nuevas evidencias sobre eventos históricos y genéticos importantes como la domesticación y cruzamientos.

Tanto la historia y las relaciones entre las poblaciones de cerdos es compleja y no muy bien descritas para algunas poblaciones como las existentes en América. En esta tesis, usando la información de SNPs evaluamos la diversidad genética de poblaciones de cerdos criollos de América, su relación con poblaciones de cerdos distribuidas en el mundo y su origen genético Ibérico, mediante el análisis independiente del cromosoma X (capitulo 3) y los autosomas (capitulo 4).

Estos estudios mostraron una alta diferenciación entre las poblaciones de cerdos de Europa y Asia, siendo especialmente pronunciada para la región no pseudoautosómica del cromosoma X. A pesar del origen ibérico de los primeros cerdos introducidos en América, se observó una reducción sustancial de éste origen genético en el casi todas las poblaciones Americanas de cerdos estudiadas. El origen genético observado en las actuales poblaciones de cerdos en América se debió primero al los cerdos Ibéricos en el siglo XV y la reciente introgresión de las razas porcinas comerciales como Duroc, Landrace, Large White o Hampshire. Además se detectó origen genético proveniente de Asia, probablemente por causa de la introgresión de las razas comerciales que llevan haplotipos asiáticos, aunque no se puede descartar el cruzamiento con razas Chinas. Debido a la gran diversidad de condiciones del medio ambiente en América, se compararon las frecuencias alélicas observadas entre las poblaciones para estimar huellas de selección en el genoma, detectándose algunos genes relacionados con el sistema cardiovascular y la conformación de las extremidades.

El ADN antiguo proporciona una valiosa información a cerca de los acontecimientos históricos que han modelado el genoma de los individuos modernos. En el capítulo 5 se analizó una parte del genoma de un cerdo que vivió en el siglo XVI en el noreste de España, junto a tres nuevos genomas de individuos modernos, un cerdo Ibérico, un jabalí de España y un cerdo criollo de Guatemala. Los genomas fueron obtenidos por métodos de secuenciación masiva. Los datos

arqueológicos y genómicos sugirieron que el cerdo antiguo era domestico y estrechamente relacionado con los cerdos Ibéricos actuales y el jabalí Europeo, y con señales genéticas de cruzamiento entre ellos. Sorprendentemente, la comparación del cerdo antiguo y el cerdo Ibérico moderno con el genoma del cerdo criollo de Guatemala, mostró que ellos son igualmente cercanos a los cerdos criollos de América, y podría apoyar la hipótesis de la reducción de origen ibérico en cerdos Americanos causado por la introgresión de otras razas. Por último, entre los genes altamente diferenciadas se encontraron aquellos que participan en el color de la capa y el aumento del rendimiento reproductivo, ambas funciones asociadas con el primeros procesos de domesticación.

Una de las estrategias de análisis para describir las relaciones de la población de cerdos utilizados en esta tesis, y ampliamente utilizado en estudios similares, es el análisis de componentes principales (PCA). Esta técnica es muy útil en presencia de gran cantidad de datos (como en los genotipos de chips de SNPs) y computacionalmente muy eficiente. Sin embargo, las proyecciones de PCA son sensibles a tamaño de muestra desbalanceado. En el capítulo 6 se evaluó una corrección en el PCA que tenga en cuenta el tamaño de la muestra de las poblaciones evaluadas ó su F_{ST} para corregir el sesgo en las proyecciones de los individuos. Las simulaciones sugieren que el método propuesto mejora las proyecciones de los individuos en dos dimensiones y en algunos caso, recupera los patrones de relaciones de la población, incluso cuando el tamaño de muestra es tan bajo como n = 1. El método ponderado de PCA puede recuperar una estructura más realista de los datos que la inferida con el PCA tradicional en poblaciones genéticamente diferenciadas.

CONTENTS

-

LIST OF TABLES

-

-

LIST OF FIGURES

-

-

LIST OF ABBREVIATIONS

-

aDNA	Ancient Deoxyribonucleic acid
BAC	Bacterial Artificial Chromosome
bp	base pair
EHH	Extended haplotype homozygosis
F_{ST}	Fixation index
He	Expected heterozigosity
Kya	Thousand year ago
LD	Linkage disequilibrium
Mb	mega base (10e6 bp)
mtDNA	Mitochondrial Deoxyribonucleic acid
Mya	Million year ago
Ne	Population size
NGS	Next generation sequencing
NPAR	Non Pseudoautosomal region
PAR	Pseudoautosomal region
PCA	Principal Component Analysis
RNA-seq	Ribonucleic acid sequencing
SNP	Single nucleotide polymorphism
SSCX	*Sus scrofa* chromosome X
WGS	Whole genome sequence
wPCA	Weighted Principal Component Analysis

Chapter 1

-

Introduction

-

1.1 The pig (*Sus scrofa*)

The origin of pig

The pig (*sus scrofa*) is an *Eutherian* mammal, member of the family Suidae that includes domestic pig and its ancestor the wild boar. All suids are endemic of Eurasian and Africa, whereas a second well established suborder of cetartiodactyla, the Tayassuidae (peccaries), lives nowadays in America (Ruvinsky and Rothschild, 1998).

Mitochondrial DNA (mtDNA) and whole genome sequence data (WGS) indicate the emerging of Eurasian *Sus scrofa* from Southeast Asia around 5.3–3.5 Myr ago (Larson, Cucchi, *et al.*, 2007; Groenen *et al.*, 2012; Frantz *et al.*, 2013). In subsequent million years, climate changes and tectonic activities in Island Southeast Asia (ISEA, Bird et al. 2005) allowed the migration of *Sus* species in contiguous sundaland resulting in the differentiation of others suids like the ancestor of *Sus cebifrons* in Philippines (Frantz *et al.*, 2013). During the first Pleistocene stage (2.4 -1.6 Mya) *Sus scrofa* East Asia mainland populations diverged from Sumatra. Nevertheless, direction of migrations (ISEA to Mainland of *vice versa*) that allowing the divergence between *S. scrofa* sub species remains unclear. Thus, the mid-Pleistocene witnessed the migrations and successful spread of *S. scrofa* through Asia. Phylogenomic analyses of complete genome sequences of wild boars from Asia and Europe locate the split of them approximately 1.6-0.8 Myr ago (Groenen *et al.*, 2012). A divergence of *S. scrofa* in discrete populations started with west migration approximately 1.2 Myr ago, reaching Europe around 0.8 Myr ago, as estimated by analyses of wild boar remains found at

Atapuerca (van der Made, 2001), and estimated genetically using molecular clock analyses (Frantz *et al.*, 2013).

Northwards migration of *S. scrofa* also took place in the same period leading the split between northern and southern Chinese populations around 0.6 Myr ago, observed from WGS (Frantz et al. 2013) and previously suggested by mtDNA (Larson *et al.*, 2010). The successful adaptation of *S. scrofa* to extremely different environments was a determinant factor for the species spreads. However, strong climatic conditions as reductions of temperature, long glacial intervals and forests contractions shaped the variability and divergence of the populations (Hewitt, 2000) including *S. scrofa.*

Tracing pig domestication

According to effective population size (***Ne***) estimated by Groenen et al. (2012), after colonization of Eurasia the *S.scrofa* population developed and kept partially constant along thousand of years, emerging a set of new distinctive subspecies through Europe and Asia (Ruvinsky and Rothschild, 1998). Prior to the domestication, two factors modelled the wild boar distribution across Eurasia. The first was the Human-mediated translocations during the last 40,000 years (Larson *et al.*, 2005; Ji *et al.*, 2011; Frantz *et al.*, 2013) in ISEA and more recently in Near East (Ottoni *et al.*, 2013). The second was the environmental pressures caused by Last Glaciation Maximum resulting in isolation, bottlenecks and population differentiation of *S. scrofa* populations being stronger in Europe than Asia (Scandura *et al.*, 2008; Alexandri *et al.*, 2012). A particular example of the effects of environmental pressure was the number of chromosomes among Southern refuges, individuals from Western Europe, originated from the Iberian refugium (Cantabrian region) have 2n = 36, whereas Balkan refugium wild boars from Eastern Europe have chromosome number 2n = 38 (Fang *et al.*, 2006). In the figure 1.1 we show the actual distribution of wild boar and *Sus scrofa* relatives and graphically represent some migration routes estimated from archaeological and genetic data.

Domestication in Asia

The evidence collected from archaeological and modern DNA data collected across Asia locate China as the main domestication centre (Larson *et al.*, 2010), and multiple ancestral origins of indigenous pigs in different areas across Asia (Wu *et al.*, 2007; Tanaka *et al.*, 2008; Luetkemeier *et al.*, 2010). According to Larson et al. (2010) after migrations across Kra Isthmus into Mainland Asia from ISEA, pigs spread reaching Japan, Ryukyu chain, Taiwan and Lanyu in subsequent years.

The transition of wild to farm pigs occurred about 8000 years ago, resulting in pigs with domestic phenotypes at early times and traditional breeds that in some regions currently persists (White, 2011). Moreover, China is one of the countries with more abundant pig breeds in the world (Yang *et al*. 2003, DAD-FAO). Some regions have been suggested as domestication centres; Yellow River drainage basin, in the northern China, and in the downstream Yangtze River region (Larson *et al.*, 2010; Figure 1.1). Likewise, The Mekong region and the middle and downstream regions of the Yangtze River in China has also been proposed as domestication centres (Wu *et al.*, 2007). Other studies call for an independent

domestication event and/or great gene flow between wild and domestic pigs in the foot of Himalaya (Tanaka *et al.*, 2008), the Tibetan highlands (Yang *et al.*, 2011; Li *et al.*, 2013) and South China (Yu *et al.*, 2013).

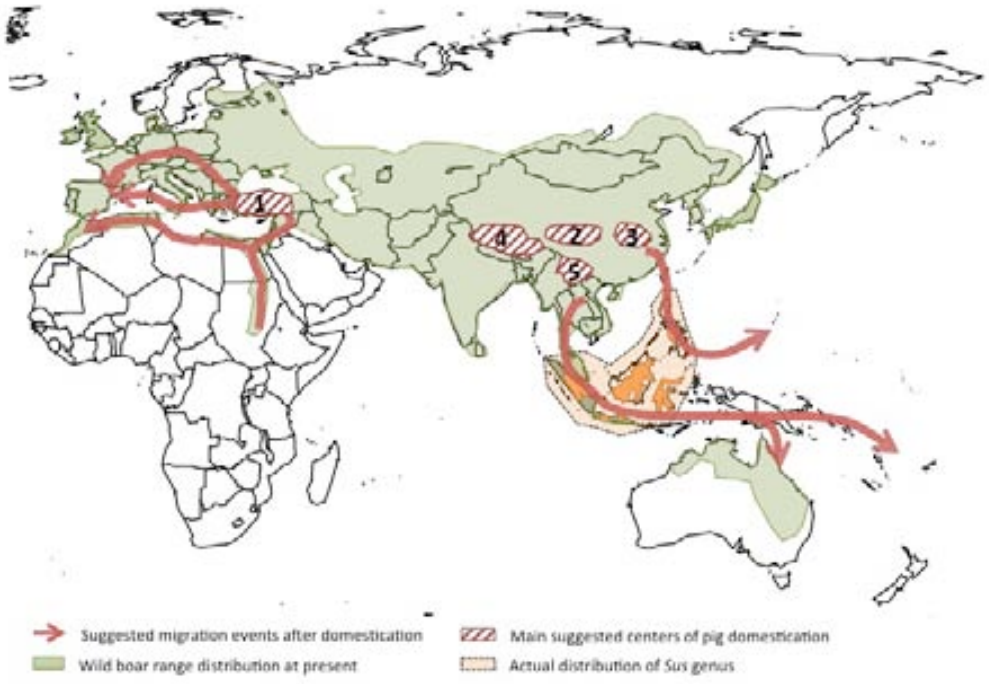


Figure 1. Current distribution of wild boar and its relatives, main domestication centres of domestication and migratory routes of pigs after domestication. Domestication centre 1 correspond to Anatolia, Near east; centres 2, 3, 4 correspond to Yellow River drainage basin in northern China; and centre 5 the downstream Yangtze River region (modified from Ramos-Onsins *et al*. submitted).

The modern Chinese domestic breeds are direct descendants from ancient domestic pigs in the region and wild boars seem not to have incorporated haplotypes from feral pigs (Larson *et al.*, 2010). Furthermore, other centres of domestication in Asia like Southeast Asia and Indus Valley Civilization derived from genetic analyses, but the lack of morphological data do not support totally these observations (Tanaka *et al.*, 2008; Larson *et al.*, 2010).

Domestication in the west: Near East and Europe

The conjugation of archaeology and molecular evidences allowed locating the earliest centre of pig domestication in Eastern Anatolia (Larson, Albarella, *et al.*, 2007; Figure 1.1). This site represents the phylogeographic boundary of Near East and European wild boar haplotypes. Using mtDNA sequences from ancient Anatolian wild and domestic pigs (dated ~6000 BC), the authors found Y1 haplotypes, one Near East linage, corroborating that earliest domestic pigs in Europe originated from populations originally domesticated in Near East and further introduced to Europe, possibly by Mediterranean routes (Larson, Albarella, *et al.*, 2007; Ottoni *et al.*, 2013). The same authors studied the current absence of Near East haplotypes in Europe populations (Larson *et al.*, 2005; Manunza *et al.*, 2013). Once Anatolian domestic pigs were introduced into Europe as far as Paris Basin, they were mixed with European

Individuals, the Near East haplotype Y1 frequency was dramatically reduced. The domestic pigs of European wild boars were introduced in Anatolia and DNA signatures of Near East were completely replaced (Ottoni *et al.*, 2013).

The Western domestication seems to be more complete considering the well documented and accurate history related with human settlement and migrations. Several events in the subsequent years modelled the history of pig populations in Europe and Near East. The human-mediated translocations and hunter-gatherer activities have shaped the populations in principal (Vigne *et al.*, 2009; Krause-Kyora *et al.*, 2013; Meiri *et al.*, 2013). Moreover, wild boar domestications or hybridizations with domestic pigs also established the distribution and differentiation of pig populations (Scandura *et al.*, 2008; Goedbloed *et al.*, 2013).

More recently, between 18th and 19th centuries, the changes in the population and economical conditions draw tremendous changes in the pig production in Eurasia. One of the earliest farming practices, the mast feeding was almost entirely replaced by intensive confinement production increasing in herd and flock sizes (Wealleans, 2013). Facing the new conditions, strong selective process took place in England modifying notably the relations of European local breeds. The arrival of Chinese breeds to Europe and intercrossing with England local breeds finished in high performance of pig breeds (Giuffra *et al.*, 2000; Megens *et al.*, 2008; Ai *et al.*, 2013). Subsequent intercrosses and selective strategies resulted in a wide variety of improved breeds including some actual international commercial breeds appeared in the mid 19th century. Each of these breeds has phenotypic differences and diversifying selection for desirable traits like coat color, ear, leanness meat quality or immunity (Amaral *et al.*, 2011; Wilkinson *et al.*, 2013) that in combination with globalized markets makes it successful in the world.

"New pigs" for a "New world"

Sixteenth century and onwards had profound implications in pig populations that were well established in Europe. From this point in history until nineteenth century, pig populations were introgressed with Chinese pigs, expanded to the Americas and underwent new selective pressures for lard and meat production and higher reproductive performance.

As mentioned previously, the pig was absent in the Americas, and only a related suids family inhabit this territory. With the arrival of Spaniards to America, and especially in the second trip of Columbus in 1493 eight pigs, horse, goat, sheep, and plants were introduced (Rodero et al. 1992) first in La Española (actual Dominican republic and Haiti). The pig settlement in America originated from importations of Iberian pigs from southwest of Iberian Peninsula (Extremadura y Andalusia) and the Canary Island (Rodero *et al.*, 1992). They adapted quickly to Caribbean environment, constituting a relevant source of meet in the very early days of conquest. Further, the domestic animals were quickly dispersed in Antilles (La Española), after on the continent though Cuba (Martínez *et al.*, 2005), Dominica, Guadalupe, the Virgin Isles and Puerto Rico, discovered afterwards (Rodero *et al.*, 1992).

During the Conquest and the Colonial periods, pig populations quickly reached a high number of individuals, due to their high prolificacy and adaptability, and were the genetic basis of

almost American porcine (i.e. feral, naturalized or creole) populations. In this sense, the Caribbean islands were the "colonizing" founders of pig in Americas during subsequent years. From here, three main routes were followed to achieve the extant presence of pigs in America. Northward route: from Veracruz in Mexico to Florida, New Mexico and California; the second one from Panama to Central America up to northern Peru and Venezuela; southward route from Rio de la Plata to Uruguay, south of Brazil and Peru and Bolivia (Revidatti, 2009).

Gradually, the pig was covering the "New world", grew up in wild and semi-wild state, showing a range of distinctive phenotypic trends (Figure 2) based on adaptation of a wide range environmental conditions (from the sea level in the coast to the high land of the Andean). Currently, we found a large number of pig populations throughout the American continent, with phenotypic particularities as Hairless in *Pelón Mexicano*, from Mexico; mule foot (Casco de Mula) from Colombia, Uruguay and North America; wattled pig (Mamellado), from Uruguay; high parasite resistance like Feral pigs from Northern Argentina; highland well adapted like Peruvian creole pig, from Peru and miniature like *Cuino*, from Mexico or *Yucatan pig*, from Central America.

Further pig introductions were carried out by Portuguese and other Europeans like England settlers. Portuguese brought soon after the discovery of America (in particular Brazil) new pigs belonging to Mediterranean breeds (Mariante and Cavalcante, 2006). Currently Brazilian pig populations like Moura, Piau or Monteiro, are characterized by their resistance to diseases and low management requirements and feeding as well as a high adaptability (Sollero *et al.*, 2009). Because the Portuguese commercial trade with Asian and African colonies, it is plausible that Asian individuals entered to South America directly or caused for European individual carrying Asian haplotypes (Ramírez, Ojeda, *et al.*, 2009; Souza *et al.*, 2009). England settlers introduced pigs to British colonies in America from British Isles. The most notable case was John Smith's settlement of Jamestown (Virginia, USA) in 1607, where introduced pigs multiplied and in very few years raising around seven hundred heads (Gade, 2012). Pigs were left in the forest and in late autumn rounded up and slaughtered, although some of them became wild and nowadays the lack of control in these populations caused substantial effects on the community (Campbell and Long, 2009).

Introgression of improved breeds in native worldwide populations

As described above, multiple events and hundred of thousand of years of natural selection has modelled the observed variability of pigs around the world. However last millennium witnessed the human mediated translocation of pigs and their interactions with new environments, leading to develop of a wide variety of phenotypes adapted to new stress conditions around the world.

1.2 Using the DNA polymorphisms to explore population genetic signatures

Selective processes in the populations can promote rapid changes in phenotype of individuals. It is expected that these phenotypic changes correlates with changes in the genome of the individuals, for instance the coat color in pigs, which is caused by mutations and structural variations (Johansson *et al.*, 1996; Kijas *et al.*, 1998). The analysis of polymorphisms in the pig genome provided us information about their relationships and variability and estimate the demography events occurred during development of populations (Larson *et al.*, 2005; Stoneking and Krause, 2011).

In the early 1980s, a large battery of molecular markers, like microsatellites, have been specially designed to identify either mutations or special motifs in the genome (Schlötterer, 2004). Each marker has different properties and advantages for the evaluation of the variations in the genome according the objectives of research (Sunnucks, 2000; Chenuil, 2006). In the 90's, microsatellite markers were widely used in genetics, because their high polymorphism, reproducibility and automation for detection. Microsatellite markers were widely used in pig genetics, from studies in population genetics (Megens *et al.*, 2008) and quantitative genetics (Andersson *et al.*, 1994; Rothschild *et al.*, 2007). However, disadvantages such as time consuming, elevated cost to obtain and difficulties in accurate detection of repetition length limit its usefulness. Currently, microsatellite and other molecular markers are used to evaluate specific questions, and availability or new genome-wide polymorphism detection technologies based on single nucleotide polymorphisms have became more popular (Liu *et al.*, 2005; Xing *et al.*, 2005).

Single Nucleotide Polymorphisms (SNPs)

The single nucleotide polymorphism or so-called SNP is a single nucleotide position in the genome that shows a mutation between sequences derived from a population (Vignal *et al.*, 2002). Despite limited polymorphism of a single SNP compared with microsatellite (Helyar *et al.*, 2011), the large number of SNPs that can be detected in the genome account for the same or more information for population or quantitative genetic parameter inference (Wakeley *et al.*, 2001; Goddard *et al.*, 2010). For instance, estimates of genetic distance or parentage verification suggest that around one hundred SNPs are as informative as 20 microsatellite markers (Kalinowski, 2002; Baruch and Weller, 2008).

Currently, SNPs are highly preferred for genotyping because they are abundant, stable and the process can be highly automated. Indeed, there are SNP genotyping technologies commercially available that allowing obtain genotypes for over four million of SNPs in humans, or ~62000 SNPs in pigs (http://www.illumina.com). For humans, it is possible detect on average one SNP in 700 nucleotides, whereas in pig we can obtain one SNP per ~40000 nucleotides. In pigs, the low SNP genotyping cost per sample, has provided invaluable information of the allele frequencies across genome and allowed the genetic characterization of several populations around the world (Ramos *et al.*, 2009; Ramayo-Caldas *et al.*, 2010; Badke *et al.*, 2012; Manunza *et al.*, 2013; Yang *et al.*, 2014).

Advantages of SNPs are evident because the cost effective method to explore variations in a large number of positions across the genome. The number of SNPs genotyped in commercial arrays continuously increases: from 3000 (3k) SNPs initially in cattle, then 50k SNPs and now 770k SNPs (Matukumalli *et al.*, 2009; Hulsegge *et al.*, 2013). In humans, HapMap and 1000 genomes projects (Altshuler *et al.*, 2010; Abecasis *et al.*, 2012) have reported more than 7 million of SNPs in worldwide populations and large part of them are now available in SNPs arrays. However, the SNP arrays suffer ascertainment bias caused by the process to estimate and select the SNPs included in it and the sample used to SNP discovery (Clark *et al.*, 2005). Thus, ascertainment bias effect will tend to ignore SNPs in low frequency (ancestral or derived) in populations not related to the discovery panel, affecting the allele frequency spectrum and hence diversity, differentiation and even recombination (Albrechtsen *et al.*, 2010).

In pigs, the Porcine 60K SNP array was discovered based on six populations (Duroc, Pietrain, Landrace, Large White, and Wild Boars from Europe and Japan) but the majority of SNPs represent mainly the variation in commercial breeds from Europe and United States (Ramos *et al.*, 2009). The authors make warnings about ascertainment bias in this SNP chip because extant LD in Chinese breeds and low SNP density on the chip to explore this population. This effect has been observed and evaluated in Chinese and American village pigs (Ai *et al.*, 2013; Burgos-Paz *et al.*, 2013) showing an increase in SNPs in low frequency in these populations.

A large number of approaches have been proposed to correct genetic parameters estimates for this effect (Ramírez-Soriano and Nielsen, 2009; Guillot and Foll, 2009; Albrechtsen *et al.*, 2010) attempting to consider how samples and SNPs were collected and selected respectively. To reduce this effect, a worldwide survey will be efficient, however the expensive cost (e.g. in livestock) limit its application. Comparison of allele frequencies in the populations of interest with an outgroup or ancestral population allows discover low frequency alleles for approximately unbiased inferences of population parameters (MacEachern *et al.*, 2009; Wang and Nielsen, 2012).

Next Generation Sequencing (NGS)

Next generation sequencing technologies has revolutionized biology, making it possible to estimate without bias almost any variation in the individual genome, at least in theory. NGS methods represent an extraordinary advance in research because the ability of processing millions of sequence reads in parallel, reducing notably the time and cost per nucleotide obtained (Mardis, 2008). This methodology produces a large amount of short sequences (reads) with a respective quality. Further, this reads could be analyzed by means of computational algorithms according to the research purposes (Metzker, 2010). Many applications of NGS in diverse fields of biology are currently available (Shendure and Aiden, 2012).

Some limitations come to light due mainly to low quality of sequences and low coverage of samples. NGS data revealed bias associated to the GC content and non-random distributions of errors in reads (Minoche *et al.*, 2011; Lou *et al.*, 2013). Additionally, in studies based on low coverage sequencing of samples it is possible that some chromosome it is not sequenced

for a specified site, simply by chance (Gravel *et al.*, 2011; Gautier *et al.*, 2013). NGS can be performed in a single pool of individuals (each sequenced <1X) that represent the variation of the populations. This strategy its very cost effective and nowadays several methods have been proposed to evaluate this data (Kofler *et al.*, 2011; Ferretti *et al.*, 2013; Korneliussen *et al.*, 2013; Nevado *et al.*, 2014).

NGS has been used to study the genome of livestock species like cattle (Van Tassell *et al.*, 2008; Stothard *et al.*, 2011), chicken (Kerstens *et al.*, 2011) and other mammals like horse (Doan *et al.*, 2012) or dogs (Freedman *et al.*, 2014). In pigs, NGS has motivated multiple applications by the recently publication of the porcine genome (Groenen *et al.*, 2012). NGS allowed the study of pig genome from multiple points o view like variability and SNP discovery (Amaral *et al.*, 2009; Esteve-Codina *et al.*, 2013), estimation of selection signatures in the genome (Amaral *et al.*, 2011; Rubin *et al.*, 2012), copy number variation (Paudel *et al.*, 2013), speciation (Frantz *et al.*, 2013) or *de novo* assembly of Chinese pig genomes (Fang *et al.*, 2012; Li *et al.*, 2013). Additionally, several works using RNA sequencing (RNA-seq) has been published focused on deciphering the functionality of pig genome for meat quality traits (Ramayo-Caldas *et al.*, 2012; Corominas *et al.*, 2013) and differential expression among pig breeds (Esteve-Codina *et al.*, 2011). Despite several advances in the pig genome draft, some of the works mentioned above evidence the incomplete annotation of pig genome and problems in the BAC order in the pig genome still remains.

The combination of both strategies for SNP detection, SNP array and NGS, allowed us study pig populations from different points of view. SNPs array allowed us to evaluate a large survey of pig samples and make inferences about the population relationships whereas NGS allowed us to evaluate, besides population relatedness, the patterns of variability that the array cannot detect because the density of SNPs or the ascertainment bias effect.

1.3 Statistical tools for the analysis of molecular data

Molecular tools have given a valuable opportunity to detect and analyze the variation of individual genome and infer the relatedness and demographic history of livestock populations. Nowadays, we dispose of highly efficient technologies to explore the whole genome complemented statistics, bioinformatics and computational resources to make complex inferences with this very large amount data. Allele frequencies of SNPs derived from arrays or NGS genotyping are suitable to estimate many population genetic parameters. Comparisons of allele frequencies, estimation of ancestries and relations of genetic and geographic data are strategies commonly used to evaluate genetic data in humans (Li *et al.*, 2008; Novembre *et al.*, 2008; Moreno-Estrada *et al.*, 2013) and animals (Vonholdt *et al.*, 2010; Kijas *et al.*, 2012; McTavish *et al.*, 2013). Additional analysis like positive selection, haplotype diversity and ancestral admixture are used for more specific questions. Next, we describe some statistical approaches used in this thesis, those for population clustering and then those related with differentiation, history and selective signatures.

Clustering methods: Multivariate and model-based methods

In genetics, multivariate analyses have been applied routinely in recent years because the large amount of markers and samples requires exploration and consequent interpretation of data in an efficient way. The main advantage of these methodologies is the dimensionality reduction of data in new synthetic variables (Jombart *et al.*, 2009). Additionally, the lack of underlying genetic model assumption and efficient computational algorithms (Reich *et al.*, 2008) make multivariate methods more attractive than model based clustering methods like STRUCTURE (Pritchard *et al.*, 2000).

As reviewed e.g., in Jombart et al. (2009), there are many multivariate methodologies applied to genetic data that includes non-metric dimensional scaling (NMDS), Principal component Analysis (PCA) or discriminant analysis (DA). In brief, NMDS is obtained by reduction of data into a distance matrix of the elements (samples) and then eigenvectors and eigenvalues for the distance matrix are estimated. In case of linearity of data, NMDS using Euclidean distance produces approximately similar results than PCA. In the DA prior group assignations must be known and subsequent maximization of variance in PCA between groups is estimated. Here, we focussed on principal components, its application and pitfalls.

Principal component analysis (PCA)

For decades, PCA has been a commonly multivariate methodology used to explore and visualize the genetic structure of populations (Menozzi *et al.*, 1978; Novembre and Stephens, 2008). The PCA popularity has increased in the last years because of the high throughput genotyping technologies (e.g., SNP arrays), where it is necessary to retaining only the linear combinations of variables explaining the maximum variance of the dataset (dimensionality reduction) in a computationally efficient manner (Paschou *et al.*, 2007). PCA advantages in population genetics include the lack of a historical model to interpretation because representation depends of data itself, usefulness in correcting for stratification in disease studies (Price *et al.*, 2006) and recent migrations and ancestry (Drineas *et al.*, 2010). Many studies have detected relevant geographical information of populations from PCA projections and expectations of PCA eigenvalues could easily associated as coalescence rates between individuals (McVean, 2009).

Despite several advantages of PCA in the population genetics field, this technique is very sensitive to the choice of the dataset and the distortion of the PCA plots due to a biased or unequal sampling leading to misinterpretation of population structure (McVean, 2009). A natural framework for this is correcting by means of a weighted PCA (wPCA), where the variables measured for each sample can be assigned a different weight (Kriegel *et al.*, 2008). In this thesis, we evaluate this methodology for first time under population genetics view (Chapter 6) and demonstrate that, in some cases, it is possible to recover the most realistic population structure projections into two-dimensional axes compared with traditional PCA.

Model-based clustering methods

This category of methods is also commonly used to evaluate population structure and, together with PCA, provide a summary of population features from genetic data. However, model-based methods involve an explicit model for the data, allowing attempt for reconstruct

historical events and ancestry estimates (Manel *et al.*, 2005). For admixture-based model, the probability of data P(X|K) for a given *K* cluster is estimated. This information is used to estimate the inherited proportion of each k-cluster individual ancestry.

Different algorithms with different drawbacks have been implemented and widely used. STRUCTURE (Pritchard *et al.*, 2000) uses a Bayesian framework to jointly modelling population membership and allele frequencies in each population. STRCUTURE allows a straightforward assessment of the statistical uncertainty in each estimate assuming linkage between loci (Falush *et al.*, 2003). ADMIXTURE (Alexander *et al.*, 2009), similarly to STRUCTURE, model the probability of the observed allele frequencies belongs a expected ancestry and population. Nevertheless, ADMIXTURE maximizes the likelihood instead of sampling the posterior distribution via MCMC like STRUCTURE. Maximum likelihood approach in ADMIXTURE can accommodate many more markers and further bootstrap is used in bock relaxation algorithm to estimate standard error of parameters.

Similar estimations of ancestry parameters and K-cluster are obtained in ADMIXTURE and STRUCTURE but large differences in computing-time make ADMIXTURE faster and therefore suitable for large databases analysis. Both approaches are well suited in presence of population structure, but misleading results appear in continuous genetic differentiation of populations (Engelhardt and Stephens, 2010). This issue has been observed in pig populations (Goedbloed *et al.*, 2013; Manunza *et al.*, 2013) highlighting the underestimation of a more accurate K-cluster value. ADMIXTURE model do not account for linkage disequilibrium (LD) whereas STRUCTURE does in the latest versions. However, caution in LD is highly recommended, especially when admixture events are recent, and pruning high LD SNPs improved estimates of ancestry. Another methods like FRAPPE (Tang *et al.*, 2005) and ChromoPainter or fineSTRUCTURE (Lawson *et al.*, 2012) that take into account haplotype and phase inferences into the model could be considered. Finally, has been released a two orders of magnitude faster version of STRUCTURE called fastSTRUCTURE (Raj *et al.*, 2013).

Genetic signatures and history of populations

The new technologies of genotyping and genome sequencing detect a large amount of genetic polymorphism data that could be used to estimate with great detail how the selective or adaptive processes, bottlenecks, migrations or simple genetic drift, could affect the variation in different regions of the genome. We used comparisons of allele frequency to specifically estimate differentiation in populations and gene flow between them.

Allele frequencies and F_{ST}

Comparison of allele frequencies between populations could be used to infer the history of populations. Wright's F-statistics (Wright, 1949) allows driving these estimations, in special the fixation index F_{ST} when SNP data is used (Helyar *et al.*, 2011). As described by (Holsinger and Weir, 2009) several fixation index estimators have been proposed but it has also been demonstrated that several issues can decrease the performance of estimators like ascertainment bias in the SNP arrays design or rare variants from NGS (Bhatia *et al.*, 2013). Some issues must be considered for F_{ST} estimations. If sample size in evaluated populations are different, F_{ST} proposed by Weir and Cockerham, (1984) can inflate single SNP estimates

(Willing *et al.*, 2012; Bhatia *et al.*, 2013). Reich et al. (2009) proposed a F_{ST} correction for sample size differences showing a better performance in very low sample size compare to Weir and Cockerham, (1984). Moreover, results of Bhatia et al. (2013) showed that F_{ST} estimator proposed by (Hudson *et al.*, 1992) is not sensitive to the ratio of sample sizes and does not systematically overestimate F_{ST} when pair of populations are evaluated.

The F_{ST} not only provide information about the history of populations. If selection modified the allele frequency at specific locus, the F_{ST} at that locus will be larger that the result of random genetic drift effect. Thus, a genome scan of F_{ST} value estimated for SNP can give pattern of population differentiation. It is expected than outlier F_{ST} values in the empirical distribution were affected by natural selection (Akey *et al.*, 2002). Once the outlier are detected it is possible explore the region and evaluate how looking for closely related genes and its function. This approach showed interesting signals of diversifying selection for morphological traits in European pigs (Wilkinson *et al.*, 2013) or differential allele frequencies between northern an southern Chinese pigs for SNPs close to estrogen receptor gene (ESR1) associated to litter size (Yang *et al.*, 2014). However, cautions in outlier of F_{ST} values as selection signature must be considered (Vilas *et al.*, 2012).

Test of admixture between populations

Admixture between populations has been evaluated by model-based clustering methodologies with the limitation that they only identify recent mixture between populations. Because the relations of populations could be occurred in the past, some extensions of population relationships between pairs of populations (i.e F_{ST}) have been proposed to explore ancient population mixture events.

Initially, Reich et al. (2009) suggested the *3-population* and *4-population* (*D-statistics*) tests to evaluate admixture, even if the gene flow events occurred hundreds of generations ago (Patterson et al. 2012). In particular, the *D-statistics* (Green *et al.*, 2010; Durand *et al.*, 2011) is a formal test for admixture, which can not only provide evidence for admixture but also provide some information about the directionality of the gene flow. In this test we consider 4 populations W, X, Y and Z with an expected topology (W, X) and (Y, Z). If the expected relation is not affected by unexpected higher gene flow between a population pair, allele frequencies differences between W, X are not correlated with those between Y and Z, thus *D-statistic* = D(W, X; Y, Z) = 0. This statistic is summarized across SNPs and normalized. If Statistic is significant deviated from zero, we can conclude asymmetry from the expected uncorrelated allele frequencies. If gene flow occurred between W and Y (or X and Z), the statistic is expected to be positive, whereas a negative value is associated to gene flow between W and Z (or X and Y).

D-statistic has been used to estimate admixtures in the human past (Green *et al.*, 2010) but also has been used to explore admixture in the pig genome (Groenen *et al.*, 2012; Frantz *et al.*, 2013). For instance, Groenen et al. (2012) found admixture between North Chinese and European wild boars, using *Potamochoerus Africanus* as outgroup, indicating migrations across Eurasia during later stage of Pleistocene. Additionally this test showed strong signals for admixture from Asia into European breeds likely due to importations of Chinese breeds into Europe in 18th and 19th century.

1.4 Analysis of ancient genomes

The history of pigs has been largely written from the genetic signatures observed in ancient samples, in principal using mtDNA (Larson *et al.*, 2005). Subsequent works have both corroborated the Larson et al. (2005) results and incorporate new insights in the history of pigs (Larson, Cucchi, *et al.*, 2007; Larson *et al.*, 2010; Meiri *et al.*, 2013). It is undisputable the usefulness of ancient mtDNA to explore history of populations (Thalmann *et al.*, 2013), however the information contained in this genome sometimes could be limited (Vilstrup *et al.*, 2013), especially in terms of variability and demography because e.g. the maternal inheritance. Genomic studies of nuclear ancient genomes will be useful to understand the complex relationships of pig populations (Ramírez, Ojeda, *et al.*, 2009; Frantz *et al.*, 2013). Several considerations must be done when ancient genomes are used including the reduced quality and biochemical changes post-mortem. However, the unknown patterns of population-state in the moment that sample lived supposes a challenge for population genetic estimates.

Ancient DNA quality and sequencing issues

Some features of ancient DNA (aDNA) make it difficult its analysis. First, although aDNA molecules can survive few hundred thousand years under favourable environmental conditions, they undergo fragmentation and post-mortem chemical changes (Stoneking and Krause, 2011; Orlando *et al.*, 2013). Once aDNA is recovered, it is needed to evaluate the percentage of endogenous DNA because microorganism DNA introduced during fossil deposition and collection. Moreover, human DNA can also contaminate remains especially during excavation and laboratory processing (Shapiro and Hofreiter, 2014). A method to estimate sample contamination can be addressed identifying sequences that differ between species of interest and the source of contamination (Green *et al.*, 2010). Additionally computational methods can be used to estimate contamination levels in ancient sequences (Skoglund *et al.*, 2014). After validating aDNA, low quantities of aDNA are characterized by very short length fragments (shorter than 500 bp) are usually recovered (Pääbo *et al.*, 2004). Finally, aDNA tends to be affected by chemical damages that modify the structure of DNA allowing nucleotide misincorporations during library preparation and sequencing. Cytosine deamination transform cytosine to uracil and subsequent sequencing detect thymine (Stoneking and Krause, 2011).

The very first studies in ancient samples basically performed a polymerase chain reaction (PCR) of several short and overlapping fragments, followed by production and sequencing of clones for each fragment and a final alignment and comparison of sequences from different clones in order to construct a consensus sequence (Rizzi *et al.*, 2012). This methodology has been widely used for amplification of ancient mtDNA because the several molecule copies found in aDNA remains in comparison with nuclear DNA. Traditional analyses of ancient samples have low efficiency (obtaining several thousand bp of DNA) as well as time-consuming. In this way, NGS offers several advantages to sequencing ancient samples (Stoneking and Krause, 2011) at very low cost and efficiency. First there is no targeted amplification step and aDNA is turned directly into a sequencing library by adding artificial adaptor sequences to both ends of each fragment. Conversely to modern DNA sample, aDNA

is naturally fragmented allowing an efficient PCR amplification previous high-throughput sequencing. Further, because the short fragment length of aDNA it can be sequenced completely from both strands reducing sequencing errors. NGS technology has been used in recent works to study aDNA samples of Neanderthal and other human ancestors successfully (Green *et al.*, 2010; Rasmussen *et al.*, 2010, 2014; Olalde *et al.*, 2014) as well as in ancient mammals (Miller *et al.*, 2008; Ramírez, Gigli, *et al.*, 2009; Orlando *et al.*, 2013).

Considerations in sequence assembly and variant calling

All NGS technologies provide a large amount of reads that are mapped against modern reference genome or used for *de novo* assembly in order to reconstruct a draft of ancient genome. In any case, some of aDNA features mentioned above make difficult the genome reconstruction. First, short reads must be removed to avoid either high divergent fragments or those that include adapter from library preparation (Schubert *et al.*, 2012). Validation of endogenous DNA could be done comparing reads to reference genome using a mapper software like BWA (Li and Durbin, 2009) or BLAST. We must take into consideration that variations in the reads sequence could derive from real nucleotide variation in the aDNA, contamination or sequencing error. Read mapping must attempt for estimation of quality of reads, estimation of endogenous DNA proportion and aDNA damages (Green *et al.*, 2010; Rizzi *et al.*, 2012). Schubert et al. (2012) reviewed different options for read mapping using BWA, for instance the seed region or gap opens, and demonstrated that ancient samples could not be aligned to modern reference genomes with the same efficiency and parameters. Other important consideration is the variant calling, because is highly influenced by the sample coverage (Nielsen *et al.*, 2012), which is usually low from ancient samples. Assuming high coverage, another issue arises from aDNA damage. In this case, it is highly encouraged to evaluate those regions with higher coverage and excluding transitions C-T and G-A. SNP comparisons with modern samples attempt for variant control in case of unexpected allele frequencies is observed.

The analysis of ancient genomes supposes a challenge for population genetics, but the results obtained provide a proxy of the variability and genetic status in the past, and largely contributes in the understanding of the patterns observed in extant populations.

Chapter 2

-

Objectives

-

The main objectives of this thesis are to evaluate the effects of a rapid colonization and selection for a new environment by examination of the patterns of genetic diversity, as well as understanding the evolutionary mechanisms involved in those processes. This thesis studies the pig colonization across the American continent.

To achieve this general objective, we propose the following specific objectives:

1. To identify the population structure and the relationships of pig populations (worldwide and within American continent) from genomic perspective, using samples from extant populations and evaluating autosomal and chromosome X.

2. To compare the present genome patterns of diversity in pigs with genome patterns previous to the colonization of America using an Iberian pig genome from the 16th century.

3. To detect signatures of adaptive selection across genome of pig populations.

4. To develop a methodology that improves the visualization of the relationships among populations considering unbalanced sampling of populations.

Chapter 3

-

Worldwide genetic relationships of pigs as inferred from X chromosome SNPs

-

W. Burgos-Paz[1], C.A. Souza[1,2], A. Castelló[1], A. Mercadé[1], N. Okumura[3], I.N. Sheremet'eva[4], L.S. Huang[5], I.C. Cho[6], S. R.Paiva[2], S. Ramos-Onsins[1], M. Pérez-Enciso[1,7]

Affiliations described in the manuscript

Animal Genetics (2013). 44(2): 130-138

Supplementary information:
http://onlinelibrary.wiley.com/store/10.1111/j.1365-2052.2012.02374.x/asset/supinfo/age2374-sup-0001-TableS1.pdf

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

doi: 10.1111/j.1365-2052.2012.02374.x

Worldwide genetic relationships of pigs as inferred from X chromosome SNPs

W. Burgos-Paz*, C. A. Souza*,†, A. Castelló*, A. Mercadé*, N. Okumura‡, I. N. Sheremet'eva§, L. S. Huang¶, I. C. Cho, S. R. Paiva†, S. Ramos-Onsins* and M. Pérez-Enciso*,††**

*Center for Research in Agricultural Genomics (CRAG), Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain. †EMBRAPA, Recursos Genéticos e Biotecnologia (Embrapa Genetic Resources and Biotechnology), Brasília, DF, 02372, Brasil. ‡STAFF Institute, Society for Techno-Innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, 305-0854, Japan. §Institute of Biology and Soil Science FEB RAS, 159, Stoletia str, Vladivostok, 690022, Russia. ¶Key Laboratory for Animal Biotechnology of Jiangxi province, Ministry of Agriculture of China, Jiangxi Agricultural University, Nanchang, 330045, China. **Subtropical Animal Experiment Station, National Institute of Animal Science, R.D.A., 175-6, O-Deung Dong, Jeju, 690-150, South Korea. ††Institut de Recerca i Estudis Avançats de Catalunya (ICREA), Pg. Lluis Companys 23, Barcelona, Spain

Summary

The phylogeography of the porcine X chromosome has not been studied despite the unique characteristics of this chromosome. Here, we genotyped 59 single nucleotide polymorphisms (SNPs) in 312 pigs from around the world, representing 39 domestic breeds and wild boars in 30 countries. Overall, widespread commercial breeds showed the highest heterozygosity values, followed by African and American populations. Structuring, as inferred from F_{ST} and analysis of molecular variance, was consistently larger in the non-pseudoautosomal (NPAR) than in the pseudoautosomal regions (PAR). Our results show that genetic relationships between populations can vary widely between the NPAR and the PAR, underscoring the fact that their genetic trajectories can be quite different. NPAR showed an increased commercial-like genetic component relative to the PAR, probably because human selection processes to obtain individuals with high productive parameters were mediated by introgressing boars rather than sows.

Keywords animal genetic resources, phylogeography, pig, sex chromosome, single-nucleotide polymorphism, wild boar.

Introduction

Archeological and genetic evidence shows that the pig was repeatedly domesticated from wild boars starting in the Neolithic period, initially in eastern Asia and later in other regions of Asia and of Europe (Larson *et al.* 2005; Fang & Andersson 2006). Our knowledge of the origin of the different pig populations subsequent to domestication is still incomplete. In Asia, differences between northern and southern wild boar populations have been found, and multiple ancestors for domesticated Asian pig populations have been reported (Fang *et al.* 2005; Luetkemeier *et al.* 2010). In Europe, the wild boar was domesticated in the Neolithic period in different places and eventually formed European local breeds (Giuffra *et al.* 2000; Ramirez *et al.* 2009). An introgression of Asian pigs into European breeds during the 18th and 19th centuries is well documented (Porter 1993; Giuffra *et al.* 2000). This event resulted in today's commercial European breeds, most of which are now widespread internationally. Other populations have been much less widely studied. In the Americas, the species was introduced during Columbus's second trip to the Hispaniola Island (today, Dominican Republic and Haiti) in 1493. The current main genetic component of these animals is supposed to be Iberian, with a possible subsequent introgression of other breeds. As for African populations, Ramirez *et al.* (2009) reported an Asian component in pigs in East African countries that was not observed in the western part of the continent.

Several approaches have been used to study population genetic relationships in pigs, including mtDNA sequences (Giuffra *et al.* 2000; Larson *et al.* 2005; Fang & Andersson 2006), the Y chromosome (Ramirez *et al.* 2009), and

Address for correspondence

M. Pérez-Enciso, Center for Research in Agricultural Genomics (CRAG), Universitat Autònoma de Barcelona, Bellaterra 08193, Spain.
E-mail: miguel.perez@uab.es

Accepted for publication 22 March 2012

autosomal genetic markers (Fan *et al.* 2002; Fang *et al.* 2005; SanCristobal *et al.* 2006). In contrast, X chromosome markers have been neglected so far. Yet, some characteristics make the X chromosome an interesting resource for the study of variability between populations (Schaffner 2004). Despite the structural differences between the X and Y chromosomes, they have a region that allows for segregation in meiosis, called the pseudoautosomal region (PAR). This region, located in the telomeric regions of the X chromosome arms, can recombine with the Y chromosome (Schaffner 2004). The rest of the chromosome constitutes the non-pseudoautosomal region (NPAR). Males carry only one copy of this chromosome, making its effective population size three-quarters that of the autosomes and increasing genetic drift for the NPAR-linked loci. Also, differences between effective population sizes in each gender can influence differences between populations depending on the migration sex ratio (Pool & Nielsen 2009; Casto *et al.* 2010).

