Universitat Autònoma de Barcelona

Facultat de Veterinària

Departament de Ciència Animal i dels Aliments

Using genomewide polymorphisms to explore demography and feralization in the pig species

Ph. D Thesis

Erica Bianco

Supervisor: **Dr. Miguel Pérez-Enciso**

Bellaterra, 2015

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

CERTIFICA:

Que **Erica Bianco** ha realitzat sota la seva direcció el treball de recerca:

"Using genomewide polymorphism to explore demography and feralization in the pig species"

Per a obtenir el grau de Doctor en Producció Animal 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 i al Departament de Genòmica de Plantes i Animals del Centre de Recerca en Agrigenòmica (CRAG).

Bellaterra, a 28 de Setembre de 2015

Dr. Miguel Pérez Enciso

“It is not the strongest of the species that survives,

nor the most intelligent,

but the one most responsive to change”

Charles Darwin

To life, that is amazing.

Abstract

The advent of next generation sequencing technologies has revolutionized the study of livestock genomics, such as in pigs. For instance, using genomic data it is possible to better understand wild boars demography and its impact on domestication. Moreover, the analysis of feral pigs improves the knowledge on the dynamics of feralization, and serves as yardstick against which to compare modern breeds. In this thesis we provide insights about the demography and the feralization process in the pig species using genomewide polymorphisms.

In the first part of this work, we analyzed 128 complete pig and 5 outgroup genomes, in order to obtain the first genomewide and worldwide catalog of SNPs of the pig. We were able to assess the ancestral allele of ~39M out of the ~48M variants found. The number of unique derived variants in European breeds (>6M) is smaller than in Asian breeds (>13M), in agreement with the Asiatic origin of *Sus scrofa*. Moreover, we found a marked correlation in allele frequencies between domestics and wild boar within Asia and within Europe. This correlation was absent across continents, due to the large evolutive distance between pigs in both continents (~1 MYA).

In the second part of this work, we tried to disentangle the demographic history of wild boars using the polymorphisms found in the previous study. We analyzed the joint site frequency spectrum of ~2M SNPs of 9 European and 8 Asian wild boars using coalescence and the analytical approach of ∂a∂i. Using coalescence we evaluated whether a split between the two populations was enough to explain the observed spectrum, but only when migration was included in the model, we found a joint spectrum coherent with the observed data. Using ∂a∂i, we analyzed 6 models that differed in the number of bottlenecks and migration events. Only the simplest model seemed to converge, whereas this was not clear for more complex scenarios. Despite this convergence issue, both methods pointed to migration events after the split as a demographic factor shaping wild boars variability. Further analyses are needed to improve wild boars' demographic history inference.

Finally, in the third part of this thesis, we focused on the analysis of the dynamics of feralization. We analyzed the genome of the feral population of *Isla del Coco* (Costa Rica), which have been isolated since its foundation in 1793, and is therefore an excellent model to study feralization dynamics. We confirm that English domestic pigs were already hybrids between Asian and European breeds in late 17th century. Interestingly, despite the bottleneck suffered, Cocos pigs average variability is comparable to those of current commercial pig breeds such Large White or Duroc. Yet, we also found a 10-Mb region with a marked decrease in variability across all sampled tested, which was previously identified as highly differentiated between wild boars and domestic breeds.

Domestication and feralization are symmetric events of pig history. The analysis of wild boars demography will serve as null model to study the dynamics before domestication. On the other side, the analysis of the feral pigs of *Isla del Coco* improved the knowledge on the last 200 years of breeding managements on domestic breeds. Moreover it will help to understand the dynamics after domestication when a hybrid animal becomes feral. All these studies are now possible only thanks to the evolution of sequencing techniques, which resulted in an increasing number of public worldwide whole genome sequence data.

Table of contents

List of Figures

List of Tables

CHAPTER I

General introduction

1. General introduction

1.1. Pig natural history

The pig (*Sus scrofa*, Linnaeus 1758) is one of the most common large mammals in the world. This even-toed ungulate belongs to the Suidae family. Wild boars are present across all Eurasia and its domestic counterparts can be found all over the world, both in captivity and in feral status. *S. scrofa* diverged from the other *Sus* species ~5.3 - 3.5 million years ago (MYA), at the beginning of Pliocene in Island South East Asia (ISEA) ecoregion (Larson *et al.* 2007b; Groenen *et al.* 2012; Frantz *et al.* 2013), already recognized by Alfred R. Wallace as a key region for speciation (Wallace 1855). During the Calabrian period (mid-Pleistocene, ~1.8 - 0.7 MYA), wild pigs spread from ISEA and colonized all Eurasia, reaching Europe around 1.2 - 0.8 MYA, according both to fossil records (van der Made 1999) and to molecular data based estimates (Giuffra *et al.* 2000; Larson *et al.* 2005; Groenen *et al.* 2012; Frantz *et al.* 2013). Approximately 0.6 MYA, a second split within Eastern wild boars occurred, and Northern and Southern Chinese *S. scrofa* split in two populations (Frantz *et al.* 2013).

The wide range of distribution for wild pigs has resulted in a large variety of shapes and colors throughout Eurasia. The size of a wild boar ranges from 50kg and 90 cm long of the Taiwanese wild boar up to 200 kg and 180 cm long of the European wild boar. For a long time, there had been a debate between authors about Eurasian wild boars being divided into different subspecies or even different species because of the difference in color, size and skull shape. Charles Darwin himself classified Asian and European wild boars in two different species: *Sus indicus* and *S. scrofa* (Darwin 1868). In general, there is a cline of phenotype variations throughout the continent, and the differences are not so noticeable, so today wild boars are all considered as a single species (*S. scrofa;* Porter 1993).

Domestication and pig breeds

Archeological evidences suggest that pig domestication from its wild ancestor started ~ 9,000 years before present (Larson *et al.* 2007b). In all species, the targets of domestication comprise milder behavior and changes in appearance traits in order to obtain better phenotypes to satisfy human needs. In dogs, for example, today we have a wide variety of dog's shapes and coat colors, resulting in several hundred different breeds actually recognized by the Federation Cynologique Internationale (www.fci.be). In the case of the pig, fat and growth have played a key role in breed selection, especially in the last 300 years (White 2011).

Unlike ruminants, which were as nomads as their owners, domestic pigs were more difficult to roam. Pigs were then animals for settled farmers rather than nomads (Porter 1993). For pigs, domestication *in situ* was easier than to move already domesticated individuals, and multiple centers of pig domestication existed across Eurasia, but still some migration occurred. Mitochondrial DNA studies (Larson *et al.* 2005) found at least 5 main domestication centers across Eurasia, from where pig migrated after domestication (Larson *et al.* 2005, 2007a; Ottoni *et al.* 2013) as reviewed in Ramos-Onsins *et al.* (2014, Figure 1.1).

Around 300 years ago, mostly in England, pig management and breeding techniques changed to more intensive one (White 2011). The main focus of pig breeders was initially fatness, because human consumption demand was for energy-rich food. In the last 60 years, however, society demanded a more reduced caloric intake so that breeds were selected towards lean meat.

In the late 18th and early 19th century, domestic breeds improvement occurred as a consequence of globalization. European pig breeders introduced Asian germplasm into domestic breeds to improve growth and litter size (Porter 1993; Giuffra *et al.* 2000). Modern commercial pig breeds are the result of these introgression events that resulted in ~20% of Asian germplasm into European individuals (Bosse *et al.* 2014b, 2015).

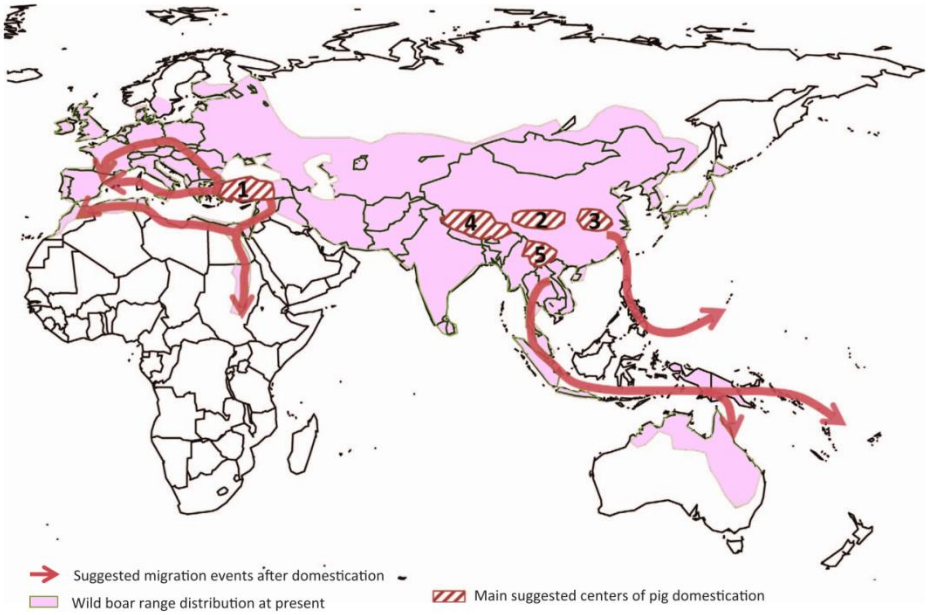
1
2
3
4
5
→ Suggested migration events after domestication
Wild boar range distribution at present
▨ Main suggested centers of pig domestication