The PAR X chromosome structure has been described in bovine (Das *et al.* 2009), horse (Raudsepp & Chowdhary 2008), dog and sheep (Toder *et al.* 1997). In pigs, the first PAR genes (*PRKX, KAL1,* and *STS*) were mapped to SSCXp/Yp by Quilter *et al.* (2002). Also, the porcine pseudoautosomal boundary (PAB) was recently mapped by qPCR between *SHROOM2* and *CLCN4* (Raudsepp *et al.* 2012). According to the Sscrofa9 assembly (Archibald *et al.* 2010), the *Sus scrofa* X chromosome (SSCX) is 125.8 Mb long and contains 701 coding genes, 32 pseudogenes, and 178 noncoding RNAs (www.ensembl.org).

Here, we report on the analysis of worldwide pig population relationships using SSCX SNP polymorphisms in both the NPAR and the PAR, and we refine the localization of both regions in the porcine genome.

Materials and methods

Samples

A total of 312 DNA samples from 224 males, 87 females, and one unknown-sex individual from 39 pig breeds collected in 30 countries around the world were genotyped in this study (Table S1). These breeds include local breeds and wild boars from Asia and Europe, local creole and feral pigs from the Americas, and a smaller sample from African countries. Information about breed was not available for the latter region. Within commercial breeds, we included the most widely used breeds in the modern pig industry, mostly of European origin, as well as some British local breeds that have been introgressed with Asian germplasm (Table S1).

SNP selection, genotyping, and quality control

We selected 96 candidate SSCX SNPs from sequences described in the study of Wiedmann *et al.* (2008), chosen according to their position (identified in the PAR) and to fulfill Illumina design requirements. The selected SNPs were genotyped using a Veracode Golden Gate Genotyping Assay Kit and analyzed in a Bead Xpress Reader (Illumina, Inc.). All genotypes were assigned using GENOME STUDIO software (Illumina, Inc.) and subsequently checked manually. Database pruning was conducted with PLINK (Purcell *et al.* 2007), excluding those SNPs with call frequency <90% and minimum allele frequency lower than 0.01 as well as those individuals with >10% missing genotypes. SNPs were annotated using Ensembl's biomart tool in the Sscrofa10 assembly (http://www.biomart.org/).

Statistical analysis

With the aim of localizing the PARs and NPARs, observed (*Ho*) and expected (*He*) heterozygosities were obtained for each locus within sex using ARLEQUIN 3.5 software (Excoffier & Lischer 2010). Successive analyses were conducted for the PARs and the NPARs independently.

To test the hypothesis of genetic structuring in this sample assay, populations were grouped into seven meta-populations according to genetic, historic, and geographic relationships: Asia, Asian wild boar, local European breeds (Europe), European wild boar (which includes Tunisian wild boar), Africa, and American creole and feral pigs (Ramirez *et al.* 2009). Analysis of molecular variance (AMOVA) implemented in ARLEQUIN software was used to measure the degree of structuring between and within meta-populations. In males, we converted the NPAR haploid genotypes into diploid homozygous genotypes (Casto *et al.* 2010). We also explored alternative approaches such as reconstructing female haplotypes and diploidizing each, but the results did not change; *P*-values were calculated by performing 10 000 permutations.

Two classification approaches, supervised and unsupervised, were performed to characterize genetic relationships between populations. First, clustering of populations was performed with discriminant analysis of principal components (DAPC) implemented in the ADEGENET package (Jombart 2008). Second, probabilistic assignment to *K* groups with a Bayesian method in STRUCTURE software (Pritchard *et al.* 2000) was used. In this analysis, two population structures were evaluated: the first including the seven meta-populations and the second considering the 53 populations (country–breeds combination) independently. For each population structure, 10 simulations with *K* values ranging from 2 to 12, considering an admixture model and allele frequencies correlated, were evaluated. For each simulation, a burn-in period of 50 000 iterations was followed by 500 000 final iterations. To infer the optimal *K* value, STRUCTURE results were analyzed according to the delta *K* method described by Evanno *et al.* (2005).

Results

Quality control

A total of 56 SNPs (Table S2) passed quality control tests and were included in the analysis. The ratio of successfully genotyped SNPs (56/96 = 0.58) was much lower than that expected with this technology. The main cause of SNP discard was the high number of missing values per sample, but we also found nine non-segregating SNPs. Data are available in PLINK format upon request.

Localization of the presumable PAR in SSCX

To estimate the localization of the PAR, heterozygosity for each SNP was calculated within sex. The first 33 SNPs, located between the positions of 1.51 and 6.77 Mb, showed heterozygosities higher than zero in males ($Ho = 0.21$) and comparable to those in females ($Ho = 0.20$), whereas the remaining 23 SNPs, localized between 7.86 and 116.43 Mb, showed values equal to zero. In contrast, females showed *Ho* values higher than zero for those SNPs ($Ho = 0.15$). This suggests that the PAR likely is located between 0 and 7.8 Mb in the SSCX. The PAR and NPAR were analyzed separately.

Allelic frequencies and population structure

The heterozygosities in the PAR and NPAR for each of the seven meta-populations are shown in Table 1. Overall Africa, commercial and American populations showed the highest values of *Ho* in the PAR and *He* in NPAR. However, these also were the most extensively sampled populations. In contrast, local breeds and wild boars from Europe showed the lowest *Ho* and *He*. This is in agreement with previous results (e.g., Ramirez *et al.* 2009), which also described low variability levels in European pigs, especially in wild boars. Also, it can be noticed that the NPAR heterozygosities are lower than that of the PAR, except in European local pigs. Nevertheless, the reduction in heterozygosity is lower than expected, that is, 75%.

To gain further insight into population structuring and the effects of each potential structuring level (i.e., breed, country, and chromosome region), we performed several AMOVAS (Table 2). First, we considered all breed–country populations and second, the seven meta-populations described; finally, we analyzed each meta-population individually. Irrespective of classification, the same trends were consistently observed: the F_{ST} values were higher in the NPAR than in the PAR, and the among-population variance component also was higher in the NPAR than in the PAR—even if the within-population component dominates (Table 2). Population structure analyses suggest that 29.7 and 40.5% of the genetic variance was explained by the among-population variance component in the PAR and the NPAR, respectively, when all breed–country groups were considered (Table 2). When pigs were grouped by each of the seven meta-populations, the among-population variance component decreased, explaining only 22.7% and 25.0% of the total variance in the PAR and the NPAR, respectively. This suggests that the meta-population classification was less biologically meaningful than the breed–country arrangement.

Pairwise F_{ST} values among the seven meta-populations showed high differentiation in both the PAR and the NPAR between Asian and other populations (Table 3). Furthermore, pairwise F_{ST} values suggest that each SSCX region tells different, albeit similar, stories about the relationship between populations. In agreement with previous results (Ramirez *et al.* 2009), we found very low differentiation between local European breeds and wild boars; yet again, it was higher in the NPAR than in the PAR, 0.06 vs. 0.05, respectively. In contrast, differentiation between Asian domestic and wild boars was much higher, especially in the NPAR ($F_{ST} = 0.60$). As for derived African and American populations, our analyses indicate that, on average, the closest related breeds were the commercial breeds, followed by the European local breeds. This result agrees with a primary Iberian origin of the American populations followed by an important international breed introgression that has blurred the initial Iberian origin of American creole pigs.

Table 1 Sample size (*n*), missing values (NA), and observed (*Ho*) and expected (*He*) heterozygosities for each region of SSCX.

			Pseudoautosomal regions		Non-pseudoautosomal region	
Population	*n*	%NA	*Ho*	*He*	*He*	He_{NPAR}/ He_{PAR}
Asia	6	1.48	0.17 ± 0.22	0.27 ± 0.19	0.30 ± 0.21	1.11
Asia WB[1]	37	2.99	0.14 ± 0.16	0.21 ± 0.19	0.15 ± 0.17	0.71
Europe local	19	0.75	0.14 ± 0.12	0.17 ± 0.15	0.16 ± 0.20	0.94
Europe WB	42	1.19	0.14 ± 0.17	0.16 ± 0.19	0.14 ± 0.20	0.87
Commercial	66	0.81	0.28 ± 0.11	0.36 ± 0.11	0.25 ± 0.20	0.69
Africa	28	1.53	0.26 ± 0.15	0.31 ± 0.14	0.25 ± 0.19	0.80
America	114	1.20	0.21 ± 0.13	0.25 ± 0.15	0.20 ± 0.20	0.80
Total	312	1.33	0.21 ± 0.10	0.32 ± 0.13	0.25 ± 0.18	0.78

[1]WB, wild boar.

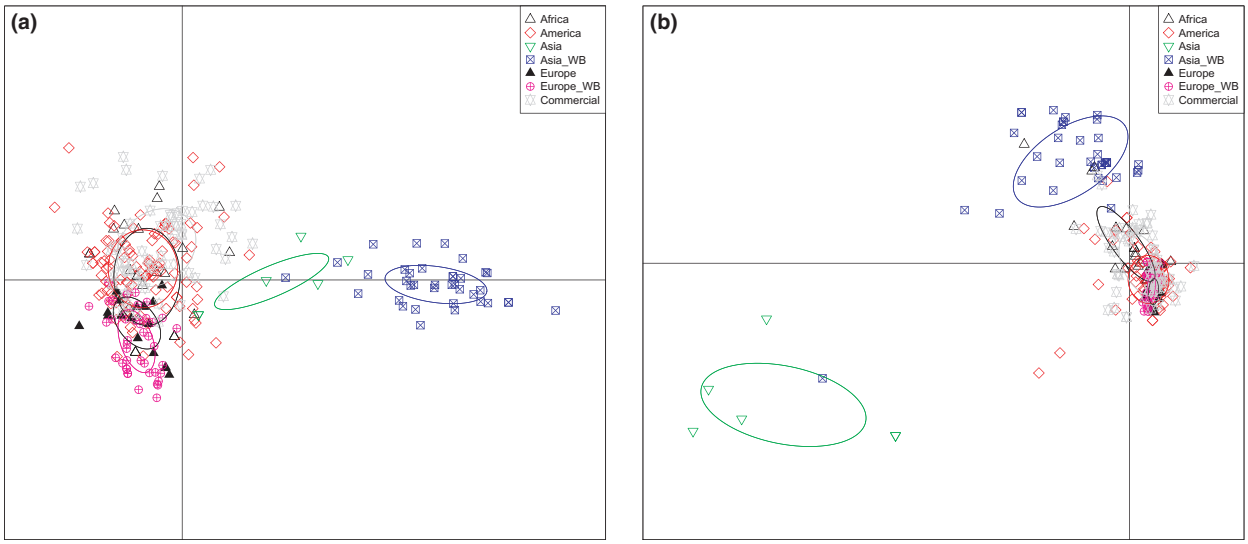


Figure 1 Scatterplot of DAPC analysis among meta-populations in the pseudoautosomal regions (a) and the non-pseudoautosomal region (b).

Probabilistic assignation to genetic groups

Differentiation of Asian populations from other groups was confirmed with DAPC in both SSCX regions (Fig. 1). Principal components retained explained 91% of the genetic variance in both the PAR and NPAR. This analysis (Fig. 1) corroborates the similarity in allelic frequencies between European local pigs and wild boars and also among the non-Asian populations. Asian local populations showed a cline-like relationship between Asian wild boars and commercial populations in the PAR (Fig. 1a). However, the discriminant analysis was more difficult to interpret regarding the NPAR (Fig. 1b). In this case, DAPC showed a higher similarity of European and commercial pigs with Asian wild boars than with Asian domestic pigs. We hypothesize that demographic effects in Asian domestication together with ascertainment bias effect and limited sampling could explain these results.

We also performed an unsupervised Bayesian probabilistic assignation of individuals to genetic groups, as implemented in STRUCTURE software. The most likely number of clusters was $K = 2$ when pigs were grouped by meta-population. In agreement with numerous previous results, Asian populations clustered in a separate group from the remaining populations (Fig. 2). Moreover, European domestic and wild boar pigs clustered in a second genetic group with a more than 97% membership proportion. In addition, commercial, African and American populations showed different degrees of admixture of European and Asian genetic components, the latter genetic component being in lower proportion.

To assess structuring in more detail, we also ran STRUCTURE with pigs grouped by breed and country. Here, $K = 3$ was the most supported value. Figure 3 shows the average cluster membership of pigs by breed and by chromosome

Table 2 AMOVA results for the different population structures in each SSCX region.

		Pseudoautosomal region			Non-pseudoautosomal region		
Group	Populations	Within populations	Among populations	F_{ST}	Within populations	Among populations	F_{ST}
Breed–country	53	3.26	1.37	0.297	1.50	1.02	0.405
Meta-populations[1,2]	7	3.89	1.14	0.227	1.99	0.66	0.250
Asia WB[3]	5	2.31	0.54	0.189	1.11	0.95	0.461
Europe	5	2.38	0.24	0.092[4]	0.92	0.66	0.417
Europe WB	8	2.33	0.42	0.154	0.99	0.52	0.345
Commercial	16	4.83	0.76	0.135	1.98	1.01	0.338
Africa	5	4.54	0.31	0.065[4]	2.35	0.11	0.046
America	13	3.01	0.60	0.167	1.79	0.57	0.242

[1]Includes the seven meta-populations Asia, Asia wild boar, Europe, Europe wild boar, Commercial, Africa, and Americas.
[2]Asia was excluded because all individuals were considered one population.
[3]WB, wild boar.
[4]No significant statistical differences were found. The rest of the F_{ST} values were highly significant.

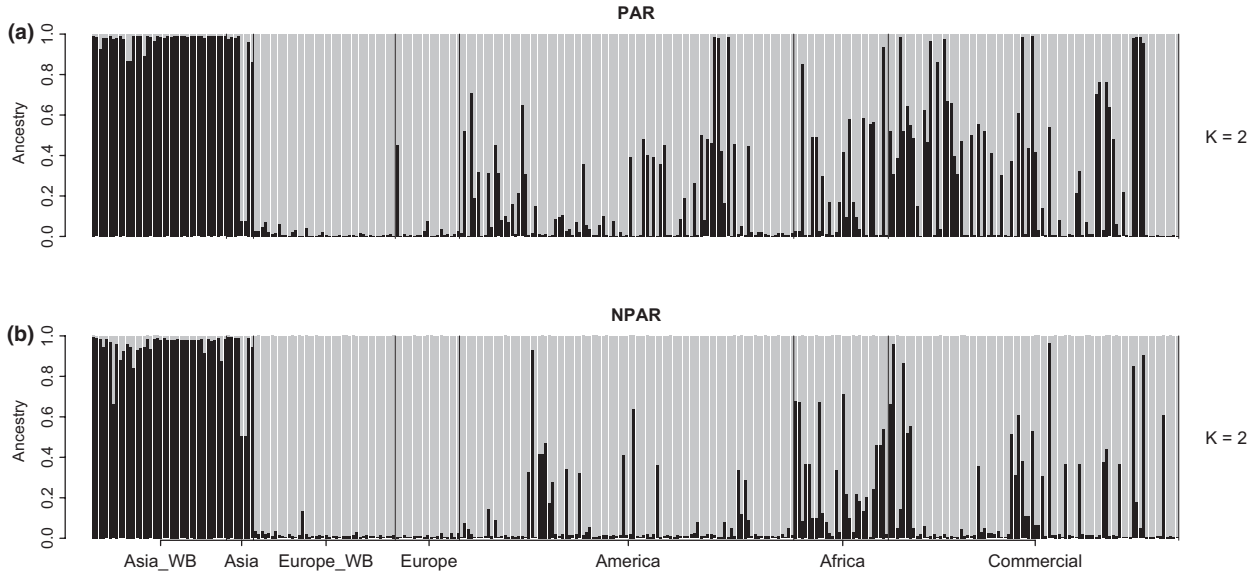


Figure 2 Bayesian probabilistic assignation of individuals considering meta-population structure in the pseudoautosomal regions (a) and the non-pseudoautosomal region (b). Breeds included in each group are described in Table S1.

region. Overall, Asian and European wild boars and Chinese pigs were rather homogeneous, especially for the PAR. Interestingly, Japanese wild boars from Ryukyu and the main islands were slightly different in the NPAR, in parallel with results from mtDNA (Wu *et al.* 2007) which suggest different founder origins. The native Korean pigs showed differences in allelic frequencies to commercial breed introgression signals. In fact, the origin of native Korean pigs is mixed, with accredited introgression of commercial pigs within the original Korean germplasm. Interestingly, the composition of the PAR and NPAR was quite distinct, with a larger proportion of European clusters in the NPAR, again in agreement with a primary European origin via males.

European populations were made up of two genetic clusters in different proportions depending on the SSCX region. In the PAR, one cluster represented 80.7% of the

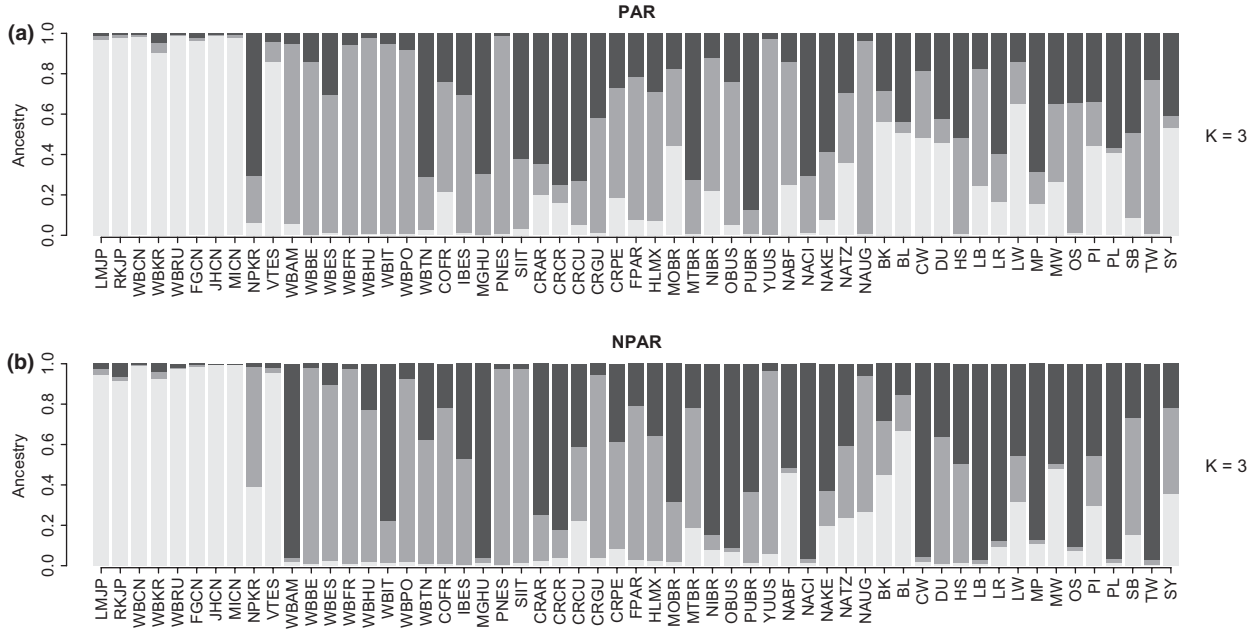


Figure 3 Average Bayesian probabilistic cluster assignation by breed and country in the pseudoautosomal regions (a) and the non-pseudoautosomal region (b). Breed and country codes as in Table S1.

Table 3 Pairwise F_{ST} values in pseudoautosomal regions (below diagonal) and non-pseudoautosomal region (above diagonal).

	Africa	America	Asia	Asia WB[1]	Europe	Europe WB	Commercial
Africa		0.10	0.52	0.27	0.16	0.21	0.02
America	0.03		0.64	0.45	0.05	0.15	0.04
Asia	0.18	0.28		0.60	0.64	0.69	0.57
Asia WB	0.44	0.50	0.22		0.51	0.53	0.34
Europe	0.05	0.03	0.36	0.58		0.06	0.07
Europe WB	0.09	0.08	0.41	0.61	0.05		0.13
Commercial	0.02	0.06	0.11	0.32	0.10	0.15	

[1]WB, wild boar.
All F_{ST} values were significant ($P < 0.05$).

genetic component in wild boars; this value decreased to 69.3% in the NPAR, influenced by high differences in allelic frequencies of Italian and Armenian individuals. Mangalitsa pigs showed allelic frequencies similar to that of Iberian pigs in the PAR, whereas in the NPAR, they were similar to Armenian pigs. As expected, commercial breeds were mosaics of European and Asian-like clusters in different degrees. However, there were differences between the PAR and NPAR, maybe as a result of differences in the sex ratio of the original breed genetic contribution to each population. For instance, in Duroc and in Pietrain, the Asian-like component decreased in the NPAR. Further, clustering was very different between the PAR and NPAR for breeds like Tamworth, Large Black, and Chester White. As in commercial breeds, three clusters with different proportions for each SSCX region were detected in America populations. Yucatan, Feral Argentinean, and Peruvian pigs were primarily European in origin, whereas an Asian-like component appeared in the NPAR for Brazilian Monteiro and Cuban pigs. In a previous work on the Y chromosome in African populations, we observed a larger Asian component in Eastern than in Western countries; however, this trend is not apparent from our results for SSCX.

Discussion

This study is a first approach to evaluate the usefulness of SNP polymorphisms in the pig X chromosome to detect population structure in a worldwide pig sample. Of 96 candidate SNPs, we retained only 56 polymorphic SNPs. These SNPs were initially obtained from pools of seven pig breeds, namely Duroc, Landrace, Yorkshire, Large White, Hampshire, Berkshire, and Pietrain, so they represent a limited variability in the whole species (Wiedmann *et al.* 2008).

Using indirect evidence of male heterozygosity, we determined that the PAR spans the first 7.8 Mb of the SSCX short arm. Quilter *et al.* (2002) mapped the *STS, KAL1,* and *PRKX* genes to the PAR using fluorescent in situ hybridization (FISH); in the Sscrofa9 assembly, the first two genes are located at 2.89 and 4.00 Mb, which does suggest a bigger PAR length in pig than in humans although smaller than in cattle (Ross *et al.* 2005; Das *et al.* 2009). Intriguingly, gene *PRKX* (ENSSSCG00000012833) is located on the Xq tail at 125.63–125.80 Mb. Further, Rohrer *et al.* (1994) reported three markers presumably located in the PAR; one of the markers, namely *SW949*, was aligned against Sscrofa9 assembly and matched two positions: 125.67 (*P*-value 1.8e−146) and 125.77 Mb (*P*-value 6.3e−144), again on the Xq tail. These data point to possible errors in the X chromosome Sscrofa9 assembly, given that a single PAR is suggested by FISH data.

The *SHROOM2* gene has been associated with PABs in pigs (Raudsepp *et al.* 2012). In the Sscrofa9 assembly, *SHROOM2* (ENSSSCG00000012104) is located at 5.26 Mb and within the bounds of the PAR suggested in this study. The *amelogenin* (*AMEL*) gene has been reported as the ancient Pseudoautosomal Boundary (PAB) in mammals (Iwase *et al.* 2003). This gene is located between 6.89 and 6.90 Mb in the pig genome (ENSSSCG00000012113), and male heterozygosity was zero in this region. Therefore, a comparative map of PAR in humans, cattle, goat, sheep, horse, and dog (Das *et al.* 2009), and the results reported here suggest that the PAR size in pigs is ~7 Mb.

The main feature of the X chromosome in population genetics studies is its low effective size, which theoretically reduces the genetic variability to three-quarters that of autosomes. Our data do show a reduction in the variability in the NPAR vs. PAR (Table 1), but this was much smaller than expected, except in commercial breeds—where the SNPs were ascertained. Although deviations from theoretical variability ratio have been reported (e.g., Gottipati *et al.* 2011), the most likely explanation for our results is SNP ascertainment bias. The analyses from the few published sequence data suggest that, in pigs, the ratio of X/A variability is actually much lower than 0.75 (Amaral *et al.* 2011; Esteve-Codina *et al.* 2011). Moreover, Ma *et al.* (2010) found a large (~31 Mb) recombination cold spot adjacent to the centromere of the pig X chromosome. Here, we found a strong reduction in *Ho* values in SNPs located between 61.52 and 92.32 Mb, in a position similar to that

in the study of Ma *et al.* Detailed analyses for this region showed a single haplotype in all samples except for two native Korean sows.

Overall, commercial populations showed the highest *Ho* and *He* values, followed by African and American populations. Given that the populations from these two continents are derived, a lower variability should be expected; these data then would suggest an important introgression of commercial breeds into the 'creole' populations. In general, *Ho* and *He* values were higher in Asian than in European populations, as reported in other studies using autosomal microsatellite markers (Fan *et al.* 2002; Luetkemeier *et al.* 2010). Moreover, as also reported by Ramirez *et al.* (2009), both domestic and European wild boar populations showed similar heterozygosity values for both SSCX regions and low differentiation, suggesting a possible gene flow between populations. In Asia, *Ho* and *He* values were higher in domestic populations than in wild boar, in contrast to results from Luetkemeier *et al.* (2010). However, these *Ho* and *He* values should be interpreted with caution, given the small sample size of some populations and differences in allele frequencies with commercial populations (ascertainment bias effect).

The X chromosome patterns of variability are strongly influenced by the sex ratio: genetic drift is very sensitive to the number of males, whereas the recombination rate depends on that of females. As revealed by STRUCTURE, a variety of different patterns between the PAR and NPAR, especially in commercial breeds, can be seen (Fig. 3). In some breeds, such as Berkshire or British Lop and most wild boars, the PAR and NPAR exhibited similar cluster compositions, but in most breeds, there were large differences between the PAR and NPAR, showing that their genetic trajectories are distinct.

Discriminant analysis of principal components and F_{ST} results showed a low differentiation between Africa, America, and commercial populations. Historical records document a strong international breed introgression in Africa, especially from the maternal side, from Asia and British breeds during the 18th and 19th centuries. Ramirez *et al.* (2009) found a primary European influence in western Africa, whereas eastern Africa exhibited an Asian genetic component. Our results cannot confirm these results; Eastern African populations (Kenya, Tanzania, and Uganda) share a large proportion of a European-like genetic group in both SSCX regions (Fig. 3). American population origins are European, mainly Iberian, with recent commercial breed introgression (Ramirez *et al.* 2009; Souza *et al.* 2009). Our results are consistent with this, showing low differentiation with the commercial population; in addition, all breeds showed a proportion of an Asian-like cluster (Fig. 3). Some populations (Yucatan, Argentinean feral pigs, and Guatemalan pigs) still preserve a high proportion of a European-like genetic cluster proportion in both regions, whereas others (Costa Rican or Argentinean creoles) show strong influences of a commercial-like genetic component. A particularly interesting case was found in the Brazilian Moura and Monteiro breeds, where the Asian-like component appears in both populations but in different SSCX regions. The origin of the Monteiro breed is possibly European but has been crossbred with Asian breeds to improve reproductive parameters (Grossi *et al.* 2006). Overall, it is difficult to quantify the introgression impact of commercial breeds in American populations.

Conclusions

Thus far, the X chromosome has been poorly studied in the pig species. In this genotyping study, our findings are in agreement with a reduction in the variability in the NPAR, although attenuated because of SNP ascertainment bias. We also find indirect evidence that the PAR comprises approximately ~7 Mb of SSCX. Two main population clusters were detected, corresponding to Asian and European origins. Nevertheless, our results show that genetic relationships between populations can vary greatly between the NPAR and the PAR, underscoring the fact that their genetic trajectories can be quite different. The NPAR showed an increased commercial-like genetic component relative to the PAR, probably because of the fact that human selection processes to obtain individuals with high productive parameters were mediated by introgressing boars rather than sows. Further studies with a much denser SNP panel should allow the detection of selective sweeps in this important chromosome.

Acknowledgements

We thank the many groups and people who helped us to collect samples, in particular the Conbiand network and Embrapa Suínos e Aves, Embrapa Pantanal, Empresa Baiana de Desenvolvimento Agrícola, Universidade Estadual do Sudoeste da Bahia, and Universidade de Brasília. Thanks also to referees for their comments. WBP is funded by COLCIENCIAS (Departamento Administrativo de Ciencia, Tecnología e Innovación, Francisco José de Caldas fellowship 497/2009, Colombia). CAS was funded by a PhD grant from CAPES (Coordenaçao de Aperfeiçoamento de Pessoal de Nível Superior, Brazil) and Universidade Catolica de Brasilia, Brazil. This work is funded by grants AGL2010-14822 (MICINN, Spain) to MPE, CGL2009-09346 (MICINN, Spain) to SERO, and Consolider project (MICINN, Spain) to Centre of Research in Agricultural Genomics.

References

Amaral A.J., Ferretti L., Megens H.J., Crooijmans R.P., Nie H., Ramos-Onsins S.E., Perez-Enciso M., Schook L.B. & Groenen M.A. (2011) Genome-wide footprints of pig domestication and selec-

tion revealed through massive parallel sequencing of pooled DNA. *PLoS ONE* **6**, e14782.

Archibald A.L., Bolund L., Churcher C. *et al.* & Swine Genome Sequencing Consortium (2010) Pig genome sequence–analysis and publication strategy. *BMC Genomics* **11**, 438.

Casto A.M., Li J.Z., Absher D., Myers R., Ramachandran S. & Feldman M.W. (2010) Characterization of X-linked SNP genotypic variation in globally distributed human populations. *Genome Biology* **11**, R10.

Das P.J., Chowdhary B.P. & Raudsepp T. (2009) Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions. *Cytogenetic and Genome Research* **126**, 139–147.

Esteve-Codina A., Kofler R., Himmelbauer H., Ferretti L., Vivancos A.P., Groenen M.A., Folch J.M., Rodriguez M.C. & Perez-Enciso M. (2011) Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. *Heredity* **107**, 256–264.

Evanno G., Regnaut S. & Goudet J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology* **14**, 2611–2620.

Excoffier L. & Lischer H.E. (2010) ARLEQUIN suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564–567.

Fan B., Wang Z.G., Li Y.J. *et al.* (2002) Genetic variation analysis within and among Chinese indigenous swine populations using microsatellite markers. *Animal Genetics* **33**, 422–427.

Fang M. & Andersson L. (2006) Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. *Proceedings of the Royal Society. B, Biological Sciences* **273**, 1803–1810.

Fang M., Hu X., Jiang T., Braunschweig M., Hu L., Du Z., Feng J., Zhang Q., Wu C. & Li N. (2005) The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Animal Genetics* **36**, 7–13.

Giuffra E., Kijas J.M., Amarger V., Carlborg O., Jeon J.T. & Andersson L. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785–1791.

Gottipati S., Arbiza L., Siepel A., Clark A.G. & Keinan A. (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nature Genetics* **43**, 741–743.

Grossi S.F., Lui J.F., Garcia J.E. & Meirelles F.V. (2006) Genetic diversity in wild (*Sus scrofa scrofa*) and domestic (*Sus scrofa domestica*) pigs and their hybrids based on polymorphism of a fragment of the D-loop region in the mitochondrial DNA. *Genetics and Molecular Research* **5**, 564–568.

Iwase M., Satta Y., Hirai Y., Hirai H., Imai H. & Takahata N. (2003) The *amelogenin* loci span an ancient pseudoautosomal boundary in diverse mammalian species. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5258–5263.

Jombart T. (2008) ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405.

Larson G., Dobney K., Albarella U. *et al.* (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**, 1618–1621.

Luetkemeier E.S., Sodhi M., Schook L.B. & Malhi R.S. (2010) Multiple Asian pig origins revealed through genomic analyses. *Molecular Phylogenetics and Evolution* **54**, 680–686.

Ma J., Iannuccelli N., Duan Y., Huang W., Guo B., Riquet J., Huang L. & Milan D. (2010) Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. *BMC Genomics* **11**, 159.

Pool J.E. & Nielsen R. (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719.

Porter V. (1993) Pigs. A Handbook to the Breeds of the World. Helm Information

Pritchard J.K., Stephens M. & Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Purcell S., Neale B., Todd-Brown K. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575.

Quilter C.R., Blott S.C., Mileham A.J., Affara N.A., Sargent C.A. & Griffin D.K. (2002) A mapping and evolutionary study of porcine sex chromosome genes. *Mammalian Genome* **13**, 588–594.

Ramirez O., Ojeda A., Tomas A. *et al.* (2009) Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Molecular Biology and Evolution* **26**, 2061–2072.

Raudsepp T. & Chowdhary B.P. (2008) The horse pseudoautosomal region (PAR): characterization and comparison with the human, chimp and mouse PARs. *Cytogenetic and Genome Research* **121**, 102–109.

Raudsepp T., Das P.J., Avila F. & Chowdhary B.P. (2012) The pseudoautosomal region and sex chromosome aneuploidies in domestic species. *Sexual Development* **6**, 72–83.

Rohrer G.A., Alexander L.J., Keele J.W., Smith T.P. & Beattie C.W. (1994) A microsatellite linkage map of the porcine genome. *Genetics* **136**, 231–245.

Ross M.T., Grafham D.V., Coffey A.J. *et al.* (2005) The DNA sequence of the human X chromosome. *Nature* **434**, 325–337.

SanCristobal M., Chevalet C., Haley C.S. *et al.* (2006) Genetic diversity within and between European pig breeds using microsatellite markers. *Animal Genetics* **37**, 189–198.

Schaffner S.F. (2004) The X chromosome in population genetics. *Nature Reviews Genetics* **5**, 43–51.

Souza C.A., Paiva S.R., Pereira R.W., Guimaraes S.E., Dutra W.M. Jr, Murata L.S. & Mariante A.S. (2009) Iberian origin of Brazilian local pig breeds based on *Cytochrome b* (*MT-CYB*) sequence. *Animal Genetics* **40**, 759–762.

Toder R., Glaser B., Schiebel K., Wilcox S.A., Rappold G., Graves J. A. & Schempp W. (1997) Genes located in and near the human pseudoautosomal region are located in the X-Y pairing region in dog and sheep. *Chromosome Research* **5**, 301–306.

Wiedmann R.T., Smith T.P. & Nonneman D.J. (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics* **9**, 81.

Wu G.S., Yao Y.G., Qu K.X., Ding Z.L., Li H., Palanichamy M.G., Duan Z.Y., Li N., Chen Y.S. & Zhang Y.P. (2007) Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome Biology* **8**, R245.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Breed codes and ISO country codes employed and breed x country sample size.

Table S2 Submission name, dbSNP accession number and position on X chromosome for each SNP.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Chapter 4

-

Porcine colonization of the Americas: A 60k SNP story

-

W. Burgos-Paz[1*], C.A. Souza[1,2*], H.J. Megens[3], Y. Ramayo-Caldas[1,4], M. Melo[5], C. Lemús-Flores[6], E. Caal[7], H.W. Soto[8], R. Martínez[9], L.A. Álvarez[10], L. Aguirre[11], V. Iñiguez[12], M.A. Revidatti[13], O.R. Martínez-López[14], S. Llambi[15], A. Esteve-Codina[1], M.C. Rodríguez[16], R.P.M.A. Crooijmans[3], S.R. Paiva[2], L.B. Schook[17], M.A.M. Groenen[3], M. Pérez-Enciso[1,18]

Affiliations described in the manuscript

Heredity (2013). 110(4): 321-330.

Supplementary information:
http://www.nature.com/hdy/journal/v110/n4/extref/hdy2012109x1.pdf

Heredity

Heredity (2013) 110, 321–330

www.nature.com/hdy

ORIGINAL ARTICLE

Porcine colonization of the Americas: a 60k SNP story

W Burgos-Paz[1,19], CA Souza[1,2,19], HJ Megens[3], Y Ramayo-Caldas[1,4], M Melo[5], C Lemús-Flores[6], E Caal[7], HW Soto[8], R Martínez[9], LA Álvarez[10], L Aguirre[11], V Iñiguez[12], MA Revidatti[13], OR Martínez-López[14], S Llambi[15], A Esteve-Codina[1], MC Rodríguez[16], RPMA Crooijmans[3], SR Paiva[2], LB Schook[17], MAM Groenen[3] and M Pérez-Enciso[1,18]

The pig, *Sus scrofa*, is a foreign species to the American continent. Although pigs originally introduced in the Americas should be related to those from the Iberian Peninsula and Canary islands, the phylogeny of current creole pigs that now populate the continent is likely to be very complex. Because of the extreme climates that America harbors, these populations also provide a unique example of a fast evolutionary phenomenon of adaptation. Here, we provide a genome wide study of these issues by genotyping, with a 60k SNP chip, 206 village pigs sampled across 14 countries and 183 pigs from outgroup breeds that are potential founders of the American populations, including wild boar, Iberian, international and Chinese breeds. Results show that American village pigs are primarily of European ancestry, although the observed genetic landscape is that of a complex conglomerate. There was no correlation between genetic and geographical distances, neither continent wide nor when analyzing specific areas. Most populations showed a clear admixed structure where the Iberian pig was not necessarily the main component, illustrating how international breeds, but also Chinese pigs, have contributed to extant genetic composition of American village pigs. We also observe that many genes related to the cardiovascular system show an increased differentiation between altiplano and genetically related pigs living near sea level.
Heredity (2013) **110,** 321–330; doi:10.1038/hdy.2012.109; published online 19 December 2012

Keywords: pig; adaptation; Americas; phylogeography; altitude; selection

INTRODUCTION

The pig, *Sus scrofa*, originated in Southeast Asia *ca* 5.3 −3.5 MYA (Groenen *et al.*, 2012) the species subsequently colonized the rest of Eurasia and North Africa (Larson *et al.*, 2005) but was absent from America before European colonization. Pigs, together with other livestock species like sheep, cattle or goats, were first introduced by Spaniards and Portuguese from the very beginning of colonization. Actually, the first recorded event of pig import into the new continent dates as early as the second Columbus trip (Crossby, 2003). On the Portuguese side, the first historical evidence of pig introduction dates from 1532 by Martim Afonso de Souza (Mariante and Cavalcante, 2006). According to Crossby (2003), 'the pigs adapted the quickest to the Caribbean environment', and the relevance of the pig as a source of meat from the very early days of conquest is well acknowledged (Elliot, 2007).

Nowadays, the porcine species is made up of a few highly specialized and widespread internationally breeds, well known for their leanness and high fertility. Although these international pig breeds have been replacing or intermixing with local American populations, numerous populations of direct descent from Iberian populations, so called 'creole', still are reported to exist. Currently, village pigs with a putative Iberian ancestry are common among many rural communities in most American countries. These animals are important to local communities not only because they provide food, but also because they are used as savings: they are sold when cash is needed. Normally, village pigs behave as commensal animals and feralization is also common, either because animals escape or because some areas were repopulated on purpose. This has occurred, for example, for hunting purposes in the USA, Argentina or South Brazil (Merino and Carpinetti, 2003). Therefore, although the original pigs introduced in the Americas should have been related to Iberian pigs and in particular to those of the Canary islands, the phylogeny and phylogeography of extant village and creole pigs that now populate the continent is likely to be very complex.

The study of village pigs is not only relevant from a social or historical perspective. America harbors a wide diversity of environments ranging from hot tropical climates to altitude (*altiplano*) dry

[1]Centre for Research in Agricultural Genomics (CRAG)—Universitat Autònoma de Barcelona (UAB), Bellaterra, Spain; [2]Embrapa Recursos Genèticos e Biotecnologia—CENARGEN, Brasilia DF, Brazil; [3]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands; [4]Facultad de Medicina Veterinaria, Universidad de Granma, Bayamo, Cuba; [5]Facultad de Medicina Veterinaria y Zootecnia, Universidad Nacional del Altiplano, Puno, Peru; [6]Unidad Académica de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Nayarit, Tepic, Mexico; [7]Colegio de Médicos Veterinarios y Zootecnistas de Guatemala, Guatemala; [8]Escuela de Zootecnia, Universidad de Costa Rica, San Pedro, Costa Rica; [9]Corporación Colombiana De Investigación Agropecuaria (Corpoica), Centro de investigaciones Tibaitatá, Bogotá, Colombia; [10]Departamento de Ciencia Animal, Universidad Nacional de Colombia, Palmira, Colombia; [11]Centro de Biotecnologia Reproductiva Animal, CEBIREA-Universidad Nacional de Loja, Ecuador; [12]Instituto de Biologia Molecular y Biotecnologia, Universidad Mayor de San Andrés, La Paz, Bolivia; [13]Universidad Nacional del Nordeste, Corrientes, Argentina; [14]Centro Multidisciplinario de Investigaciones Científicas y Tecnológicas, Universidad Nacional de Asunción, Asunción, Paraguay; [15]Facultad de Veterinaria, Universidad de la República, Montevideo, Uruguay; [16]Departamento de Mejora Genética Animal, INIA, Madrid, Spain; [17]University of Illinois, Urbana-Champaign, Urbana-Champaign, IL, USA and [18]Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona, Spain
[19]These authors contributed equally to this work.
Correspondence: Dr M Pérez-Enciso, Centre for Research in Agricultural Genomics (CRAG), Universitat Autònoma de Barcelona (UAB), Bellaterra, Barcelona 08193, Spain.
E-mail: miguel.perez@uab.es
Received 2 July 2012; revised 10 October 2012; accepted 12 November 2012; published online 19 December 2012

climates. Pigs thrive in all these areas, except in the very dry ones, resulting in animals adapted to extreme environments, quite distinct from those of temperate Europe. On a long, evolutionary scale, adaptation is usually characterized by an accelerated rate of non synonymous changes in protein coding regions, or in regulatory regions. Nevertheless, adaptation in the context of domestic species must stand primarily on standing variants, because of the short period of time considered. Pigs were brought into the Americas a few hundred years ago, a very short time on an evolutionary scale. Despite this, dramatic phenotypic changes have occurred. For instance, feral pigs develop much larger resistance to parasites or lack of food than pigs from international highly productive breeds. Some environments like high altitude in the *altiplano* or extreme and continuous heat in Cuba or North East Brazil also poses serious physiological challenges. The fact that adaptation must have occurred in a short time span suggests that rapid changes in allelic frequencies must have occurred, and also that excess of differentiation (for example, F_{ST}) can be a good proxy to detect these events (Akey *et al.*, 2002; Vaysse *et al.*, 2012).

Although some studies of American local pigs (Ramirez *et al.*, 2009; Souza *et al.*, 2009) or in other species like creole cattle (Delgado *et al.*, 2011; Gautier and Naves, 2011) have been reported, they concern a small number of populations and/ or a few markers. In this work, by using a 60k single-nucleotide polymorphism (SNP) chip (Ramos *et al.*, 2009), we provide the first comprehensive genomic analysis of village pigs from a wide sample of American countries, ranging from Cuba to North Argentina. This work was motivated by our interest in answering the following broad questions: (1) What is the origin of American village pig populations and their structure? Although admixing has certainly occurred, it is important to quantify its extent, for example, how much fraction of Iberian germplasm still exists, if any? (2) Is there any relationship between geographic and genetic distance, at least in the most isolated areas where admixing with modern breeds is likely to be rare? (3) And last but not least: is there any signal affecting the distribution of genotypic frequencies as a result of adaptation to extreme environments? All these questions bear relevance to both genetic and historical issues, and answering them will improve our understanding of how organisms adapt rapidly to extreme environments.

MATERIALS AND METHODS

Samples

We focused on sampling village pigs, for example, pigs living in a feral or semi-feral status from rural communities or assigned a 'creole' status, that is, thought to be of Iberian ancestry (Elliot, 2007). Sampling of relatives, for example, sibs, and animals showing evidence of intercrossing with international breeds was avoided. Our results showed that this was not always accomplished, as discussed below. A total of 206 animals from 14 countries were genotyped: the USA, Cuba, Guadeloupe, Mexico, Guatemala, Costa Rica, Colombia, Ecuador, Peru, Bolivia, Paraguay, Uruguay, Argentina and Brazil. These animals showed a wide variety of phenotypes, they lived outdoors, often in extreme climates and environments (Table 1, samples are described with more detail in Supplementary File 1).

Genotypes were also used from a wide hapmap catalog that are either potential founders of the American populations or outgroups (Table 1). These included local Mediterranean pigs from Spain (Iberian and Canary Islands pigs), Portugal (Bisaro) and Sicily (*Nero Siciliano*), international breeds (Duroc, Landrace, Large White, Hampshire) plus four breeds from East China, the most likely origin of pigs exported to other continents: Meishan, Jiangquhai, Jinhua and Xiang pig. Chinese pigs were genotyped because of the accredited partial Asian ancestry of international breeds and to assess whether there is any evidence of direct introgression of Chinese germplasm into the Americas. Finally, we genotyped Western wild boars.

Table 1 Pigs genotyped in this study

Country	*Population/ breed*	*Location*	*Code*	N (n)
Village pigs				
USA	Ossabaw pig	Ossabaw island	USOB	7
	Yucatan	Indiana	USYU	10
	Guinea hog	Several locations	USGH	15
Mexico (MX)	Cuino	Nayarit	MXCU	7
	Hairless	Several locations	MXHL	9
Cuba (CU)	Creole	Pinar del Río (West)	CUWE	5
		Sancti Spiritus (Center)	CUCE	1
		Granma (East)	CUEA	12
Guadeloupe (GP)	Creole	Guadeloupe	GPCR	4
Guatemala (GU)	Creole	Baja Verapaz, Salamá	GUCR	14
Costa Rica (CR)	Creole	Guanacaste, Alajuela	CRCR	12
Colombia (CO)	Zungo	Cereté (Córdoba)	COZU	10
	Creole	Alto Baudó (Chocó)	COCR	11
Ecuador (EC)	Creole	Loja	ECCR	5 (1)
Peru (PE)	Creole	Titicaca area	PECR	16
Brazil (BR)	Moura	Concórdia	BRMO	9
	Monteiro	Poconé	BRMT	10
	Piau	Bahia	BRPU	9
	Nilo	Goias	BRNI	2
Bolivia (BO)	Creole	Oruro	BOCR	6 (3)
Paraguay (PY)	Feral pig	San Pedro	PYFP	3 (3)
Argentina (AR)	Creole	Misiones	ARMS	9
	Feral pig	Esteros del Iberá	ARFP	6
	Semi feral	Formosa	ARFO	10
	Creole	Salta	ARNW	3 (3)
Uruguay (UY)	Cerdo pampa	Rocha	UYCP	1 (1)
Outgroup pigs				
Spain	Iberian	Several locations	ESIB	16
	Canarian	Canary islands	ESCN	4
Portugal	Bisaro	Several locations	PTBI	14
Italy	Black sicilian	Sicily	ITSI	4
Poland, Hungary, Tunisia	Wild boar	Several locations	WB	13
Denmark, Holland, USA	Duroc	Several locations	DU	20
Denmark, Holland, USA	Landrace	Several locations	LR	20
Denmark, Holland, USA	Landrace	Several locations	LW	20
UK, USA	Hampshire	Several locations	HS	14
China	Jiangquhai	Jiangsu	JQ	11
China	Jinhua	Zheijiang	JH	17
China	Xiang pig	Guizhou	XP	13
China	Meishan	Jiangsu	MS	17

N = total sample size; n = number of samples with a high percentage of missing values (<20%) and removed from F_{ST} and ADMIXTURE analyses.