Figure 1.1*:* Sus scrofa *range. Suggested domestication centers are indicated by dashed circles. Red arrows indicate the suggested migration events. Adapted from Ramos-Onsins* et al. *2014.*

Creole pigs

Pigs were absent in the American continent before European colonization. The first recorded pig import in America is dated at the end of 1493, in the second Columbus trip to the Caribbean (Rodero *et al.* 1992; Crosby 2003; Zadik 2005). It was customary to bring some pigs during long ship trips, as source of meat. Moreover, pigs have a high capacity to adapt to all types of environments, except the driest, so that they quickly and easily settle in the Caribbean first, and in the rest of the American continent later (Elliott 2006). Today, village pigs of Iberian ancestry (creole pigs) are common in many American countries, although international pig breeds have been replacing and intermixing with local populations. Burgos-Paz *et al.* (2013) showed that America's creole pigs are the result of multiple colonization and introgression events rather than a single pig introgression followed by a stepping stone colonization of America. The Iberian ancestors signature is still present, even if subsequent admixture with commercial or Asian breeds occurred throughout time.

1.2. Pig genomics

The first pig complete genome was sequenced in 1999 by Lin *et al.* It was a mitochondrion complete sequence, extracted from heart tissue of a Landrace pig. Since that first sequence, mitochondrial genome (or just a part of it such as the D-loop control region or the *CytB* gene) was used to disentangle questions about pig evolution, biogeography, breed ancestry and domestications (Giuffra *et al.* 2000; Alves *et al.* 2003; Larson *et al.* 2005; Fang & Andersson 2006; Scandura *et al.* 2008; Ramírez *et al.* 2009; Fernández *et al.* 2011; Ottoni *et al.* 2013; Vilaça *et al.* 2014; Bianco *et al.* 2015b; Noce *et al.* 2015).

In 2009, Ramos *et al.* developed a high density 60k SNP chip, the PorcineSNP60 Beadchip, uniformly distributed along pig genome. This SNP chip has been utilized in numerous studies (225 citations). Genome wide association studies (for example Amaral *et al.* 2011; Ojeda *et al.* 2011; Ramayo-Caldas *et al.* 2012), copy number variants analyses (such as in Ramayo-Caldas *et al.* 2010; Wang *et al.* 2014), recombination maps (Tortereau *et al.* 2012; Muñoz *et al.* 2012; Fernández *et al.* 2014), population differentiation and population genetic studies (Herrero-Medrano *et al.* 2013; Burgos-Paz *et al.* 2013) are some of the topics that have been addressed with this resource.

As sequencing technologies quickly improved, the eighteen autosomal chromosomes, the X chromosome and a small part of chromosome Y were sequenced and assembled in 2012 (Sscrofa10.2; Groenen *et al.* 2012). The availability of a reference genome revolutionized the questions that can be tackled. It made possible to understand the genomic level of variability and uncover the region of homozygosity in domestic breeds (Bosse *et al.* 2012; Groenen *et al.* 2012; Veroneze *et al.* 2013; Bianco *et al.* 2015b), to detect possible region under selection between domestic pigs and wild boars (Amaral *et al.* 2011; Rubin *et al.* 2012; Moon *et al.* 2015), to analyze ancient genome of a Iberian pig and compared it with modern samples (Ramírez *et al.* 2014); and to study variability and discover SNPs (Ai *et al.* 2013, 2015; Esteve-Codina *et* *al.* 2013; Bianco *et al.* 2015a). Sscrofa 10.2 is not the only *de novo* assembled pig sequence, Fang *et al.* (2012) and Li *et al.* (2013) assembled *de novo* a Wuzhishan and a Tibetan pig respectively, which were used as reference sequence for mapping in studies that focused on Chinese breeds, because of better mappability (Li *et al.* 2013; Ai *et al.* 2015).