Genotyping and quality control

Samples were genotyped with the Illumina's porcine SNP60 BeadChip (Ramos *et al.*, 2009). Raw data were visualized and analyzed with the Genome Studio software (Illumina, San Diego, CA, USA). Among the 62 163 SNPs initially present on the chip, 46 259 were finally selected using PLINK (Purcell *et al.*, 2007) by pruning monomorphic SNPs or SNPs with an allele frequency below 0.05, SNPs located on the sex chromosomes, SNPs with more than 5% missing genotypes, SNPs not mapped on the Sscrofa10.2 assembly or SNPs for which the ancestral allele could not be identified. The ancestral allele was estimated

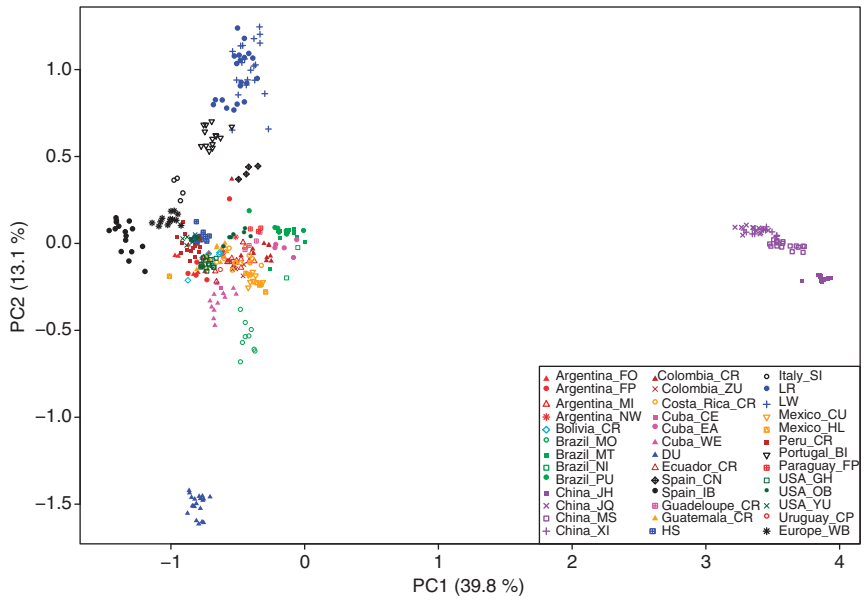


Figure 1 Principal component analysis using all samples.

based on *S. verrucosus* genotypes (Groenen *et al.*, 2012). Raw data had high-genotyping quality (call rate > 0.95) except for a few samples from Paraguay, Bolivia and Uruguay that were retained for their interest but not used in all analyses. Specifically, they were removed from the admixture and F_{ST} analyses.

Analysis

To visualize genetic distances between populations, principal component analyses (PCA) were obtained with smartpca program from EIGENSOFT (Price *et al.*, 2006). A complete relationship between individuals was drawn via a Neighbor Joining algorithm and visualized with DENDROSCOPE v. 2.7.4 software (Huson *et al.*, 2007) using pairwise identity-by-state genetic matrix distance (1-IBS) obtained with PLINK v. 1.07. To examine potential origins of each population, the Maximum Likelihood approach implemented in ADMIXTURE v 1.20 (Alexander *et al.*, 2009) was employed. First, ADMIXTURE was run in an unsupervised manner with a variable number of clusters $K = 2$–20. Lowest 10-fold cross-validation values were used to choose an optimum *K*-value, as suggested by the authors. Default termination criteria were used. We also considered a partial supervised approach where some samples were assumed to be of known ancestry K. Both PCA and ADMIXTURE were run by pruning markers in high linkage disequilibrium using the option –indep in PLINK. A total of 18 499 markers were selected for these analyses. To determine the relation between genetic and geographical distances of American pig populations, Mantel tests were performed in ADEGENET R package v. 1.3 −1 (Jombart, 2008) using exact sampling-site GPS coordinates and 1-IBS genetic distance matrix. The genetic differentiation between populations was assessed by the F_{ST} fixation index. Following Akey *et al.* (2010), we also considered a standardized F_{ST} measure. For each SNP and population k, we computed

$$d_k = \sum_{j \neq k} \frac{F_{ST}^{kj} - E[F_{ST}^{kj}]}{\mathrm{sd}[F_{ST}^{kj}]}, \qquad [1]$$

where $E[F_{ST}^{kj}]$ and $\mathrm{sd}[F_{ST}^{kj}]$ denote the average value and s.d. of F_{ST} between populations *k* and *j*, respectively, over all SNPs. Statistics *d* was obtained either summing across all pairs of populations, that is, a global measure of differentiation, or between population *k* and their three nearest populations in terms of lowest F_{ST}. This latter statistics is similar to that proposed by Yi *et al.* (2010), and should be more powerful to identify selection than is Akey's statistics (Equation (1)) as it provides a direction to the allele frequency trajectory and reduces noise relative to the global test, where all population pairs are averaged. All populations with $N > 4$ were analyzed individually. Finally, some groups of populations, namely American populations—excluding Brazil—vs European and international populations were also evaluated. In this case, we used Equation (1) as

$$d = \sum_k \sum_j \frac{F_{ST}^{kj} - E[F_{ST}^{kj}]}{\mathrm{sd}[F_{ST}^{kj}]},$$

where subscripts *k* and *j* refer to populations in groups 1 (for example, America) and 2 (for example, Europe). The average *d* statistics over SNPs in non-overlapping windows of 1 Mb were plotted. Windows with an average *d* value above 2.0 s.d. (empirical distribution corresponding to the 1% extreme windows) in each population containing at least five SNPs were considered as candidate regions for selection. To complement the differentiation analyses, we also applied a selection test based on homozygosity extent (iHS). In this case, haplotypes were inferred with fastPHASE v. 1.4.0 (Scheet and Stephens, 2006) using subpopulation label information. Haplotype frequencies were then used to evaluate the presence of selective patterns for each SNP across the pig genome as described (Voight *et al.* (2006), and inferred using the rehh R-package v. 1.0 (Gautier and Vitalis, 2012). The 1 Mb windows with extreme average |iHS| scores across SNPs in that window were retained for further analysis.

Gene annotations within candidate regions were obtained by using the preliminary annotation of assembly 10.2 provided by ensembl (Groenen *et al.*, 2012). Overrepresentation of GO categories was determined with the DAVID database (Huang *et al.*, 2009), and pathway analyses were carried out with IPA, the ingenuity system (www.ingenuity.com).

Simulations

Given the difficulty of interpreting some of the results because of SNP ascertainment bias in the chip, we used coalescence simulation under a simplified model. We assumed four populations (Asia, International, Iberian and Creole, Supplementary File 2). Asian pigs diverged from European pigs 1 MYA (assuming one generation every two years), European pigs split into International and Iberian pigs ~ 500 years ago. Both Iberian and International pigs contributed to creole pigs in approximately equal proportions, international pigs were introgressed with Chinese pigs (10%), whereas Iberian remained isolated. We studied variable Chinese contribution to creole pigs: 0, 1 and 10%. We ran coalescence simulations with mlcoalsim v. 1.9 (Ramos-Onsins and Mitchell-Olds, 2007). Out of the 10 000 independent loci simulated, we randomly selected 1000 such that the frequency spectrum in the International population was approximately flat, as observed in our data, in order to mimic ascertainment bias. Unsupervised and partially supervised

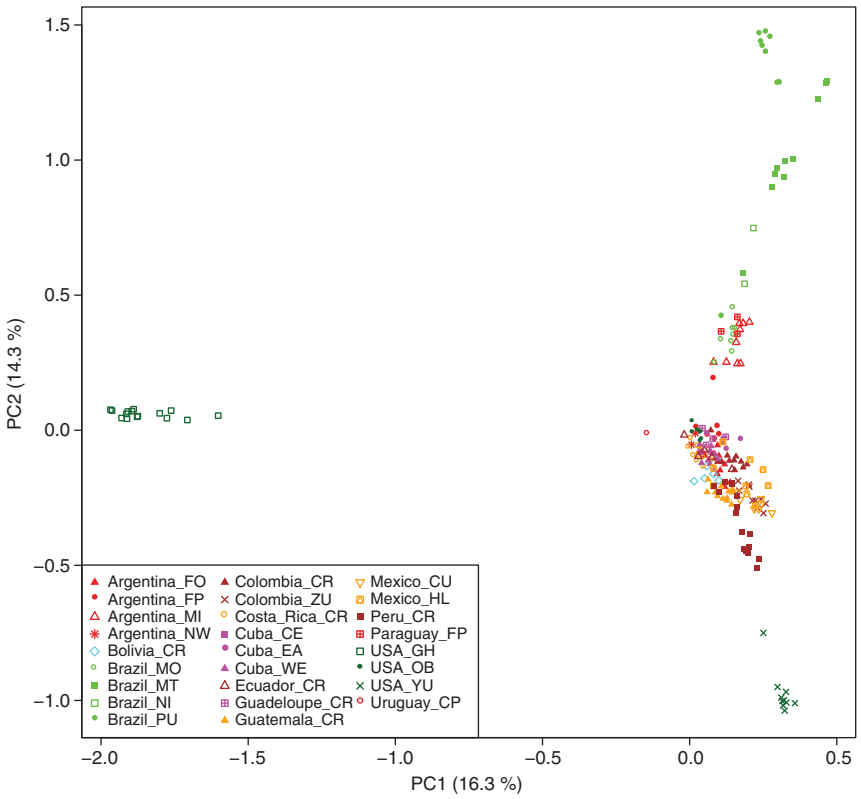


Figure 2 Principal component analysis using American samples only.

ADMIXTURE ($K=3$) was applied to the simulated data, and we evaluated the bias in estimating the Chinese contribution.

RESULTS

A wide continent with shrunken genetic variation

We know from historical and genetic evidence that American pigs descend primarily from European pigs (Ramirez *et al.*, 2009; Souza *et al.*, 2009). The original flow began with pigs from the Iberian Peninsula and the Canary Islands, followed by a more recent intercrossing with international breeds. The PC analysis (Figure 1) partially agrees with this initial hypothesis. The first axis explains ~40% of total variance and is predominantly geographical: It reflects the dramatic genetic distance between Asian and European populations. Chinese breeds and the Mediterranean Iberian breed represent both extremes on this axis. Large White and Landrace, international breeds known to have been introgressed with Chinese pigs, lie closer to Asia than do the Iberian pigs or European wild boars, which have remained isolated and unmixed with Asian germplasm. Nevertheless, these international breeds fall clearly within the 'European' neighborhood. Some Iberian pigs seem to be outliers. Although there is good evidence of sub-structuring among Iberian pigs (Alves *et al.*, 2006), we show later that this is caused by introgression from Duroc. Quite interestingly, the second axis, explaining a much lower fraction of variance (13%), primarily reflects the effects of artificial selection, with Landrace/ Large White vs Duroc breed representing the two extremes of the axes. The Iberian pig, an unimproved breed, lies broadly at the same level as wild boar on the second axis. This great distance between Duroc and other international or Mediterranean breeds is somewhat unexpected, as the original Duroc-Jersey breed was created in the USA with pigs of several ancestries, including Iberian and African animals (Porter, 1993).

As for the American populations, these lie in a relatively wide area in between Iberian, Bisaro, Canary, Landrace and Large White breeds, a symptom of their predominant European descent. American pigs, nonetheless, do form a complex conglomerate of their own that is both explained by both PCA axes, the likely contribution of Iberian)) but also of Duroc, Landrace and Large White (the second axis). Therefore, American populations are clearly admixed. Interestingly, some American populations, like Brazilian Piau or Monteiro or East Cuban pigs, are closer to the Chinese cluster than other American populations. Similarly, Brazilian Moura is closer to Duroc than the rest of the American populations (See also Supplementary Files 3 and 4). An interesting observation is that Portuguese Bisaro and Canarian pigs cluster distantly from Spanish Iberian pigs, despite being from the same geographical or national origin. The traditional view (Porter, 1993) of porcine phylogeography is the presence of two main clades among European pigs: the Mediterranean clade represented, for example, by Iberian pigs, and the Celtic clade from Northern areas, represented by Landrace or Bisaro. Nevertheless, original Canarian pigs should not cluster with these Celtic breeds because they are supposed to represent primigenious pigs, maybe with African ancestry. We hypothesize that the modern Canary pigs we genotyped here are actually introgressed with international and/or Asian breeds. This interpretation agrees with historical records (García-Dory *et al.*, 1990) as well as with the report of Asian lineages in the mitochondrial DNA of Canary pigs (Clop *et al.*, 2004). It is plausible that Asian germplasm was introduced into Canary pigs by the British, who were influential in Canarian agriculture development during late nineteenth century (García and Capote, 1982).

Next, to gain in refinement and to focus on the main goal of this work, the PC analysis was run with American village pigs only (Figure 2). From a strict American point of view, the extreme breeds are Guinea Hog, Yucatan and Brazilian Piau. Our data support a distinct origin of Guinea Hogs from the rest of village pigs in the Americas and from either Yucatan or Ossabaw pigs. In terms of F_{ST}, the closest populations to Guinea Hog were Costa Rican and Formosa (Argentina) pigs, although both relatively high: 0.13 and 0.14, respectively. A point worth mentioning is that Ossabaw and Yucatan pigs were clearly differentiated (average $F_{ST}=0.16$), despite an assumed shared Iberian ancestry. Yucatan was the closest breed to Spanish Iberian, whereas Ossabaw clustered among other American village pigs, and was in the same clade as Guadelupe pigs in the dendrogramme (Supplementary File 3). Our genotypic data support a clear separation between these breeds.

But perhaps the most noticeable observation from Figure 2 is that the second axis separates Brazilian from the rest of American pigs, with the exception of Moura. Although this partitioning is also seen in Figure 1, it is not so evident when all breeds are analyzed jointly. There also exists variability within Brazilian populations though. Piau was the most distantly related population to the rest of American village pigs, whereas Moura was the closest to, for example, Paraguayan feral pigs or Argentinean Misiones. Although it is tempting to interpret this as two separate routes of colonization, the Portuguese and the Spanish routes, this is not the sole explanation. We shall return to this point later.

A complementary view to that of the PCA is the dendrogramme pictured in Supplementary File 3. Although most pigs from the same population or breed tend to cluster together, exceptions are an Ossabaw pig within the Duroc clade or a Costa Rican pig mixed among Large Whites, both of these are probably recent admixtures with these international breeds. These animals, together with two outlier Iberian pigs, were removed to compute F_{ST} analyses. An interesting outlier is MXHL0140. This is a hairless Mexican pig from Veracruz province that clusters with Yucatan pigs, instead of with the rest of hairless pigs, which are positioned near the Duroc clade. Given the Mexican origins of Yucatan pigs, a plausible explanation is that this pig is actually a survivor of the ancient Mexican pigs currently perpetuated by US Yucatans, whereas extant Mexican 'traditional' breeds have been crossed with Duroc or other alien breeds. The results shown in Table 2, discussed below, suggest that the main source of introgression in Mexican pigs has been the Duroc breed.

Table 2 Predicted cluster composition using partly supervised ADMIXTURE ($K=6$)

Population code	*IB*	*LR*	*LW*	*DU*	*HS*	*CN*
USGH	0.00	0.00	0.00	0.00	1.00	0.00
USOB	0.37	0.20	0.11	0.12	0.13	0.07
USYU	0.99	0.00	0.00	0.00	0.01	0.01
CUCE	0.32	0.09	0.07	0.25	0.22	0.05
CUEA	0.41	0.13	0.09	0.13	0.08	0.16
CUWE	0.36	0.13	0.03	0.29	0.16	0.04
GPCR	0.36	0.09	0.23	0.16	0.07	0.09
MXHL	0.49	0.06	0.07	0.20	0.07	0.11
MXCU	0.52	0.00	0.12	0.18	0.06	0.11
GUCR	0.60	0.10	0.03	0.11	0.10	0.05
CRCR	0.36	0.12	0.13	0.19	0.14	0.07
COCR	0.50	0.13	0.09	0.11	0.06	0.12
COZU	0.72	0.02	0.02	0.10	0.04	0.11
ECCR	0.44	0.11	0.09	0.19	0.12	0.04
PECR	0.67	0.06	0.11	0.11	0.04	0.02
BOCR	0.50	0.07	0.10	0.18	0.14	0.01
ARFP	0.47	0.18	0.07	0.20	0.07	0.02
ARFO	0.56	0.11	0.07	0.16	0.07	0.02
ARMS	0.34	0.37	0.02	0.16	0.05	0.06
BRMT	0.18	0.69	0.00	0.03	0.00	0.09
BRMO	0.13	0.30	0.02	0.45	0.04	0.05
BRNI	0.16	0.55	0.00	0.17	0.01	0.12
BRPU	0.02	0.93	0.01	0.02	0.00	0.02
Average	0.41	0.19	0.06	0.15	0.12	0.06

Abbreviations: CN, China; DU, Duroc; HS, Hampshire; IB, Iberian; LR, Landrace; LW, Large White.
Population codes as in Table 1.

Geography and genetic structuring

Neither PC analysis nor dendrogrammes (Figures 1 and 2, Supplementary File 3) reveal any broad clustering by geographic origin. For instance, Peruvian populations were positioned between Yucatan and Guatemalan pigs. Northeast Argentinean and Cuban pigs were scattered among other geographically distant pigs. In some cases, though, geography and genetics correlated: Paraguay feral pigs clustered with nearby Misiones pigs and Bolivian pigs were close to Peruvian ones. The two Colombian populations belonged to the same clade (Supplementary File 3), yet their F_{ST} was 0.19. In general, we did not observe that genetic distance or average F_{ST} was a proxy for geographic distance. To test the relation between geographic and genetic distances, a Mantel test was performed. Figure 3a shows the results for all samples. Except for pigs sampled in the same location, geographic distance explains very little of the variation in genetic distance. The coefficients of determination (r^2) were 0.09 and 0.04, respectively, when pigs from the same location were considered or not. Notice that a reduced genetic distance among pigs in the same site can be due simply to sampling close relatives in the same or nearby villages.

Given the historical complexity of American colonization and because a shorter geographical distance does not necessarily imply a more active trade route, we circumscribed the analyses to a narrower, hopefully simpler space. Two regions were reanalysed separately. First, the North Argentinean pigs (Misiones, Corrientes, Formosa and Salta provinces) together with nearby Paraguay feral pigs; and second, Central America (Mexico, Guatemala and Costa Rica). It can be seen, again, that correlation vanishes and even becomes slightly negative when pigs from the same spot are removed (Figures 3b and c). In Argentina, the r^2 was 0.16, but vanished ($r^2<10^{-3}$) when pair of pigs with a geographic distance of zero were removed. Similarly, the r^2 in Central America were 0.23 and 0.15, respectively, in each of the two analyses. This suggests that pigs from nearby locations are genetically related, maybe because local communities exchange animals, but also that pigs can be imported from different remote or foreign locations. Overall, a classical stepping-stone model is not applicable to this human-mediated livestock colonization, where geographic distance explains only a tiny fraction of total genetic variability. Note that this pattern could also reflect an incipient pattern of breed formation. In fact, except for Brazil, legislation on local breeds or populations is very recent in Latin American countries and in general not strictly enforced.

Next, ADMIXTURE was used to characterize genetic structure across American village pigs and their putative ancestral breeds. The unsupervised method detects $K=14$ clusters as an optimum partition number (Figure 4a). This suggests an underlying highly complex genetic structure, despite the apparent uniformity within American village pigs portrayed by PC (Figures 1 and 2). In Figure 4a, a number

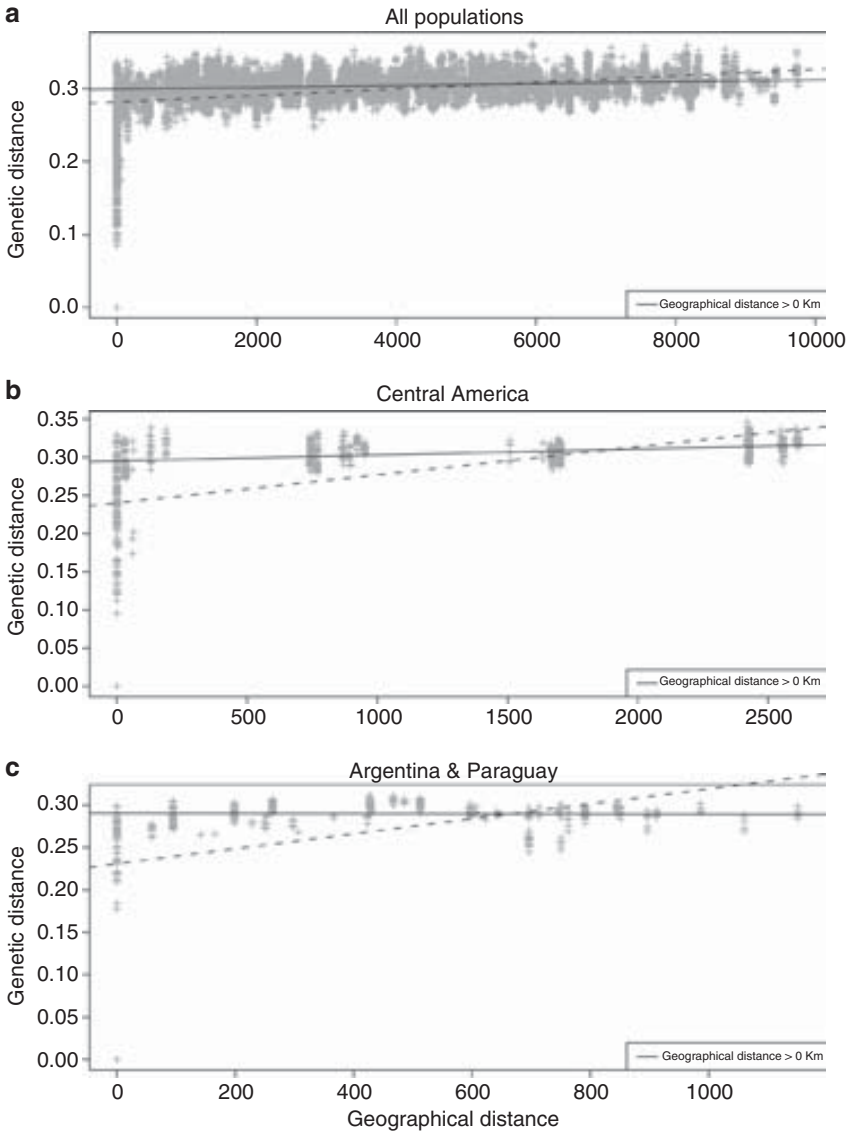


Figure 3 Correlation between geographic and genetic distances: (**a**) all American samples; (**b**) Central American samples; (**c**) North East Argentina and Paraguay. Continuous (dashed) line is regression not including (including) samples from the same location (geographic distance 0).

of populations are identified as homogeneous, that is, Iberian, Duroc, Hampshire, Guinea Hog, Yucatan, Cuino, Piau and Chinese breeds, and, to a lesser extent, Landrace, Large White and Colombian Zungo. Other populations, primarily American, but also Bisaro and Canary, are admixed. In agreement with previous results, the method does not detect a strong structuring between Iberian and European wild boar (Ramirez *et al.*, 2009; van Asch *et al.*, 2012). If we take a uniform cluster assignment as a signature of recent isolation, Figure 4 suggests that Guinea Hog, Yucatan, Cuino, Colombian Zungo and Piau would be the American populations that show less or no degree of recent introgression. Except for Cuino, for which there are no official records, this agrees with the fact that these are established breeds with their own breeding programmes.

It is also illuminating to consider a partially supervised analysis. In this case, some pigs were assigned a predefined cluster. We ran cases $K=13$ and $K=6$. With $K=13$, a predefined cluster was assigned to those pigs from uniform breeds as suggested by the unsupervised analysis (Figure 4a). A value $K=13$ was used instead of $K=14$ because no population was assigned fully to a fourteenth cluster. This analysis (Figure 4b) suggests a putative Brazilian Piau cluster to be predominant among Brazilian breeds, primarily in Monteiro, and where Moura is largely introgressed with Duroc. Similarly, a hypothetical Colombian Zungo cluster would be present among many American village populations. A problem with this supervised analysis is that the large number of clusters assumed, without considering historical processes, makes interpretation difficult. To simplify matters, we considered a smaller number of clusters ($K=6$) that represent all known major origins of American village pigs: Iberian, Landrace, Large White, Duroc, Hampshire and Chinese pigs. Therefore, we make the simplifying, but reasonable, assumption that the genetic make-up of American pigs can be largely explained in terms of these six origins. The analysis (Figure 4c) still shows that American populations are clearly admixed but to different degrees; heterogeneity within populations is also evident. Assuming the hypothesis of these six clusters representing the main ancestral populations of American village pigs, the Iberian pig represents an

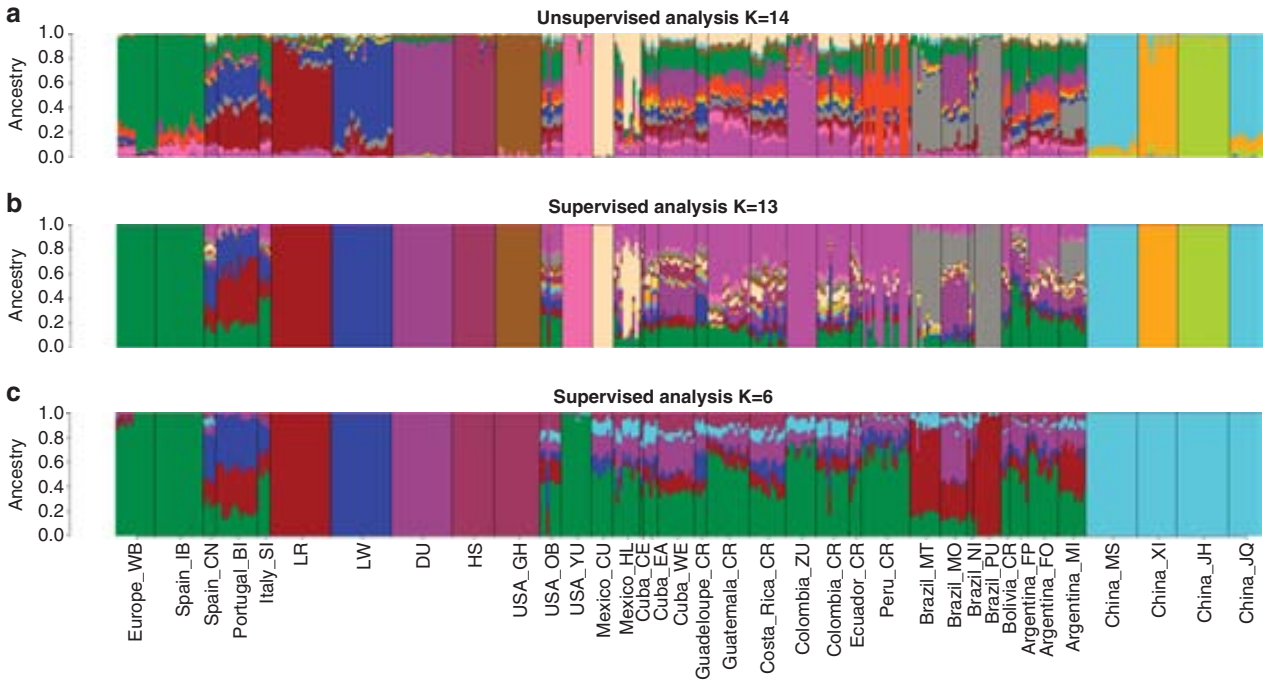


Figure 4 ADMIXTURE analyses: (**a**) unsupervised, $K=14$; (**b**) supervised, $K=13$; (**c**) supervised, $K=6$.

important component, especially in Yucatan, Peruvian and Colombian Zungo pigs.

Nevertheless, this Iberian component varies largely in importance across populations. In fact, PCA analyses (Figure 1) suggests that a pure Iberian ancestry is unlikely. A more specific analysis with ADMIXTURE (Figure 4 and Table 2) confirms that American pigs are partly of Iberian origin, but that this origin is not necessarily predominant, except Yucatan or perhaps Peru and Colombian Zungo. The inferred average Iberian contribution to American village pigs is ~40%, ranging from Yucatan (~99%) to Brazilian Moura or Piau (~0%). Supplementary File 4 shows the F_{ST} between the putative main founders (Iberian and international breeds) and the genotyped American populations. Except for a few populations studied, namely Yucatan, Peruvian altiplano, feral Argentinean pigs and Colombian Zungo, Iberian was not the closest breed. Overall, American populations were equidistant between Landrace, Large White and Iberian breeds, whereas Duroc is the most distant one.

ADMIXTURE also suggests that an Asian component cannot be ruled out for several populations (Table 2 for supervised $K=6$ and Supplementary File 5, unsupervised $K=14$). European wild boar is our negative control, and ADMIXTURE does report <1% of Chinese assignment, as in Iberian and Sicilian pigs. Also in agreement with records, Large White and Landrace have variable levels of introgression from Chinese breeds. Bisaro and Canary pigs are likely to be admixed recently with international breeds, the latter displaying a considerable influence of Chinese pigs, in agreement with previous mitochondrial DNA results (Clop *et al.*, 2004). Within the Americas, the breeds with little or no inferred Chinese introgression are Yucatan, Ossabaw, Mexican hairless, Bolivian, Peruvian and some Argentinean pigs. In contrast, Eastern Cuba, Pacific Colombian creole and some Brazilian pigs (Nilo predominantly) may have a non negligible percentage of Chinese germplasm. The closest Chinese breed, in terms of F_{ST}, was consistently the Jiangqhuai breed (Supplementary File 4). This breed is originally from the Taihu lake area, the origin of the most prolific Chinese pigs, and is also renowned for its good meat quality. In agreement with reports (Porter, 1993), this supports the belief that Chinese pigs were imported to improve upon the characteristics of local European pigs.

Three levels of Chinese migration into the Americas were compared via simulation. Chinese contribution was overestimated with the unsupervised ADMIXTURE, whereas values are better estimated with partially supervised ADMIXTURE, unless migration is very small (1%, Supplementary File 6). For instance, in the unsupervised analysis, the Chinese contributions were estimated to be 8.7, 10.5 and 18.2% when true migration rates were 0%, 1% and 10%, respectively. The equivalent supervised estimates were 4.1%, 4.5% and 11.6%, respectively. In contrast, the contributions of Iberian and International pigs were reasonably well estimated. The simulated site frequency spectra, together with the observed spectra from some populations in our data is in Supplementary File 7, and shows that the simulated model reproduces, approximately, the observed data.

Signals of adaptation: size and altitude

First, we investigated whether there is evidence for any common selective signature between American village pigs, excluding Brazilian samples and minipigs (Yucatan, Cuino and Guinea Hogs) and their European and international ancestors. Supplementary File 8 shows over-represented GO categories ($P<0.01$) within genes in 1 Mb windows with average d statistics greater than 2 s.d. over the mean, that is, ~1% extreme windows. Despite the apparent heterogeneity among breeds and populations, it is noteworthy that a few GO categories were highly over-represented. These ontologies are related to development (specifically limb morphogenesis), vitamin A metabolism and behavior. Therefore, this may suggest that a common response among American populations has involved modifying their pattern of development and, perhaps, also by how they respond to external stimuli.

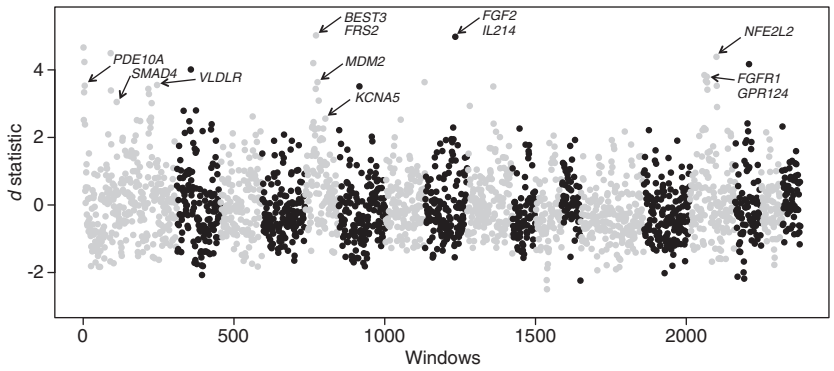


Figure 5 *d* profile in the Peruvian population showing the position of some relevant genes. Each dot represents a 1 Mb window containing at least five SNPs.

Adaptation to altitude was specifically explored. Among the environmental challenges posed by the American continent to livestock life, the Andean altiplano is probably one of the harshest. Figure 5 contains the profile of the *d* statistics; there were 87 extreme windows ($d > 2$ s.d. over the mean) that contained 301 annotated genes. The most significant enriched category was the peptidyl-citrulline biosynthetic process (Supplementary File 8); interestingly, citrulline has been reported to relax blood vessels and may improve adaptation of blood circulation to altitude. It is also remarkable that, among the genes in extreme F_{ST} windows, we found several genes known to be involved in response to hypoxia (*SMAD4, MDM2, VLDLR, KCNA5*) although their corresponding GO categories were not significantly enriched. A detailed inspection showed that a total of 54 out of the 301 annotated genes are also involved in the cardiovascular system phenotype and physiological characteristics of the mammalian heart and blood vessels (Supplementary File 9), and IPA analyses showed that over 70 of the 301 genes were involved in cardiovascular or hematological diseases (Supplementary File 10).

The alternative statistics iHS resulted in far fewer outlier windows, may be because of detection of homozygosity requires denser SNP spacing than that employed here. Only three windows (Supplementary File 11) were over 1.4 s.d. and only the most significant window, that on SSC2, overlapped with the differentiation analysis (Figure 5). There are no reported genes in current porcine assembly for this window. Yet, analysis using our own unpublished RNAseq data allowed us to identify several unannotated genes. A subsequent annotation with blast2go, Gotz *et al.* (2008) identified gene *EMR1*, which is involved in respiratory diseases. The second most extreme window (SSC9) contained three genes, *TBP12, GNG11* and *GNGT1*, which are involved in blood coagulation.

DISCUSSION

We present the most extensive genomic analysis of American creole livestock species to date. The samples genotyped represent a comprehensive overview of the extant genetic variability in American village pigs; these pigs are, importantly, adapted to a wide array of climates and environmental conditions, for example, heat, altitude or diseases. With data at hand, most American populations showed a high degree of admixture, greater than their parental populations, that is, Iberian, Large White, Landrace or Duroc, together with a putative direct Chinese influence. The genetic landscape that we observe is that of a complex conglomerate, in contrast to similar analyses in other species with a much more marked structure, such as dogs. Nevertheless, the analyses of village dogs have also proven to be much more complex than those of well-established breeds (Boyko *et al.*, 2009). In particular, we did not observe that genetic distance or average F_{ST} was a proxy for geographic distance, likely because livestock populations have a great mobility and corresponding complex genetic histories.

There are two potential problems regarding the interpretation of results. First, the limited number of individuals sampled and second, SNP ascertainment bias. While small samples may not be so relevant when the number of markers is high (Willing *et al.*, 2012), the consequences of SNP ascertainment bias are, however, much more difficult to assess. Theoretical and simulation work have shown that 'PCA projections from genotype data will be similar to PCA projections from resequencing data, but will typically be larger in magnitude' (McVean, 2009), that is, distances will be biased, although the topology will be conserved. To explore, even if tentatively, this issue we ran coalescence simulations. Although our goal was not to comprehensively analyze all potential models, the simulations suggest: (i) that a partially supervised approach is more reliable than an unsupervised method, and (ii) that the estimate of Chinese influence can be biased upwards when the true migration is zero or very small ($\sim$1%) but are more accurate as migration rate increases. The supervised ADMIXTURE estimates of Chinese influence with are reasonably large in some populations, notably in Eastern Cuban, Guadeloupe, Mexico, Pacific Colombian and Brazil's Nilo. Therefore, a Chinese contribution in these cases would not be an artefact. Although there is evidence of direct introgression from Asia into the Americas (Ramirez *et al.*, 2009; Lemus and Ly, 2010), this Asian influence might also be indirect, mediated by international breeds. Complete resequencing and comprehensive simulations will help to elucidate this issue.

The term 'creole' (Spanish *criollo*, Portuguese *crioulo*) is used to refer to descendants from the Iberian Peninsula (Elliot, 2007). As with humans, the traditional view is that 'creole' pigs are descendants of pigs imported from the Iberian Peninsula. However, the actual ancestry of the many breeds termed 'creole' throughout the Americas is unknown. Our data suggest that this contribution has been dramatically attenuated in current village pigs. If the contribution of the Spanish Iberian pig to American creoles is smaller than anticipated, we can speculate whether creole pigs have undergone a dramatic introgression with international-breed pigs or whether extant Iberian pigs are different from those of several centuries ago. We favor the first hypothesis: (i) there is little structuring between European wild boar and Iberian pigs (Ramirez *et al.*, 2009; van Asch *et al.*, 2012), (ii) a greater Iberian contribution is ascribed by a supervised analysis to the most preserved or isolated populations

(Yucatan, Peru) than to other populations (Table 2), and (iii) the introduction of international breeds all over the world replacing local livestock is well known. As a result, village pig populations are far from being static genetic pools. In fact, the presence of outliers in some of these populations, rather than being simply 'noise' or errors in sampling, illustrates that village pigs are dynamic populations whose genetic structure can change quickly and deserve conservation. In fact, the ancestral Mexican population of Yucatan pigs is now almost extinct, so current Yucatan mini-pigs should actually more faithfully reflect the ancestral genetic variability of Mexican pigs than do modern *cuino* or *pelón* pigs. In all likelihood, international breeds will continue to be introgressed into American village pig populations, whereas the flow of Iberian pigs was interrupted long ago.

Historical records, mitochondrial DNA data (Souza *et al.*, 2009) and our data support that Brazilian pigs are mostly related to European local pigs, as are the rest of American village pigs. Nevertheless, Brazilian pigs clustered separately at the continent level (Figure 3). Although this result should be considered cautiously, given that the American principal components explain a small fraction of worldwide variance where the Asia—Europe axis is predominant (Figure 1), it seems to be a general trend that Brazilian pigs are closely related among each other (See also the dendrogramme in Supplementary File 3). Can this be explained by different histories from the early days of colonization or is it due to more recent events? Certainly, Portuguese and Castilians divided their area of influence in America from the very beginning due to the Treaty of Tordesillas, in 1493. Empirical support for this hypothesis is also provided by the fact that Bisaro pigs, a Portuguese breed, are genetically closer in terms of F_{ST} to Brazilian populations than are Iberian pigs. Yet, it is worth noting as well that Portugal was ruled by the Spanish Hapsburg dynasty during a large initial period of the colony (1581–1640), therefore increasing trade between and within Iberian kingdoms and their colonies in the Americas. There were also intermittent periods of Dutch rule in NE Brazil, for example, 1624–1654 in Pernambuco. F_{ST}'s also show that Bisaro pigs are nearer to many American populations than are Iberian pigs, which would suggest a predominant Portuguese 'pig colonization' America-wide. Similarly, Canary pigs are also close to American pigs. However, as Figure 4 suggests, there is evidence that both ancient Bisaro and Canary pigs have been intermixed with modern breeds. What is the cause, therefore, of a specific Brazilian signature? First, note that Moura is somewhat separate from the rest of Brazilian pigs, and they exhibit an increased Duroc component. Mariante and Cavalcante (2006) do report that local Brazilian pigs were crossed to Duroc-Jersey to make up Moura. As for the rest of Brazilian breeds, the explanation is not so clear. A Chinese contribution cannot be ruled out, at least in Nilo and in Monteiro. Further, classical studies (Vianna, 1956) mention that Portuguese imported pigs from their colony Macau in China. Interestingly, some pigs in Misiones, Argentina are still called Macau. The ADMIXTURE supervised analysis suggests a strong Landrace component in Piau with $K = 6$ (Table 2 and Figure 4c), whereas larger K suggests a cluster of its own and shared with other Brazilian populations (Figures 4a and b). The Piau breed originated in the states of Goias, São Paulo and Minas Gerais, likely a result of crosses between local and other breeds like Poland China or Duroc, among others (Mariante and Cavalcante, 2006). All in all, it can be hypothesized that the difference between Brazil and Spanish America that we see today is caused by distinct introgression patterns, rather than by distinct initial colonization processes.

A major task in order to understand adaptation at the molecular level is to characterize the genes that have responded to selection, either artificial selection or natural selection as a result of adapting to extreme environments. Our results bear special relevance regarding the adaptation to altitude. Our study identified ~300 highly differentiated genes. Remarkably, about 54 has a role in blood circulation and four of them (*SMAD4, MDM2, VLDLR, KCNA5*) were *a priori* functional candidates in human studies (Simonson *et al.*, 2010). Among those, a few merit special attention. *FGF2* and *FGFR1* are involved in phenotypic modulation of vascular smooth-muscle cells (Chen *et al.*, 2009). *NFE2L2* has a role in the coordinated upregulation of genes in response to oxidative stress, whereas *GPR124* regulates angiogenesis in the central nervous system (Kuhnert *et al.*, 2010). Additional genes include *BEST3, PDE10A*, *PDE11A* and *IL21*. *BEST3* is expressed in smooth-muscle cells and is important for regulation affecting vasomotion. *PDE10A* and *PDE11A* are expressed in components of the trigeminovascular pain signaling system (Kruse *et al.*, 2009). *PDE10A* is also involved in progressive pulmonary vascular remodeling, increasing its expression in some pulmonary diseases (Tian *et al.*, 2011). Finally, interleukin 21 signaling has a critical role in promoting the lung inflammatory response to acute pneumovirus infection (Spolski *et al.*, 2012). Adaptation to altitude has received attention in humans (see Cheviron and Brumfield, 2012 for a review), and physiological differences caused by altitude have been studied in cattle (Wuletaw *et al.*, 2011). However, to our knowledge, this is the first report of indirect evidence of genetic adaptation to altitude in livestock. It should be noted that, given the relatively low density of markers and the large window used (1 Mb), the selective footprints described are probably among the most extreme ones and other indirect evidence of selective events are waiting to be identified with more data and with more refined tools.

CONCLUSION

To conclude with a paraphrase of Novembre *et al.* (2008): creole porcine genes in the Americas do not mirror geography. They look rather like a blur of history. Genetic evidence supports the belief that creole pig populations are relatively homogeneous within a short geographic radius, a shared ancestry likely due to the exchange of pigs between nearby communities. Aside from that, geographic distance explains just a tiny fraction of variation in coancestry. Across the Americas, the genomic patterns observed are not compatible with a classical stepping-stone colonization model, reminding us that livestock is highly mobile, especially in the case of pigs. Modern village pigs in the Americas are the result of many independent colonization and introgression events, including may be a direct Chinese introgression. Importantly, these data also confirm our initial hypothesis regarding adaptation: extreme climates have posed important challenges to pigs.

DATA ARCHIVING

Data have been deposited at Dryad: doi:10.5061/dryad.t1r3d.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all people and institutions who helped to collect samples, provided genotypes or comments: network CONBIAND, coordinated by JV Delgado, J Capote, M Amills, T Strumia, S Quisbert, S de la Rosa, M Montenegro, M Sturek. We thank A Mercadé and A Castelló for genotyping part the animals and SE Ramos-Onsins for help with coalescence simulations. We are grateful to M Teran-Garcia (UIUC) and colleagues for her helpful assistance with running IPA's system at the Department of Food Sciences and Human Nutrition,

(license 11904312, obtained with an equipment grant from the Office of Research, College of ACES). WBP is funded by COLCIENCIAS (*Francisco José de Caldas* fellowship 497/2009, Colombia), CAS thanks grants from CAPES and EMBRAPA (Brazil), YRC is recipient of a PhD studentship from MICINN (Spain, ref. AP2008-01450), AEC is recipient of a PhD studentship from MICINN (Spain). Work funded by Consolider CSD2007-00036 'Center for Research in Agrigenomics' and AGL2010-14822 grants (Spain) to MPE, EU SABRE project FOOD-CT-2006-01625, USDA project 2007-04315 (USA), *Facultad de Ciencias Agrarias, San Pedro* (UNA), *Unión de Gremios de la Producción* (UGP) and *Empresa San Rafael Agricola y Ganadera SRL* (Paraguay), *Universidad Técnica de Oruro* (Bolivia), *Programa de Conservación de los Bancos de Germoplasma, Instituto Colombiano Agropecuario* (grant 048-2011) and *Ministerio de Agricultura y Desarrollo Rural* (Colombia), and *Centro de Validación de Tecnologías Agropecuarias* (CEDEVA, Formosa, Argentina).

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J *et al.* (2010). Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci USA* **107**: 1160–1165.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002). Interrogating a high-density snp map for signatures of natural selection. *Genome Res* **12**: 1805–1814.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Alves E, Fernández AI, Barragán C, Ovilo C, Rodríguez C, Silió L (2006). Inference of hidden population substructure of the Iberian pig breed using multilocus microsatellite data. *Span J Agric Res* **41**: 37–46.

Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, Corey L *et al.* (2009). Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proc Natl Acad Sci USA* **106**: 13903–13908.

Chen P-Y, Simons M, Friesel R (2009). FRS2 via fibroblast growth factor receptor 1 is required for platelet-derived growth factor receptor $\hat{i}^2$-mediated regulation of vascular smooth muscle marker gene expression. *J Biol Chem* **284**: 15980–15992.

Crossby A (2003). *The Columbian Exchange*. Praeger Publishers: Connecticut.

Delgado JV, Martinez AM, Acosta A, Alvarez LA, Armstrong E, Camacho E *et al.* (2011). Genetic characterization of Latin-American Creole cattle using microsatellite markers. *Anim Genet* **43**: 2–10.

Cheviron ZA, Brumfield RT (2012). Genomic insights into adaptation to high-altitude environments. *Heredity (Edinb)* **108**: 354–361.

Clop A, Amills M, Noguera JL, Fernandez A, Capote J, Ramon MM *et al.* (2004). Estimating the frequency of Asian cytochrome B haplotypes in standard European and local Spanish pig breeds. *Genet Sel Evol* **36**: 97–104.

Elliot JH (2007). *Empires of the Atlantic World: Britain and Spain in America 1492-1830*. Yale University Press.

García M, Capote J (1982). *El Cerdo Negro Canario*. Cabildo Insular de La Palma: Excmo.

Gautier M, Naves M (2011). Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol Ecol* **20**: 3128–3143.

García-Dory MA, Martínez-Vicente S, Orozco-Piñán F (1990). *Guía De Campo De Las Razas Autóctonas Españolas*. Alianza Editorial: Madrid.

Gautier M, Vitalis R (2012). Rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**: 1176–1177.

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ *et al.* (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**: 3420–3435.

Groenen M, Archibald A, Uenishi H, Tuggle C, Takeuchi Y, Rothschild M *et al.* (2012). Pig genomes provide insight into porcine demography and evolution. *Nature* **491**: 393–398.

Huang D, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44–57.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 460.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Kruse LS, Muller M, Tibak M, Gammeltoft S, Olesen J, Kruuse C (2009). PDE9A, PDE10A, and PDE11A expression in rat trigeminovascular pain signalling system. *Brain Res* **1281**: 25–34.

Kuhnert F, Mancuso MR, Shamloo A, Wang H-T, Choksi V, Florek M *et al.* (2010). Essential regulation of cns angiogenesis by the orphan G protein-coupled receptor GPR124. *Science* **330**: 985–989.

Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J *et al.* (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**: 1618–1621.

Lemus C, Ly J (2010). Estudios de sostenibilidad de cerdos mexicanos pelones y cuinos. La iniciativa nayarita. *Revista Computadorizada de Producción Porcina* **17**: 89–98.

Mariante AS, Cavalcante N (2006). *Animals of the Discovery*. Embrapa: Brasilia.

McVean G (2009). A genealogical interpretation of principal components analysis. *PLoS Genet* **5**: e1000686.

Merino M, Carpinetti B (2003). Feral pig Sus scrofa population estimates in bahía Samboronbón conservation area, Buenso Aires Province, Argentina. *J Neotrop Mammal* **10**: 269–275.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al.* (2008). Genes mirror geography within Europe. *Nature* **456**: 98–101.

Porter V (1993). *Pigs: A Handbook to the Breeds of the World*. Comstock Publishing.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.

Ramirez O, Ojeda A, Tomas A, Gallardo D, Huang LS, Folch JM *et al.* (2009). Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol Biol Evol* **26**: 2061–2072.

Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE *et al.* (2009). 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.

Scheet P, Stephens M (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629–644.

Ramos-Onsins SE, Mitchell-Olds T (2007). MIcoalsim: multilocus coalescent simulations. *Evol Bioinform Online* **3**: 41–44.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ *et al.* (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**: 72–75.

Souza CA, Paiva SR, Pereira RW, Guimaraes SE, Dutra Jr WM, Murata LS *et al.* (2009). Iberian origin of Brazilian local pig breeds based on Cytochrome b (MT-CYB) sequence. *Anim Genet* **40**: 759–762.

Spolski R, Wang L, Wan C-K, Bonville CA, Domachowske JB, Kim H-P *et al.* (2012). IL-21 promotes the pathologic immune response to pneumovirus infection. *The Journal of Immunology* **188**: 1924–1932.

Tian X, Vroom C, Ghofrani HA, Weissmann N, Bieniek E, Grimminger F *et al.* (2011). Phosphodiesterase 10a upregulation contributes to pulmonary vascular remodeling. *PLoS ONE* **6**: e18136.

van Asch B, Pereira F, Santos LS, Carneiro J, Santos N, Amorim A (2012). Mitochondrial lineages reveal intense gene flow between Iberian wild boars and South Iberian pig breeds. *Anim Genet* **43**: 35–41.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S *et al.* (2012). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet* **7**: e1002316.

Vianna AT (1956). *Os Suínos: Criação prática e econômica*, 2nd edn. Serviço de Informação Agrícola, Ministério da Agricultura: Rio de Janeiro.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006). A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.

Willing ME, Dreyer C, van Oosterhout C (2012). Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. *PLoS One* **7**: e42649.

Wuletaw Z, Wurzinger M, Hol tT, Dessie T, Sölkner J (2011). Assessment of physiological adaptation of indigenous and crossbred cattle to hypoxic environment in Ethiopia. *Livest Sci* **138**: 96–104.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE *et al.* (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**: 75–78.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)

Chapter 5

-

Genome data from a 16th century pig illuminate modern breed relationships and suggest specific selection targets

-

O. Ramírez[1†], **W. Burgos-Paz[2,3†]**, E. Casas[2], M. Ballester[2,3], E. Bianco[2,3], I. Olalde[1], V. Novella[4], M. Gut[5], C. Lalueza-Fox[1], M. Saña[4], M. Pérez-Enciso[2,3,6*]

Affiliations described in the manuscript

Manuscript submitted to Heredity

Supplementary information in Annexes.

Genome data from a 16th century pig illuminate modern breed relationships and suggest specific selection targets

O. Ramírez[1†], W. Burgos-Paz[2,3†], E. Casas[2], M. Ballester[2,3], E. Bianco[2,3], I. Olalde[1], V. Novella[4], M. Gut[5], C. Lalueza-Fox[1], M. Saña[4], M. Pérez-Enciso[2,3,6*]

1 Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), PRBB, 08003 Barcelona, Spain.
2 Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain.
3 Departament de Ciència Animal i delsAliments, Universitat Autònoma de Barcelona, 08193 Bellaterra,Spain.
4 Departament de Prehistòria, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.
5 Centro Nacional de Análisis Genómico (CNAG), PCB, 08028 Barcelona, Spain.
6 Institut Català de Recerca i Estudis Avançats (ICREA), Carrer de Lluís Companys 23, Barcelona, 08010, Spain.
† Equal contribution

*Correspondence
Miguel Pérez-Enciso
miguel.perez@uab.es
+34 93 935636600
Centre for Research in Agrigenomics (CRAG),
Campus UAB,
08193 Bellaterra, Spain

ABSTRACT

Ancient DNA (aDNA) provides direct evidence of historical events that have modeled the genome of modern individuals. In livestock, resolving the differences between the effects of initial domestication and of subsequent modern breeding is not straightforward without aDNA data. Here, we have obtained shotgun genome sequence data from a 16th century pig from Northeastern Spain, together with three new modern genomes from Iberian pig, Spanish wild boar and a Guatemalan Creole pig. Comparison with both, mitochondrial and genome data shows that the ancient pig is closely related to extant Iberian pigs and to European wild boar. Although the ancient sample was clearly domestic, admixing with wild boar also occurred. Specific differentiation analyses allowed us to pinpoint genes that have been plausibly affected by initial domestication. Among those, we found genes involved in coat color and an increase the reproductive performance, both known functions associated with early domestication process.

INTRODUCTION

Ancient DNA is a powerful tool to unravel the complex history of domestic species (Larson et al. 2007; Ottoni 2012; Krause-Kyora et al. 2013; Thalmann et al. 2013). Most ancient DNA in livestock studies have focused on very early events, such as domestication, and have provided very limited evidence based primarily on a single locus like mitochondrial DNA (mtDNA). Although mtDNA is useful for phylogeographic analyses due to its high substitution rate, using a single locus that only reflects the matrilineal history does not help in resolving the complete demographical history of a population. Nor is it useful to study changes fuelled by selection, be it natural or artificial. Artificial selection, via the processes of breeding, has dramatically sculpted livestock genome diversity in a very short time frame. The history of livestock breeds comprises vivid examples of accelerated evolution. The largest selection intensities for livestock have been exerted only since the last century, whereas domestication was presumably a process that preceded such events by many more centuries. Disentangling the effects of domestication from those of modern breeding on the genome by simply comparing wild and domestic specimens (say wild boar vs. pig) is difficult because domestics carry signatures of modern breeding and selection. To that end, the sequencing of ancient domestic genomes predating the advent of breeds and modern artificial selection era is unavoidable.

Fortunately, paleo-population genetics has become feasible with the advent of new sequencing technologies (Wall and Slatkin 2012). In the case of pigs, our knowledge of ancient genomes is currently limited to short mtDNA sequences (e.g., Larson et al. 2007; Meiri et al. 2013) and a fragment of the MC1R gene (Krause-Kyora et al. 2013), which is involved in coat color determination. Despite this limited evidence, the history of the pig is unfolding to be much more complex than anticipated. Although the domestication of the pig in the Near East (NE) at least by 8,500 BC is well documented (Conolly et al. 2011), recent investigations have shown that early domestic pigs in Europe carried a distinctive NE mitochondrial lineage, which is was gradually replaced by a local European wild boar signature (Larson et al. 2007; Ottoni 2012; Manunza et al. 2013). Pig meat was a key component in the Neolithic diet and it was kept continuously throughout Prehistory, with different management strategies. It was during Roman times when pig farming, throughout the whole of Western Europe, underwent one of the most significant transitions. Several

ancient texts on stock-breeding published by Latin authors (Cato, Varro, Columella and Palladius) advised on the practise of selective breeding as a way to contribute to increasing the productivity of the species (MacKinnon 2004).

During the late medieval and early Modern era, another key turning point in pig breeding occurred. Archaeo-zoological analyses of faunal assemblages dated to this time highlight important changes in the health and size of domestic pigs. These changes are thought to be the product of new farming strategies and selection criteria (Albarella 1997; Thomas 2005; Albarella et al. 2009). In this sense, one of the aspects that has most often been emphasised is a change from a system of extensive farming to rearing in confinement (Ervynck et al. 2007), allowing for a more intensive control of the animals and their nutrition (Thomas et al. 2013). This system would have directly influenced the speed of the development after birth (Albarella 1997). At the same time, keeping the animals permanently in stables isolated the domestic population from wild animals, thus decreasing the possibilities of hybridisation and gene flow between populations (Albarella et al. 2009).

The 16th century was an important milestone for pig history. In addition to changes in breeding practices mentioned, it predates the introgression of Asian germplasm that occurred from the 17th century onwards, when porcine colonization of the Americas was beginning in earnest, and three centuries before the creation of modern breeds and of ensuing intense selection for growth and leanness that continues today. Therefore, pigs from this period would represent the original European genome and can serve as a yardstick against which to compare selective and introgression events that were to happen later in time. It is also of particular interest to ascertain whether extant modern Iberian pigs are representative of past porcine populations, because these local Mediterranean pigs are thought not to be introgressed with Chinese pigs (Alves et al. 2003). This will be of utmost importance to identify the Asian footprints in European pigs and its relationship with selective events. Also of historical interest is to characterize the genetic legacy of the ancient pig in modern American Creole pigs. Creole pigs have been thought to be direct descendants of 16th century pigs but its actual history is seemingly much more complex (Burgos-Paz et al. 2013).

To understand better these issues, we present here the partial genome sequence of a 16th century pig from the Montsoriu castle in North East Spain (province of *Girona*). Montsoriu Castle is at present one of the most representative examples of social and economic organization in medieval and early modern times. The continuous and large archaeological sequence allows for the dynamics and changes in livestock husbandry from the 10th to the 16th century to be traced back. This is one of the few examples where it is possible to assess the evolution of pig breeding practices and their impact on the species (Font et al. 2010). Pig remains could therefore be representative of the general improvements in agronomic techniques and the application of new selective pressures during late Middle Ages (e.g., Ervynck et al. 2007).

In addition to the ancient pig, we also sequenced three new genomes pertaining to a wild boar from the same Spanish region, an Iberian pig from the highly inbred strain Guadyerbas (Toro et al. 2008), and an American Creole pig from Guatemala. These modern samples

provide evidence on important historical and genetic events like domestication, admixing and the relationship with Creole pigs.

MATERIALS AND METHODS

Archaeological Sampling and Context

Montsoriu Castle is located in the province of Girona, in the north-east of the Iberian Peninsula (41°46'58''N, 2°32'30''E, at 632.5 m.a.s.l.). It is one of the most representative examples of social and economic organization in medieval and early modern times (10-16th centuries). During the 2007 season, an abandoned cistern was excavated. It corresponds to the last stable occupation phase at the castle, and it yielded an extremely well-preserved assemblage (UE 10955). This assemblage is a unique, very varied and complete finding, and provides a full panorama of daily life in a castle in Renaissance time (Font et al. 2008).

A total of 1,729 pig remains were retrieved from UE10955. The slaughtering patterns reveal that the specimens, mostly males, were systematically consumed towards the end of their growth stage (88%), mainly between 12 and 18 months of age (40%). Some adult females are also present and these would have been slaughtered at the end of their breeding life. Interestingly, osteometric analyses have shown that the remains of this species correspond to animals larger than those recorded in earlier centuries at the same site.

The sample selected for sequencing was a tibia (diaphysis and distal epiphyses) of an adult without any apparent pathology and aged over 3.5 years. Age was estimated according to fusion stage (Silver 1969). Bone surface characteristics demonstrate that the bone buried quickly, which inhibited weathering and deterioration. Measurements, taken according to (Von Den Driesch 1976), were SD=19.5 / Bd=28.8 / Dd=25.5 and ensure that the bone corresponds to a domestic animal. The sample was dated to between 1520 and 1570, as determined accurately with the combination of stratigraphic and cultural criteria, integrating historical documentary sources, numismatic and archaeological artifacts date production(Font et al. 2008; Font et al. 2010).

DNA extraction and sequencing

DNA extractions of the ancient and modern samples were performed at different times and in different laboratories. All experimental procedures on ancient samples were performed in a dedicated ancient DNA laboratory (IBE-PRBB, Barcelona), where no previous work with modern pigs had been conducted. DNA extraction was performed for each of the three best preserved ancient pig samples. DNA was isolated by a conventional phenol–chloroform precipitation protocol and microcolumn concentration (Millipore), as described elsewhere (Lalueza-Fox et al. 2007; Sánchez-Quinto et al. 2012). The extract was purified with a gene clean silica method using a DNA extraction Kit (Fermentas, USA). Following extraction we amplified and sequenced a 77 bp fragment of the mitochondrial cytochrome b (MT-CYB) gene to test the quality of the samples. Amplification was performed using a two-step PCR protocol(Krause et al. 2006). Amplified products were purified with a gene clean silica method (Fermentas, USA) and cloned using the Topo TA cloning kit (Invitrogen, The Netherlands). White colonies were subjected to 30 cycles of PCR with M13 universal primers and subsequently sequenced with an Applied BioSystems 3100 DNA sequencer, at the sequencing service of the *Universitat Pompeu Fabra* (Barcelona).The partial sequence of the

MT-CYB gene was obtained from two of the three samples. Of these two samples we selected the individual that, according to bone size, was the most likely domestic specimen.

From this individual DNA, three single end lanes of 100 bp length reads were sequenced at Fasteris (www.fasteris.com, Plan-les-Ouates, Switzerland) using HiSeq2000. The library was prepared with the TruSeq DNA sample preparation kit from Illumina following the instructions of the manufacturer.

Modern samples were sequenced in *Centro Nacional de Análisis Genómico* (CNAG, www.cnag.cat) also using HiSeq2000 Illumina platform. The library preparation of each modern sample was performed according to the Illumina paired-end sequencing protocol with minor modifications. Ancient and modern samples were sequenced in different institutions to avoid contamination as much as possible.

Ancient Data Alignment and Quality Control

To process raw ancient data, we first removed stretches of N´s and stretches of consecutive bases with 0, 1, or 2 quality scores from the 3' and 5' ends of the reads. Reads shorter than 30 nucleotides were discarded for further analyses. Post-morten degradation results in a short length of ancient DNA sequences. As a result, adapter sequences ligated during library preparation can be present at the end of the reads. This can affect the correct mapping to the reference genome and it can also bias the SNP calling. Therefore, we used AdapterRemoval (Lindgreen 2012) to remove adapter sequences from the reads, discarding sequences shorter than 30 bp after adapter trimming. We found 13% of the reads containing adapter sequence, keeping a total of 408,912,560 reads with an average length of 93 bp, which were aligned to the pig reference genome. We mapped reads to the current pig genome assembly (Sscrofa10.2) using BWA (Li and Durbin 2009) with the quality trimming parameter set to a Sanger quality score of 15. Furthermore, to improve the ancient DNA read mapping against modern reference genomes the edit distance parameter was set to 0.02, and the seed region (the first 32 nucleotides) was disabled following recommendations in (Schubert et al. 2012). Finally, we removed duplicates with SAMtools rmdup option. In order to assess the level of human contamination, we mapped all the reads to the human reference assembly (GRCh37/hg19) using BWA.

For allele determination in the ancient sample, we considered only reads with minimum mapping quality of 20, and base quality (Phred score) of at least 30 if there was a single read covering that position or 20 with depths 2-5x. Positions covered with >5x were discarded, as being most likely caused by repetitive or copy number variant regions. To avoid post mortem DNA damages that lead to increased C→T and G→A transitions, we only retained those positions where the ancient allele was also observed in any of the eight modern pig genomes used for this study (Table S1). This filtering should decrease dramatically the number of post mortem changes accepted as true variants by a factor of ~ $1/2\ a_{16}q$. This is the probability of finding a base in the eight modern samples that coincide with a given post mortem damage, with $a_{16}q$ being the expected number of polymorphisms to be found in eight diploid samples or 16 chromosomes, a_n being Ewens' constant (a_{16} = 3.4) and q, the effective size (or ≈ 0.002 in pigs), constant 1/2 occurs because half of the potential errors will be accepted, when they match any of the two alleles found in the modern population. The expected percentage of

polymorphisms that are singletons in the ancient sample and are therefore discarded even if being true variants, is also obtained from Ewen's sampling term, assuming a standard neutral model, as $1\text{-}a_{16}/a_{17} \sim 2\%$ (because we ascertain 16 modern chromosomes but only one ancient allele) so the bias due to removing singletons is expected to be small.

Modern Data Alignment and Genotype Calling

Modern sample reads were aligned with BWA (Li and Durbin 2009) allowing for 7 mismatches. Sites between 5x and double average depth + 1 were considered for analyses. Sample details are in Table S1. Genotypes were called using the SAMtools mpileup option and filtered with vcfutils.pl varFilter, all modern samples were analyzed together setting a minimum depth to 5x and a maximum depth of twice the average sample's depth plus one, minimum map quality of 20 and minimum base quality of 20. Setting a maximum depth was done to minimize risks of wrongly called SNPs caused by copy number variants or repetitive regions, as in Groenen et al. (2013) or Esteve-Codina et al. (2013). The resulting vcf file was merged with the ancient reads. For further analyses, we retained only the positions without any missing data where the ancient reads were compatible with the modern genotypes. Genotypes were stored and managed as plink (Purcell et al. 2007) files, using custom perl and shell scripts as needed.

Mitochondrial Analysis

Complete mitochondrial sequences were downloaded from Genbank (accessions AF486866, EU117375, FJ236991, FJ236992, FJ236993, FJ236994, FJ236995, FJ236996, FJ236997, FJ236998, FJ236999, FJ237000, FJ237003, NC_012095). In addition, aligned bam files were obtained for project ERP001813 (Genbank accession,(Groenen et al. 2012)), from Wuzhishan Chinese mini pig ((Fang et al. 2012), AJKK00000000) and from Iberian genome (Esteve-Codina et al. 2013) (SRX245748); mtDNA consensus sequences were obtained from these complete genomes and from the modern samples sequenced here using SAMtools (Li et al. 2009). All sequences were aligned with MUSCLE v3.8.31 (Edgar 2004) using options diags and maxiters 2. A neighbor-joining NJ tree was obtained with Mega 5.1 (Tamura et al. 2011) using pairwise deletion, maximum composite likelihood and homogeneous rates model.

Array Genotyping Data

In order to position the 16th century pig among worldwide samples, we combined 60k SNP array genotypes from two biodiversity panels. The first panel is a wide sample (n = 379) of international, Chinese and American Creole breeds and European and Tunisian wild boar (Burgos-Paz et al. 2013), whereas the second panel (n = 40) comprised Near East wild boars (Turkey, Iran and Armenia) and Romanian Mangalitza, a central European local pig (Manunza et al. 2013). These data were combined with the genotypes inferred from sequence in the ancient pig. As before, only positions with enough quality and alleles compatible with modern samples were retained. Given that SNP alleles in the array do not directly correspond to actual sequenced bases, we employed the following procedure:

1. SNP alleles were coded from forward to TOP/BOTTOM using usingGenGen pipeline (Wang et al. 2007).

2. We identified the set of chip SNP positions that were represented in the ancient sequence, filtered by quality criteria described (map quality ≥ 20, base quality ≥ 30 if depth 1, BQ ≥ 20 if depth 2-5).
3. We checked, in several modern pigs that were genotyped with the 60k chip and sequenced, the actual polymorphisms found at those positions. We considered only those SNPs with at least two copies per allele.
4. Monomorphic and triallelic SNPs were discarded.
5. Additionally, for every SNP, we verified the strand orientation and allele with the probe flanking regions provided by the International Pig Sequencing Consortium during the development of the Illumina's array.
6. Allele coding was verified with several public datasets of our own (Burgos-Paz et al 2013), Badke and Steibel (https://www.msu.edu/~steibelj/JP_files/SNP_chip.html) and from M. Groenen et al. (pers. comm.).
7. We discarded SNPs where the ancient allele did not match any of the two modern sample alleles.

We performed principal component analysis (PCA) with prcomp R package (R Development Core Team 2011). Given that only one ancient allele can be recovered for most of positions, we duplicated the ancient allele to build a homozygous genotype. This is equivalent to oversampling to reduce bias in PC, which is very sensitive to unequal sampling (McVean 2009). To verify the robustness of this procedure, we also sampled one random allele from each SNP in the modern samples. A few of these plots are shown for comparison.

To visualize relationships across populations, we computed the average Euclidean distances between all pairs of individuals from two different populations using the first four principal components, those explaining most of variability. Suppose individual i has principal component value PC_{ik} for component k. The Euclidean distance with the first Q components is $d_{ij} = \sqrt{\sum_{k=1,Q}(PC_{ik} - PC_{jk})^2}$. As in Burgos-Paz et al. (2013), we ran a partially supervised admixture (Alexander et al. 2009) analysis using K=7 clusters corresponding to origins Iberian, wild boar, Duroc, Landrace, Large White, Chinese breeds and Hampshire. Pig origins from those breeds were assumed to be known without error, whereas those of the remaining individuals were inferred from these K=7 'pure' origins.

Pairwise Allele Differences

If only one allele of a genotype is ascertained, precise allele differences with a diploid genotype cannot be measured. They can nevertheless be bounded between maximum and minimum values, or weighted assuming Hardy-Weinberg equilibrium and allele frequency of A as f:

Ancestral	Genotype	Min	Max	Weighted
A-	AA	0.0	0.5	0.5 (1-f)
	AB	0.5	0.5	0.5
	BB	0.5	1.0	0.5 (1+f)

In practice, all three distances are highly correlated. Unless otherwise stated, here we employed the weighted measures because they were directly comparable with differences obtained between two diploid genotypes from modern samples.

Complete genome sequence data

We considered the genome of the ancient pig in comparison to eight modern genomes that are publicly available or were shotgun sequenced for this study. Specifically, we re-sequenced a wild boar from the same area of NE Spain (WB), an Iberian pig (IB) from the highly inbred strain Guadyerbas (Toro et al. 2008), and an American Creole pig (CR) from Guatemala. In addition, we used one publicly available genome from each of Duroc (DU), Landrace (LR), Large White (LW), Hampshire (HS), and Pietrain (PI) breeds (Table S1). These samples represent all main modern international pig breeds together with potentially closest extant relatives. As above, we retained bases of the ancient pig only if they were found in any of the eight modern samples. For the analyses, we used only the bi allelic variable positions without missing data in any of the samples.

To quantify similarity with wild boar vs. domestic, we extracted positions in the ancient pig sequence corresponding to SNPs with extreme frequency differences between wild boar and domestic pig, according to Table S3 from Rubin et al. We extracted genotypes for those positions from four sequenced individuals of several breeds and European wild boar (Groenen et al. 2012), and we computed allele frequencies per breed. In the ancient pig only one allele can be ascertained so frequencies were 0 or 1. Suppose f_D and f_W are, respectively, allele frequencies in domestic and wild boar as reported by Rubin et al., and f_S is the frequency obtained in our sample; we computed $p_D = f_D f_S + (1- f_D)(1- f_S)$ and $p_W = f_W f_S + (1- f_W)(1- f_S)$, the probabilities of a 'domestic' or 'wild' allele being equal to the sample allele as an assignment probability to the sample being domestic or wild boar. Standard errors were computed with bootstrap using library boot from R package.

To test for admixture, we calculated the D statistics and their corresponding normalized values (z-scores) using ADMIXtools' qpDstat (Patterson et al. 2012). To compute the z-score, jackknife was used as recommended by the authors, with the number of blocks set to 496. This statistic provides information about the direction of the gene flow. Having four populations W, X, Y and Z, if Z-score is positive then the gene flow occurred between either W and Y or X and Z; if negative, either between X and Y or W and Z. We considered different quartets containing the ancient, Iberian, Hampshire, European wild boar and a Sumatran wild boar (accession ERX149139) as outgroup. As for European wild boars, we used the Spanish wild boar sequenced here and a publicly available French specimen (accession ERX149180).

Non-standard-neutral regions

The ancient pig provides a unique opportunity to contrast potential selection targets that may have been operational in modern pigs. We focused in populations closely related to the ancient pig to avoid distortion caused by Chinese introgression: wild boar, Iberian and Duroc. Given the small time span occurred since its divergence, we searched for regions of extreme differentiation (F_{ST}), as this criterion is more sensible to recent events. We computed F_{ST} among wild boar vs. the rest (WB-ANIB). Extreme regions of this comparison should be enriched in genes targeted by domestication (WB-ANIB). We only analyzed those positions

where the ancient allele could be recovered, following the same criterion as in the rest of analyses above. Next, we identified genotype frequencies in four sequenced pigs from each of Iberian and European wild boar. We estimated F_{ST} from $\{E(f^2)-[E(f)]^2\}/\bar{f}(1-\bar{f})$, where the expectation is taken between average allele frequency in the first group (say WB) and the second comparison group (say AN and IB), i.e., $E(f^2) = [\,(\frac{f_{WB}}{2})^2 + (\frac{f_{IB}+f_{AN}}{2})^2\,]/2$, and $\bar{f}$ is average allele frequency over populations.

We carried out the differentiation analyses by genes and by windows. In the former case, we computed the F_{ST} parameter using all SNPs for each gene +/- 10 kb upstream and downstream. We used average F_{ST} across all SNPs in each gene as the criterion for selection. Within each of the four comparisons, we selected the genes with average F_{ST} + 3 SD and with SNP density higher than the mean SNP density. We did the same analysis by windows of 100 kb, selecting the windows with $F_{ST} > \overline{Fst}$ + 3 SD and with at least five SNPs. For the final gene list, we merged genes present in at least one of the gene or window analyses. Often, genes were in both lists. Note that both analyses are complementary; a gene can be selected in the window analysis even if it has few SNPs within the CDS provided nearby SNPs are in enough disequilibrium. Besides, a by-gene analysis will be sensitive to small genes holding differentiated SNPs within the CDS even if surrounded by SNPs that are not in disequilibrium.

The list of outlier genes was analyzed with PANTHER (Protein Analysis THrough Evolutionary Relationships) (Mi et al. 2013) using default options. PANTHER provides a functional analysis combining GO. For this, we employed the mouse annotation, because is far more complete than that of the pig.

RESULTS

Ancient Sequencing and Quality Control

Out of three single read lanes on HiSeq2000 total of 414,198,109 reads of 101 nucleotides were generated. After trimming, filtering and removing duplicates (see methods), 3,594,543 aligned reads were retained. This is equivalent to a shotgun efficiency of 0.85%, similar to those reported in other ancient samples from the Iberian Peninsula (García-Garcerà et al. 2011; Sánchez-Quinto et al. 2012). When the alignment was carried out against the human genome, 60,488 reads were mapped, indicating that human contamination was ~0.34% (Figure S1), also in concordance on bone material handled by archaeologists (Ramírez et al. 2009; García-Garcerà et al. 2011). Nevertheless, only 1.68% of the reads that mapped in the pig genome also mapped in human (Figure S1). These reads are likely to originate from highly conserved regions, and therefore expected to show low levels of variability. Given that our analyses considered only SNPs also present in the pig modern samples and the implausibility of the same SNP appearing in two distant lineages, it is unlikely that human contamination affects the results reported here.

Although the DNA from the ancient sample was extracted in a dedicated ancient DNA laboratory (IBE-PRBB, Barcelona), where no previous work on pigs had been carried, there is still a small probability of contamination with other pig samples. We calibrated the possibility of this event by checking for heterozygote positions the mitochondrial sequence. We obtained a low depth-of-coverage (2.4×) in the ancient mtDNA genome and we found 41 heterozygote

positions; 21 (51%) of these were C/T or G/A changes that are likely attributable to postmortem damage. To determine whether the rest of heterozygote sites could be due to contamination from other pigs, and not to sequencing errors, we analyzed if these position are polymorphic in the panel of 41 complete mtDNA sequences used in this study. Only 6 out of 20 heterozygous positions were also segregating in at least one modern complete mtDNA sequence. The same analysis in a low depth-of-coverage mtDNA genome (1.9×) from a modern pig (Duroc) rendered very similar results, 37 heterozygote positions and 6 of these segregating in the panel of the 41 complete mtDNA. In all, it seems that contamination from other porcine samples is unlikely to bias the results presented here.

After alignment, 9% of the *Sus scrofa* 10.2 assembly was covered with average depth of 2x (equivalent to a genome wide average depth ~ 0.11×); the percentage of genome aligned was uniform across chromosomes except sex chromosome X (Figure S2). The pig sequenced was a sow, as evident from uniform depth along chromosome X, and equal to autosomal average (Figure S3). For polymorphism analyses, sites with depth over 5x were discarded in order to minimize artifacts caused by duplicated regions. To avoid spurious C to T and G to A substitutions attributable to post mortem DNA damage (Briggs et al. 2007), we retained ancient alleles only if also found in any of eight modern samples also analyzed (see methods, Table S1). Although this filtering will cause a bias by discarding true SNPs found in the ancient pig only, this bias is likely to be very small but is expected, in turn, to dramatically reduce the false discovery rate (see methods). In fact, this strategy allowed us to retrieve the same mutation profile as in modern samples (Figure S4). The numbers of polymorphic sites before and after filtering were 250,622 and 208,628, respectively, i.e., an estimation of post mortem damage of 16.7%. Note that this is an upper bound because some true SNPs are filtered out if not found in the modern samples (this percentage should be small, though, as shown in the methods section). This value was close to that found by PCR in an analysis of a fragment of mitochondrial cytochrome b gene MT-CYB (13.1%).

As for modern sequences, the numbers of reads were 347,750,566 (Guatemalan Creole), 342,150,846 (Iberian), and 375,306,190 (Spanish wild boar), resulting in average depths12-13x after filtering by base and map quality (Table S1).

Mitochondrial Phylogeography

Complete ancient mtDNA sequence was aligned with published sequences and the three modern samples sequenced in this study (Figure 1). As observed with shorter mtDNA fragments like the control region or cytochrome b, e.g., (Larson et al. 2005) European wild boar and domestic breed haplotypes were not split into distinct clades but were rather intermixed. Note also that some European domestic pigs harbor Asian haplotypes, as a result of Chinese introgression. As for the ancient pig, unsurprisingly, it is within the European clade. The nearest sequences (d = 0.00023 ± 0.00014) were found in black hairless Iberian Guadyerbas strain and in *Lampiño de Guadiana* Iberian strain (Figure 1). The nearest wild boar mtDNA sequences were from a Spanish wild boar (accession FJ237000, d=0.0061 ± 0.00024). Note that the Guatemalan Creole haplotype was clearly of Iberian origin as well and was positioned next to Iberian sequence FJ236995 (d=0.00068, H6 haplotype code from (Alves et al. 2003).

Worldwide Context Inferred from SNP Arrays

We combined the ancient sample genotypes with those from two porcine diversity panels, (Burgos-Paz et al. 2013; Manunza et al. 2013), both genotyped with the 60k porcine array (Ramos et al. 2009). These panels comprised 419 samples from 38 populations and breeds including wild boars of European, North African, and Near East origin, international breeds, local breeds from Asian and Europe and American Creole pigs. American Creole pigs are putative descendants of the ancient sample's relatives (Table S2). A total of 4,090 autosomal SNPs from the 60k array could be retrieved from the ancient sample.

First, to position the ancient sample and to investigate whether a Near East legacy could still be detected, we ran an unsupervised Admixture (Alexander et al. 2009) analysis excluding the Creole pigs, well known to have been admixed. Preliminary analyses suggested K = 12 as the optimum number of components. Results with this K value (Figure 2) suggest that the NE component is completely absent from the ancient sample. The admixture analysis strongly supports a 100% Iberian component to the ancient pig.

A Principal Component Analysis (PCA) of those SNPs (Figure 3) broadly agrees with the original analysis that included the complete SNP dataset from (Burgos-Paz et al. 2013; Manunza et al. 2013), showing that the 4,090 SNPs used here are a representative set - although always subject to SNP ascertainment. Figure 3 was drawn using a randomly sampled allele from each genotype, to match the fact that only one ancient allele is generally observed. As can be seen in Figure S5, sampling has a very small effect in the PC projection.

The first principal component (PC1) explains a much larger fraction of variance (17.7%) than the second axis (3.7%); and that PC1 axis is primarily geographical, separating Asian from European populations. The Near East (NE) wild boars are closer to European than to Asian pigs, and NE genetic structure grossly coincides with their geographic origin. International breeds, well known to be admixed with Chinese pigs (Giuffra et al. 2000), are closer to Chinese pigs than are Iberian pigs, not known to have been admixed. Also, all Creole populations show evidence of admixture as found in Burgos-Paz et al (2013). In the PCA plot, the ancient sample was located within the modern Iberian pig cluster; and it does not show evidence of Asian admixing either.

Overall, PC based distances showed very similar values between Iberian pigs, of both ancient and modern origin, and American Creole populations (Figure S6). As in Burgos-Paz et al (2013), we found that Yucatan minipigs (originally from Mexico), Peruvian and some North Argentinean village pigs were the closest populations to both the ancient and the Iberian pigs. To confirm this and as a complement to PC-based distances, we computed average pairwise allele differences between the ancient pig and Creole pigs (Figure 4). For comparison, those with the Iberian and Duroc breeds are also shown. Given that only one ancient allele can be recovered for most positions, genotypic distances can be approximated under different assumptions (methods). Assuming that 'true' frequencies in the unobserved ancient allele can be approximated by average frequencies across porcine populations, divergence is consistently higher between the ancient and Creole populations than with Iberian pigs (Figure 4). However, if we take Iberian pig as a better proxy for ancient frequencies, the

predicted allele differences between the ancient and Creole populations are very similar to those between Creole and modern Iberian pigs.

Genomewide Analysis

To gain a more faithful view of genetic relationships than with ascertained array SNPs and to extend the study beyond the mitochondrial lineage, the complete ancient sequence available was combined with eight additional modern sequences (Table S1). Three of these are new samples that were sequenced for this study: an Iberian pig from the Guadyerbas strain (Toro et al. 2008), a Creole pig from Guatemala and a Spanish wild boar from the same region as the ancient sample. The five other samples are public sequences, and comprised one sample from each of the most widespread international breeds: Duroc, Landrace, Large White, Hampshire and Pietrain. After SNP calling and filtering (see methods) we retained 794,514 autosomal SNPs without any missing value across samples. Figure 5 shows the PCA and a neighbor-joining tree with distances between samples. The figures show the result of random sampling one of the two alleles in the modern samples; for comparison, other replicates are shown but the effect of allele sampling was, again, negligible (Figure S7). The PCA (Figure 5) has the first axis bounded by the wild boar and Large White, which is the international breed with the largest Chinese component. The second axis primarily explains divergence with Duroc. In agreement with the array SNPs (Figures2, 3) and mtDNA data (Figure 1), the ancient sample is closest to the Iberian pig and wild boar.

We computed autosomal divergence (% of allele differences) between the ancient and the eight modern sequences (Table S3), which again shows that the Iberian pig is the closest sample to the ancient pig, followed by Spanish wild boar, Hampshire and Creole. The length of ancient homozygous stretches (IBS blocks) shared with the Iberian was also the largest, followed at distance by wild boar. All other samples, including Creole pig, were less similar to the ancient pig. Use of other publicly available sequences from European wild boar led consistently to similar results (not presented).

Mitochondrial and genomic data (Figures1 and 2, and Table S3) suggest, as the archaeological data, that the ancient pig is domestic. The ancient pig, though, is also close to wild boar (Figure 5). To test this, we identified 24 positions in the ancient genome that were among the 227 SNPs described by (Rubin et al. 2012) as highly differentiated between wild boar and domestic pigs. For those positions, we also determined the genotypes from a subset of sequenced modern animals and we computed the probability that the sample originates from either wild boar or domestic. Results (Table S4) indicate that the ancient sample is much more likely to be a domestic pig than a wild boar ($P_D = 0.72 \pm 0.07$ vs. $P_W = 0.27 \pm 0.07$). These probabilities are comparable to other domestic pigs (Duroc, Large White, Creole), and somewhat higher than the Iberian pigs ($P_D = 0.65 \pm 0.05$). As control, note that wild boar probabilities are reversed ($P_D = 0.32 \pm 0.03$ and $P_W = 0.70 \pm 0.04$).

Despite genetic differentiation between wild boar and domestics, wild boar admixing with domestic pigs has been repeatedly suggested, based both on genetic and historical evidence (Thomas 2005; Ramírez et al. 2009). The availability of an animal from five centuries ago may help in resolving whether this admixing occurred predominantly within the last centuries or predate that time. To investigate this, we applied the D-statistics as implemented in

ADMIXtools (Patterson et al. 2012). This statistic was first used by Green et al. (2010) to detect admixing between human and Neanderthal genomes, and is very powerful to detect admixture between ancient populations, even if they are closely related. The results strongly suggest admixing, both in the ancient and in the modern Iberian pig (Table S5). Results were very similar when the wild boar was from Spain or from France. Taken together, the D-statistics suggest gene flow levels of equal intensity between wild boar and both the ancient and Iberian pigs, but that admixing did not occur frequently enough to wipe out genetic differences between them, as shown by the discriminant SNPs in Table S4 and genetic distances in Table S3.

Potential Selection Signatures

A comprehensive catalogue of functional mutations in the pig - or in any other livestock species - is still in its early infancy. Nevertheless, indirect methods based on excess of differentiation or of homozygosis can be used to detect positive selection (Amaral et al. 2011; Rubin et al. 2012). Since excess of homozygosity cannot be quantified within the ancient sample, we looked for regions of extreme differentiation using the F-statistic (F_{ST}). We computed F_{ST} across regions where the ancient pig was sequenced and using four additional samples from both wild boar and Iberian breed (methods). We focused on the closest populations to the ancient sample in order to avoid as much as possible distortions caused by Asian introgression. Here we discuss some of the most biologically relevant genes and their gene ontologies.

We sought to investigate specific changes that may have occurred early in domestication by selecting windows and genes showing extreme differentiation in wild boar vs. the ancient and Iberian breeds (WB vs. ANIB); 157 genes were within the regions of *F*st larger than genome-average plus 3 SD (methods). Interestingly, after the Bonferroni correction for multiple testing, we observed a significant enrichment in genes related with "carbohydrate metabolic process" (P = 0.0031) and "disaccharide metabolic process" (P = 0.0013). Three important genes related with galactosidase activity were identified (*Q0Q237, Glb1l2* and *Glb1l3*). Among the top most differentiated regions analysis (Table S6) we found the Follicle-stimulating hormone (*FHSB*) and Tyrosinase-related protein 1 (*TYRP1*) genes. Other interesting genes that also appeared in our top 100 most differentiated regions analysis (Table S6) are v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (*KIT*) and Cryptochrome 2 (*CRY2*).

DISCUSSION

Ancient genomic data is needed to resolve the intricacies in the history of domestic species and to characterize the timing of selective events occurred between domestication and the modern breeding era. Here, we provide genome data from a female pig that lived in the Iberian Peninsula during the mid 16th century, before Asian introgression and contemporary to the beginnings of American colonization. Despite the shallow coverage attained, extensive comparison with modern genome sequence and with a large genotyped diversity panel allows us to draw relevant conclusions concerning pig genetic history.

We did not find any evidence of Near East wild boar legacy in the ancient sample, although this might be possibly due to the low resolution attained with the SNP array; a complete

genome could resolve this matter. In contrast, it seems clear that the 16th century pig was domestic (Tables S3, S4, Figures 1 and 2). This was not perhaps unexpected, given that the sample was chosen to avoid sampling a wild boar as much as possible, but it is reassuring that genetic data reflect this. It agrees as well with the fact that animal, and specifically pig breeding, was an important activity in Montsoriu castle (Novella 2013). All data suggest a close relationship to extant Iberian pigs; this is interesting as it demonstrates that the Iberian pigs, at least the traditional strains analyzed here, have not been admixed with Asian pigs. The next closest population to the ancient pig was European wild boar (Table S3), indicating a low differentiation between wild boar and Iberian pigs and in agreement with previous works (e.g., Ramírez et al. 2009). Our data also suggest that admixing between wild boar and both ancient and Iberian pigs have occurred, and that admixing levels were very similar in either the Iberian or the ancient sample (Table S5). Note, though, that the degree of admixing was not enough to wipe out differences between wild and domestic pigs. We found a larger number of haplotype blocks shared between ancient and Iberian than between ancient and wild boar (1351 vs. 865, Table S3); it can be tentatively hypothesized, then, that gene flow wild-boar domestic occurred primarily before the 16th century rather than during modern ages but more coverage is needed to resolve this definitely.

Among Iberian strains, the ancient sample seems directly related to extant black hairless strains, as follows from the mtDNA haplotype (Figure 1). Modern Iberian pigs are red or black, but never white. Unsurprisingly, we found no evidence of the KIT gene (ENSSSCG00000008842) duplication, which is responsible for the white color (Giuffra et al. 2002): in the ancient sample, 5,126 bp within the bounds of KIT gene (SSC8:43550236-43602062) were covered with average depth 1.02, almost identical to the average depth in that chromosome (1.07). In contrast, depth in the KIT gene for the Large White sample, known to carry the duplicated gene, was 20.02 or about twice average depth along SSC8 (10.41), whereas depth in the Iberian sample, which has a single copy of the gene, was the same in the KIT gene and along SSC8, 11.97 and 13.03. Figure S7 shows the distinct patterns in an individual with and without the duplication, and the plots strongly suggest that the ancient sample lacks the duplication. Furthermore, all historical depictions from Iberian pigs in the epoch shows predominantly black pigs and none white (Martín-Rivas 2012). Unfortunately, there were no reads aligned to the MC1R gene in the ancient pig, which would have allowed us to confirm either red or black coat color, but all evidence points to a non white individual.

Pig introduction from Spain in the Americas started with Columbus' second trip, in 1493 (Rodero et al. 1992; Zadik 2005), and pigs adapted quickly to the new environments (Elliot 2007). A matter of historical interest, therefore, is to disclose whether American Creole pigs are more related to the ancient sample than to modern Iberian pigs. The availability of a contemporary pig from the initial American colonization period helps to illuminate this issue. However, our genotypic (Figures 3, 4 and S6) and sequence data (Figure 3, Table S3) in fact show that the ancient pig and modern Iberian pigs are equally close to American Creole pigs. This suggests, as pointed out in our previous work (Burgos-Paz et al. 2013), that American Creole pigs have lost much of its Iberian origin by admixing with other breeds.

Genome wide data from an ancient pig, prior to modern intense selection for lean and growth traits, also provides us with an opportunity to understand selection at the gene level and separate them from those brought about by domestication and by Asian introgression. Among the highly differentiated genes between the ancient and Iberian vs. wild boar, we found genes related to coat color (*KIT, TYRP1*) and to reproductive performance (Galactose metabolism, *FHSB, CRY2*), in all likelihood among the first traits selected during domestication (Osadchuk 2006; Larson and Burger 2013; Linderholm and Larson 2013). Galactose is involved in lactation processes that are important for a sow ability to mother or nurse her young. *FHSB* enables ovarian folliculo genesis to the antral follicle stage and is essential for Sertoli cell proliferation and maintenance of sperm quality in the testis (Fan and Hendrickson 2005). *TYRP1* is involved in the synthesis of Eu-Melanin (Singh et al. 2013) and a mutation in this gene is associated with the brown color in some pigs (Ren et al. 2011). *KIT* plays key roles in melanogenesis, erythropoiesis, spermatogenesis and T-cell differentiation (Mithraprabhu and Loveland 2009). In pigs, several alleles of this gene were associated with different color variants (Fontanesi et al. 2010). CRY2, has a critical role in tuning the circadian period that regulate the seasonal breeding (Ye et al. 2011). While wild boar is a short-day which mates annually during the transition period from autumn to winter (Mauget 1982), domestic pigs almost completely lack breeding seasonality.

Here we provide the first - to our knowledge - genome-wide data from an ancient domestic pig. More ancient genomes from different epochs and geographic areas will be needed to validate the results presented here, among them, the timing and extent of admixing with the wild boar. More data will also help to clarify the selective events that have occurred from domestication until the creation of modern breeds and those ongoing as a result of current industrial selection programs. Our data suggest that most of divergence between the ancient pig and modern international pig breeds is caused by Asian introgression rather than by selection itself, because the divergence among modern Iberian pigs, wild boar and the ancient sample is much smaller than with breeds known to have been admixed with Asian germplasm.

DATA ARCHIVING

Ancient, Iberian, Spanish wild boar and Guatemalan reads have been submitted to SRA (accession XXX), aligned mitochondrial fasta file, plink files with genotypic data have been deposited in Dryad (accession XXXX).