During the last years, porcine genomics then shifted from genotype data to whole genome resequencing data, and today there are almost 300 pig genomes publicly available (Fang *et al.* 2012; Groenen *et al.* 2012; Li *et al.* 2013, 2014; Esteve-Codina *et al.* 2013; Ramírez *et al.* 2014; Molnár *et al.* 2014; Ai *et al.* 2015; Bianco *et al.* 2015b; Kim *et al.* 2015; Moon *et al.* 2015).

Next Generation Sequencing methods

In the last fifteen years, there was a remarkable evolution in genomics technology, with a consequent decrease in their costs. Since 2008, when the next generation (or new or second generation) of DNA sequencing technologies (NGS) became available, the cost per genome dropped from ~70,000 USD using Sanger sequencing to ~4,500 USD for a human size genome (~3Gb) at coverage of 30X (Wetterstrand 2014). In the near future, a new generation of sequencers is expected to appear, and would be even cheaper to obtain a whole genome sequence. This new technology will make possible, for example, personalized medicine in humans. Population genetics also had, and will have, a substantial benefit due to the reduction of sequencing costs. If obtaining the sequence of an individual could be done at a reasonable price, it would be possible to obtain the whole genome sequence of many individuals from different populations. This allows studying variability and divergence between populations in a detail not yet possible, from a genomic point of view, both in natural populations and in domestic breeds.

The process from sequencing to variant calling is visualized in the flow chart in Figure 1.2 and described in the following paragraphs.

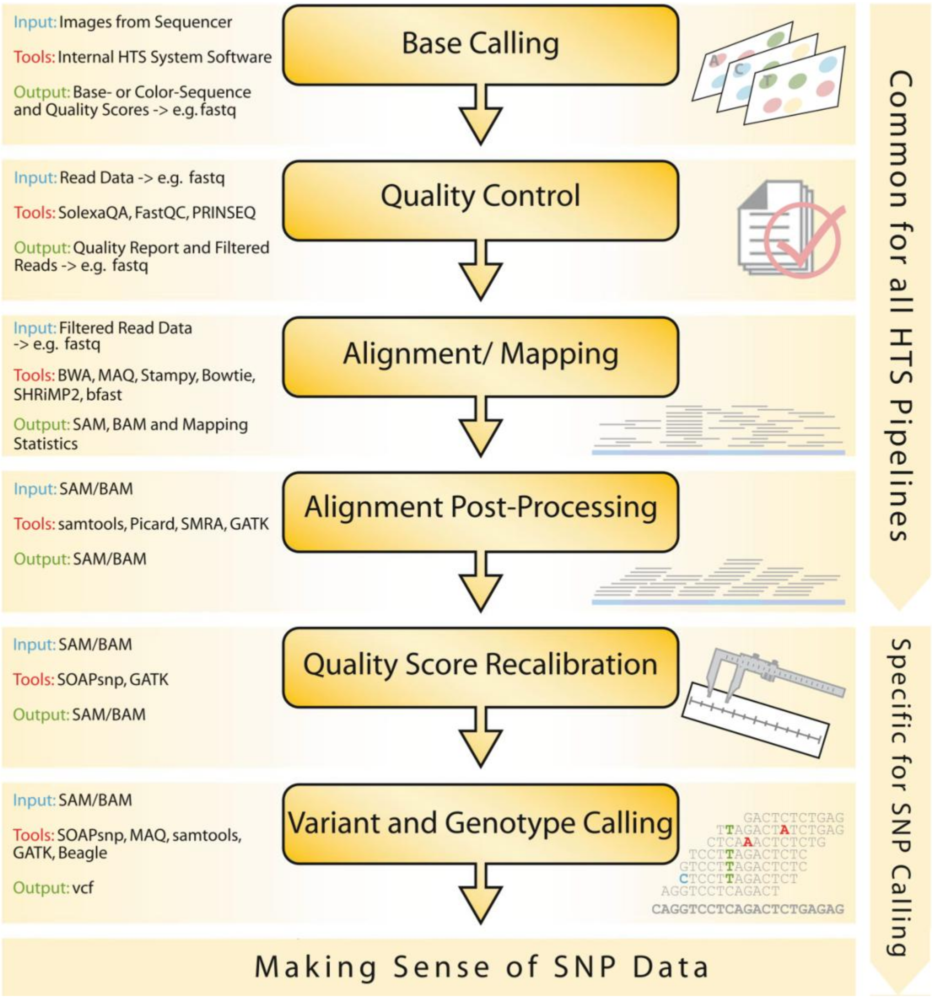
Input: Images from Sequencer
Tools: Internal HTS System Software
Output: Base- or Color-Sequence and Quality Scores -> e.g. fastq
Base Calling
A
C
T
Input: Read Data -> e.g. fastq
Tools: SolexaQA, FastQC, PRINSEQ
Output: Quality Report and Filtered Reads -> e.g. fastq
Quality Control
Input: Filtered Read Data -> e.g. fastq
Tools: BWA, MAQ, Stampy, Bowtie, SHRiMP2, bfast
Output: SAM, BAM and Mapping Statistics
Alignment/ Mapping
Input: SAM/BAM
Tools: samtools, Picard, SMRA, GATK
Output: SAM/BAM
Alignment Post-Processing
Common for all HTS Pipelines
Input: SAM/BAM
Tools: SOAPsnp, GATK
Output: SAM/BAM
Quality Score Recalibration
Input: SAM/BAM
Tools: SOAPsnp, MAQ, samtools, GATK, Beagle
Output: vcf
Variant and Genotype Calling
GACTCTCTGAG
TTAGACTATCTGAG
CTCAAACTCTCTG
TCCTTAGACTCTC
GTCCTTAGACTCTC
CTCCTTAGACTCT
AGGTCCTCAGACT
CAGGTCCTCAGACTCTGAGAG
Specific for SNP Calling
Making Sense of SNP Data