ACKNOWLEDGMENTS

We thank J. Tura, J. Mateu, G. Font, S. Pujadas, J.M. Llorens and *MuseuEtnològic del Montseny* for the ancient sample, A. Luarca and E. Caal for providing the Guatemalan Creole sample, L. Silió and M.C. Rodríguez for the Iberian sample, M. Amills for genotype data in Manunza et al. (2013), L.A. Frantz for discussions on the D statistic, and D.A. Hughes for proof correcting the English. Archaeological excavation funded by *Direcció General d'Arxius, Biblioteques, Museus i Patrimoni and Museu d'Arqueologia de Catalunya*. Work funded by 2010ACOM00030 (AGAUR) to MS, FEDER and BFU2012-34157 grant (Spain) to CLF, Consolider CSD2007-00036 “Centre for Research in Agrigenomics” and AGL2010-14822 grants (Spain) to MPE. OR is a postdoctoral Researcher from the JAEDOC program cofounded by ESF, EB is recipient of a FPI grant from ministry of Research (Spain), IO of a predoctoral fellowship from the Basque

Government (DEUI, Spain), WBP is funded by COLCIENCIAS (*Francisco José de Caldas* fellowship 497/2009, Colombia).

REFERENCES

Albarella U, Dobney K, Rowley-Conwy P. 2009. Size and shape of the Eurasian wild boar (*Sus scrofa*), with a view to the reconstruction of its Holocene history. *Environ Archaeol* 14: 103

Albarella U. 1997. Size , power , wool and veal : zooarchaeological evidence for late medieval innovations. In: De Boe G, Verhaegue F, editors. Environment and Subsistence in Medieval Europe. Vol. 9. IAP Rapporten. p. 19–30.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19: 1655–1664.

Alves E, Ovilo C, Rodríguez MC, Silió L. 2003. Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Anim Genet* 34: 319–324.

Amaral AJ, Ferretti L, Megens H-J, Crooijmans RPM a, Nie H, Ramos-Onsins SE, Perez-Enciso M, Schook LB, Groenen M a M. 2011. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One* 6: e14782.

Briggs AW, Stenzel U, Johnson PLF, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* 104: 14616–14621.

Burgos-Paz W, Souza CA, Megens HJ, et al. 2013. Porcine colonization of the Americas: a 60k SNP story. *Heredity* 110: 321–330.

Cheng R, Qiu J, Zhou X-Y, et al. 2011. Knockdown of STEAP4 inhibits insulin-stimulated glucose transport and GLUT4 translocation via attenuated phosphorylation of Akt, independent of the effects of EEA1. *Mol Med Rep* 4: 519–523.

Conolly J, Colledge S, Dobney K, Vigne J-D, Peters J, Stopp B, Manning K, Shennan S. 2011. Meta-analysis of zooarchaeological data from SW Asia and SE Europe provides insight into the origins and spread of animal husbandry. *J Archaeol Sci* 38: 538–545

Von Den Driesch A. 1976. A guide to the measurement of animal bones from archaeological sites. Havard: Harvard University Press

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.

Elliot J. 2007. Empires of the Atlantic World: Britain and Spain in America 1492-1830. Yale University Press

Ervynck A, Lentacker A, Müldner G, Richards M, Dobney K. 2007. An investigation into the transition from forest dwelling pigs to farm animals in Medieval Flanders, Belgium. In: U. Albarella, K. M. Dobney, A. Ervynvk PR-C, editor. Pigs and Humans:10,000 years of interaction. Oxford: Oxford University Press. p. 171–196.

Esteve-Codina a, Kofler R, Himmelbauer H, Ferretti L, Vivancos a P, Groenen M a M, Folch JM, Rodríguez MC, Pérez-Enciso M. 2011. Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. *Heredity* 107:256–264.

Esteve-Codina A, Paudel Y, Ferretti L, et al. 2013. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. *BMC Genomics* 14:148.

Fan QR, Hendrickson W a. 2005. Structure of human follicle-stimulating hormone in complex with its receptor. *Nature* 433: 269–277.

Fang X, Mou Y, Huang Z, et al. 2012. The sequence and analysis of a Chinese pig genome. *Gigascience* 1: 16.

Font G, Llorens J, Mateu J, Pujadas S. 2010. L'abandonament del castell de Montsoriu. Segles XVI-XVII. *Monogr del Montseny* 18:207–215.

Font G, Mateu J, Pujadas S, Tura J. 2008. Síntesi històrica del castell de Montsoriu. *Monogr del Montseny* 23: 109–134.

Fontanesi L, D'Alessandro E, Scotti E, Liotta L, Crovetti a, Chiofalo V, Russo V. 2010. Genetic heterogeneity and selection signature at the KIT gene in pigs showing different coat colours and patterns. Anim Genet 41: 478–492.

García-Garcerà M, Gigli E, Sanchez-Quinto F, Ramirez O, Calafell F, Civit S, Lalueza-Fox C. 2011. Fragmentation of Contaminant and Endogenous DNA in Ancient Samples Determined by Shotgun Sequencing; Prospects for Human Palaeogenomics. *PLoS One* 6: 7.

Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT, Andersson L. 2000. The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* 154: 1785–1791.

Giuffra E, Törnsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, Anderson SI, Archibald AL, Andersson L. 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. *Mamm Genome* 13: 569–577.

Green RE, Krause J, Briggs AW, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328: 710–722.

Groenen M a M, Archibald AL, Uenishi H, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491: 393–398.

Krause J, Dear PH, Pollack JL, et al. 2006. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 439: 724–727.

Krause-Kyora B, Makarewicz C, Evin A, et al. 2013. Use of domesticated pigs by Mesolithic hunter-gatherers in northwestern Europe. *Nat Commun* 4: 1–7.

Lalueza-Fox C, Römpler H, Caramelli D, et al. 2007. A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science* 318: 1453–1455.

Larson G, Albarella U, Dobney K, et al. 2007. Ancient DNA , pig domestication , and the spread of the Neolithic into Europe. *Proc Natl Acad Sci USA* 104: 15276–15281.

Larson G, Burger J. 2013. A population genetics view of animal domestication. *Trends Genet* 29: 197–205.

Larson G, Dobney K, Albarella U, et al. 2005. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* 307: 1618–1621.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

748 Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA
749 polymorphism data. *Bioinformatics* 25: 1451–1452.
750 Linderholm A, Larson G. 2013. The role of humans in facilitating and sustaining coat colour
751 variation in domestic animals. *Semin Cell Dev Biol* 24: 587-593
752 Lindgreen S. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC*
753 *Res Notes* 5: 337.
754 MacKinnon M. 2004. Production and consumption of animals in Roman Italy: integrating the
755 zooarchaeological and textual evidence. *J Rom Archaeol Supplement*: 264.
756 Manunza A, Zidi A, Yeghoyan S, et al. 2013. A high throughput genotyping approach reveals
757 distinctive autosomal genetic signatures for European and Near Eastern wild boar. PLoS
758 One 8:e55891.
759 Martín-Rivas S. 2012. El cerdo ibérico y el Arte en España. Universidad Internacional de
760 Andalucía
761 Mauget R. 1982. Seasonality of reproduction in the wild boar. In: DJA C, Foxcroft G, editors.
762 Control of Pig Reproduction. London: Butterworths. p. 509–526.
763 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the
764 consequences of genomic variants with the Ensembl API and SNP Effect Predictor.
765 *Bioinformatics* 26: 2069–2070.
766 McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet*
767 5: e1000686.
768 Meiri M, Huchon D, Bar-Oz G, et al. 2013. Ancient DNA and population turnover in southern
769 levantine pigs--signature of the sea peoples migration? *Sci Rep* 3: 3035.
770 Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene
771 function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids*
772 *Res 41*: D377–86.
773 Mithraprabhu S, Loveland KL. 2009. Control of KIT signalling in male germ cells: what can we
774 learn from other systems? *Reproduction* 138: 743–757.
775 Novella V. 2013. La dieta avícola en el siglo XVI: conservación y consumo de aves en el Castillo
776 de Montsoriu (Montseny). In: Lopez B, editor. LA producción de alimentos. Arqueología,
777 historia y futuro de la dieta mediterranea. Murcia: Universidad de Mazarrón. p. 109–
778 119.
779 Osadchuk L V. 2006. Reproductive potential of male silver foxes Vulpes vulpes after long
780 selection for the domesticated behavior type. *J Evol Biochem Physiol* 42: 182–189.
781 Ottoni C , Flink LG, Evin A, Geörg C, De Cupere B, Van Neer W, Bartosiewicz L, Linderholm A,
782 Barnett R, Peters J et al. 2013. Pig domestication and human-mediated dispersal in
783 western Eurasia revealed through ancient DNA and geometric morphometrics. *Mol Biol*
784 *Evol* 30: 824-832.
785 Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich
786 D. 2012. Ancient admixture in human history. *Genetics* 192: 1065–1093.
787 Purcell S, Neale B, Todd-brown K, et al. 2007. PLINK : A Tool Set for Whole-Genome
788 Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81: 559–575.
789 R Development Core Team. 2011. R: A language and environment for statistical computing.
790 Vienna: Foundation for Statistical Computing

Ramírez Oscar, Gigli E, Bover P, Alcover JA, Bertranpetit J, Castresana J, Lalueza-Fox C. 2009. Paleogenomics in a temperate environment: shotgun sequencing from an extinct Mediterranean caprine. *PLoS One* 4: e5670.

Ramírez O, Ojeda A, Tomàs A, et al. 2009. Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol Biol Evol* 26: 2061–2072.

Ramos AM, Crooijmans RPM a, Affara N a, et al. 2009. 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.

Ren J, Mao H, Zhang Z, Xiao S, Ding N, Huang L. 2011. A 6-bp deletion in the TYRP1 gene causes the brown colouration phenotype in Chinese indigenous pigs. *Heredity* 106: 862–868.

Rodero A, Delgado J V, Rodero E. 1992. Primitive Anadalusian livestock and their implications in the discovery of America. *Zootecnia* 41: 384–400.

Rubin C-J, Megens H-J, Martinez Barrio A, et al. 2012. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci USA* 109: 19529–19536.

Sánchez-Quinto F, Schroeder H, Ramirez O, et al. 2012. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Curr Biol* 22:1494–1499.

Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid K a S, Willerslev E, Krogh A, Orlando L. 2012. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 13: 178.

Silver A. 1969. The ageing of domestic animals. In: Brothwell DR, Higgs ES, editors. Science in Archaeology. London: Thames and Hudson. p. 283–302.

Simonsen A, Lippe R, Gaullier J, Brech A, Callaghan J, Toh B, Murphy C, Zerial M. 1998. EEA1 links PI(3)K function to Rab5 regulation of endosome fusion. *Nature* 394: 2–6.

Singh S, Malhotra AG, Pandey A, Pandey KM. 2013. Computational model for pathway reconstruction to unravel the evolutionary significance of melanin synthesis. *Bioinformation* 9: 94-100.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Bol Evo*l 28: 2731–2739.

Thalmann O, Shapiro B, Cui P, et al. 2013. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science* 342: 871–874.

Thomas R, Holmes M, Morris J. 2013. “So bigge as bigge may be”: tracking size and shape change in domestic livestock in London (AD 1220–1900). *J Archaeol Sci* 40: 3309–3325.

Thomas R. 2005. Animals, economy and status: the integration of Zooarchaeological and historical evidence in the study of Dudley Castle, West Midlands (c.1100-1750). Archaeopre. Oxford: BAR British Series 392

Toro M a., Rodrigañez J, Silio L, Rodriguez C. 2008. Genealogical Analysis of a Closed Herd of Black Hairless Iberian Pigs. *Conserv Biol* 14: 1843–1851.

Wall JD, Slatkin M. 2012. Paleopopulation genetics. Annu Rev Genet 46: 635–649.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 81: 1278–1283.

Ye R, Selby CP, Ozturk N, Annayev Y, Sancar A. 2011. Biochemical analysis of the canonical model for the mammalian circadian clock. *J Biol Chem* 286: 25891–25902.

Zadik B. 2005. The Iberian Pig in Spain and the Americas at the time of Columbus. University of California.

Figure Legends

Figure 1: Complete mtDNA NJ tree. The upper clade corresponds to the Asian clade, with five sequences, and the bottom is the European clade. The first two letters represent the breed: AN, ancient; CR, Guatemalan Creole; DU, Duroc; HS, Hampshire; IB, Iberian; LR, Landrace; LW, Large White; PI, Pietrain; WB, wild boar, followed by the accession number. Samples AN, IB, CR and WB were those sequenced here. The eight samples with solid bullets were used to compare with the ancient sample (Table S1).

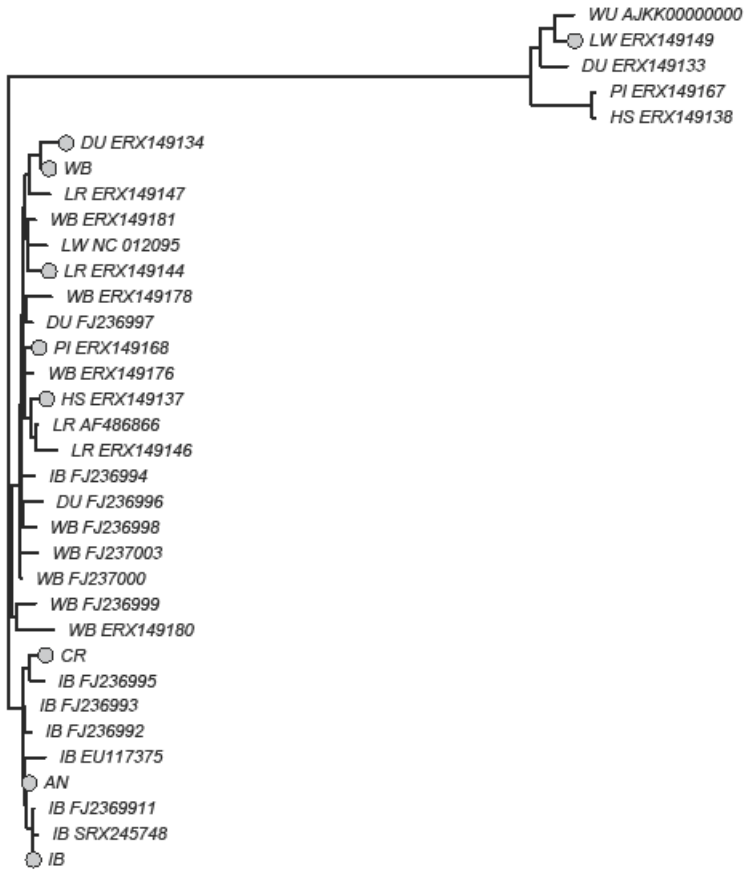
Figure 2: Unsupervised Admixture analysis using the 4090 SNPs recovered in the ancient (AN) sample. The breed codes are: MS, Meishan; XI, Xian;JH, Jinhua; JQ, Jiangquhai; WB, wild boar; IB, Iberian; AN, ancient sample; LR, Landrace; LW, Large White; DU, Duroc; HS, Hampshire. Data from (Burgos-Paz et al. 2013) and Manunza et al. (2013).

Figure 3: First and second principal component representation of the porcine diversity panel fully described in (Burgos-Paz et al. 2013) and in Manunza et al. (2013)using the 4090 SNPs recovered in the ancient sample. Populations are grouped by color. The breed codes are: AN, ancient; BI, Bisaro; CE, Central Cuba; CR, Creole; CU, Cuino; EA, East Cuba; FO, Formosa; FP, feral pig; GH, Guinea Hog; HL, Hairless; IB, Iberian; JH, Jinhua; JQ, Jiangquhai; LR, Landrace; LW, Large White; MO, Moura; MI, Misiones; MS, Meishan; MT, Monteiro; MUL, mulefoot; NI, Nilo; OB, Ossabaw; PU, Piau; SI, Black Sicilian; WB, wild boar; WE, West Cuba; XI, Xian; YU, Yucatan minipig.

Figure 4: Average allele differences across 4,090 SNPs between the ancient pig (AN), Iberian (IB), Duroc (DU) and American Creole populations. Population codes are as in Figure 3. As only one allele can be usually recovered from the ancient population, allele sharing was obtained assuming the general frequency of the allele across populations (AN) or taking the frequency from the Iberian population (AN_IB). For comparison, allele divergence with Iberian and Duroc are shown. Standard deviations, obtained by bootstrapping, are about 0.004.

Figure 5: Left: PCA using all autosomal positions recovered from sequence data. AN, ancient; CR, Creole; DU, Duroc; HS, Hampshire; IB, Iberian; LR, Landrace; LW, Large White; PI, Pietrain; WB, wild boar. Right: Neigbor-Joining tree using mdist function from plink. The figure represents one random sample of one allele per SNP for each modern sample.

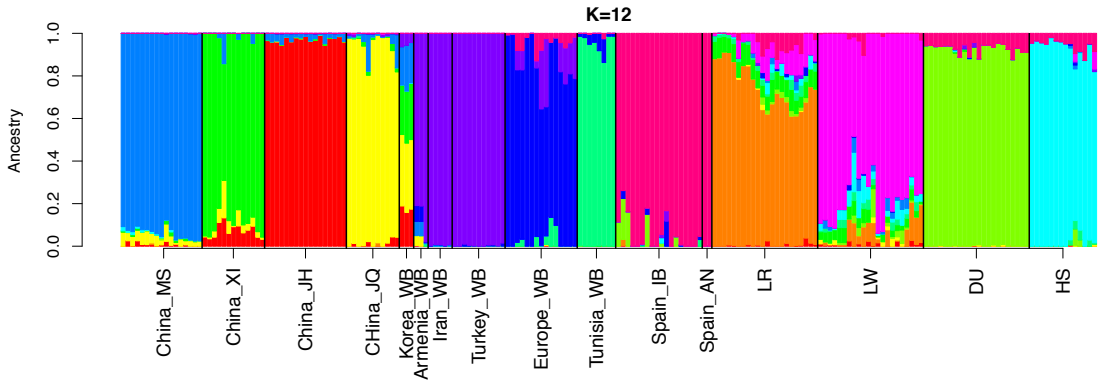
876 **Figure 1**



877

878 **Figure 2**

879

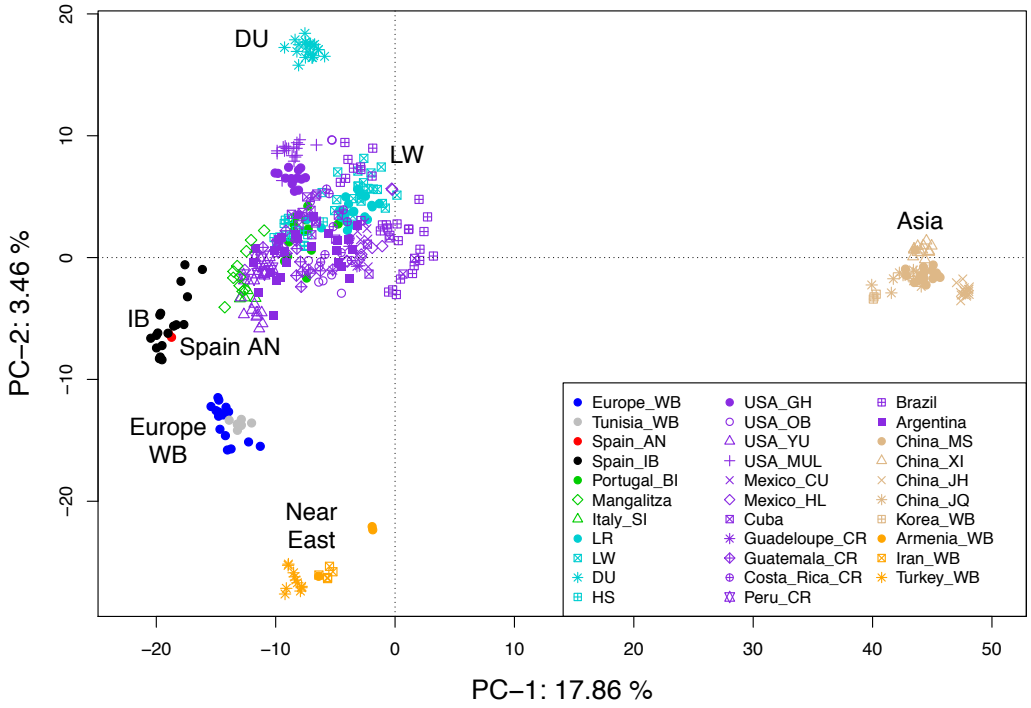


880

881

882

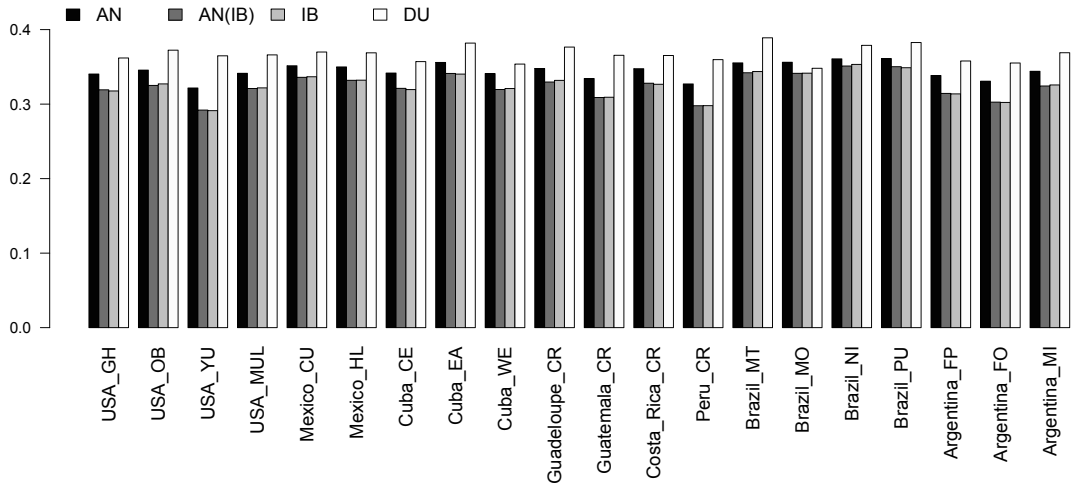
883 **Figure 3**

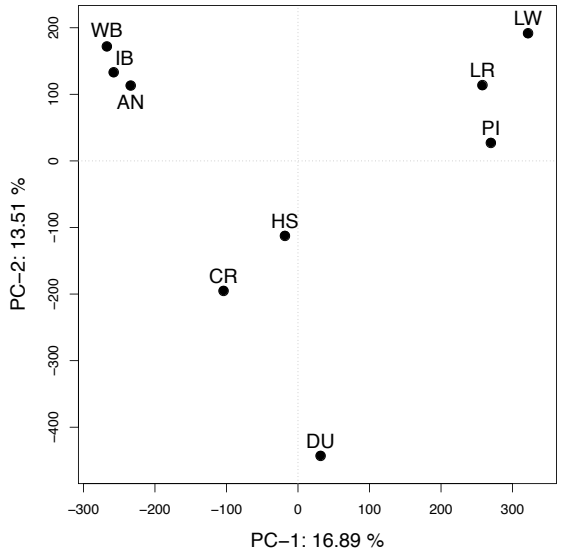


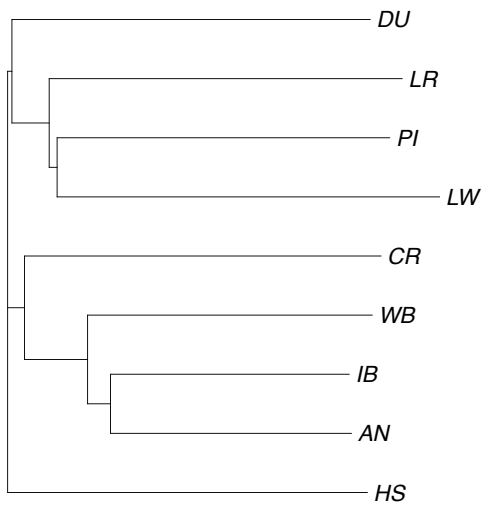
884

885

886


AN
AN(IB)
IB
DU
0.4
0.3
0.2
0.1
0.0
USA_GH
USA_OB
USA_YU
USA_MUL
Mexico_CU
Mexico_HL
Cuba_CE
Cuba_EA
Cuba_WE
Guadeloupe_CR
Guatemala_CR
Costa_Rica_CR
Peru_CR
Brazil_MT
Brazil_MO
Brazil_NI
Brazil_PU
Argentina_FP
Argentina_FO
Argentina_MI


WB
IB
AN
LW
LR
PI
HS
CR
DU
PC−2: 13.51 %
PC−1: 16.89 %
200
100
0
−100
−200
−300
−400
−300
−200
−100
0
100
200
300


DU
LR
PI
LW
CR
WB
IB
AN
HS

Chapter 6

-

Correcting for unequal sampling in Principal Component Analysis of genetic data

-

William Burgos-Paz [1,2], Sebastián E. Ramos-Onsins [1,2], Miguel Pérez-Enciso [1,2,3], Luca Ferretti [1,2]

[1] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, 08193 Bellaterra, Spain.
[2] Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.
[3] Institut Català de Recerca i Estudis Avancats (ICREA), Barcelona, Spain.

Manuscript in preparation

Correcting for unequal sampling in Principal Component Analysis of genetic data

William Burgos-Paz[1,2], Sebastián E. Ramos-Onsins[1], Miguel Pérez-Enciso[1,2,3], Luca Ferretti[1,2]

[1] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, 08193 Bellaterra, Spain

[2] Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

[3] Institut Català de Recerca i Estudis Avançats (ICREA), Carrer de Lluìs Companys 23, Barcelona, 08010, Spain

Correspondence

William Burgos-Paz: william.burgos@cragenomica.es

Sebastián E. Ramos-Onsins: sebastian.ramos@cragenomica.es

Miguel Pérez-Enciso: miguel.perez@uab.cat

Luca Ferretti**:** luca.ferretti@gmail.com

ABSTRACT

Principal component analysis (PCA) is one of the most widely used tools to explore variability of high dimensional data for population and quantitative genetics. Its popularity has recently increased due to the huge amount of molecular markers available in datasets worldwide. However, a common issue in population genetics is uneven sampling of populations due to external constraints, limiting the usefulness of PCA because of well-known sample size sensitivity and two-dimensional projection bias. Here we evaluated the use of weights in PCA (wPCA) for genetic data in order to correct for sampling bias, and proposed two ways to build weights according to differentiation of the populations evaluated. Simulations suggest that wPCA improves the two-dimensional projections of PCA data and, in some cases, recovers population relationships patterns even when sample size is as low as $n = 1$. Using real data sets we are able to recover a more realistic population structure than inferred with traditional PCA, and demonstrated the usefulness of wPCA in relating individual projections into a particular population.

Keywords
Covariance matrix, weight, single nucleotide polymorphism, graphical representation

INTRODUCTION

The Principal Component Analysis (PCA) is a widely used tool for intuitive visualization of the structure of a data set in many fields, including ecology and genetics. The idea behind this technique is to reduce the complexity of the data by retaining only the combinations of variables that describe most of the structure of the data set. Individual data is then projected onto the space of these combinations and visualized as a low-dimensional plot. Since visualization is often the primary interest of this method, PCA plots are usually two-dimensional or at most three-dimensional, therefore only the 2-3 most relevant combinations are usually retained.

The PCA is frequently applied to visualize the genetic structure of samples of natural populations. Since the first application in genetic data (Menozzi *et al.*, 1978), numerous reports have shown its usefulness in population genetics, especially in the detection of graphical patterns of population relationships, estimation of variability along geographical axes (Novembre and Stephens, 2008), or populations processes like migration and admixture (McVean, 2009). This technique allows also to discover the genetic origin of individuals by comparison with the position of known populations in a PCA plot. An example is the Sardinian origin of the Tyrolean Iceman found in the Italian Alps (Keller *et al.*, 2012) or, more recently, the La Braña 1 human sample, where PCA comparisons showed that the approximately 7,000-year-old individual was more related to extant northern European populations (Olalde *et al.*, 2014).

Popularity of PCA has notably increased in the last years because the large amount of molecular markers available (i.e SNP) in different species and populations. The PCA extracts the fundamental structure of the dataset (dimensionality reduction) in a computationally efficient manner (Patterson *et al.*, 2006; Paschou *et al.*, 2007). Despite several advantages of

PCA in the population genetics field, this technique is highly sensitive to sampling. In ecological and genetic studies, the data are usually sequences or genotypes of individuals sampled in the field. Sampling can follow different criteria depending on the aim of the study: it could be a complete representation of the variability in some phenotypes, a uniform sampling of geographical areas, it could be based on specific locations or on known population structure or ecological niches. In many studies, geography, ecology and phenotype are intertwined and play a joint role in the choice of the sample. No universal rule exists for an optimal sampling, and different sampling criteria result in a different PCA. In Figure 1 we show an example of four simulated populations with three phenotypes where PCA plots differ if individuals are sampled uniformly according to population structure (Figure 1b) or to phenotype (Figure 1c).

In practice, it is often not even possible to sample individuals according to some criteria chosen *a priori*. First, the choice of sampling could be biased by incomplete knowledge of the population structure or cryptic variations. For some species, there could also be ethical and conservation issues. Second, many factors like budget constraints, physical or political accessibility of geographical regions, availability of samples and technical problems in their conservation and sequencing limit sampling in the field. The distortion of the PCA plots due to a biased sampling is a known problem (Novembre and Stephens, 2008; McVean, 2009). In the example above, the PCA of an unbalanced sample from different populations is represented in figure 1d. The apparent representation of the data is quite different from the one that would result from a fair sampling (Figure 1b) and it is not informative about the actual structure of the data but rather about intertwine between the genetic structure and the sampling itself.

In this paper, we suggest a simple method to correct for unequal sampling. First we illustrate the method of weighted PCA (wPCA) and propose two simple forms for the weights. Then we discuss some examples of the effect of unequal sample size and we illustrate how the correction works for different standard population genetic models using coalescent simulations. Finally we show results from the application of wPCA to publicly available datasets from *C. elegans* population and human populations.

RESULTS

Unequal sampling and weighted PCA: Theoretical framework

In the context of population genetics, the basic data for PCA are allele frequencies obtained from genetic markers. Here we will focus on biallelic polymorphisms obtained by n sequences. These variants are most commonly single nucleotide polymorphisms (SNPs) obtained with microarrays or by direct sequencing. For sequences, the genotype for a given position can be indicated by 0 and 1 (depending on the derived allele being absent or present in the sequence), whereas for microarrays the genotype of each individual can be 0, 1 or 2, depending on the number of derived alleles in the individual in that position. We denote these data by x_s^i, where i indicates the sequence and s the SNP. Given a set of genetic data of n sequences/genotypes and n_{SNP} biallelic SNPs for each sequence, the first step of the PCA is to

compute the covariance $\boldsymbol{Cov}\,(\boldsymbol{x_s}, \boldsymbol{x_{s'}}) = \boldsymbol{E}(\boldsymbol{x_s}, \boldsymbol{x_{s'}}) - \boldsymbol{E}\,(\boldsymbol{x_s})\boldsymbol{E}(\boldsymbol{x_{s'}})$, which is usually computed as:

$$\boldsymbol{Cov}\,(\boldsymbol{x_s}, \boldsymbol{x_{s'}}) = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{ij}} \boldsymbol{x_s^i x_{s'}^j} - \left(\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i}} \boldsymbol{x_s^i}\right)\left(\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{j}} \boldsymbol{x_{s'}^j}\right) \quad \textbf{(1)}$$

Then the principal axes of the covariance matrix (i.e. the eigenvectors corresponding to the largest eigenvalues) are obtained by standard linear algebra and the data are centered and orthogonally projected on the subspace formed by these axes. The resulting low-dimensional representation of the data explains the highest fraction of the variance. For visualization purposes, only the first two axes are often plotted and used to interpret the data, however statistical tests can also be used to estimate the number of significant principal components to retain (Shriner, 2011).

PCA is therefore a projection from the high-dimensional space of the data to the most informative linear subspace. The structure of genetic data in this space is informative because the distance in the whole PCA space has a simple geometric interpretation in terms of evolutionary distances. In particular, the distance d between two sequences in PCA space is related to their Hamming distance d_H (Bornberg-Bauer, 1997; Li and Huang, 2007) by the relation $d = \sqrt{d_H}$ (Bornberg-Bauer and Chan, 1999). Under neutral evolution, the average Hamming distance is related to the splitting time t (Tang *et al.*, 2002) by the relation $d_H = 1 - e^{-2\mu t} \simeq 2\mu t$, where μ is the mutation rate ($\mu \ll 1$). Therefore, the distance is related to the splitting time

$$d \simeq \sqrt{2\mu t} \quad (2)$$

A similar relation exists for genotype data. Other connections between coalescence times and PCA have been proposed in (McVean, 2009).

A common assumption in equation 1 is that all individuals should be equally important in the PCA (that is, the weight of each individual is *1/n*). This is usually correct since there is often no prior information about the origin and grouping of individuals. However, this is not true for samples taken from structured populations such as natural populations in different habitats or geographical locations, since these individuals can usually be separated in different populations or subpopulations depending on the place where the sample was taken or on phenotypic traits.

There is no clear-cut answer about how the different populations or subpopulations should be weighted. However, as discussed before (Figure 1), the usual approach of equal weights for all individuals suffers from a clear bias due to the sampling process. If some populations have been extensively sampled while only a few individuals have been sampled from other populations (e.g. technical difficulties), the resulting plot will be mostly informative about the structure of the first populations but will say nothing or, worse, misinterpret the others.

To correct for this sampling bias, two ingredients are needed:

i. All individuals should be classified (tentatively, at least) into populations or subpopulations of interest

ii. The relative weight of each population should be chosen.

The classification of individuals into populations can be done either by expert knowledge, exploiting the phenotype or geographical/ecological information available for the samples, or in an automated way using clustering algorithms on the genetic data. However, there is not a unique definition of what a population is, since this depends on the ecological and genetic question studied, and both ways for classification described above have their own advantages and disadvantages. Expert knowledge can take into account the specific focus of the study, but it could be biased towards well-known populations or specific phenotypic traits, therefore overestimating some components of the genetic diversity. Moreover, clustering algorithms are unbiased with respect to the data, but they could not also resolve correctly a complex population structure resulting from migration, admixture and introgression events. For the rest of the paper, we assume that populations have been specified in a meaningful way and focus on the choice of the weights.

Since there is often no previous knowledge of the relation between populations, our initial suggestion is to *weight all the populations (or subpopulations) equally*. This choice has been proposed in the past in the form of resampling a smaller number of individuals for each population in order to reach equal representation (McVean, 2009). This is undesirable in terms of reduced amount of data, particularly if some populations have been poorly sampled and only a few individuals are available for the analysis.

The natural framework for this correction is the use of weights in PCA (Kriegel *et al.*, 2008) where each sequence can be assigned a different weight $\boldsymbol{w}_i$. In particular, the covariance can be rewritten naturally as

$$\boldsymbol{Cov}\,(\boldsymbol{x_s}, \boldsymbol{x_{s'}}) = \sum_i \boldsymbol{w_i x_s^i x_{s'}^i} - \left(\sum_i \boldsymbol{w_i x_s^i}\right)\left(\sum_j \boldsymbol{w_j x_{s'}^j}\right) \quad (3)$$

where the weights are numbers between $0 < \boldsymbol{w_i} < 1$ with $\sum_i \boldsymbol{w_i} = \boldsymbol{1}$. This reduces to the usual PCA when $w_i = {}^{1}/_{n}$.

We propose here an analogous strategy to inform in the covariance matrix the relevance of the observed data in presence of unequal sample size for different populations. For this, we denote the number of populations by *npop* and the number of sequences in population A by n_A. Since the weight of population A is the sum of the weights of all the sequences belonging to A, the corrected weight for each sequence of A is

$$\boldsymbol{w_i} = \frac{\boldsymbol{1}}{\boldsymbol{n_{pop}}}.\frac{\boldsymbol{1}}{\boldsymbol{n_A}},\ \boldsymbol{i} \in \boldsymbol{A} \quad (4)$$

Overcorrection for small sample sizes

Despite the wPCA correction, a residual bias due to sampling persists. In fact, for a population with only 1 or 2 individuals sampled, the internal variability of the population cannot be represented correctly even after the weighting, since this variability has not been sampled at all.

More in general, very small sample sizes do not usually represent the genetic background of the population. The effect on the wPCA is an overcorrection that tends to distort the PCA plot if weighted according to eq. 4. The reason for this effect lies in the misclassification of intrapopulation and interpopulation variability, which is inherent to small samples. Related to that, we observe that wPCA not only recovered the most realistic population structure projected in PC axes, additionally, the variance proportion explained in the firsts PC also increased (see Figure 3). Therefore, we explored other genetically meaningful weights for covariance matrix estimation and correct for the uneven sample. Here, the genetic differentiation of populations F_{ST} could be a useful and simple estimator that could potentially decrease the overcorrection caused by $w = {}^{1}/_{n}$ used previously.

In the case of well-differentiated populations (high F_{ST} values) the overcorrection is small because most of the internal variability comes from differences between populations. However, for low F_{ST} most of the internal variability is shared between populations and the differentiation comes from differences in allele frequencies observed. Consider a heterozygosity π_{int} within the population and π_{diff} between populations, with a set of *npop* populations. This population should contribute a component $\propto \pi_{int} + \pi_{diff}$ to the total variance, of which only the component π_{diff} will contribute to the structure of the PCA. However, if only n_A individuals are sampled from the population, the sample heterozygosity will be reduced by a factor $f_s < 1$ with respect to the population heterozygosity π_{int}. Since this reduction cannot be recovered by the wPCA, in order to balance the contributions to the total variance, the weight Ω of the population should be reduced proportionally to the reduction of its contribution to the total variance:

$$\Omega/\left(\pi_{int} + f_s\,\pi_{diff}\right) \simeq {}^{1}/_{n_{pop}}\ /\left(\pi_{int} + \pi_{diff}\right) \tag{5}$$

Therefore the weight of the population should be

$$\Omega \simeq \frac{1}{n_{pop}}\left[F_{ST} + f_s(1 - F_{ST})\right] \tag{6}$$

And for an individual, the weight should be

$$w \simeq \frac{1}{n_{pop}\cdot n_A}\left[F_{ST} + f_s(1 - F_{ST})\right] \tag{7}$$

where $F_{ST} \simeq \pi_{diff}/\left(\pi_{int} + \pi_{diff}\right)$. If the sampling inside the population is representative, we have $f_s = 1 - {}^{1}/_{n_A}$. We denote by wPCA-F_{ST} the PCA corrected according to the factors in equation (7).

We will now evaluate the behavior of the wPCA in presence of commonly models found in population genetics.

The effect of unequal sampling on PCA

Distortions in the PCA 2-dimensional plot can appear already with three populations. Consider the simple scenario of three populations *A, B* and *C* (Figure 2a): population *A* splits from *B* at a past time *T* (in number of generations from the present), then population *C* splits from *B* more recently at time t from the present. We assume that there is no migration between populations and that the substitution rate per base μ is constant in time and equal for all populations, so the genetic distances are $d_{AB} = d_{AC} = 2\mu LT$ and $d_{BC} = 2\mu Lt$ respectively, where L is the length of the sequence. We also assume that there is some internal variability (per base) of size $\pi_A \sim \pi_B \sim \pi_C \sim \pi \ll \mu t$.

In this model, given that the internal substructure is negligible ($\pi \ll \mu T, \mu t$), the three populations *A, B, C* can be well approximated by three points in the PCA space, therefore should lie in an isosceles triangle in the plane formed by the two principal axes (Figure 2b). Basic geometry tells that the height of the triangle is $\sqrt{2\mu L(T-t)}$ and the base is $\sqrt{2\mu Lt}$.

We consider the PCA plot given by the first two principal axes for a sample composed by $n_B = n_C$ individuals from *B* and *C* and a different number of individuals n_A from *A*. The PCA plot could change abruptly because of unbalanced sampling. In particular, if *A* is sampled much less than other populations, i.e. $n_A < \frac{2\pi_B}{4\mu(T-t)-\pi_B} n_B$, then the first axis of the PCA becomes the *BC* direction and the second axis is a combination of the internal variability of *B* and *C*, while *A* remains in the middle of the plot (Figure 2c). The reason is that the weight of the small component associated with *B,C* internal variability is enhanced by a factor n_B/n_A with respect to the larger component associated with *A-BC* differentiation.

Therefore, in this example, a strong unbalance between n_A and n_B, n_C will cause a strong change in the arrangement of the populations on the PCA plot. This change has a strong effect on the interpretation of the plot. In fact, according to the sample projections, one could think of *A* as a mixture of *B* and *C*, although *A* is an independent population that is actually an outgroup for *B,C*. Similarly, individuals that are admixtures between *B* and *C* could be easily misclassified as belonging to *A*. Alternatively, if *A* is sampled much more than the other populations, i.e. for $n_A > \frac{\mu t}{\pi_A} n_B$, the first axis remains the one between *A* and the other two populations which collapse to the same point, while the second axis is dominated by the internal variability of *A* (Figure 2d). Similar results would be obtained for a strong between n_C and n_A, n_B.

To illustrate this, we performed coalescence simulations of three populations with 20 individuals following the model in Figure 2a. The effective population size was similar in the populations. Then we simulated a reduction in the sample size was performed and only one individual in pop2 was evaluated. As expected, a triangle-like projection for the PCA of simulated data (Figure 3) was obtained. The population with lowest sample size (pop2) was

shrunken towards the outgroup population (pop1) but, always, in the middle of the plot. For pop1, sparse points were obtained mainly related with higher variability. When wPCA was used, we recovered the expected triangle-like relationships, although an increase in variance explained in two axes was also found likely caused by overcorrection mentioned previously.

Thus for, we have only considered particular cases of low complexity of population structure where wPCA seems to have important contributions to represent in a more realistic way the demography of populations.

wPCA and low sample size: Application in simulated datasets

We validate the benefits of the use of weights in PCA by simulation to address in a systematic way the effect of wPCA in populations considering demographic structure and migration. Six populations in three different migration scenarios (IM = Island model; MM = hierarchical structure and SS = Stepping Stone) were simulated to test the applicability of wPCA in presence of sample size reductions. Both, inverse population size and wPCA-F_{ST} weights were used considering the population of origin as grouping criterion. However, similar results were observed in the sample projections and only the results for inverse sample size weight are showed. As expected, the PCA projections for the entire simulated data set could be easily associated to the simulated model and used to suggest graphical relationships among populations (Figure 4).

We performed a systematic sample size reduction in one or three populations for each simulated model ranged from n = 1 to n = 75 whereas for the others was n = 100, then we compared the observed projections with PCA and wPCA. In agreement with McVean, (2009), the distortion in PCA projections was noticeable even though, e.g. for IM, the sample size of the poorly sampled populations was n = 75 (Figure 5). The PCA projections obtained for each model did not allow reconstructing the expected relationships compared to entire data set, especially for IM and MM. In fact, 2-dimensional plots showed the pattern observed in previous section i.e. populations with low representation in the sample are projected in the middle of the plane.

The use of sample size information as weight in the wPCA resulted in a better performance in the sample projections for almost all simulated scenarios in comparison with the observed in the PCA. The sample projections of MM (Figure 5, middle) and SS (Figure 5, right) were the closest to the expected; interestingly, wPCA in the MM model can recover the most realistic true structure even if the lowest sample size was n = 1 (Figure 5). The PCA and wPCA projections had the worst behavior in IM (Figure 5, left). The inherent complexity of population relationships in the model alongside low differentiation can produce misleading interpretations. However, comparison of PCA and wPCA projections showed that the latter required lower sample sizes to obtain the expected projection of data. The sample size reduction in one population resulted in remarkable distortions for all models evaluated here, and especially in those where populations show a low genetic differentiation (e.g IM). In this case, the information of allelic frequencies is not enough to discriminate their real position regarding the well-represented populations. Therefore the contribution in the PC variance is

low. Meanwhile, the effect of sample size is lower in well-structured populations (e.g. MM and SS models), although they are not exempt of misinterpretations.

Robustness of wPCA in 2-dimensional sample projection

The distribution of the projection in PC axes is highly dependent of the sample used (i.e. allele frequencies of the individuals). In order to evaluate the robustness of wPCA estimations for each simulated scenario we generated 1000 random data sets for each sample size reduction. For each dataset we calculated the mean of Euclidean distances in the projection (***d***) within and between populations and estimate the difference with the ***d*** estimated for the entire simulated population. Values closer to zero suggest a projection closer to the obtained from entire dataset.

In all scenarios, PCA projections showed very similar d values among populations through different sample sizes, and ***d*** was close to the expected when the simple size of poorly sampled population was higher than $n = 50$ (Figure 6). In contrast, ***d*** values in wPCA were closer to the expected even if lower sample size was $n = 1$ or $n = 3$ for MM and SS respectively, thus, corroborating the better behavior with respect to PCA. As observed in (Figure 6), IM model showed large differences in the sample projections. The distribution of Euclidean distances calculated from wPCA projections had higher variance compared to PCA, only for IM model when sample sizes was lower than $n = 7$.

The PC projections are highly associated with the samples used. In both PCA and wPCA, projections can change in terms of axis (positive or negative) and ***d*** between sample points, mainly due to how similar allele frequencies are. However, the use of inverse of sample size as weight of PCA result in improved projections of data, and better interpretations of population relatedness from the graphical representation.

Applications of wPCA to real population data

To further assess the usefulness of wPCA we apply this approach to three real genotype datasets publicly available and containing populations with low sample representation. The SNP genotypes dataset included *C. elegans* from Andersen *et al.*, (2012), human populations from Li *et al.*, (2008) and the ancient sample La Braña 1 from Olalde *et al.*, (2014). We performed PCA and wPCA estimations using both weights $w = {}^{1}/_{n}$ and wPCA-F_{ST} described here to show the advantages (and pitfalls) of both.

wPCA of worldwide *C. elegans* populations

The genetic diversity of a worldwide collection of 200 wild strains of *C. elegans* was analyzed using a dataset of 41,188 SNPs derived from RAD sequencing. This collection represents one of the most comprehensive survey of *C. elegans* used to date for studying the genetic relationships of populations in this species. The initial results obtained by Andersen *et al.*, (2012) suggested a low genetic variation and non well-established population subdivision of *C. elegans.* In fact, PCA projections not allowed a clear distinction of the populations despite the large variability harbored showed by the Pacific Rim and Hawaii strains.

Because the aim is to compare the graphical representation of data using PCA or wPCA considering the sample size, we followed the methodology proposed in Andersen *et al.*, (2012) selecting only the so-called "Isotypes" and the SNPs according to MAF and linkage disequilibrium. The PCA projection showed very little differences to the original (Supplementary Figure. 3 in Andersen *et al.*, (2012)), but general representation and population distribution was properly projected (Figure 7, left). Then, we performed the wPCA estimations using the inverse population sample size and the F_{ST} as weighting approaches. The grouping criterion was the geographical origin of samples. The wPCA showed similar sample projections in the two first PCs, but now, the differentiation of CB4856, DL238, JU775 and QX1211 is more prominent than in the unweight projection (Figure 7 middle and right). Moreover, the variance proportion explained by wPCA (~37%) is higher than in PCA (~29%).

Weighted PCA resulted in a more clearly geographic structure than in the original plot, e.g. the differentiation of European and North American strains is now observed. Indeed, almost the all strains from Europe were clustered in the bottom left corner of the plane whereas almost North American samples were clustered in the top right corner. Additional information (e.g. continent of origin) for population clustering resulted in similar projections (Figure not presented). Finally, wPCA highlighted the low differentiation of these samples since European, African and American populations are essentially clustered together.

wPCA of modern human and La Braña 1 ancient sample

Finally, we used the wPCA to analyze SNP genotype data of human populations. First, we consider the Human Genome Diversity Project (HGDP) data (Li *et al.*, 2008) because differences in sample size and the well-described human population relationships. Here, the main purpose was to evaluate whether wPCA allows extract information that PCA does not. Second, we used genotypes of La Braña 1 Mesolithic sample (Olalde *et al.*, 2014) to evaluate whether the use of weights helps us to estimate properly the projection of a sample regarding other population samples.

In the HGDP dataset, we performed wPCA using the sample size of each population to build the weighted covariance matrix. The lowest sample size was African San population (n = 5) whereas the largest sample size was Middle East Palestinians (n = 46). In all cases we performed standard estimation of eigenvectors and eigenvalues to compare in similar conditions the estimations of wPCA and PCA. Therefore, basic projections of data are obtained here and further corrections of the results e.g. procrustes (Wang *et al.*, 2010), could be performed.

The sample projections for PCA are similar to those reported in the original works (Jakobsson *et al.*, 2008; Li *et al.*, 2008), but the projection observed from wPCA showed some differences, in particular for Oceania and America (Figure 8, left). The use of different grouping criteria to weight the covariance matrix (e.g. fifty-one levels of population or five levels for geographical regions) resulted in differences of the graphical representation of data (Figure 8, middle and right). This suggested that alongside the sample size of populations, the internal

differentiation of the geographical region (e.g. Africa) produces changes in the relatedness of populations and therefore, changes in the projections of data.

We already noticed that in a population with complex structure like humans, the sample size could affect the projection of PCA. Moreover, wPCA retained a higher variance proportion explained in two PCs than the retained in traditional PCA. This suggested that a proportion of variance still remain to be deciphered, maybe related with the internal diversity of populations within regions. Therefore, we performed wPCA within geographic regions, because each geographic region has a different history and population relationships. Unexpectedly, when wPCA was applied to European population using the population as grouping criterion, the population with small sample size (Tuscan, $n = 8$) abruptly altered the sample projection in the plane compared to PCA projection (Figure 9a) probably because the overcorrection. Tuscan samples were projected in the outer of the plane whereas other populations were collapsed in the middle (Figure 9b). Since we used each population as weight information, we then included Tuscan samples within Italians (Figure 9c). This modification in the weights resulted in better graphical representation of data. Additionally, the variance proportion explained was higher than explained in PCA without signals of overcorrection (Figure 9d).

In this case, wPCA can be also used to explore an empirical assignation of samples to a particular population or phenotype. This information potentially leads to detect overcorrection when similar data is considered in different class. The changes in projections when Tuscan populations was weighted apart to Italians suggested that differentiation of Tuscans is not so evident, in agreement with observations from microsatellites (Rosenberg *et al.*, 2002) and SNPs (Di Gaetano *et al.*, 2012).

Next, taking advance of potential detection of outliers of wPCA and prior assignation of weights, we explored the assignation of a sample into a particular population. In this case, the La Braña 1 sample from the Mesolithic period is a perfect candidate. Olalde *et al.*, (2014) explored the genetic data of this ancient sample and observed its higher genetic affinity with Northern Europe populations than Southern population, considering the geographical location where remains were found. Here, we reconstructed the PCA estimations and performed wPCA using two criteria for weights, clustering the La Braña 1 sample with Iberian population and alternatively with northern Europe populations.
First, we estimate PCA without procrustes or outlier correction as in the original work (Olalde *et al.*, 2014), in order to evaluate how effective is the wPCA to discriminate the sample assignation. The PCA projection pattern coincides with the reported in the study, although La Braña 1 is more distant from the populations (Figure 10a). When La Braña 1 sample was clustered with Iberian population (or assuming an independent origin), wPCA caused a strong distortion of projections (Figure 10b). In agreement with the authors, this suggests that there is no genetic affinity of the Mesolithic sample with Southern populations like Iberian. However, when La Braña 1 was clustered with Finns the wPCA projections notably improved and the ancient sample was projected closer to Finns (Figure 10c). This result suggests the higher allele sharing between La Braña 1 and Northern population than with Iberian populations.

DISCUSSION

Principal component analysis has been widely used in genetics providing a suitable tool to explore the genetic relationship among populations and individuals. Classical PCA performs dimensionality reduction of data (i.e. tens of hundreds of SNPs) in a lower number of components, starting with the first that recovers the large fraction of variance of the data, and then other orthogonal components with smaller variance for each. In the presence of atypical data, PCA dimensionality reduction could be unreliable because of its sensitivity to the choice of data and the variance captured by the first component variance capture may be not represent the variance of normal data (Hubert *et al.*, 2005). To correct for this effect, in recent years some papers have showed the beneficial use of weighting functions in PCA estimations (Yue and Tomoyasu, 2004; Tan and Chen, 2005; Pinto da Costa *et al.*, 2011), although in a different setting that the one explored here.

In the case of population genetic data, restriction in the number of samples to analyze could lead to biased estimations because lack representation of genetic background of populations or relationships among them. In fact, the problem of sample size depends of many factors widely discussed (Chakraborty, 1992; Gordon *et al.*, 2002; Leberg, 2002; Kalinowski, 2005).

In this paper, we presented a thorough discussion of the bias inherent in sampling. It could be intuitively thought that there is no need to correct the PCA provided sampling is random, even if unbalanced. However, this is not the case, as we showed with several examples. The distortion of the PCA projections depends on the sample size and especially on the relative sample size of different populations. The differential representation of samples leads to changes in the covariance of observed genotypes between individuals and therefore generates a bias in the estimations of eigenvalues and eigenvectors. The application of weighted covariance in the estimation of PCA supposes a promising strategy to correct the effect of uneven sample size in genetic data, showing that population structure could be recovered in populations with low sample size (Figure 5). The use of F_{ST} provides an improvement to the overcorrection observed in extremely low sample size or very low genetic differentiation. Both weights were highly efficient as observed in *C. elegans* (Figure 7). For simple analyses the wPCA using the inverse of sample size is suggested for correction, but in presence of low population differentiation, the combination of F_{ST} and sample size (eq. 7) can improve data interpretations.

Of course it is possible to apply the same strategy at multiple levels, for example populations and subpopulations, weighting equally all the subpopulations in the same population. However, the effectiveness of this depends on the amount of prior information available. More generally, it would be possible to weight according to other features, like population density, population size, phenotypes and so on. While these alternative weighting schemes have their drawbacks, as they are not directly related to the genetic structure of the sample, they could nevertheless be useful to answer different biological questions. For example, the concordance of relatedness of populations and individual grouping to build the weights can improve graphical representations, whereas misallocation or misunderstanding of populations relationships lead to abrupt modifications of projections without possibilities of interpretation (e.g. Figure 9b, Figure 10b).

Using SNP genotype data from publicly databases and simulations, we demonstrate the changes in the data representation in the PCA, and that the use of weights in PCA has a potential benefit in the interpretation of data without altering the results. For instance, the results observed in *C. elegance* using PCA limited some interpretation about signals of differentiation of the populations, which are more evident in wPCA projections. Furthermore, the use of weights can contribute to distinguish graphically whether assignation of clusters by specific criteria is in agreement with underlying genetic model. The wPCA for La Braña 1 sample and Tuscan population are examples of this advantage. However, it is highly recommended to evaluate different criteria to build the weights. Notably, if the population subdivision used for weights in the wPCA does not correlate with the actual population structure, the data projections tend to distortion.

Finally, the proposed method could contribute to understand the relationships in high dimensional data because it extracts additional information of the samples relationship. The joint interpretation of wPCA projections with other statistics of population genetics benefits the interpretation of relatedness for unequally sampled populations.

MATERIAL AND METHODS

Simulated dataset

To estimate the robustness of wPCA in genetic data, six populations in three different migration scenarios, island model (IM), hierarchical model (MM) and stepping stone model (SS), were simulated by coalescence using MLCOALSIM.v2 (available at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html) . The effective population size was Ne = 1000 for all populations and models. For IM and SS models, the migration rate among populations was M = 1. For the MM, we simulated three subpopulations with one, two and three populations each. The migration rate within subpopulations was M = 1 and between subpopulations was M = 0.1. All migration rates were scaled by the effective population size M = *4Nem*. Each simulated population consisted of five hundred individuals with thousand independent SNPs each. Further, for each population we randomly selected a hundred individuals (n = 100) for downstream analyses.

Comparison of PCA and wPCA

First, we performed the PCA in the entire dataset (n = 500) to determine the expected population structure of the simulated dataset. The effect of sample size in PCA projections was evaluated performing a sample reduction, ranged from n=1 to n=75, for one or three populations of each sampled dataset (n=100) whereas the sample size of the other populations was n=100. Further, for each migration scenario and sample size reduction we estimated the PCA scores using *prcomp* package and for wPCA the *cov.wt()* function implemented in R (R Development Core Team, 2011). Several R functions were written to build the weights, estimation of F_{ST} and wPCA and are available upon request. To estimate the robustness of the PCA and wPCA, one thousand bootstrap samples for each sample size reduction and migration scenarios were performed.

Differences in the sample sizes between populations result in a distortion in the space projection, especially in the first two principal components. To measure this distortion, we compared the average projected position of the individuals in the two first PCs in PCA and wPCA datasets against the projected position of individuals in the entire dataset with Euclidian distance (***d***). This measure allowed us to establish which of either PCA or wPCA points (individuals) projected was closer to the expected value (entire population). Bootstrap sampling and Euclidian distance estimations were carried out using custom scripts in R.

Analysis of empirical data

We analyzed three publicly available genotype datasets. The Cuban and Brazilian pig dataset consisted of 18,308 autosomal SNPs of 48 samples analyzed in (Burgos-Paz *et al.*, 2013). The *C. elegans* dataset consisted of 200 samples with 41,188 SNPs (Andersen *et al.*, 2012). Following the proposed methodology, we selected for PCA only 97 samples so-called "Isotypes", but three out 97 samples with unknown origin were removed for further analyses. After quality control and SNP LD pruning we retained 3,968 SNPs for PCA and wPCA estimations. Finally for Humans, first we used the SNP genotype data from 938 unrelated individuals obtained from the Human Genome Diversity Project HGDP-CEPH (Li *et al.*, 2008). We pruned the database selecting only autosomal SNPs with no missing genotypes retaining 488,919 SNPs. Finally, we request to Carles Lalueza-Fox, the SNP genotypes for La Braña 1 Mesolithic sample analyzed in Olalde *et al.*, (2014) to reproduce the PCA projection in Figure 3 in their paper. After quality control and SNP LD pruning we retained 1,400,167 SNPs for 760 samples. For management and quality control of databases we used PLINK (Purcell *et al.*, 2007).

ACKNOWLEDGMENTS

We would like to thank to Marie-Anne Félix for comments about wPCA in *C. elegans.* We also thank to Carles Lalueza-Fox and Iñigo Olalde for La Braña 1 sample genotypes. WBP is funded by COLCIENCIAS (Francisco José de Caldas fellowship 497/2009, Colombia). Work funded by AGL2010-14822 grant (Spain) to MPE, CGL2009-09346 grant (Spain) to SERO.

APENDIX

Implementation of wPCA

The weighted PCA can be implemented in a couple of ways, either modifying directly the code used to compute the PCA or modifying the input data for the analysis. Both ways are straightforward. The direct modification of the code requires the substitution of the covariance matrix with a weighted covariance, plus the implementation of the weights according to the formula (4), with a minimal modification of existing code. For example, codes written in R can be adapted to the wPCA simply by using the function *cov.wt()* instead of *cov().* Further, estimation of PCA scores and PC variance contributions can be extracted from *Eigen* estimations as follows:

Considering data as an array of [nind * nsnps], and w a vector of [nind*1] where each value corresponds to the weight estimated per individual in eq. 4 or eq.7:

```
# Estimation of weigthed covariance matrix
      cwt<-cov.wt(data, wt=as.vector(w))

# Estimation of eigenvectors and eigenvalues
      eigenvals<- eigen(cwt$cov)

# Obtaining wPCA scores
      PCA_cov.wt<- data %*% eigenvals$vectors

# Obtaining wPCA variance per PC
      PCA_variance<- eigenvals$values/sum(eigenvals$values)
```

An alternative approach is to add multiple copies of the individuals of each population until their contribution to the sample is approximately equal, and then perform the standard PCA. For example, if the numbers of individuals sampled from populations *A*, *B,C* are n_A =15, n_B = 22 and n_C = 5 respectively, we can add other two copies of A, one copy of B and eight copies of C to reach a balanced representation of individuals (n'_A = 3×15 = 45; n'_B = 2×22 = 44, n'_C = 9×5 = 45). This is essentially equivalent to the use of the weights (eq. 4) and does not require any modification of software code (e.g. EIGENSOFT (Patterson *et al.*, 2006)). Therefore it is a convenient implementation for most users.

Singular Value Decomposition and genealogical interpretation of wPCA

In this section we review how the Singular Value Decomposition (SVD) approach to PCA works in the weighted case. A review of the SVD formalism in the unweighted case can be found in (McVean, 2009); we borrow the notation from the same paper.

We denote by *X* the matrix of centered genetic data of size *L* × *n,* with components $X_{si} = x_s^i - \sum_j w_j x_s^j$ (i.e, each column contains the data of a single SNP for all individuals in the sample, with mean centered at 0). We use matrix formalism throughout the section.

Standard (unweighted) PCA is based on the diagonalization of the covariance matrix $C = \frac{1}{n}XX^T$ by a transformation *P* such that PCP^T is a diagonal matrix with diagonal elements in decreasing order. Similarly, weighted PCA is based on the covariance matrix

$$C = \frac{1}{n}XWX^T \quad (5)$$

Where *W* is a diagonal matrix with $W_{ii} = w_i$

We consider the SVD of the matrix $X\sqrt{W}$, where the square root is defined to be a positive definite matrix. We have

$$X\sqrt{W} = U\Sigma V^T \quad (6)$$

with U and V orthogonal matrices of size $L \times n$ and $n \times n$ respectively, and Σ a diagonal $n \times n$ matrix with positive elements in decreasing order. Then it is easy to show that the wPCA transformation is $P = U^T$.

The SVD decomposition is obtained from the smaller n × n matrix

$$M = \frac{1}{L} \sqrt{W} X^T X \sqrt{W} \quad (7)$$

In fact, V and Σ are the solutions of the linear problem

$$MV^T = \frac{1}{L} V^T \Sigma^2 \quad (8)$$

that reduces to standard eigenvalues problems by taking each column separately. Then U is easily obtained as

$$U = X \sqrt{W} V \Sigma^{-1} \quad (9)$$

In the unweighted case, the matrix M is directly related to the average coalescent times of pairs of individuals (McVean, 2009). This extends to the weighted case, with some differences. For simplicity, we assume that the data are DNA sequences. Denoting the average coalescence time between individuals i and j by $\bar{t}_{ij}$ and the average sum of branch lengths by $\bar{T}$, the expected value of M is

$$E(M_{ij}) = \frac{1}{\bar{T}} \sqrt{w_i w_j} (\bar{t}_i + \bar{t}_j - \bar{t} - \bar{t}_{ij}) \quad (10)$$

where $\bar{t}_i = \Sigma_j w_i \bar{t}_{ij}$ and $\bar{t} = \Sigma_{ij} w_i w_j \bar{t}_{ij}$. Therefore, the relation between coalescence times and PCA outlined in McVean, (2009) is still valid for wPCA.

REFERENCES

Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, *et al.* (2012). Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.* **44**: 285–90.

Bornberg-Bauer E (1997). How are model protein structures distributed in sequence space? *Biophys. J.* **73**: 2393–2403.

Bornberg-Bauer E, Chan HS (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. U. S. A.* **96**: 10689–10694.

Burgos-Paz W, Souza CA, Megens HJ, Ramayo-Caldas Y, Melo M, Lemús-Flores C, *et al.* (2013). Porcine colonization of the Americas: a 60k SNP story. *Heredity (Edinb).* **110**: 321–30.

Chakraborty R (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Hum. Biol.* **64**: 141–59.

694 Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, *et al.* (2012). An Overview
695 of the Genetic Structure within the Italian Population from Genome-Wide Data. *PLoS One* **7**.

696 Gordon D, Finch SJ, Nothnagel M, Ott J (2002). Power and Sample Size Calculations for Case-
697 Control Genetic Association Tests when Errors Are Present: Application to Single Nucleotide
698 Polymorphisms. *Hum. Hered.* **54**: 22–33.

699 Hubert M, Rousseeuw PJ, Vanden Branden K (2005). ROBPCA: A New Approach to Robust
700 Principal Component Analysis. *Technometrics* **47**: 64–79.

701 Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C, *et al.* (2008). Genotype,
702 haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–
703 1003.

704 Kalinowski ST (2005). Do polymorphic loci require large sample sizes to estimate genetic
705 distances? *Heredity (Edinb).* **94**: 33–6.

706 Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, *et al.* (2012). New insights into
707 the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat.*
708 *Commun.* **3**: 698.

709 Kriegel H-P, Kröger P, Schubert E, Zimek A (2008). A General Framework for Increasing the
710 Robustness of PCA-Based Correlation Clustering Algorithms. In: *Scientific and Statistical*
711 *Database Management*, Vol 5069, pp 418–435.

712 Leberg PL (2002). Estimating allelic richness: effects of sample size and bottlenecks. *Mol.*
713 *Ecol.* **11**: 2445–2449.

714 Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, *et al.* (2008).
715 Worldwide human relationships inferred from genome-wide patterns of variation. *Science*
716 **319**: 1100–1104.

717 Li J, Huang S (2007). Evolving in Extended Hamming Distance Space: Hierarchical Mutation
718 Strategy and Local Learning Principle for EHW. In: Kang L, Liu Y, Zeng S (eds) *Evolvable*
719 *Systems: From Biology to Hardware*, Lecture Notes in Computer Science. Springer Berlin
720 Heidelberg Vol 4684, pp 368–378.

721 McVean G (2009). A genealogical interpretation of principal components analysis. *PLoS Genet.*
722 **5**: e1000686.

723 Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in
724 Europeans. *Science (80-.).* **201**: 786–792.

725 Novembre J, Stephens M (2008). Interpreting principal component analyses of spatial
726 population genetic variation. *Nat. Genet.* **40**: 646–9.

727 Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, *et al.* (2014).
728 Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European.
729 *Nature*.

730 Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, *et al.* (2007).
731 PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS*
732 *Genet.* **3**: 1672–86.

733 Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet.* **2**:
734 e190.

735 Pinto da Costa JF, Alonso H, Roque L (2011). A weighted principal component analysis and its
736 application to gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**: 246–52.

737 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007). PLINK: a
738 tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum.*
739 *Genet.* **81**: 559–575.

740 R Development Core Team R (2011). R: A Language and Environment for Statistical
741 Computing. *R Found. Stat. Comput.* **1**: 409.

742 Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky L a, *et al.* (2002).
743 Genetic structure of human populations. *Science* **298**: 2381–5.

744 Shriner D (2011). Investigating population stratification and admixture using eigenanalysis of
745 dense genotypes. *Heredity (Edinb).* **107**: 413–20.

746 Tan K, Chen S (2005). Adaptively weighted sub-pattern PCA for face recognition.
747 *Neurocomputing* **64**: 505–511.

748 Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW (2002). Frequentist estimation of
749 coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**:
750 447–59.

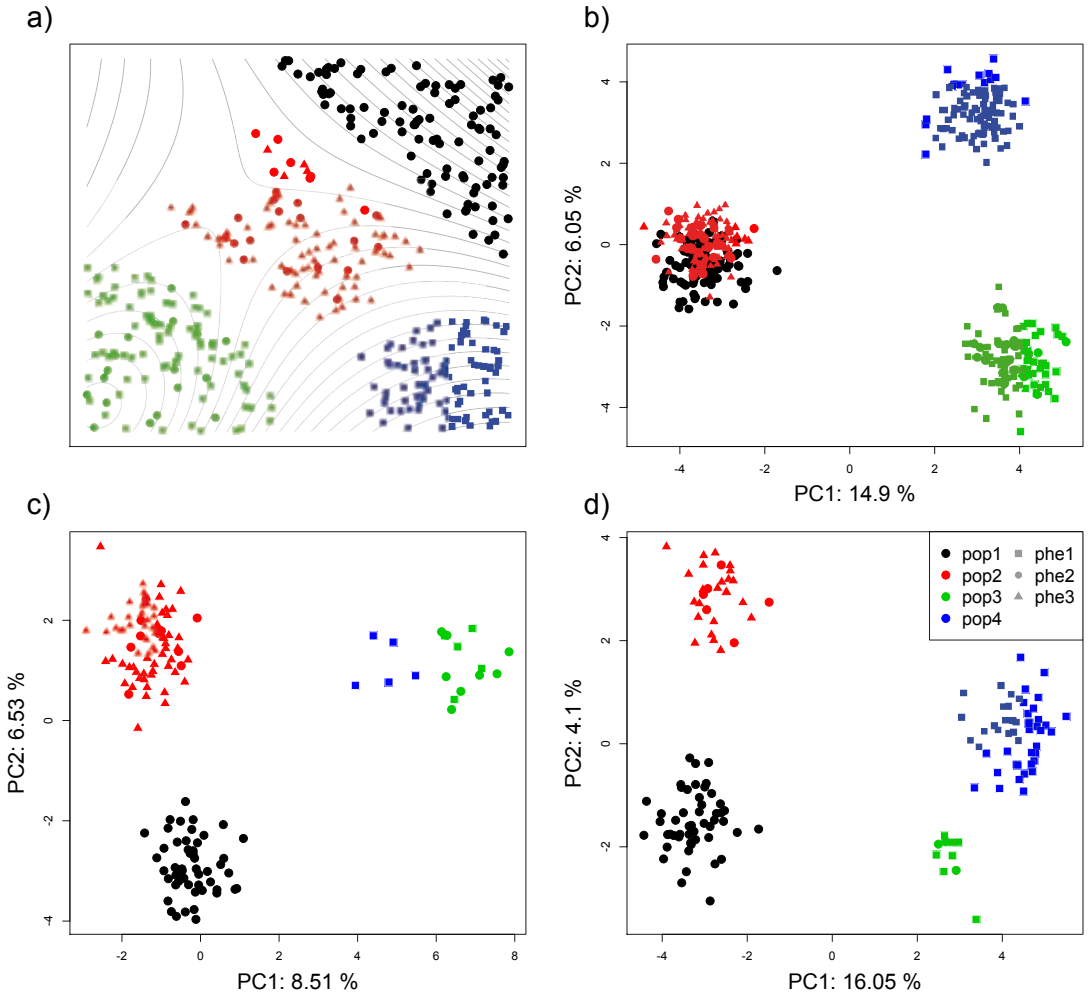
751 Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, *et al.* (2010).
752 Comparing spatial maps of human population-genetic variation using Procrustes analysis.
753 *Stat. Appl. Genet. Mol. Biol.* **9**: Article 13.

754 Yue HH, Tomoyasu M (2004). Weighted principal component analysis and its applications to
755 improve FDC performance. *2004 43rd IEEE Conf. Decis. Control (IEEE Cat. No.04CH37601)* **4**.

756

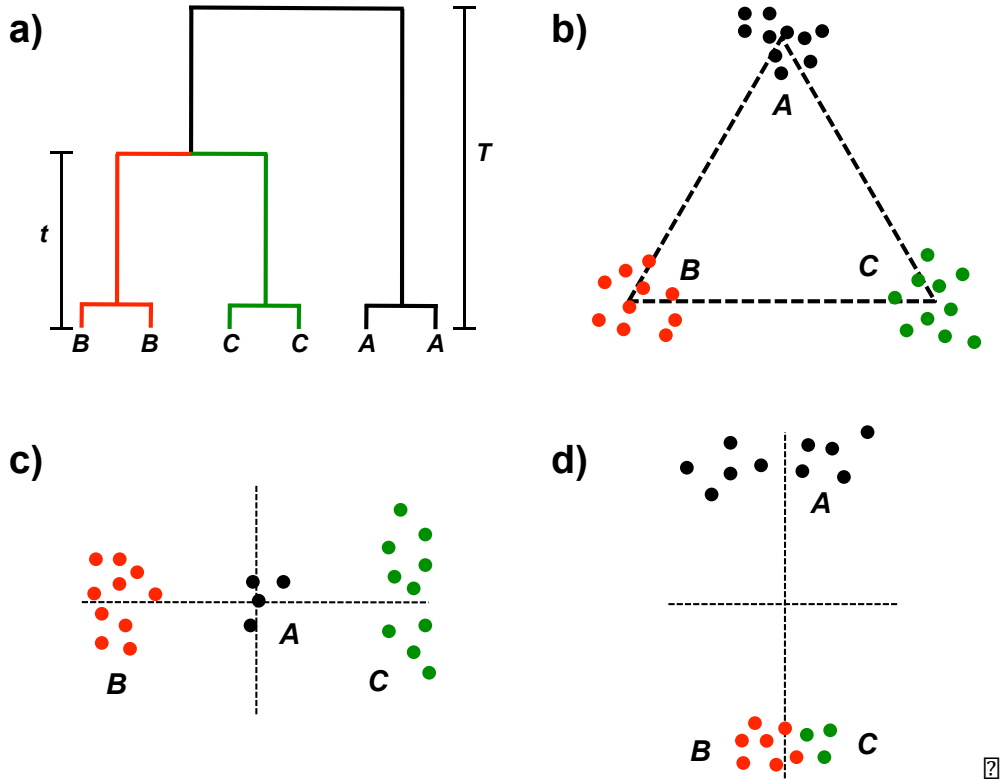
757

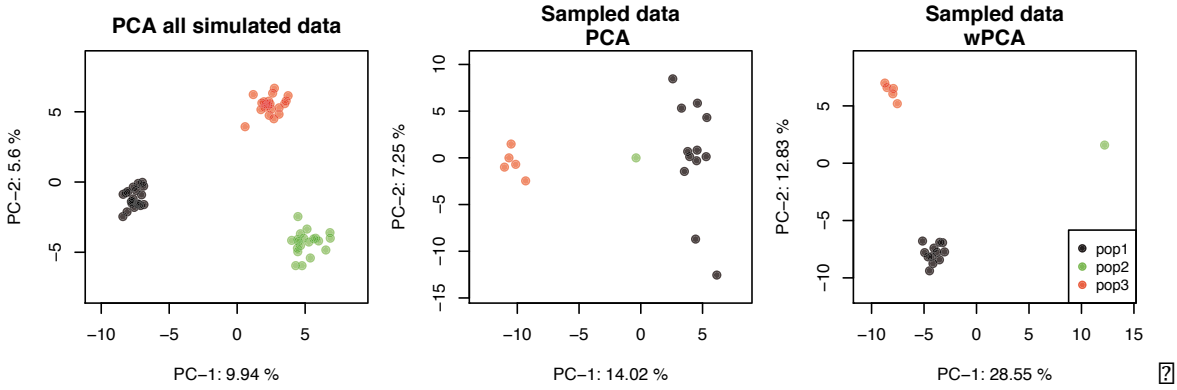
mxp
mxl
ma0
ma3
maA
man
mae
max

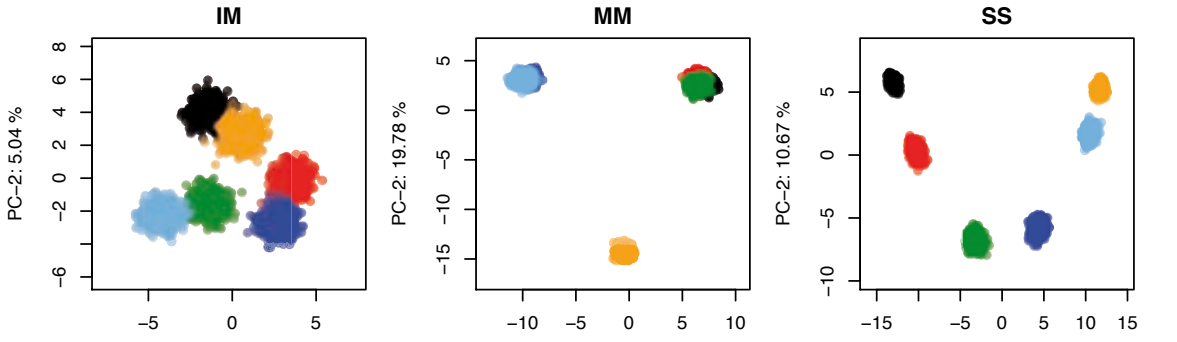


maa
mam

30n


a)
T
t
B
B
C
C
A
A
b)
A
B
C
c)
A
B
C
d)
A
B
C


PCA all simulated data
PC–2: 5.6 %
PC–1: 9.94 %
Sampled data
PCA
PC–2: 7.25 %
PC–1: 14.02 %
Sampled data
wPCA
PC–2: 12.83 %
PC–1: 28.55 %
pop1
pop2
pop3


IM
MM
SS
PC−2: 5.04 %
PC−2: 19.78 %
PC−2: 10.67 %
PC−1: 5.42 %
PC−1: 26.33 %
PC−1: 40.56 %

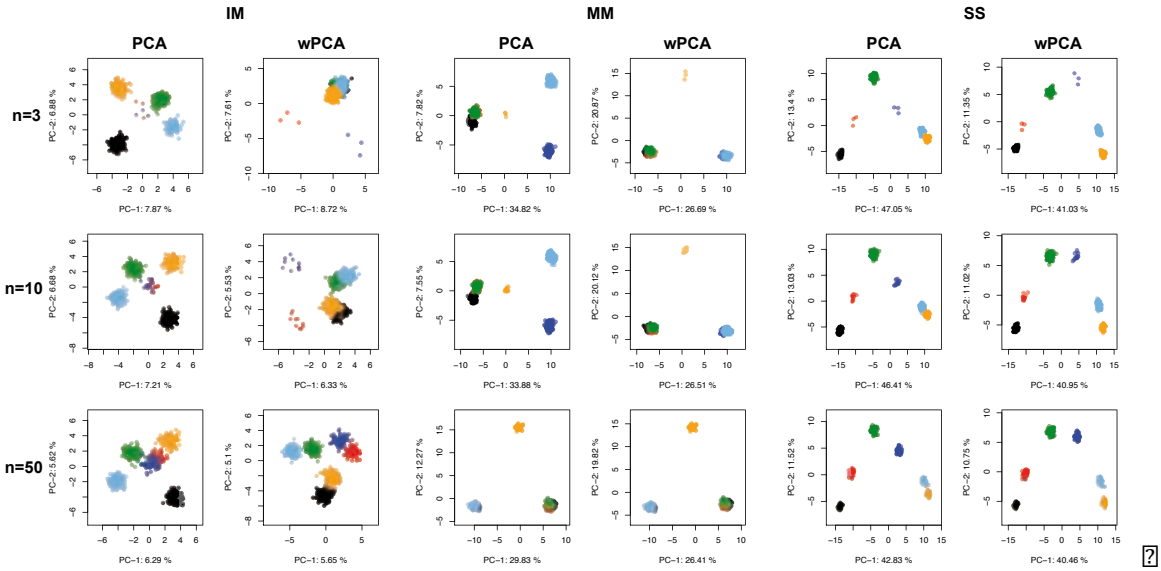

IM
MM
SS
PCA
wPCA
n=3
n=10
n=50

Figure 6. Euclidean distance differences between entire simulated data set and PCA (green) and wPCA (red) projections. MMb corresponds to similar population structure in MM but only one population was poorly sampled.

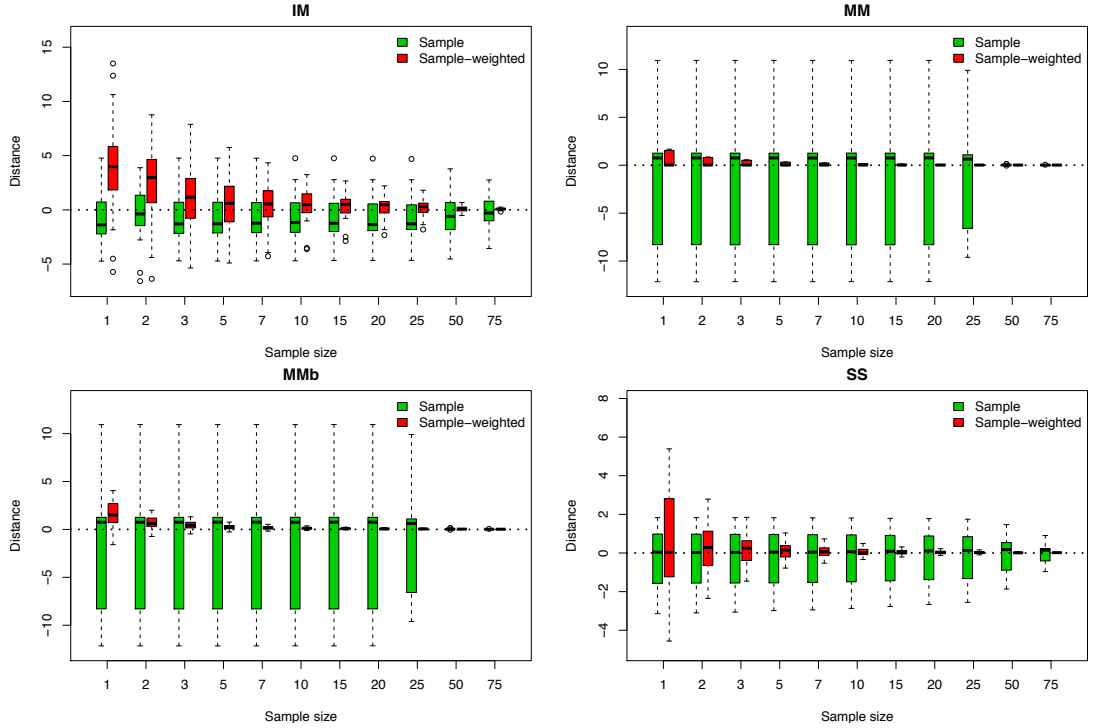
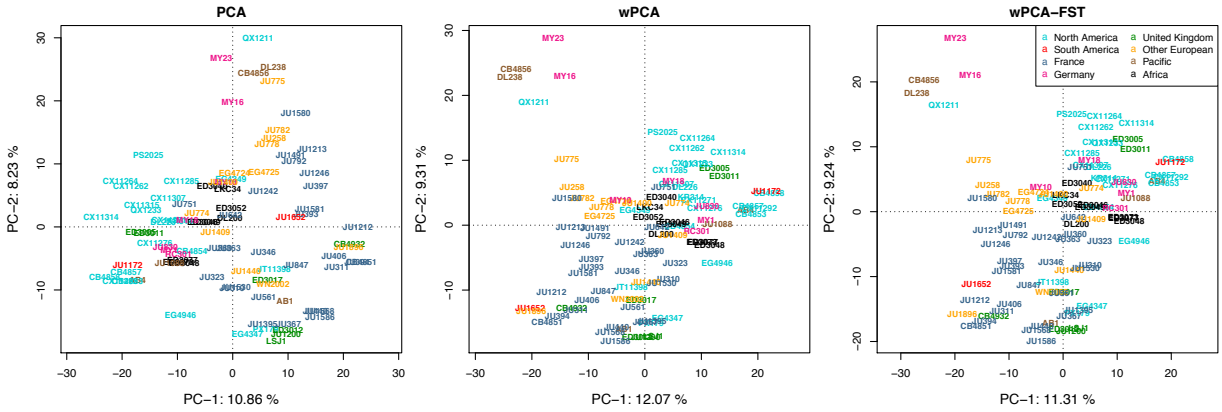
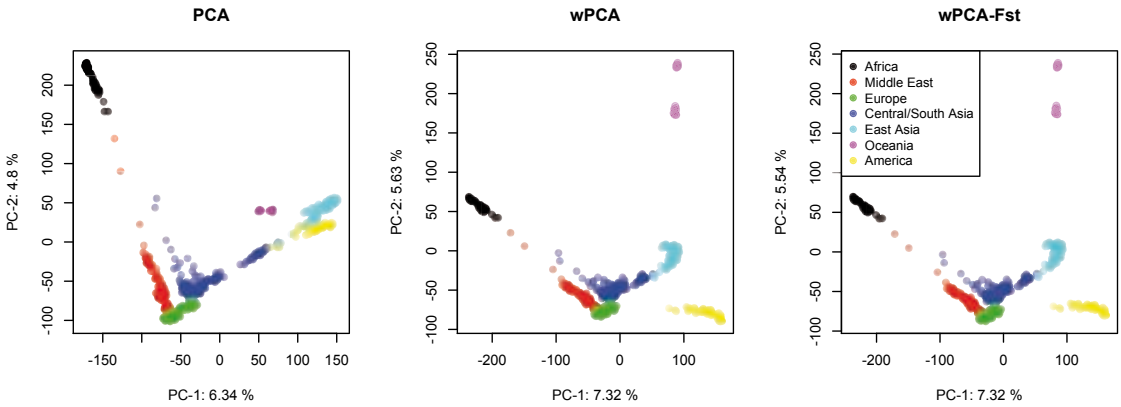
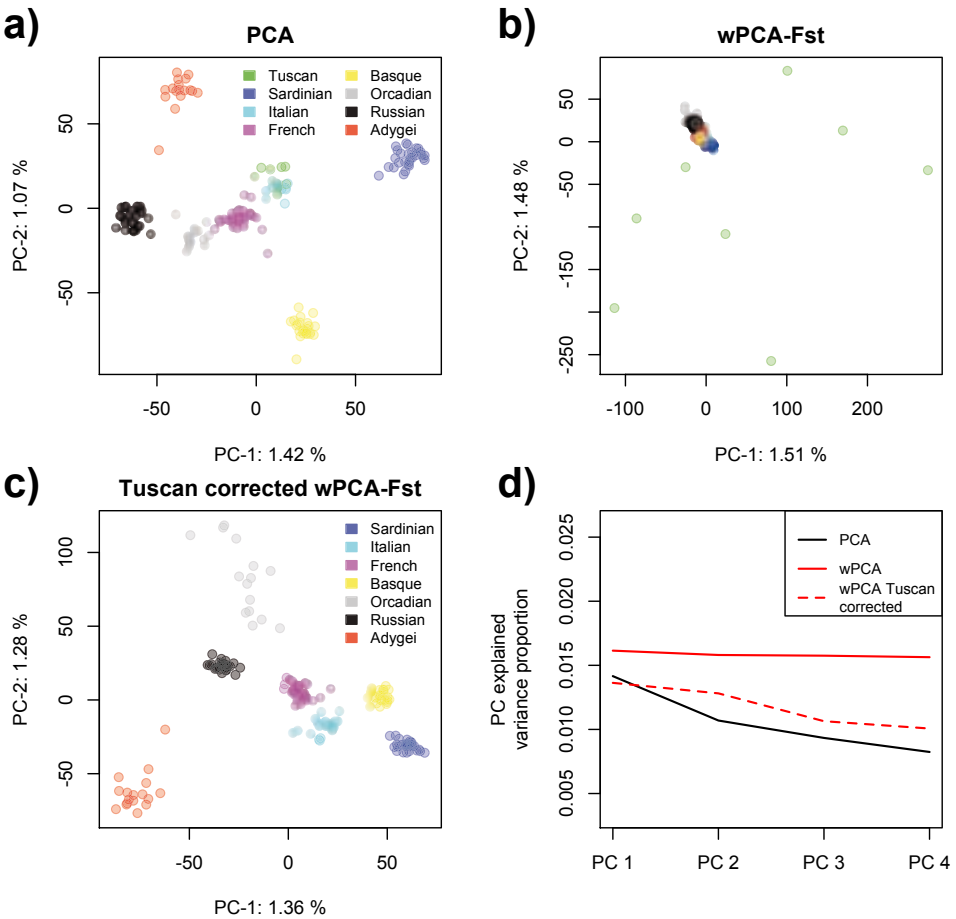
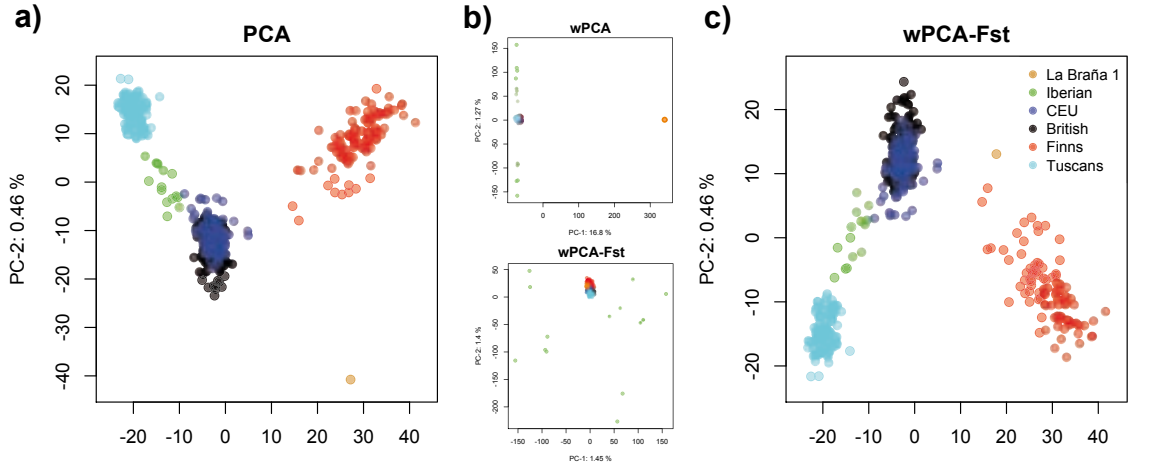


Figure 7. PCA (left), wPCA (middle) and wPCA-F_{ST} (right) for *C. elegans* dataset




PCA
wPCA
wPCA-Fst
PC-2: 4.8 %
PC-2: 5.63 %
PC-2: 5.54 %
PC-1: 6.34 %
PC-1: 7.32 %
PC-1: 7.32 %
Africa
Middle East
Europe
Central/South Asia
East Asia
Oceania
America


a)
PCA
b)
wPCA-Fst
c)
Tuscan corrected wPCA-Fst
d)
PC-2: 1.07 %
PC-1: 1.42 %
PC-2: 1.48 %
PC-1: 1.51 %
PC-2: 1.28 %
PC-1: 1.36 %
PC explained variance proportion
PC 1
PC 2
PC 3
PC 4
Tuscan
Sardinian
Italian
French
Basque
Orcadian
Russian
Adygei
PCA
wPCA
wPCA Tuscan corrected


a)
PCA
PC-2: 0.46 %
PC-1: 0.92 %
b)
wPCA
PC-1: 16.8 %
wPCA-Fst
PC-1: 1.45 %
c)
wPCA-Fst
La Braña 1
Iberian
CEU
British
Finns
Tuscans
PC-2: 0.46 %
PC-1: 0.92 %

Chapter 7

-

Discussion

-

Pig populations like those extant across American continent had not been well-characterized Similarly, the ancestry of European breeds, the further colonization and expansion across territory or the putative effects of selection over the genome have not been studied. Nevertheless the information from different genetic markers combined with archaeological remains clearly allowed to disentangle key process in the complex population history of pigs. Using SNPs information from SNP arrays and NGS, here we characterize genetically the American village pigs and their relationships with worldwide pig populations. The complementarity of these sources of information lead us to validate some hypothesis about the colonization and open a window for further exploration of this genome as a valuable genetic resource for population genetics and breeding studies.

7.1 The pig population structure in the world: Ancestry and diversity of American village pigs

The availability of new and powerful genetic tools has allowed us answering questions about pig population relationships. In particular, we are interested in give answers to the following questions: Is there a fingerprint of human mediated colonization in extant American village pigs? Can we observe evidences of adaptive processes in the recent American colonizing populations?

To solve these questions, we used genomic tools to explore the largest survey of pig populations from the American countries sampled to date. This survey included extant pig populations from different environments, feral specimens and an ancient DNA sample.

The diversity and relationships of pig populations

The studies of pig population genetics based on sexual chromosomes, especially in the X-chromosome (SSCX) are yet scarce. Little was known about the pseudoautosomal region (PAR) in SSCX (Quilter *et al.*, 2002); We considered that the reduction heterozygosity in males, as indirect measure, was a good predictor of the PAR size. The differences in heterozygosity estimates along the chromosome markers allowed us to suggest the presumable location of the pseudoautosomal region boundary (PAB) around 7Mb of SSCX*p* arm. Indeed, the recent works about SSCX structure on pigs (Skinner *et al.*, 2013; Das *et al.*, 2013) confirmed the estimations in this thesis, demarcating the PAB at X:6,743,567 bp, in intron 3-4 of *SHROOM2* and showing that *SHROOM2* is truncated in SSCY. The estimation of PAR is important for posterior considerations in population genetics of the pigs.

Altogether, the information derived from SSCX (Chapter 3), and autosomes (Chapter 4 and 5) showed a primary pattern of differentiation between Asian and European pig populations, as main centres of domestication. This result agrees with the pattern of differentiation observed previously using autosomal data (Megens *et al.*, 2008) as well as with the uniparental inherited lineages from mitochondrial DNA (Giuffra *et al.*, 2000) or Y-chromosome data (Ramírez, Ojeda, *et al.*, 2009). However, the location of the PAR and the NPAR allowed us to separate the analyses in these two chromosomal regions, and explore in detail the effect of demographic events in continental differentiation of pig populations because of the inherent characteristics of X-chromosome (e.g. the *Ne* reduction). One of the characteristics that makes the analysis of X-chromosome interesting (Schaffner, 2004) is the diversity reduction in the NPAR regarding the PAR, which is expected assuming that effective size of X-chromosome variability is ¾ of that observed in autosomes when equal number of males and females are present in population. This is true according to the observed results for commercial breeds, American village pigs and African populations, but deviations from this expected ratio were observed in Asian and European populations.