***Figure 1.2**: From reads to SNPs. Data processing workflow from base calling in the sequencer to SNPs (when a reference genome is available). Adapted from Altmann* et al. *2012.*

DNA sequencing

After DNA extraction and library construction, the first step is the base calling. In this step, image-capturing devices convert the recorded signal of DNA synthesis into nucleotide bases. To do so, different algorithms have been developed. Statistical models are used to provide error estimation of the base calling (quality). Base calling algorithms and error estimation are specific to the sequencing platform, such as Illumina, Roche or SOLiD and more recently, PacBio. The standard output format of base calling is fastq (Cock *et al.* 2009), a flat file that includes the nucleotide sequence (read) and quality scores. Using Illumina pair end sequencing platform, read length is currently at least 100 bp.

Quality control

Manufacturer normally provides a read quality measures but, to improve mappability to the reference, reads can be analyzed and filtered using tools such as FastQC (*). This tool allows checking for expected GC content, overrepresented reads and distribution of nucleotides per read positions, among others. In this way it is possible to recognize and remove bad quality reads and also to perform read trimming. A typical behavior of High Throughput Sequencing (HTS) platforms is that error probability increases with read length. Read trimming is then applied to remove those bases at the end of the read that are more prone to errors, to improve its mappability. After quality control, a filtered fastq file is used to perform mapping.

Mapping

Mapping is the process by which the sequenced reads are aligned against a reference genome, or, in other words, mapping determines the location of the reads (chromosome and position) onto the reference genome. This has to be done for each of the millions of reads generated, taking into account for sequencing errors, indels and SNPs. The most popular algorithm is today based in Burrows-Wheeler transform methods (such as BWA, Li & Durbin 2009). The final format for the mapped reads is SAM (Sequence Alignment/Map) format (Li *et al.* 2009). This can then be compressed into BAM or CRAM files, which occupy less disk space and are easily accessible.

Variant calling

Once the individual reads have been mapped onto the reference genome, it is possible to perform genotype calling. This process allows us obtaining the list of sites that are variants between the individual and the reference genome, being the individual heterozygote or homozygote for the alternative allele(s). This is

(*) http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

the variant list used to perform population studies, research for causative variants, or genome wide association studies among others.

1.3. Demographic inference

The different demographic events undergone by populations have shaped their variability, both within and between populations (Wright 1931). One of the major goals of evolutionary and population genetics is to determine the demographic history of a species based on these variability patterns. Demography inference based on DNA data complements archeological evidences, such as pre-historical events, and a proper demographic model serves as a null model in genomewide test for selection (Nielsen *et al.* 2007). The advent of NGS makes it feasible to have access to a large amount of data and variants of one or more individuals from the same population, so that demographic inference can be made on a larger and unbiased set of variants.

Different methods have been proposed to infer the demographic history using genetic information: haplotypes and linkage disequilibrium (LD) based methods, whole genome sequence data based methods, and methods based on a reduced set of genomic information, such as genomic summary statistics (Approximate Bayesian Computation methods) or site frequency spectrum based methods.