Deviations from ¾ in the ratio X-chromosome/Autosome (X/A) variability for pigs could be influenced by two main situations: 1) Changes in the population size and 2) Sex-bias demography events in founder populations. The divergence between Asian and European populations has been estimated around 1.2 Mya with a decrease of population size around 50 kya more severely in Europe than Asia. These events resulted in higher levels of variability in wild boars from Asia, followed by domestic pigs from Asian and Europe and ultimately European wild boars (Bosse *et al.*, 2012; Groenen *et al.*, 2012). Under this scenario, an expected deviation in X/A ratio is plausible (Pool and Nielsen, 2007), and two studies on pigs have showed evidences of this. First, Amaral et al. (2011) observed a reduction in variablity of X-chromosome in commercial populations and wild boars from Europe. According their results, the variability ratio ranged $0.36 < X/A < 0.8$. Further, Esteve-Codina et al. (2013) analyzing Iberain pigs noticed that reduction in the ratio reached values as lower than $X/A = 0.27$. The authors suggest a dramatic reduction in population size and selective processes in the breed.

Secondly, any deviation of the sex ratio from equality will enhance the chance for random genetic drift and reducing the effective population size (Hartl and Clark, 1997). For the study

of SSCX is important considering that migrations and human-mediated animal translocations are gender biased (Lindgren *et al.*, 2004). In this sense, the origin of pig populations has been highly influenced by male or female-biased migrations. For instance, Ramírez et al. (2009) suggested that Chinese introgression in 18th-19th centuries was fundamentally maternal because the lack of HY3 haplotype of chromosome Y (SSCY) in commercial breeds, and relatively abundant in the Chinese counterpart. Another study (Fang and Andersson 2006) pointed out this bias, showing the absence of European mitochondrial haplotypes in Chinese breeds evaluated. Despite the uneven sex sampling, we observed changes in the differentiation pattern of the populations whether SNPs are located in PAR or NPAR. The maternal introgression of commercial breeds could potentially be associated with a reduction in their Asian ancestry for NPAR, because genetic drift in males. Interestingly, Duroc breed, one of the most widely pig breeds used in the world, showed an absence of Asian ancestry in NPAR opposite to what was observed in PAR. The sex ratio may influence the ratio of variability in SSCX and autosomes, and therefore the genetic relatedness of populations. For instance, reproductive technologies (e.g Artificial insemination) accelerate the dissemination of selected boars. Using the patterns of variation in SSCX, we can extract additional information about the pig population relatedness, tracing the movement of animals and estimate demographic events. The availability of SNP data (>1200 SNPs on SCCX included in the Illumina's array) and genome sequence will be fundamental to study the pig populations, but it is very important to improve the annotation of this chromosome as well to correct the BAC order.

The autosomal diversity evaluated in chapter 4, provided more evidences about colonization and relatedness of American village pig populations. In the light of historical evidence, the Caribbean was the main centre for European conquerors and resources brought after the arrival of Columbus. Then, livestock species were dispersed from here towards North and South America by different routes and patterns of the animal introduction (Butzer, 1992). For Central and South America, the livestock species including pigs were important for the transport of material as well as a source of protein for the conquerors. These species were rapidly replacing the well-adapted species in the new world. During the first century of colonization the Spanish conquerors quickly moved across America and the movement of animals through the territory resulted in multiple founder effects (Rodero *et al.*, 1992; Revidatti, 2009).

Assuming simple genetic models of rapid colonization (Olivieri, 2009), we expected to find evidence of this founder effect in American village pigs, in terms of a reduction of variability, higher genetic differentiation or even correlation between genetic and geographical distances between populations (Olivieri, 2009). Conversely, we found similar patterns of genetic diversity across American village pig populations as well as low levels of differentiation in both SSCX and Autosomes. Nevertheless, we observed mild patterns of differentiation in Brazilian and Argentine populations observed in PCA and in Admixture analyses. This pattern is surely influenced by the Portuguese genetic origin of the pigs introduced in Brazil (initially at Southern) and supported by the close relationships of Brazilian with Argentinean populations, and both with Bisaro pig breed. Further, in the context of human populations, several immigrations events of European settlers and the trade with Africa modified notably

the demography of Brazil and Argentina and it is also likely that contributed in the population structure of pigs in these territories (Metcalf, 2005).

The low differentiation between American village pigs and the low correlation between genetic and geographic distance among them can be the result of demographic events occurred initially after colonization and more recently in the 19th century. The colonization in Central and South America carried both new agricultural techniques and animal husbandry that were adapted for the Native human populations in the continent. The range expansion of new settles also played an important role in development and growth of livestock populations in the continent. However, in the subsequent years of colonization, it was produced a drastic reduction of the number of Native American human populations for multiple circumstances (e.g. epidemic, land grant or contested between conquerors); several populations and settles disappeared and the remaining were allocated in nuclei. The native human populations inhabit America then gradually abandoned the land used for agriculture and livestock. The conditions of tropical low land caused that these zones were the first abandoned and the livestock became to be feral (Butzer, 1992). Pigs became to be feral in many regions, and considering the actual range dispersion of feral pigs as a proxy and the exceptional capability to colonize areas (e.g. Gabor *et al.*, 1999), we can suspect that after colonization they disperse rapidly and adapted to environmental conditions. Many feral individuals or populations could be overlapped and a gene flow between them could model their low structure initially.

Nevertheless, the appearance of the commercial pig breeds is the most recent historical event that altered drastically the genetic relationships of pigs in the world. In fact, when the autosomal diversity was analysed, the second major cluster of variation comes from these populations or breeds (Chapter 4, Figure 1). The introgression of Asian breeds into European domestic populations (specially Britain breeds) is interpreted as the reason to observe higher variability in commercial pigs than European Wild boar (Ojeda *et al.*, 2011; Bosse *et al.*, 2012). Further artificial selection in the introgressed breeds promoted to the formation of a well defined commercial populations, with numerous distinctive phenotypes, variability and performance (Badke *et al.*, 2012; Rubin *et al.*, 2012; Wilkinson *et al.*, 2013).

The ancestry of American village pigs: The role of commercial breeds

Regarding American populations, the observed patterns in heterozygosity and ancestry revealed two key features in their history: 1) The Iberian origin of the populations, and 2) The strong signals of introgression of commercial breeds introgression into American village populations. From a meta-population level, the SSCX showed intermediate levels of diversity and differentiation, with almost complete European genetic ancestry. Interestingly, the breed-country analysis reveals that some American populations like Yucatan and Guatemala creole still preserve the Iberian origin, but others showed high levels of ancestry from commercial breeds (Chapter 3, Figure 3).

The analysis of autosomal SNP data provided increased resolution in the ancestry of American village pig populations. First, despite well-documented Iberian origin of pigs introduced in America (Rodero et al. 1992), the currently genetic variation shows a relevant

reduction of this ancestry (Chapter 4, Figure 4). Ramírez et al. (2009) pointed out that American pig populations showed either allele or haplotypes found in different populations around the world. Our results support that observation but also reveal a complex admixture with different proportions of European or even Asian origins.

The introduction of animals from different origins, like Portugal and England primarily, played an important role in the development of populations in America by increasing its diversity. However, the generalized use of commercial breeds and intercrossing with local pigs contributed in the reduction of the Iberian ancestry in the last two centuries. To date, Iberian pigs is a reduced population with no evidence of Asian or even other European local pig introgression (Alves *et al.*, 2009), and its principal contribution to genetic background of American village pigs was relegated to the colonial period in the American continent. Likewise, the Iberian pigs were used for the formation of commercial breeds like Duroc (Porter, 1993), although the more widely commercial breeds do not carried an Iberian component (e.g Large white).

We observed the presence of Asian ancestry, which could be introduced directly from Asian animals or indirectly from commercial breeds. For example, this ancestry is particularly evident in Caribbean populations like Cuba and Central America, and probably mediated by Chinese immigrations occurred in the middle 1800's (Lai and Tan, 2010). The development of highly efficient commercial breeds in Europe and United States by the combination of European and Asian haplotypes satisfied the global requirements of food at global level. In this process, several Chinese breeds were crossed and tested with different results. For instance, the introduction of Meishan breed from China into United Stated increased the reproductive efficiency of breeds at the same time that produced a detrimental of meat quality, so this intercross was later abandoned (Blackburn and Gollin, 2009). Notably, these commercial breeds shared high levels of variability that confers a tremendous potential for phenotypic improvement in the future (Knox, 2014). The popularity of these breeds in Europe and the large pork industry in North America rapidly caused the expansion of improved and specialized breeds around the world. The exportation of boars and sows resulted in the subsequent crossbreeding with extant local pigs. The high diversity and small differences in allele frequencies of SNPs in SSCX between American village pigs (and also African) populations with commercial breeds (Chapter 3, Table 1), and the and low differentiation between them observed in autosomes showed the important role of recent introgression in the local pig populations. Therefore, the admixture and close relationship of American populations with commercial breeds likely contributed to observe Asian ancestry in the evaluated populations.

Several works have stressed the introgression of commercial breeds into American populations (e.g. Martínez et al. 2005; Silva et al. 2011; Lemus-Flores et al. 2001). In general, the American village pig populations are breed under traditional methods (e.g. backyard) using agricultural by-products or forage. As mentioned by Martínez et al. (2005), the "creole" pig is completely integrated into the history, culture and lifestyle of smallholders. Nevertheless, commercial breeds have been gradually replacing the creole pig populations mostly to increase the performance of creoles (Lemus-Flores *et al.*, 2001) and/or simply by

unawareness of their valuable genetic resource. Both can potentially endanger a genetic resource well adapted to their local environment.

Initially, the introgression of commercial breeds is not uniform. It means that several breeds in different times and routes have introgressed the American village pigs. In order to estimate some routes for this introgression, we have performed several analyses considering different SNPs their ancestry or frequency, as well as we have evaluate additional strategies to evaluate the introgression of commercial breeds (e.g. Treemix; Pickrell and Pritchard, 2012). Although the analyses support the contribution of commercial breeds in the genetic background of extant American village pigs and provided us some directions of these introgression events, the limitation in the SNPs used, in terms of ascertainment bias, limited the power of the analyses. It is likely that the use of genome sequences help us to estimate how the introgression events occurred in American village pigs.

Moreover, using the partial ancient genome of an Iberian pig (Chapter 5), we can recover part of 500 years of history of pigs, one of the most important periods for artificial selection in Europe and contemporary to the expansion to America. We were able to directly observe variation in pig genome previous to introgression of Asian populations into Europe and possibly to detect the ancient signals of ancestry. In this sense, we explored two genetic signatures of ancestry: 1) Near East ancestry in the ancient European sample and, 2) the ancestry of pigs in America. Near East pigs were introduced in Europe during Neolithic (Ottoni *et al.*, 2013), but they were rapidly replaced by European local haplotypes. Previous report support the absence of this ancestry in extant European pigs (Manunza et al. 2013). Considering the antiquity of the ancient sample, we suspected to find some evidence of ancestry although that was not observed (Chapter 5, Figure 2). The analysis of Near Eastern sequences as well as the study of older samples may help us to understand this event in the history of pigs.

Noteworthy, the SNP array genotyped (Chapter 4) data and the NGS data (Chapter 5) showed that the ancient and modern Iberian pigs are equally close to American pigs and corroborated the previous interpretation that Iberian pigs are not the primary genetic ancestry contribution in American village pigs nowadays. If we consider the ancient sample as a proxy of the ancestral Iberian pig population, the observed results suggest that Iberian pigs have not largely changed his genetic composition and variability in the recent centuries (Esteve-Codina *et al.*, 2013; Herrero-Medrano *et al.*, 2013). Moreover, we can see rapid changes in American populations along time compared to Iberian pigs; the commercial pig genetic component is relevant to extant populations, which in turn is influenced by the variability attained after introduction of Asian populations at 18th century. The founder effect of American populations was strong because the limited number of Iberian pigs firstly introduced. They were abundant in Antilles and rapidly were found in mainland of Central and South America. The environmental conditions alongside battles in the conquest period witnessed the faster growth of some pig populations than others across territory (Del Río Moreno, 1996). However, considering the genetic variability observed in both ancient and modern Iberian pigs as an initial value of the American village pig variability, the introgression of commercial breeds largely modified the genetic structure of American village

pigs. The estimation of demographic changes in American populations will be relevant for disentangling their genetic trajectory.

Considerations in the estimations of diversity and ancestry

Using a particular data set of SNPs in SSCX we described the diversity of pig populations and demonstrate their utility to detect the relatedness in pigs. Additionally we used the Illumina's porcine SNP60K BeadChip (Ramos *et al.*, 2009), to explore the autosomal diversity in pigs. However, because there is an ascertainment bias with analysed SNPs, cautions with the interpretations have been considered.

Using simulations, we observed that SNP ascertainment bias could affect the estimations of population structure and ancestry related with Asian pigs (chapter 4). In fact, the results of simulations suggested that we could find Asian ancestry in American populations, even if no gene flow occurred between Asia and America due to the Asian haplotypes carried by commercial breeds. The porcine SNP60K BeadChip was designed with SNPs obtained from six populations including four commercial breeds and two wild boars: one from Europe and one from Japan. Therefore, SNP selection is ascertained towards intermediate frequency loci in western breeds. This array has been useful to detect quantitative trait loci (e.g. Fan et al. 2011; Sanchez et al. 2014), copy number variations (e.g Ramayo-Caldas et al. 2010; Chen et al. 2012) or linkage disequilibrium (e.g Uimari and Tapio 2011; Badke et al. 2012) in commercial or crossbreeding populations. However, estimations of LD decay previously noticed of the large differences in SNP requirements when were used to explore genetically local populations from Asia and Europe (Amaral *et al.*, 2008). For example, Ai, Huang, and Ren (2013) found 3-fold more monomorphic SNPs in the Chinese animals than the western counterpart. An ascertainment bias correction is not straightforward due the complex demography of pig, but simulations in a most realistic history provide suitable information for further comparisons.

Finally, the ancient pig sample represents a valuable source of information but extremely optimal conditions are required to obtain useful endogenous DNA from ancient samples. We only were capable to obtain ~0.11x coverage genome from the pig remains, a low value compared to those obtained for example in Thistle Creek horse (Orlando *et al.*, 2013), ancient human Saqqaq remains (Rasmussen *et al.*, 2010) or Denisovan (Meyer *et al.*, 2012). Fortunately, lower values of environmental DNA contamination were found (including human DNA) allowing us obtaining a sequence with enough quality and reliability. Because the low depth in ancient sample, it is expected some constrains to detect polymorphism resulting in a decrease of heterozygous positions and new variants. Sequencing error together with spurious substitutions post mortem can affect also the polymorphism detection.

7.2 Clues on selective processes revealed from ancient and modern pig genomes

Since domestication, pigs have suffered phenotypic changes affecting behaviour (White, 2011) or morphology (Cucchi *et al.*, 2011; Ottoni *et al.*, 2013). These phenotypic variants have been traced and allowed reconstruct (up to some level) the history and evolution of

populations (Rowley-Conwy *et al.*, 2012). Analyses with autosomal markers (i.e. microsatellites or SNPs genotypes) as well as mitochondrial sequences have contributed primarily to establish the current genetic status of populations, track the ancient migration routes and identify signatures of selection in the genome. But is the release of the sequence of pig genome (Groenen *et al.*, 2012) that opened a new window of possibilities to understand the evolution (Frantz *et al.*, 2013), demographic patterns (Bosse *et al.*, 2012), structural variations (Esteve-Codina *et al.*, 2013; Paudel *et al.*, 2013) and selection (Rubin *et al.*, 2012; Li *et al.*, 2013) in pigs.

It is difficult to measure how variations arose by adaptive selection in the genome of American village pigs. The comparison of allele frequencies of variations in American samples with Iberian pigs may help to understand how adaptive selection and demography modelled the variability of American populations. Although, the gene flow from modern commercial breeds to American village pigs is also an important source of standing variants (Hedrick, 2013), especially considering that those populations carries a high proportion of Asian ancestry.

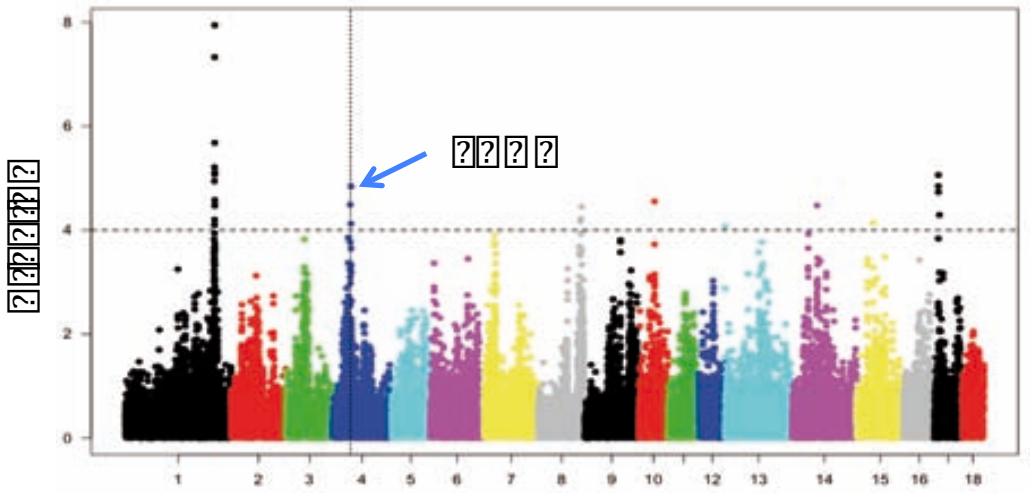
Two selective processes were evaluated in this thesis: 1) those affecting American village pig populations, and 2) those before Asian introgression to Europe. First, several generations of natural selection conferred to the American village pigs the adaptation to a wide range of environmental conditions, disease resistance and low management requirements (Linares *et al.*, 2011). For example, despite creole pigs show lower reproductive performance or growth rate compared to commercial breeds, they are capable to survive in extreme environmental conditions and lack of feed, whereas commercial breeds requires specific conditions of temperature, humidity and feed to growth. Interestingly, when we evaluated selective signatures in the genome of American populations and contrast them with the genome of commercial breeds, we observed several high-differentiated genomic regions with genes involved in processes like limb morphogenesis and development, skeletal system development and less significant genes related with locomotion or metabolic regulation (Chapter 4, Supplementary file 8). The large number of regions detected indicates that multiple loci control the local adaptation in America. In this case, the understanding of how migration and selection in these populations have modified the patterns of variation is difficult because some variants could have a favourable effect in some conditions but being deleterious in other (Savolainen *et al.*, 2013).

Nowadays is very necessary to elaborate accurate experimental designs for estimating the effect of introgression (Blanquart *et al.*, 2013). However, it is possible to compare the differentiation of commercial populations with those that showed particular phenotypes, like miniature pigs or populations adapted to high altitude. These phenotypes showed particular signatures of selection. These populations were analysed using two methods: allele differentiation based on F_{ST} (Akey *et al.*, 2002) and extended haplotype homozygosity (iHS; Voight *et al.*, 2006). The first method showed that several genomic regions have large allele frequency differences between American village pigs. It corroborates that some variations could increase (or decrease) their frequency in response to a specific environmental factor. However, the second method suggested that some variants have moved their frequencies towards the derived allele probably by increasing of their fitness. The genomic region with

signatures of selection contained some genes involved in related metabolic functions of the phenotype. For instance, the selected regions for miniature pig harboured genes involved in limb morphogenesis and development (Chapter 4; Supplementary file 8). In the case of high altitude population, genomic regions contained genes associated with respiratory disease, hypoxia and blood circulation (Chapter 4; Figure 5).

Additional statistics has been used to estimate selective signatures in American village pig populations. The most interesting result comes from comparison of Peruvian (a population adapted to the high altitude) and the rest of American populations. We inspected differential patterns of extended haplotype homozygosity (EHH) between populations (Tang *et al.*, 2007). Surprisingly, this statistic allowed us identify additional genomic regions with selective signatures (Figure 1), but in particularly one region in SSC4 that contains the gene *STK3* a protein kinase activation that presumably allows cells to resist unfavorable environmental conditions (Taylor *et al.*, 1996). Further, a gene network analyses, using ClueGO (Bindea *et al.*, 2009), with all genes located in the regions with significant signal of selection (P-value of Rsb score <0.001) showed that the *STK3* gene is highly connected with other genes related to stimulus signal processes. The information derived from SNPs analyzed here (i.e allele frequencies and EHH) provided clear evidences of selective signatures in American village pigs, despite the limited number of SNPs and sample size. The analysis of complete genome sequence could benefit the detection of selective signatures in American pig populations, and will help us to answer some important questions like: are the causative variants of the new environmental adaptation recently created? Are they originated from the ancestral founder populations? Or they are the result of recent introgression of commercial pigs?

The comparison of the genomes of extant domestic populations with the wild counterpart is currently the main source of information about domestication and selective processes. For example, Rubin et al. (2012) explored the effects of domestication comparing commercial and domestic pigs with wild boars from Europe and found strong selective signatures (by selective sweeps analyses) in regions with genes involved in the increased number of vertebrae in domestic pigs and also in the coat color. Likewise, Wilkinson et al. (2013) using a combination of SNP array and re-sequencing, identified selective signatures for coat color, ear morphology and loci related with productive traits in pigs.



uub

7.3 The use of weights in principal component analysis

One of the most common situations in genetics is the uneven sample size of populations. Their impact in genetic estimations falls in the uncertainty of allele frequencies estimated because by chance sampling might not well-represent the alleles in the population (Chakraborty, 1992; Hale *et al.*, 2012). Notably, several genetic parameters depend of allele frequencies observed in the sample and principal component analysis (PCA) is not the exception. The PCA has been largely used to explore the population structure in data because the easy computational implementation and faster estimates (Patterson *et al.*, 2006). Indeed, new contributions to increase the efficiency of PCA estimation leads to perform the analysis of thousands of individuals up to 125-fold times faster (Abraham and Inouye, 2014).

Prior to our work, the problem of uneven sampling or low sample representation in PCA had not been studied and only a particular proposal have been done to correct this bias (McVean, 2009). In fact, many researchers seem to be unaware of this PCA behaviour. We explored a strategy to solve this issue by applying weights in the allele frequency covariance matrix used for estimation of eigenvalues and eigenvectors. As observed, when sampling is performed considering a particular feature (e.g geographical or phenotypic) of populations, the resulted projection of data reflects the sampling instead of underlying relationships (Chapter 6, Figure 1). In this sense, the used weights could be considered as prior information about the relationships of populations (supervised PCA) and therefore contribute to the improvement of graphical representation. Consequently, corrections potentially favoured the interpretation of population relationships in presence of uneven or very low sample size of some populations.

The correction described in Chapter 6 provided a suitable strategy to deal with differences in sample size of studied populations. In general, if covariance matrix is weighted using, for each individual, the sample size of population that individuals come from, the projections tend to the most realistic underlying demographic model. For well-structured populations the correction is useful even in sample sizes equal than *n=1*. However, cautions in the interpretation can be considered because that lower sample size potentially does not represent the diversity of a population and ultimately the population relationships (Chakraborty, 1992; Fung and Keenan, 2014).

We noticed also the robustness of these corrections because the arbitrary use of weights resulted in dramatic changes in sample projections. It is expected that used weights correlates with history or relatedness of populations. In any case, the lack of knowledge about population relatedness can be explored using different weights. This was observed when we considered the Iberian origin of La Braña 1 sample (Chapter 6, Figure 10). The PCA projection was overcorrected lacking a clear interpretation. Nevertheless, when La Braña 1 was clustered and weighted similar to Northern samples an enhanced interpretation can be perform.

Finally, we applied this method to the American village pigs SNPs data set used in Chapter 4. The PCA performed in this study showed two axes of differentiation between populations (Chapter 4, figure 1). The first axis highlight the geographical differentiation of Chinese

breeds with European, and the second axis reflect differences between well-established commercial breeds. We observed some differences in the sample projection in the 2-dimensional space for PCA and wPCA (Figure 2). The use of the weight based on w = 1/n showed the expected pattern of differentiation described in PCA with an atomized projections for populations as result of an overcorrection for CUCE (n = 1). However, the wPCA with F_{ST} correction showed additional patterns of differentiation between populations that only was observed with other analyses like model-based clustering (Chapter 4, Figure 4). Duroc breed (DU) and other commercial breeds were projected closer to American populations and the differentiation within Chinese breeds is more evident. Interestingly, wPCA show the Brazilian populations more distant to the rest of Americans, which coincides with the different genetic contribution of Iberian pigs in their history.

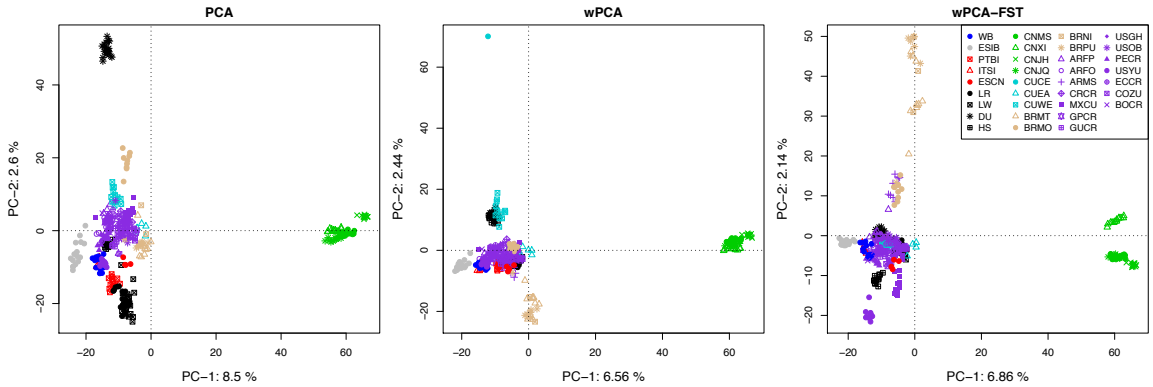


Figure 2. PCA and wPCA in American village pig populations

Additional strategies for weights selection could be used. Criteria like phenotype, sample origin or model based cluster, always with biological sense, can improve the PCA projections in presence of uneven sample size.

Chapter 8

-

Conclusions

-

1. Indirect estimation of SNP heterozygosity on males allowed us to delineate the pseudoautosomal region of SSCX, locating its boundary around 7Mb. The differential analysis of the PAR and the NPAR of SSCX provided relevant information about relatedness of worldwide pig populations and their differences caused by demography.

2. A clear genetic differentiation of Asian and western pig populations has been observed in SSCX and autosomes, as a result of a long divergence of the populations. Nevertheless, the low genetic differentiation of commercial breeds and American village pigs suggested an evident introgression of commercial haplotypes, likely motivated for the needed to increase the performance of the American village pigs.

3. The genetic diversity of American village pig populations is relatively high. However, both low genetic differentiation between them and no correlation between genetic and geographic distance support the hypothesis that multiple breeds or populations of pigs, like Iberian or Bisaro and recently for Commercial breeds, have contributed in their formation. The reduction of Iberian ancestry in American village pigs is prominent, being replaced by Commercial haplotypes. Only few populations with particular phenotypes (e.g mini-pigs) still conserve a large European ancestry.

4. We have reported for first time the partial genome sequenced of a pig ancient sample from the 16th century. The sample from Montsoriu Castle (Girona, Spain) corresponded to a female and was genetically close to the modern Iberian pig populations. This genome has provided information that confirmed ancient admixture events between Iberian pigs

and European wild boars and selection previously to introgression of Asian breeds in Europe.

5. The proposed correction for PCA can dramatically improve the graphical representation of population relationships in the presence of uneven sample size. We found this method to be effective and robust, and the performance increases when weights correlates with the history or demography of populations.

-

REFERENCES

-

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Abraham G, Inouye M (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**: e93766.

Ai H, Huang L, Ren J (2013). Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. *PLoS One* **8**.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**: 1805–1814.

Albrechtsen A, Nielsen FC, Nielsen R (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* **27**: 2534–2547.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**: 1655–1664.

Alexandri P, Triantafyllidis A, Papakostas S, Chatzinikos E, Platis P, Papageorgiou N, *et al.* (2012). The Balkans and the colonization of Europe: the post-glacial range expansion of the wild boar, Sus scrofa. *J. Biogeogr.* **39**: 713–723.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, *et al.* (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52–58.

Alves E, Fernández AI, Fernández-Rodríguez A, Pérez-Montarelo D, Benitez R, Ovilo C, *et al.* (2009). Identification of mitochondrial markers for genetic traceability of European wild boars and Iberian and Duroc pigs. *Animal* **3**: 1216–23.

Amaral AJ, Ferretti L, Megens HJ, Crooijmans RPMA, Nie H, Ramos-Onsins SE, *et al.* (2011). Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One* **6**.

Amaral AJ, Megens H-J, Crooijmans RPMA, Heuven HCM, Groenen MAM (2008). Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* **179**: 569–579.

Amaral AJ, Megens H-J, Kerstens HHD, Heuven HCM, Dibbits B, Crooijmans RPMA, *et al.* (2009). Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**: 374.

Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, *et al.* (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science (80-.).* **263**: 1771–1774.

Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP (2012). Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* **13**: 24.

Baruch E, Weller JI (2008). Estimation of the number of SNP genetic markers required for parentage verification. *Anim. Genet.* **39**: 474–479.

Bhatia G, Patterson N, Sankararaman S, Price AL (2013). Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**: 1514–21.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, *et al.* (2009). ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**: 1091–1093.

Bird MI, Taylor D, Hunt C (2005). Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: A savanna corridor in Sundaland? *Quat. Sci. Rev.* **24**: 2228–2242.

Blackburn H, Gollin D (2009). Animal genetic resource trade flows: The utilization of newly imported breeds and the gene flow of imported animals in the United States of America. *Livest. Sci.* **120**: 240–247.

Blanquart F, Kaltz O, Nuismer SL, Gandon S (2013). A practical guide to measuring local adaptation. *Ecol. Lett.* **16**: 1195–205.

Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LAF, Schook LB, *et al.* (2012). Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. *PLoS Genet.* **8**.

Burgos-Paz W, Souza CA, Megens HJ, Ramayo-Caldas Y, Melo M, Lemús-Flores C, *et al.* (2013). Porcine colonization of the Americas: a 60k SNP story. *Heredity (Edinb).* **110**: 321–30.

Butzer KW (1992). The Americas before and after 1492: An Introduction to Current Geographical Research. *Ann. Assoc. Am. Geogr.* **82**: 345–368.

Bywater K a., Apollonio M, Cappai N, Stephens P a. (2010). Litter size and latitude in a large mammal: the wild boar Sus scrofa. *Mamm. Rev.* **40**: 212–220.

Campbell TA, Long DB (2009). Feral swine damage and damage management in forested ecosystems. *For. Ecol. Manage.* **257**: 2319–2326.

Chakraborty R (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Hum. Biol.* **64**: 141–59.

Chen K, Baxter T, Muir WM, Groenen MA, Schook LB (2007). Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). *Int. J. Biol. Sci.* **3**: 153–165.

Chen C, Qiao R, Wei R, Guo Y, Ai H, Ma J, *et al.* (2012). A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics* **13**: 733.

Chenuil A (2006). Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica* **127**: 101–120.

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

Corominas J, Ramayo-Caldas Y, Puig-Oliveras A, Estellé J, Castelló A, Alves E, *et al.* (2013). Analysis of porcine adipose tissue transcriptome reveals differences in de novo fatty acid synthesis in pigs with divergent muscle fatty acid composition. *BMC Genomics* **14**: 843.

Cucchi T, Hulme-Beaman A, Yuan J, Dobney K (2011). Early Neolithic pig domestication at Jiahu, Henan Province, China: clues from molar shape analyses using geometric morphometric approaches. *J. Archaeol. Sci.* **38**: 11–22.

Das PJ, Mishra DK, Ghosh S, Avila F, Johnson G a, Chowdhary BP, *et al.* (2013). Comparative organization and gene expression profiles of the porcine pseudoautosomal region. *Cytogenet. Genome Res.* **141**: 26–36.

Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot S V (2012). Whole-Genome sequencing and genetic variant analysis of a quarter Horse mare. *BMC Genomics* **13**: 78.

Drineas P, Lewis J, Paschou P (2010). Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. *PLoS One* **5**.

Durand EY, Patterson N, Reich D, Slatkin M (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**: 2239–2252.

Engelhardt BE, Stephens M (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* **6**: e1001117.

Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M (2011). Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics* **12**: 552.

Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silió L, *et al.* (2013). Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. *BMC Genomics* **14**: 148.

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

Fan B, Onteru SK, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF (2011). Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. *PLoS One* **6**.

Fang M, Andersson L (2006). Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. *Proc. Biol. Sci.* **273**: 1803–10.

Fang M, Berg F, Ducos A, Andersson L (2006). Mitochondrial haplotypes of European wild boars with 2n = 36 are closely related to those of European domestic pigs with 2n = 38. *Anim. Genet.* **37**: 459–464.

Fang M, Hu X, Jiang T, Braunschweig M, Hu L, Du Z, *et al.* (2005). The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Anim. Genet.* **36**: 7–13.

Fang M, Larson G, Ribeiro HS, Li N, Andersson L (2009). Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet.* **5**.

Fang X, Mou Y, Huang Z, Li Y, Han L, Zhang Y, *et al.* (2012). The sequence and analysis of a Chinese pig genome. *Gigascience* **1**: 16.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013). Population genomics from pool sequencing. *Mol. Ecol.*: 5561–5576.

Frantz LA, Schraiber JG, Madsen O, Megens H-J, Bosse M, Paudel Y, *et al.* (2013). Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome Biol.* **14**: R107.

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, *et al.* (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLoS Genet.* **10**: e1004016.

Fung T, Keenan K (2014). Confidence intervals for population allele frequencies: the general case of sampling from a finite diploid population of any size. *PLoS One* **9**: e85925.

Gabor TM, Hellgren EC, Bussche RA, Silvy NJ (1999). Demography, sociospatial behaviour and genetics of feral pigs (Sus scrofa) in a semi-arid environment. *J. Zool.* **247**: 311–322.

Gade DW (2012). Hogs (pigs). In: Press CU (ed) *The Cambridge World History of Food*, Cambridge Vol 1, pp 536–541.

Galíndez R, Ramis C, Angulo L (2011). Exploración inicial de la diversidad genética del cerdo criollo venezolano usando RAPD. *Rev. la Fac. Agron.* **37**: 55–63.

Gautier M, Foucaud J, Gharbi K, Cézard T, Galan M, Loiseau A, *et al.* (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol. Ecol.* **22**: 3766–79.

Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT, Andersson L (2000). The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**: 1785–1791.

Goddard ME, Hayes BJ, Meuwissen THE (2010). Genomic selection in livestock populations. *Genet. Res. (Camb).* **92**: 413–421.

Goedbloed DJ, Megens HJ, Van Hooft P, Herrero-Medrano JM, Lutz W, Alexandri P, *et al.* (2013). Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. *Mol. Ecol.* **22**: 856–66.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, *et al.* (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108**: 11983–11988.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, *et al.* (2010). A draft sequence of the Neandertal genome. *Science (80-.).* **328**: 710–722.

Groenen M a M, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, *et al.* (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**: 393–8.

Guillot G, Foll M (2009). Correcting for ascertainment bias in the inference of population structure. *Bioinformatics* **25**: 552–4.

Hale ML, Burg TM, Steeves TE (2012). Sampling for Microsatellite-Based Population Genetic Studies: 25 to 30 Individuals per Population Is Enough to Accurately Estimate Allele Frequencies. *PLoS One* **7**.

Hartl DL, Clark AG (1997). *Principles of population genetics*. Sinauer associates Sunderland.

Hedrick PW (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* **22**: 4606–18.

Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, *et al.* (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* **11 Suppl 1**: 123–136.

Herrero-Medrano JM, Megens H-J, Groenen MAM, Ramis G, Bosse M, Pérez-Enciso M, *et al.* (2013). Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC Genet.* **14**: 106.

Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.

Holsinger KE, Weir BS (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* **10**: 639–650.

Hudson RR, Slatkin M, Maddison WP (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.

Hulsegge B, Calus MPL, Windig JJ, Hoving-Bolink a H, Maurice-van Eijndhoven MHT, Hiemstra SJ (2013). Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. *J. Anim. Sci.* **91**: 5128–34.

Ji Y-Q, Wu D-D, Wu G-S, Wang G-D, Zhang Y-P (2011). Multi-Locus Analysis Reveals A Different Pattern of Genetic Diversity for Mitochondrial and Nuclear DNA between Wild and Domestic Pigs in East Asia. *PLoS One* **6**: e26416.

Johansson MM, Chaudhary R, Hellmén E, Höyheim B, Chowdhary B, Andersson L (1996). Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. *Mamm. Genome* **7**: 822–830.

Jombart T, Pontier D, Dufour A-B (2009). Genetic markers in the playground of multivariate analysis. *Heredity (Edinb).* **102**: 330–41.

Kalinowski ST (2002). How many alleles per locus should be used to estimate genetic distances? *Heredity (Edinb).* **88**: 62–65.

Kerstens HH, Crooijmans RP, Dibbits BW, Vereijken A, Okimoto R, Groenen MA (2011). Structural variation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced representation libraries. *BMC Genomics* **12**: 94.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, *et al.* (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* **10**: e1001258.

Kijas JM, Wales R, Törnsten A, Chardon P, Moller M, Andersson L (1998). Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics* **150**: 1177–1185.

Knox R V (2014). Impact of swine reproductive technologies on pig and global food production. *Adv. Exp. Med. Biol.* **752**: 131–60.

Kofler R, Pandey RV, Schlötterer C (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**: 3435–6.

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* **14**: 289.

Krause-Kyora B, Makarewicz C, Evin A, Flink LG, Dobney K, Larson G, *et al.* (2013). Use of domesticated pigs by Mesolithic hunter-gatherers in northwestern Europe. *Nat. Commun.* **4**: 2348.

Kriegel H-P, Kröger P, Schubert E, Zimek A (2008). *Scientific and Statistical Database Management* (B Ludäscher and N Mamoulis, Eds.). Springer Berlin Heidelberg: Berlin, Heidelberg.

Lai WL, Tan C-B (Eds.) (2010). *The Chinese in Latin America and the Caribbean*. Brill Academic Publishers.

Larson G, Albarella U, Dobney K, Rowley-Conwy P, Schibler J, Tresset A, *et al.* (2007). Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc. Natl. Acad. Sci. U. S. A.* **104**: 15276–15281.

Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, *et al.* (2007). Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. *Proc. Natl. Acad. Sci. U. S. A.* **104**: 4834–4839.

Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, *et al.* (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science (80-.).* **307**: 1618–1621.

Larson G, Liu R, Zhao X, Yuan J, Fuller D, Barton L, *et al.* (2010). Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 7686–7691.

Laval G, Iannuccelli N, Legault C, Milan D, Groenen MA, Giuffra E, *et al.* (2000). Genetic diversity of eleven European pig breeds. *Genet. Sel. Evol.* **32**: 187–203.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012). Inference of population structure using dense haplotype data. *PLoS Genet.* **8**.

Lemus-Flores C, Ulloa-Arvizu R, Ramos-Kuri M, Estrada FJ, Alonso RA (2001). Genetic analysis of Mexican hairless pig populations. *J. Anim. Sci.* **79**: 3021–3026.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, *et al.* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science (80-.).* **319**: 1100–1104.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, *et al.* (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* **45**: 1431–8.

Linares V, Linares L, Mendoza G (2011). Caracterización etnozootécnica y potencial carnicero de Sus scrofa "cerdo criollo" en Latinoamérica. *Sci. Agropecu.* **2**: 97–110.

Lindgren G, Backström N, Swinburne J, Hellborg L, Einarsson A, Sandberg K, *et al.* (2004). Limited number of patrilines in horse domestication. *Nat. Genet.* **36**: 335–336.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005). Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genet.* **6 Suppl 1**: S26.

Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, *et al.* (2013). High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **110**: 19872–7.

Luetkemeier ES, Sodhi M, Schook LB, Malhi RS (2010). Multiple Asian pig origins revealed through genomic analyses. *Mol. Phylogenet. Evol.* **54**: 680–686.

Lum JK, McIntyre JK, Greger DL, Huffman KW, Vilar MG (2006). Recent Southeast Asian domestication and Lapita dispersal of sacred male pseudohermaphroditic "tuskers" and hairless pigs of Vanuatu. *Proc. Natl. Acad. Sci. U. S. A.* **103**: 17190–17195.

MacEachern S, Hayes B, McEwan J, Goddard M (2009). An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic ca. *BMC Genomics* **10**: 181.

Van der Made J (2001). The ungulates from Atapuerca: Stratigraphy and biogeography. *Anthropologie* **105**: 95–113.

Manel S, Gaggiotti OE, Waples RS (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. Evol.* **20**: 136–42.

Manunza A, Zidi A, Yeghoyan S, Balteanu VA, Carsai TC, Scherbakov O, *et al.* (2013). A High Throughput Genotyping Approach Reveals Distinctive Autosomal Genetic Signatures for European and Near Eastern Wild Boar. *PLoS One* **8**.

Mardis ER (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**: 133–141.

Mariante AS, Cavalcante N (2006). *Animals of the Discovery* (Embrapa, Ed.). Embrapa: Brasilia.

Martínez AM, Pérez-Pineda E, Vega-Pla JL, Barba C, Velázquez FJ, Delgado J V (2005). Caracterización genética del cerdo criollo cubano con microsatélites. *Arch. Zootec* **54**: 369–375.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, *et al.* (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**: e5350.

McTavish EJ, Decker JE, Schnabel RD, Taylor JF, Hillis DM (2013). New World cattle show ancestry from multiple independent domestication events. *Proc. Natl. Acad. Sci. U. S. A.* **110**: E1398–406.

McVean G (2009). A genealogical interpretation of principal components analysis. *PLoS Genet.* **5**: e1000686.

Megens H-J, Crooijmans RP, Cristobal MS, Hui X, Li N, Groenen MA (2008). Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genet. Sel. Evol.* **40**: 103–128.

Meiri M, Huchon D, Bar-Oz G, Boaretto E, Horwitz LK, Maeir AM, *et al.* (2013). Ancient DNA and population turnover in southern levantine pigs--signature of the sea peoples migration? *Sci. Rep.* **3**: 3035.

Menozzi P, Piazza A, Cavalli-Sforza L (1978). Synthetic maps of human gene frequencies in Europeans. *Science (80-.).* **201**: 786–792.

Merino ML, Carpinetti BN (2003). Feral pig Sus scrofa population estimates in Bahía Samborombón Conservation Area, Buenos Aires province, Argentina. *Mastozoología Neotrop.* **10**: 269–275.

Metcalf AC (2005). *Go-betweens and the Colonization of Brazil: 1500--1600*. University of Texas Press.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**: 31–46.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, *et al.* (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science (80-.).* **338**: 222–6.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, *et al.* (2008). Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**: 387–390.

Minoche AE, Dohm JC, Himmelbauer H (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* **12**: R112.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, *et al.* (2013). Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* **9**: e1003925.

Nevado B, Ramos-Onsins SE, Perez-Enciso M (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Mol. Ecol.* **23**: 1764–1779.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One* **7**.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, *et al.* (2008). Genes mirror geography within Europe. *Nature* **456**: 98–101.

Novembre J, Stephens M (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**: 646–9.

Ojeda A, Ramos-Onsins SE, Marletta D, Huang LS, Folch JM, Pérez-Enciso M (2011). Evolutionary study of a potential selection target region in the pig. *Heredity (Edinb).* **106**: 330–8.

Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, *et al.* (2014). Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*.

Olivieri I (2009). Alternative mechanisms of range expansion are associated with different changes of evolutionary potential. *Trends Ecol. Evol.* **24**: 289–92.

Ollivier L, Alderson L, Gandini GC, Foulley J-L, Haley CS, Joosten R, *et al.* (2005). An assessment of European pig diversity using molecular markers: Partitioning of diversity among breeds. *Conserv. Genet.* **6**: 729–741.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, *et al.* (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**: 74–8.

Ottoni C, Flink LG, Evin A, Geörg C, De Cupere B, Van Neer W, *et al.* (2013). Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics. *Mol. Biol. Evol.* **30**: 824–32.

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, *et al.* (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**: 645–679.

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, *et al.* (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* **3**: 1672–86.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, *et al.* (2012). Ancient admixture in human history. *Genetics* **192**: 1065–93.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet.* **2**: e190.

Paudel Y, Madsen O, Megens H-J, Frantz L a F, Bosse M, Bastiaansen JWM, *et al.* (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* **14**: 449.

Pickrell JK, Pritchard JK (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* **8**.

Pool JE, Nielsen R (2007). Population size changes reshape genomic patterns of diversity. *Evolution (N. Y).* **61**: 3001–3006.

Porter V (1993). *Pigs: a handbook to the breeds of the world*. Helm Information.

Price EO (1999). Behavioral development in animals undergoing domestication. *Appl. Anim. Behav. Sci.* **65**: 245–271.

Price EO (2002). *Animal domestication and behavior* (EO Price, Ed.). CABI: Wallingford.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**: 904–9.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Quilter CR, Blott SC, Mileham AJ, Affara NA, Sargent CA, Griffin DK (2002). A mapping and evolutionary study of porcine sex chromosome genes. *Mamm. Genome* **13**: 588–594.

Raj A, Stephens M, Pritchard JK (2013). Variational Inference of Population Structure in Large SNP Datasets.

Ramayo-Caldas Y, Castelló A, Pena RN, Alves E, Mercadé A, Souza CA, *et al.* (2010). Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC Genomics* **11**: 593.

Ramayo-Caldas Y, Mach N, Esteve-Codina A, Corominas J, Castelló A, Ballester M, *et al.* (2012). Liver transcriptome profile in pigs with extreme phenotypes of intramuscular fatty acid composition. *BMC Genomics* **13**: 547.

Ramírez O, Gigli E, Bover P, Alcover JA, Bertranpetit J, Castresana J, *et al.* (2009). Paleogenomics in a temperate environment: Shotgun sequencing from an extinct Mediterranean caprine. *PLoS One* **4**.

Ramírez O, Ojeda A, Tomàs A, Gallardo D, Huang LS, Folch JM, *et al.* (2009). Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol. Biol. Evol.* **26**: 2061–2072.

Ramírez-Soriano A, Nielsen R (2009). Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* **181**: 701–10.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, *et al.* (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* **4**.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford Jr TW, *et al.* (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**: 225–229.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, *et al.* (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**: 757–762.

Reich D, Price AL, Patterson N (2008). Principal component analysis of genetic data. *Nat. Genet.* **40**: 491–2.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009). Reconstructing Indian population history. *Nature* **461**: 489–94.

Revidatti MA (2009). Caracterización de cerdos criollos del Nordeste Argentino. Universidad de Córdoba.

Del Río Moreno JL (1996). El cerdo. Historia de un elemento esencial de la cultura castellana en la conquista y colonización de América (siglo XVI). *Anu. Estud. Am.* **53**: 13–35.

Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D (2012). Ancient DNA studies: new perspectives on old samples. *Genet. Sel. Evol.* **44**: 21.

Rodero A, Delgado J V, Rodero E (1992). Primitive Andalusian livestock and their implications in the discovery of America. *Arch. Zootec* **41**: 383–400.

Rothschild MF, Hu Z, Jiang Z (2007). Advances in QTL mapping in pigs. *Int. J. Biol. Sci.* **3**: 192–197.

Rowley-Conwy P, Albarella U, Dobney K (2012). Distinguishing Wild Boar from Domestic Pigs in Prehistory: A Review of Approaches and Recent Results. *J. World Prehistory* **25**: 1–44.

Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, *et al.* (2012). Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U. S. A.* **109**: 19529–36.

Ruvinsky A, Rothschild MF (1998). Systematics and evolution of the pig. *Genet. Pig*: 1–16.

Sanchez M-P, Tribout T, Iannuccelli N, Bouffaud M, Servin B, Tenghe A, *et al.* (2014). A genome-wide association study of production traits in a commercial population of Large White pigs: evidence of haplotypes affecting meat quality. *Genet. Sel. Evol.* **46**: 12.

Savolainen O, Lascoux M, Merilä J (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.* **14**: 807–20.

Scandura M, Iacolina L, Crestanello B, Pecchioli E, Di Benedetto MF, Russo V, *et al.* (2008). Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: are the effects of the last glaciation still detectable? *Mol. Ecol.* **17**: 1745–1762.

Schaffner SF (2004). The X chromosome in population genetics. *Nat. Rev. Genet.* **5**: 43–51.

Schlötterer C (2004). The evolution of molecular markers--just a matter of fashion? *Nat. Rev. Genet.* **5**: 63–69.

Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KAS, Willerslev E, *et al.* (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**: 178.

Setchell BP (1992). Domestication and reproduction. *Anim. Reprod. Sci.* **28**: 195–202.

Shapiro B, Hofreiter M (2014). A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. *Science (80-.).* **343**: 1236573–1236573.

Shendure J, Aiden EL (2012). The expanding scope of DNA sequencing. *Nat. Biotechnol.* **30**: 1084–1094.

Sierra C (2001). El cerdo cimarrón (Sus scrofa , Suidae) en la Isla del Coco , Costa Rica : Composición de su dieta , estado reproductivo y genética. *Int. J.* **49**: 1147–1157.

Silva EC da, Dutra Junior WM, Ianella P, Gomes Filho MA, Oliveira CJP de, Ferreira DN de M, *et al.* (2011). Patterns of genetic diversity of local pig populations in the State of Pernambuco, Brazil. *Rev. Bras. Zootec.* **40**: 1691–1699.

Skinner BM, Lachani K, Sargent C a, Affara N a (2013). Regions of XY homology in the pig X chromosome and the boundary of the pseudoautosomal region. *BMC Genet.* **14**: 3.

Skoglund P, Northoff BH, Shunkov M V., Derevianko AP, Paabo S, Krause J, *et al.* (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci.*: 1318934111–.

Sollero BP, Paiva SR, Faria DA, Guimarães SEF, Castro STR, Egito AA, *et al.* (2009). Genetic diversity of Brazilian pig breeds evidenced by microsatellite markers. *Livest. Sci.* **123**: 8–15.

Souza CA, Paiva SR, Pereira RW, Guimarães SEF, Dutra WM, Murata LS, *et al.* (2009). Iberian origin of Brazilian local pig breeds based on Cytochrome b (MT-CYB) sequence. *Anim. Genet.* **40**: 759–762.

Stoneking M, Krause J (2011). Learning about human population history from ancient and modern genomes. *Nat. Rev. Genet.* **12**: 603–614.

Stothard P, Choi J-W, Basu U, Sumner-Thomson JM, Meng Y, Liao X, *et al.* (2011). Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics* **12**: 559.

Sunnucks P (2000). Efficient genetic markers for population biology. *Trends Ecol. Evol.* **15**: 199–203.

Tanaka K, Iwaki Y, Takizawa T, Dorji T, Tshering G, Kurosawa Y, *et al.* (2008). Mitochondrial diversity of native pigs in the mainland South and South-east Asian countries and its relationships between local wild boars. *Anim. Sci. J.* **79**: 417–434.

Tang H, Peng J, Wang P, Risch NJ (2005). Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**: 289–301.

Tang K, Thornton KR, Stoneking M (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* **5**: 1587–1602.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, *et al.* (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* **5**: 247–52.

Tast A, Hälli O, Ahlström S, Andersson H, Love RJ, Peltoniemi OA (2001). Seasonal alterations in circadian melatonin rhythms of the European wild boar and domestic gilt. *J. Pineal Res.* **30**: 43–9.

Taylor LK, Wang HC, Erikson RL (1996). Newly identified stress-responsive protein kinases, Krs-1 and Krs-2. *Proc. Natl. Acad. Sci. U. S. A.* **93**: 10099–104.

Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, *et al.* (2013). Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science (80-.).* **342**: 871–4.

Uimari P, Tapio M (2011). Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. *J. Anim. Sci.* **89**: 609–614.

Vicente ÁM, Alés RF (2006). Long Term Persistence of Dehesas. Evidences from History. *Agrofor. Syst.* **67**: 19–28.

Vignal A, Milan D, SanCristobal M, Eggen A (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* **34**: 275–305.

Vigne J-D, Zazzo A, Saliège J-F, Poplin F, Guilaine J, Simmons A (2009). Pre-Neolithic wild boar management and introduction to Cyprus more than 11,400 years ago. *Proc. Natl. Acad. Sci. U. S. A.* **106**: 16135–16138.

Vilas A, Pérez-Figueroa A, Caballero A (2012). A simulation study on the performance of differentiation-based methods to detect selected loci using linked neutral markers. *J. Evol. Biol.* **25**: 1364–76.

Vilstrup JT, Seguin-Orlando A, Stiller M, Ginolhac A, Raghavan M, Nielsen SCA, *et al.* (2013). Mitochondrial Phylogenomics of Modern and Ancient Equids. *PLoS One* **8**.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006). A map of recent positive selection in the human genome. *PLoS Biol.* **4**: 0446–0458.

Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, *et al.* (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**: 898–902.

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001). The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. *Am. J. Hum. Genet.* **69**: 1332–1347.

Wang Y, Nielsen R (2012). Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias. *Mol. Ecol.* **21**: 974–86.

Wealleans AL (2013). Such as pigs eat: the rise and fall of the pannage pig in the UK. *J. Sci. Food Agric.* **93**: 2076–83.

Weir BS, Cockerham CC (1984). Estimating F-Statistics for the analysis of population structure. *Evolution (N. Y).* **38**: 1358–1370.

White S (2011). From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. *Environ. Hist. Durh. N. C.* **16**: 94–120.

Wilkinson S, Lu ZH, Megens HJ, Archibald AL, Haley C, Jackson IJ, *et al.* (2013). Signatures of Diversifying Selection in European Pig Breeds. *PLoS Genet.* **9**.

Willing EM, Dreyer C, van Oosterhout C (2012). Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. *PLoS One* **7**.

Wright S (1949). The genetical structure of populations. *Ann. Eugen.* **15**: 323–354.

Wu G-S, Yao Y-G, Qu K-X, Ding Z-L, Li H, Palanichamy MG, *et al.* (2007). Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome Biol.* **8**: R245.

Xing C, Schumacher FR, Xing G, Lu Q, Wang T, Elston RC (2005). Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. *BMC Genet.* **6 Suppl 1**: S29.

Yang S, Li X, Li K, Fan B, Tang Z (2014). A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC Genet.* **15**: 7.

Yang S-L, Wang Z-G, Liu B, Zhang G-X, Zhao S-H, Yu M, *et al.* (2003). Genetic variation and relationships of eighteen Chinese indigenous pig breeds. *Genet. Sel. Evol.* **35**: 657–671.

Yang S, Zhang H, Mao H, Yan D, Lu S, Lian L, *et al.* (2011). The local origin of the Tibetan pig and additional insights into the origin of Asian pigs. *PLoS One* **6**.

Yu G, Xiang H, Wang J, Zhao X (2013). The phylogenetic status of typical Chinese native pigs: analyzed by Asian and European pig mitochondrial genome sequences. *J. Anim. Sci. Biotechnol.* **4**: 9.

-

Annexes

-

SUPPLEMENTARY FIGURES

Figure S1: Scheme of reads mapping to either the pig or human genomes, and to both genomes.

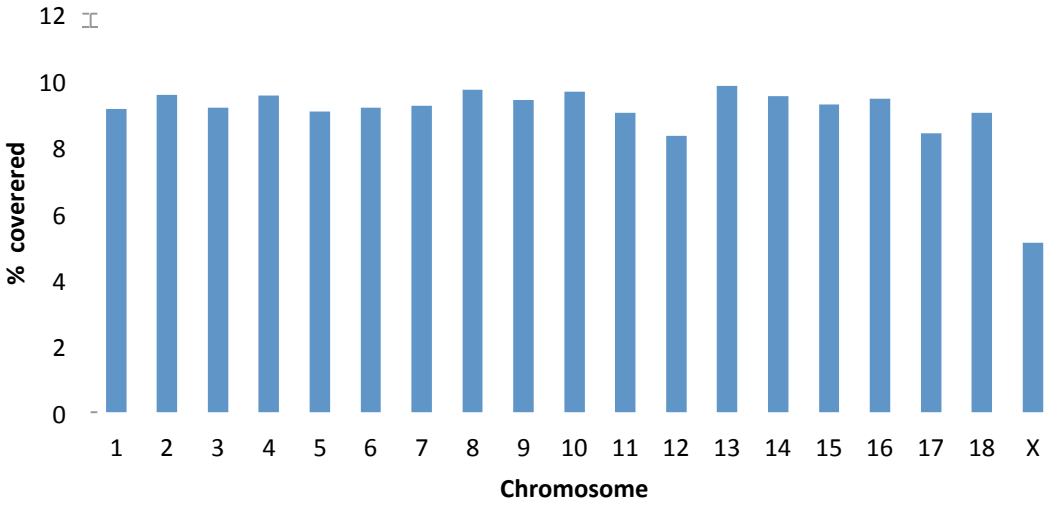
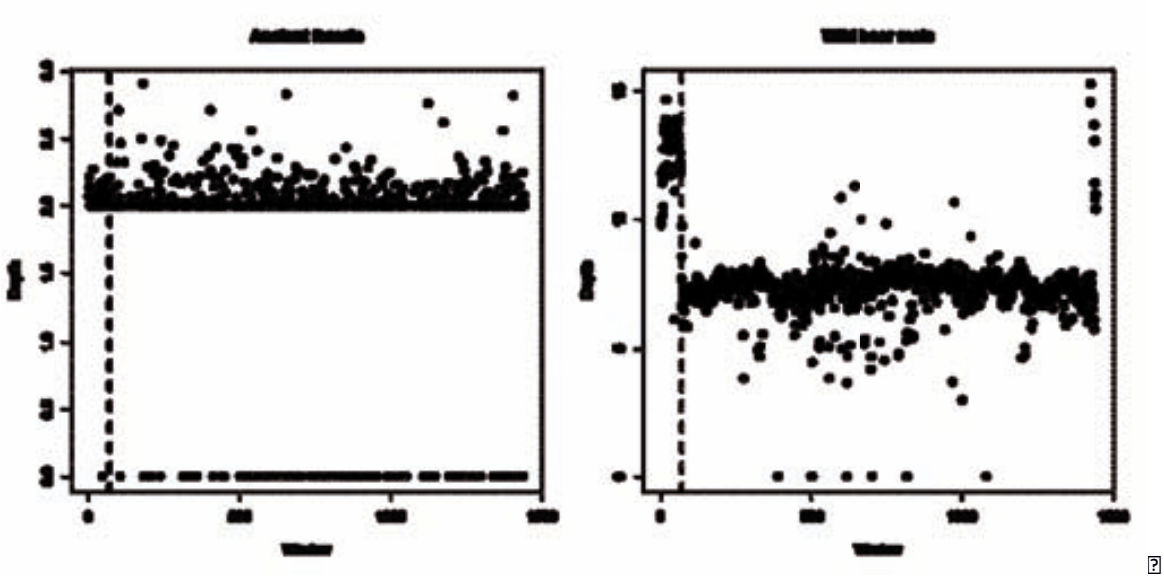
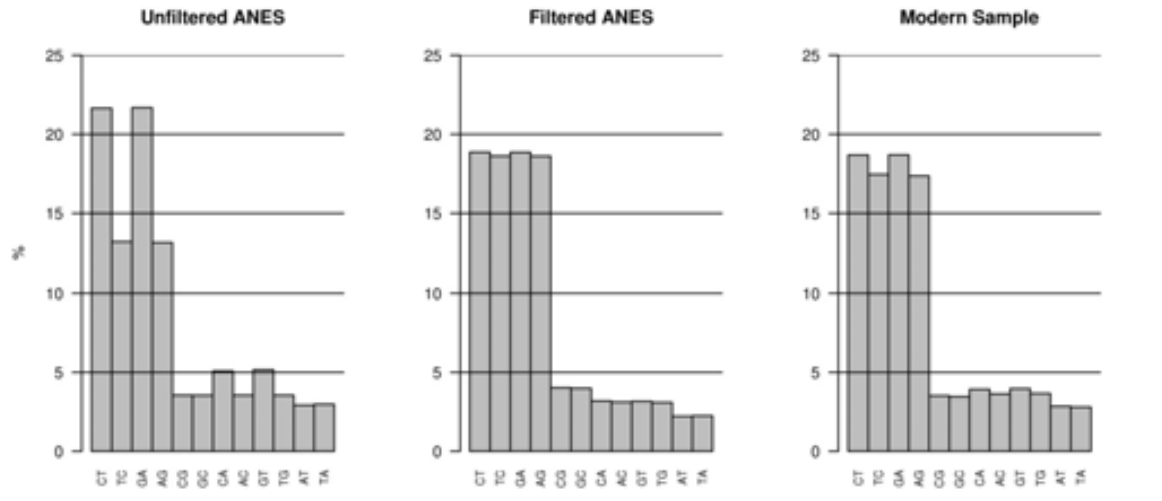


Figure S2: Number of bases with enough quality covered per chromosome divided by total number of bases per chromosome.




Unfiltered ANES
Filtered ANES
Modern Sample
%
25
20
15
10
5
0

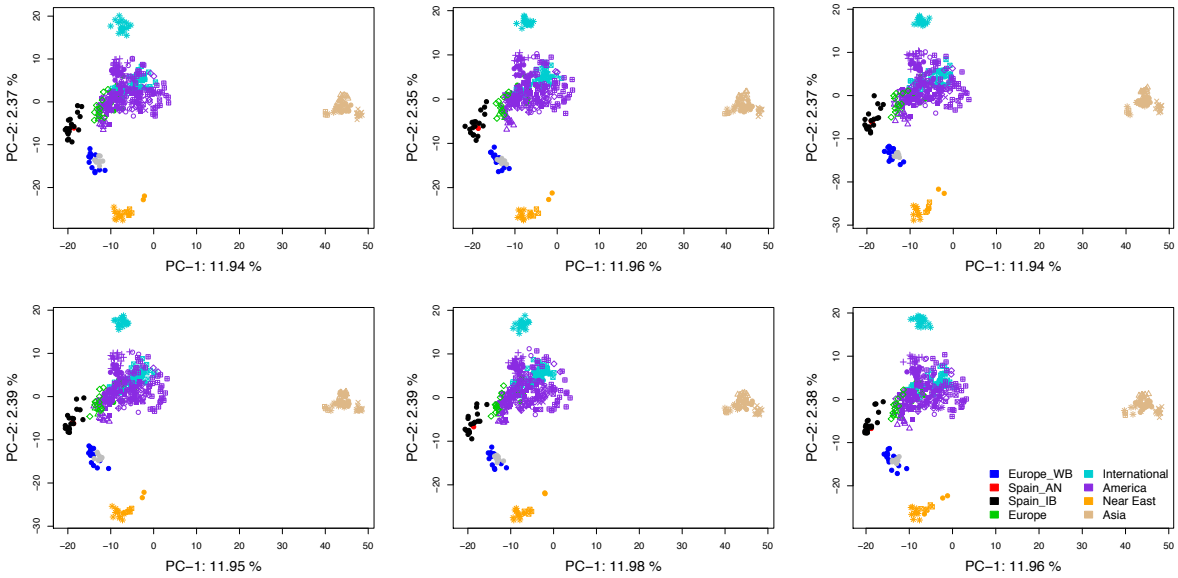


Figure S5: Six PCA graphs of the SNP chip data when only one of the two alleles is randomly sampled. Group colors as in Figure 3 in main manuscript. The ancient sample is in red, Iberian pigs in black.

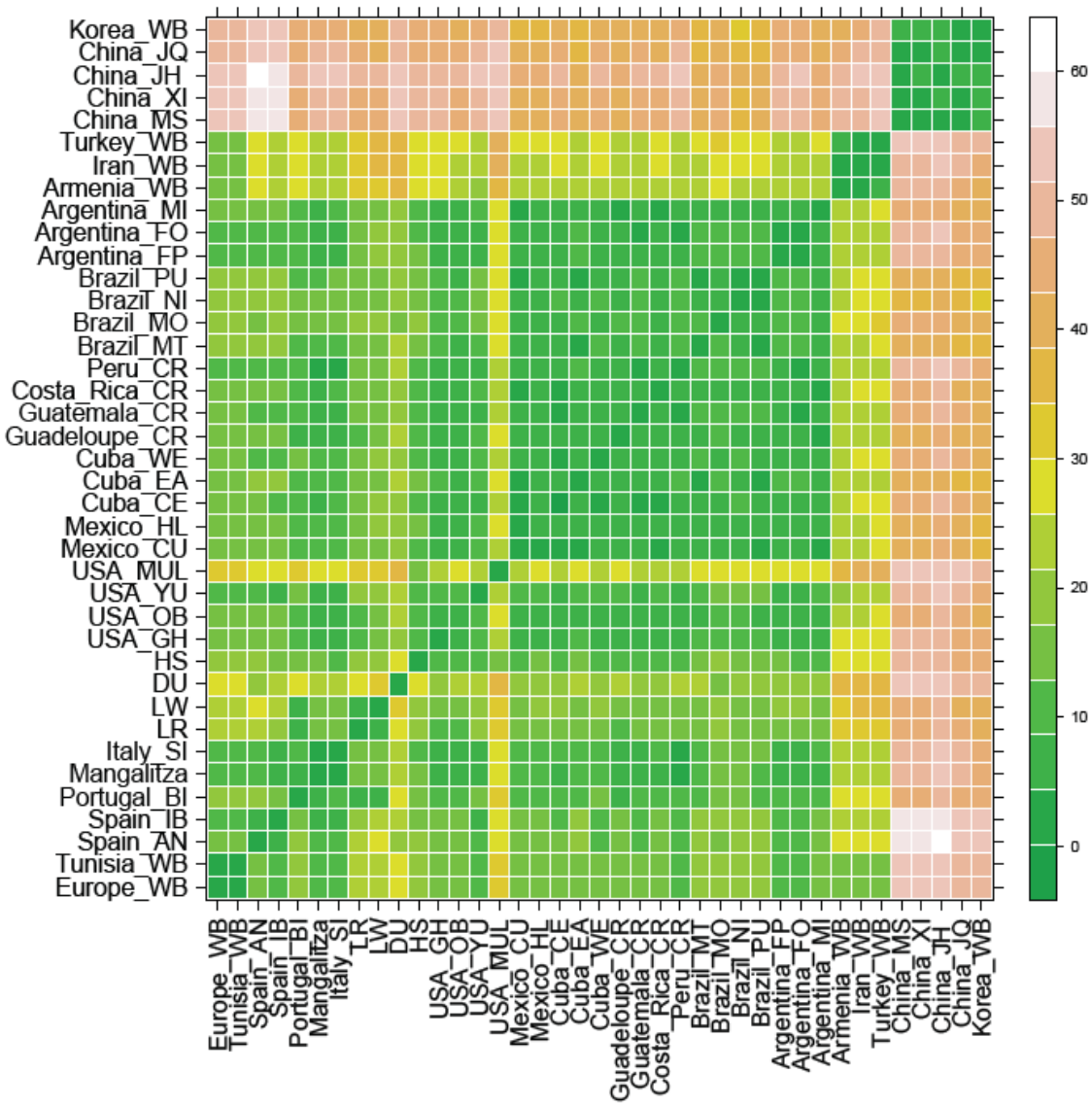
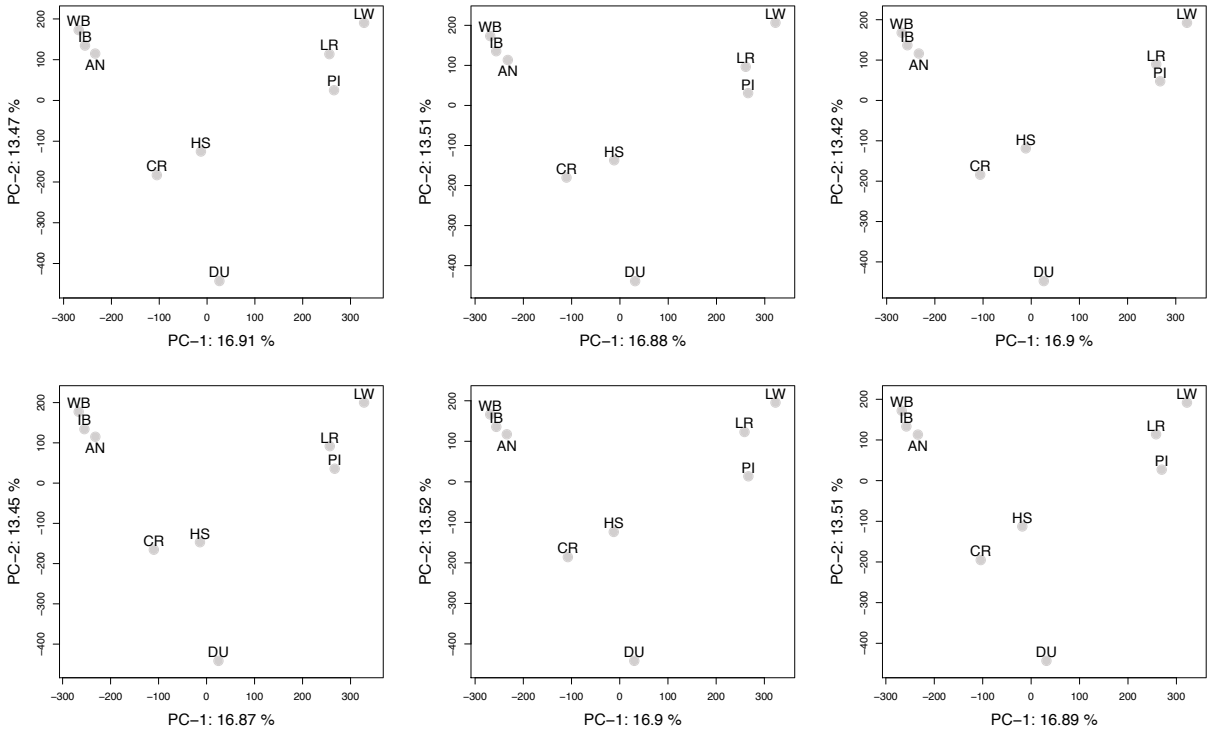
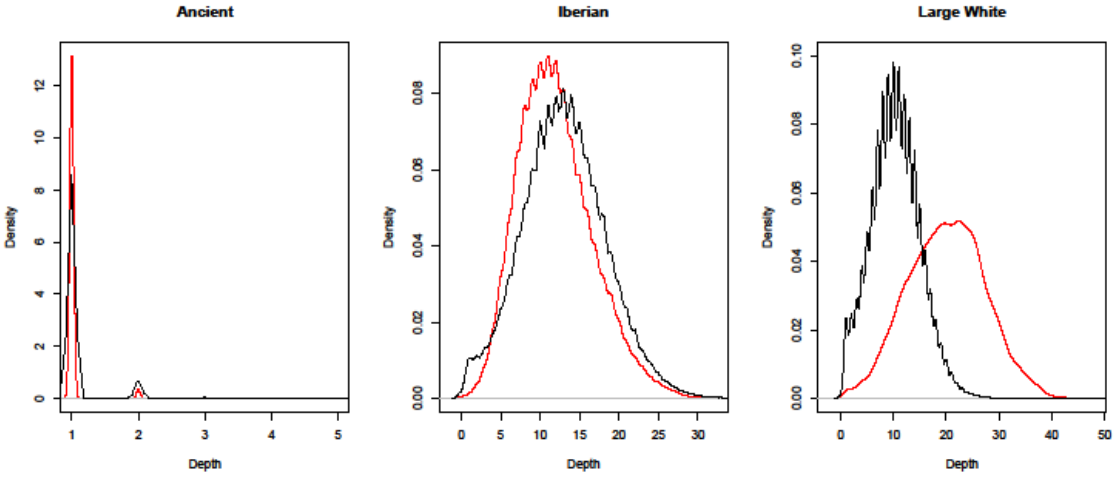


Figure S6: Euclidean distance between populations using the first four principal components. The darker the green, the closer the populations. Note that Chinese populations make a cluster of their own, and also Near East wild boars, clearly seen as well in Figure 3. US Mulefoot (USA_MUL) also looks a distinct population. Population codes are as in Figure 3.


WB
IB
AN
LW
LR
PI
HS
CR
DU
PC−2: 13.47 %
PC−1: 16.91 %
PC−2: 13.51 %
PC−1: 16.88 %
PC−2: 13.42 %
PC−1: 16.9 %
PC−2: 13.45 %
PC−1: 16.87 %
PC−2: 13.52 %
PC−1: 16.9 %
PC−2: 13.51 %
PC−1: 16.89 %


Ancient
Iberian
Large White
Density
Depth

CgA

SUPPLEMENTARY TABLES

Table S1. Modern sequenced samples used in comparison to the ancient pig.

Name	Breed	Country	Sex	Accession	Depth
CRGU1508	CR	GU	F	Unpub	12.0
DU23M02	DU	na	M	ERX149134	11.6
HA20U01	HS	na	M	ERX149137	11.5
IBGM0327	IB	ES	M	Unpub	13.0
LR24F01	LR	na	F	ERX149144	13.7
LW22F02	LW	na	F	ERX149149	10.0
PI21F06	PI	na	F	ERX149168	10.5
WBES0717	WB	ES	M	Unpub	13.0

Table S2. Modern 60k genotyped samples used in comparison to the ancient pig.

Group	Population/Breed	Country*	N
Asia	Meishan (MS)	CN	17
	Xi'ang (XI)	CN	13
	Jinhua (JH)	CN	17
	Jiangquhai (JQ)	CN	11
	Korean WB	KR	3
Near East wild boar	Armenian WB	AM	3
	Iranian WB	IR	5
	Turkish WB	TR	11
Western Wild Boar	European WB	PO, ES, HU, RU	15
	Tunisian WB	TN	8
Local European breeds	Iberian (IB)	ES	18
	Bisaro (BI)	PT	14
	Black Sicilian (SI)	IT	4
	Mangalitza (MG)	HU	11
International breeds	Landrace (LR)	DK, NL, USA	22
	Large White (LW)	DK, NL, USA	22
	Duroc (DU)	DK, NL, USA	22
	Hampshire (HS)	UK, USA	14
American Creole pigs	Guinea Hog (GH)	USA	15
	Ossabaw pig (OB)	USA	8
	Yucatan (YU)	USA	10
	Mulefoot (MU)	USA	18
	Cuino (CU)	MX	7
	Hairless (*pelón*, HL)	MX	11
	Central Cuba	CU	1
	East Cuba	CU	5
	West Cuba	CU	12
	Creole (CR)	GP	4
	Creole	GT	14
	Creole	CR	12
	Creole	PE	16
	Monteiro (MT)	BR	10
	Moura (MO)	BR	9
	Nilo (MI)	BR	2
	Piau (PU)	BR	10
	Feral pig (FP)	AR	6
	Formosa (FO)	AR	10
	Misiones (MI)	AR	9
Total	**-**	**-**	**419**

* ISO country code

Table S3: Number of IBS blocks shared with the ancient pig and genetic distances between samples.

Sample	IBS blocks	AN	IB	WB	DU	CR	LW	LR	PI
Iberian	1351	0.213							
Wild boar	865	0.235	0.220						
Duroc	479	0.271	0.274	0.294					
Creole	718	0.268	0.264	0.288	0.307				
Large White	655	0.289	0.299	0.315	0.329	0.333			
Landrace	683	0.279	0.283	0.293	0.318	0.323	0.316		
Pietrain	721	0.275	0.281	0.299	0.311	0.319	0.311	0.307	
Hampshire	786	0.266	0.266	0.287	0.306	0.312	0.325	0.314	0.311

An IBS block was defined as a 100 kb window containing at least 20 SNPs with identical genotypes in both samples and without any heterozygous SNP. Genetic distances were computed as average pairwise allele differences across SNPs when the ancient allele frequency is weighted (methods).

Table S4: Ancestral allele frequencies (f) for each of highly differentiated SNPs between wild boar and domestic in the ancient (AN), Iberian (IB), Creole (CR), Duroc (DU), Large White (LW) and European wild boar (WB).

SSC	Coordinate	f_D [1]	f_W [2]	f_{AN}	f_{IB}	f_{CR}	f_{DU}	f_{LW}	f_{WB}
1	128805851	0.19	0.89	0.00	0.00	0.00	0.00	0.25	0.75
1	201979961	0.19	0.90	0.00	0.25	0.50	1.00	0.13	0.88
1	264077520	0.19	1.00	0.00	0.75	0.00	0.00	0.25	1.00
1	275396221	0.82	0.14	0.00	1.00	0.50	1.00	0.50	0.25
2	44523890	0.89	0.17	1.00	1.00	1.00	0.66	1.00	0.75
2	73922624	0.02	1.00	1.00	0.50	0.00	0.00	0.00	1.00
2	78111017	0.02	0.88	1.00	0.00	0.50	0.00	0.00	0.50
2	78316956	0	0.86	0.00	0.00	0.00	0.00	0.00	0.75
4	84484881	0.12	1.00	0.00	0.50	0.50	0.00	0.00	1.00
5	65462758	0.83	0.11	1.00	0.88	0.50	0.63	1.00	0.38
5	66250455	0.11	0.92	0.00	0.25	0.00	0.25	0.13	0.88
5	66250462	0.88	0.19	1.00	0.75	1.00	1.00	1.00	0.13
5	98841690	0.12	0.89	0.00	0.00	1.00	0.25	0.50	0.88
7	80971392	0.16	0.86	0.00	0.00	1.00	0.38	0.13	0.88
7	100683919	0.12	0.89	1.00	0.50	0.50	0.25	0.38	0.88
7	122546941	0.14	1.00	0.00	0.88	0.50	0.13	0.63	0.88
11	12255462	0.09	0.95	0.00	0.00	0.00	0.00	0.38	0.88
11	12255600	0.17	1.00	1.00	0.75	0.00	0.25	0.38	1.00
12	20015054	0.08	0.86	0.00	0.00	0.00	0.13	0.00	0.50
13	42296157	0.09	0.88	0.00	0.75	0.00	0.00	0.13	0.50
14	124376399	0.17	1.00	0.00	0.67	0.00	0.00	0.00	0.25
15	80425831	0.98	0.17	1.00	0.75	1.00	1.00	1.00	0.13
15	95115766	0.05	0.93	0.00	0.38	0.50	0.38	0.50	0.63
15	95115793	0.92	0.19	1.00	0.38	0.50	0.38	0.63	0.63
			P_D [3]	0.72	0.65	0.68	0.75	0.74	0.32
			P_W	0.27	0.37	0.31	0.23	0.26	0.71
			SD	0.07	0.05	0.05	0.04	0.03	0.03

1 Reported frequency in domestic by Rubin et al; 2 Reported frequency in wild boar by Rubin et al.; 3 Probability of domestic (P_D) or wild boar (P_W); SD, standard deviation obtained by bootstrap.

Table S5. D statistics and associated z-scores in different quartets.

W	X	Y	Z	D	z-score
OUT	WBES	: AN	HS	-0.33	-23.3
OUT	WBES	: IB	HS	-0.34	-23.4
OUT	WBFR	: AN	HS	-0.33	-23.5
OUT	WBFR	: IB	HS	-0.33	-22.6

The D-statistic counts the difference between ABBA and BABBA patterns. 'The pattern ABBA refers to biallelic sites where X has the outgroup allele and Y and Z share the derived copy. The pattern BABA corresponds to sites where X and Z share the derived allele and Y has the outgroup allele' (Durand et al., 2011, Mol. Biol. Evol. 28:2239). Negative value indicates gene flow between X and Y or W and D. OUT is Sumatra's wild boar (accession ERX149139), used as outgroup; WBES, Spanish wild boar (Table S1); WBFR is French wild boar (accession ERX149180); AN, ancient; IB, Iberian; HS, Hampshire.

Table S6 List of top 100 most differentiated genes based on the gene or window differentiation analyses among wild boar vs. the ancient and Iberian breeds (WB vs. ANIB).

Gene	F_{ST} value	Gene	F_{ST} value	Gene	F_{ST} value
Xylb	0.851	*Gm22753*	0.57	*Rsf1*	0.471
Fshb	0.825	*Ldlrap1*	0.564	*Ccdc149*	0.469
Blcap	0.806	*Gpr126*	0.562	*Gm25987*	0.469
Extl1	0.752	*Kit*	0.561	*Sod3*	0.469
Olfr974	0.752	*Apbb2*	0.549	*Nbeal1*	0.467
Zkscan8	0.742	*Cpa6*	0.546	*Gm25657*	0.462
Timd4	0.735	*Tpd52*	0.546	*Gramd1c*	0.462
Grm7	0.723	*D630039A03Rik*	0.541	*Abcc1*	0.462
Rab17	0.719	*Mmp15*	0.54	*Pom121*	0.458
Tyrp1	0.706	*Gm25879*	0.535	*Agl*	0.458
Slc22a13	0.706	*Rc3h1*	0.535	*Frrs1*	0.458
Dcaf17	0.705	*Serpinc1*	0.535	*Grsf1*	0.456
Mettl8	0.705	*Eif3e*	0.531	*Rufy3*	0.456
Galnt11	0.69	*Kcnj4*	0.531	*Ankfy1*	0.454
Kmt2c	0.69	*Gm13306*	0.525	*Cyb5d2*	0.454
Ccdc172	0.686	*Arl1*	0.523	*Arhgap12*	0.452
Zfp706	0.684	*Utp20*	0.523	*Bend4*	0.452
Plac9a	0.682	*Lrrc42*	0.522	*Slc30a9*	0.452
Plac9b	0.682	*Hap1*	0.518	*Cry2*	0.452
Tmem254a	0.682	*Serinc2*	0.512	*Slc35c1*	0.452
Leprotl1	0.671	*Rhbdl2*	0.511		
Cntnap2	0.644	*Tmem144*	0.511		
Slc16a14	0.631	*Suclg1*	0.506		
Mll3	0.625	*Dyx1c1*	0.505		
Ap3m2	0.621	*Gm20510*	0.505		
Tlr1	0.617	*Pygo1*	0.505		
Laptm5	0.615	*Nsun7*	0.502		
Dusp5	0.614	*Mipep*	0.499		
Zbtb10	0.611	*B4galt5*	0.498		
Gja6	0.587	*Mettl8*	0.497		
Mycbp	0.587	*Gata3*	0.496		
Rragc	0.587	*Sec24b*	0.496		
Faf1	0.578	*Pak1*	0.495		
Gm23966	0.578	*Gm22541*	0.488		
Ptx3	0.577	*Gm24144*	0.486		
Veph1	0.577	*Otog*	0.485		
Adat2	0.577	*Ush1c*	0.485		
Fam211b	0.573	*Zfp64*	0.483		
Cdh7	0.572	*Rmnd1*	0.477		
Cga	0.57	*Zbtb2*	0.477		

-

Curriculum vitae

-

SUMMARY. My research interests center on analysis of populations and detection of patterns of selection and evolution in species. Additionally I have experience in quantitative genetics together with statistics, programming and application of bioinformatic tools.

1. EDUCATION

2011 **MS.** Universidad Autònoma de Barcelona-España. **Research Caracterización genética de poblaciones de credos *Sus scrofa* (ARTIODACTYLA: SUIDAE) mediante análisis de SNPs en el cromosoma X Advisors:** Miguel Pérez Enciso - Sebástian Ramos-Onsins.

2010. **MS.** Universidad de Antioquia-Colombia. **Research:** Genetic population structure of commercial lines of Guinea pigs (*Cavia porcellus*) inferred with microsatellites molecular markers and mitochondrial DNA. **Advisor:** Mario Fernando Cerón Muñoz.

2007 **BS.** Universidad de Nariño-Colombia. **Research:** Molecular characterization of three guinea pig populations using RAPDs. **Advisor:** Carlos Solarte Portilla

2. PROFFESIONAL EXPERIENCE

2012 2013 **Teaching.** Computational genetic improvement laboratory. Universidad Autónoma de Barcelona. 8 hours/año.

2011 up to date **Researcher**. Centre for Research in Agricultural Genomics CRAG. Consortium IRTA-CSIC-UAB. **Work**: Development of statistical and computational tools in animal genomics. http://www.cragenomica.es/

2006 up to date **Researcher.** Animal Breeding and Genetic program PROMEGALAC-Universidad de Nariño-Colombia. **Work:** Genetic analysis of dairy cattle populations**.** http://promegalac.udenar.edu.co/

2009. **Teaching.** Animal breeding and Genetics. **In** Corporación Universitaria Lasallista-Colombia. www.lasallista.edu.co/

3. FELLOWSHIPS AND AWARDS

2011 **TRAVEL GRANT.** COST Action EU. Course assistance: Statistical learning methods for DNA-based prediction of complex traits http://www.statseq.eu

2010 **AWARDS**. Master degree with honors. Universidad de Antioquia – Colombia

2009 **FELLOWSHIP**. Instituto Colombiano Para El Desarrollo De La Ciencia Y La Tecnología "Francisco José De Caldas" – Colciencias-Colombia. For study abroad at doctoral level. http://www.colciencias.gov.co/

2007 **AWARDS.** Undergraduate degree with honors. Universidad de Nariño-Colombia

2006 **GRANTS.** VIII Research funding for undergraduate "Alberto Quijano", Universidad De Nariño – Colombia

4. PUBLICATIONS

2014 Elizabete Cristina da Silva, Nadia de Jager, **William Burgos-Paz**, Antonio Reverter, Miguel Perez-Enciso, Eugeni Roura. Characterization of the Porcine Nutrient and Taste Receptor Gene Repertoire in domestic and wild populations across the globe**.** For submission: BMC Genomics.

2104 O. Ramírez, **William Burgos-Paz**, E. Casas, M. Ballester, E. Bianco, I. Olalde, V. Novella, M. Gut, C. Lalueza-Fox, M. Saña, M. Pérez-Enciso. Genome data from a 16th century pig illuminate modern breed relationships and suggest specific selection targets. Heredity. Submitted.

2014 Ramos-Onsins S.E., **Burgos-Paz William**, Manunza A. and Amills M. Invited review: Mining the Pig Genome to Investigate the Domestication Process**.** Heredity. Submitted.

2013 Miguel Pérez-Enciso, **William Burgos-Paz**, Sebastián E. Ramos-Onsins. On 'Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars'**.** Nature genetics. Submitted.

2012 **William Burgos-Paz,** CA Souza, HJ Megens, Y Ramayo-Caldas, M Melo, C Lemus-Flores, E Caal, HW Soto, R Martínez, LA Álvarez, L Aguirre, V Iñiguez, MA Revidatti, OR Martínez-López, S Llambi, A Esteve-Codina, MC Rodríguez, RPMA Crooijmans, SR Paiva, LB Schook, MAM Groenen, M Pérez-Enciso. Porcine colonization of the Americas: a 60k SNP story. Heredity, 110(4):321-30

2012 **Willlliam Burgos Paz**, C.A. Souza, A. Castelló, A. Mercadé, N. Okumura, I.N. Sheremet'eva, L.S. Huang, I.C. Cho, S. R.Paiva, S. Ramos-Onsins, M. Pérez-Enciso. Worldwide genetic relationships of pigs as inferred from X chromosome SNPs. Animal Genetics Apr; 44(2):130-8

2011 **William Burgos Paz**, Mario Fernando Cerón Muñoz, Carlos Solarte Portilla. Genetic diversity and population structure of the Guinea pig (*Cavia porcellus*, Rodentia, caviidae) in Colombia. Genetics and Molecular Biology 34: 327-335

2010 **William Burgos Paz**, Mario Fernando Cerón Muñoz, Manuel Moreno. Comparación de métodos para la extracción de ADN en cuyes (*Cavia porcellus* Rodentia, caviidae). Livestock Research For Rural Development *ISSN:* 0121-3784. *v.*22 *fasc.*4 p.81

2010 Carlos Eugenio Solarte Portilla, Carol Yovanna Rosero Galindo, **William Burgos Paz**, Gema Lucia Zambrano Burbano, Jhoana Meliza Eraso Cabrera, Fabio Mejia Lopez. El cuy Genético.
Livestock Research For Rural Development. *v.*22 *fasc.5* p.1

2010 **William Burgos Paz**, Mario Fernando Cerón Muóoz, Carlos Solarte Portilla. Efecto del tamaño de camada y número de parto en el crecimiento de cuyes (*Cavia porcellus* Rodentia: caviidae).
Revista Lasallista Investigación. vol.7 no.2 p.47-55

2009 Samir Julian Calvo, Edwin Martinez, Juan Fernando Tirado, Juan David Corrales Alvarez, Alba Montoya Atehortua, **William Burgos Paz**, Mario Fernando Cerón Muñoz, Manuel Moreno. Caracterización genética de las razas criollas BON y Romosinuano. Livestock Research For Rural Development *v.*21 *fasc.*4 p.54

2007 **William Burgos Paz**, Carlos Eugenio Solarte Portilla, Carol Yovanna Rosero Galindo, Heiber Cardenas Henao. Polimorfismos en la longitud de fragmentos amplificados (AFLP`s) a partir de muestras de sangre almacenadas en tarjetas FTA para la especie *Cavia Porcellus* lin. (Rodentia: caviidae). Revista Colombiana De Ciencias Pecuarias. *v.*20 *fasc.*1 p.67 – 72

2007 **William Burgos Paz**, Carlos Eugenio Solarte Portilla, Carol Yovanna Rosero Galindo, Heiber Cardenas Henao. Caracterización molecular de 3 líneas de *Cavia porcellus* mediante la aplicación de AFLP´s. Revista Colombiana De Ciencias Pecuarias. *v.*20 *fasc.*1 p.49 – 58

5. ABSTRACTS AND POSTERS

2014 **William Burgos-Paz,** Sebastián Ramos Onsins, Miguel Pérez Enciso, Lica Ferretti**.** Correcting For Unequal Sampling in Principal Component Analysis of Genetic Data. XVII Reunión Nacional de Mejora Genética Animal, Barcelona, España.

2014 **William Burgos-Paz,** Sebastián Ramos Onsins, Miguel Pérez Enciso, Luca Ferretti. Correcting For Unequal Sampling in Principal Component Analysis of Genetic Data. Accepted World Congress on Genetics Applied to Livestock production. Vancouver, Canada.

2014 **William Burgos-Paz**, O. Ramírez, E. Casas, M. Ballester, E. Bianco, I. Olalde, V. Novella, M. Gut, C. Lalueza-Fox, M. Saña, M. Pérez-Enciso. Genome data from a 16th century pig illuminates modern breed relationships. Accepted World Congress on Genetics Applied to Livestock production. Vancouver, Canada.

2012 **William Burgos-Paz**, Miguel Pérez-Enciso, AmMap consortium. The porcine colonization of the Americas: A 60k SNP story. Society for Molecular Biology and Evolution SMBE 2012, Dubín- Irlanda, Junio de 2012.

2012 **William Burgos-Paz,** Miguel Pérez-Enciso. Adaptive signals in American pig populations. Pig Diversity meeting. Menorca- España, Mayo de 2012.

2012 **William Burgos-Paz,** Sebastián Ramos-Onsins, Miguel Pérez-Enciso Relación genética de poblaciones de cerdos inferida con SNPs del cromosoma X. XVI Reunión Nacional de Mejora Genética Animal, Menorca- España, Mayo de 2012.

2009 **William Burgos-Paz**, Mario Fernando Ceron Munoz, Manuel Moreno, Carlos Eugenio Solarte Portilla. Diversidad genética de líneas comerciales de cuyes *Cavia porcellus* Lin. (Rodentia: caviidae) con marcadores moleculares microsatellites. X Encuentro Nacional y III Internacional de Investigadores de las Ciencias Pecuarias. Revista Colombiana de Ciencias Pecuarias v.22 , fasc.3 *p.*456

2009 **William Burgos-Paz**, Mario Fernando Ceron Munoz, Manuel Moreno, Carlos Eugenio Solarte Portilla. Estructura filogeográfica de poblaciones de cuyes (*Cavia porcellus*) en Colombia y su relación con poblaciones de Suramérica. X Encuentro Nacional y III Internacional de Investigadores de las Ciencias Pecuarias. Revista Colombiana de Ciencias Pecuarias v.22, fasc.3 *p.*461

2007 **William Burgos-Paz.** Evaluación de la diversidad genética de tres poblaciones de cuyes *Cavia porcellus* Lin. (Rodentia: caviidae) mediante el marcador molecular RAPD. IX Encuentro Nacional y II Internacional de Investigadores de las ciencias pecuarias. Revista Colombiana de Ciencias Pecuarias. v.20, fasc.4

-

Colophon

-

The studies included here has been supported by Consolider CSD2007-00036 “Centre for Research in Agrigenomics”, AGL2010-14822 grant (Spain), CGL2009-09346 (MICINN, Spain)

William Orlando Burgos Paz is a recipient of a fellowship for abroad studies at doctoral level from the Departamento Administrativo de Ciencia, Tecnología e Innovación COLCIENCIAS-COLOMBIA, Francisco José de Caldas fellowship 497/2009.