Linkage disequilibrium and haplotypes based methods

Linkage disequilibrium (LD) and haplotypic patterns are the results of demographic history and recombination events (Pritchard & Przeworski 2001). Bottlenecks increase LD and reduce the number of haplotypes of the population. In 2009, Lohmueller *et al.* proposed a method to infer demographic history using the joint distribution of the number of haplotype and the frequency of the most common haplotype in windows across genome. With this method, an excess of windows at high frequency of the most common haplotype would be a signature for a bottleneck (Lohmueller *et al.* 2009), so that recent demographic events can be detected (Sabeti *et al.* 2007).

Demography also shapes the length of the runs of homozygosity (ROH) and the size of long identity by descent (IBD) segments. In fact, if a bottleneck occurred, the length of a IBD segment and the log scaled frequency of IBD segments show a linear correlation (Gusev *et al.* 2012), so that it is possible to trace the number of generations that have passed since the last bottleneck (Gusev *et al.* 2012). Another possible method to evaluate the effective population size in the past is the evaluation of long ROH (Kirin *et al.* 2010; Howrigan *et al.* 2011). This is an individual-based method that compares the two homologous of a chromosome and evaluates the number and length of ROH. For example, if there is a large number of short ROH and a few long, this can be interpreted as an expansion event: recent large effective population size (low inbreeding), but a reduced effective population size in the past (Kirin *et al.* 2010).

LD and ROH demographic inference are sensible to recent population size fluctuations and may detect recent bottleneck with only considering one individual, nevertheless, migrations and split between two populations cannot be estimated.

Whole genome sequence coalescence based methods

Today whole genome sequence (WGS) data are becoming to be widely accessible. Each individual genome contains its whole demographic history. Li & Durbin (2011) developed a method to reconstruct demographic history from a single diploid genome: the pairwise sequential Markovian coalescent (PSMC) model. The principle is to recover the time since the most recent common ancestor by integrating over all possible coalescent trees of the two genome sequences of a diploid individual to obtain the estimation of the demographic parameters. Past variations in population size are then provided when fitting the hidden Markov model (Li & Durbin 2011).

More recently, Schiffels & Durbin (2014) proposed an improved approach over PSMC to include an arbitrary number of individuals in the demographic inference (MSMC: multiple sequential Markovian coalescent method). This method, as PSMC, finds the most recent coalescence between genome's haplotypes. The advantage of MSMC is that it uses each pair of haplotypes from all the genome sequences included in the analysis, increasing the details and allowing to infer effective population size variations also in the last 20,000 years, which was not possible with PSMC, when only two haplotypes were used (Schiffels & Durbin 2014).

In both cases (PSMC and MSMC), one of the major advantages of these methods is that it is not needed to establish a prior model (Li & Durbin 2011; Schiffels & Durbin 2014).

Genome data summary statistics based methods

One of the most powerful methods to infer demographic parameters is based on Approximate Bayesian Computation (ABC; Tavaré *et al.* 1997; Beaumont *et al.* 2002). Using ABC, it is not necessary to calculate the exact value of the likelihood function of a given model, which is often impossible to calculate due to its complexity. ABC allows to approximate the posterior distribution of model's parameters and to estimate the model posterior probability (Tavaré *et al.* 1997; Beaumont *et al.* 2002). The idea behind ABC's demographic inference is to calculate a distance (ε) between observed and simulated data's summary statistics and to accept those models that provide the best fit to observed data. The most used ABC method is rejection algorithm (reviewed in Beaumont 2010; Bertorelle *et al.* 2010; Sunnåker *et al.* 2013; Robinson *et al.* 2014). One of the main caveats of ABC lies in choosing the summary statistics that will be used: sufficient summary statistics to capture all relevant features of genomic data have to be chosen. On the other side, the Euclidean distance threshold must also be found to have a good balance between accuracy and efficiency of the algorithm (normally a positive small value of ε; Bertorelle *et al.* 2010).

ABC process is summarized in Figure 1.3. Summary statistics are calculated from the obtained data. For all the chosen models, the prior distribution of the parameters is given and multiple simulations are performed for each values of the prior distribution. The same summary statistics based on observed data are calculated on the results of simulations and the distance ε is calculated between simulated and observed summary statistics: those models whose ε is above the selected threshold are rejected. With the retained models, a posterior distribution of the model's parameters is approximated.

ABC is extremely flexible in the choice of models, but the main problem resides in the selection of summary statistics.

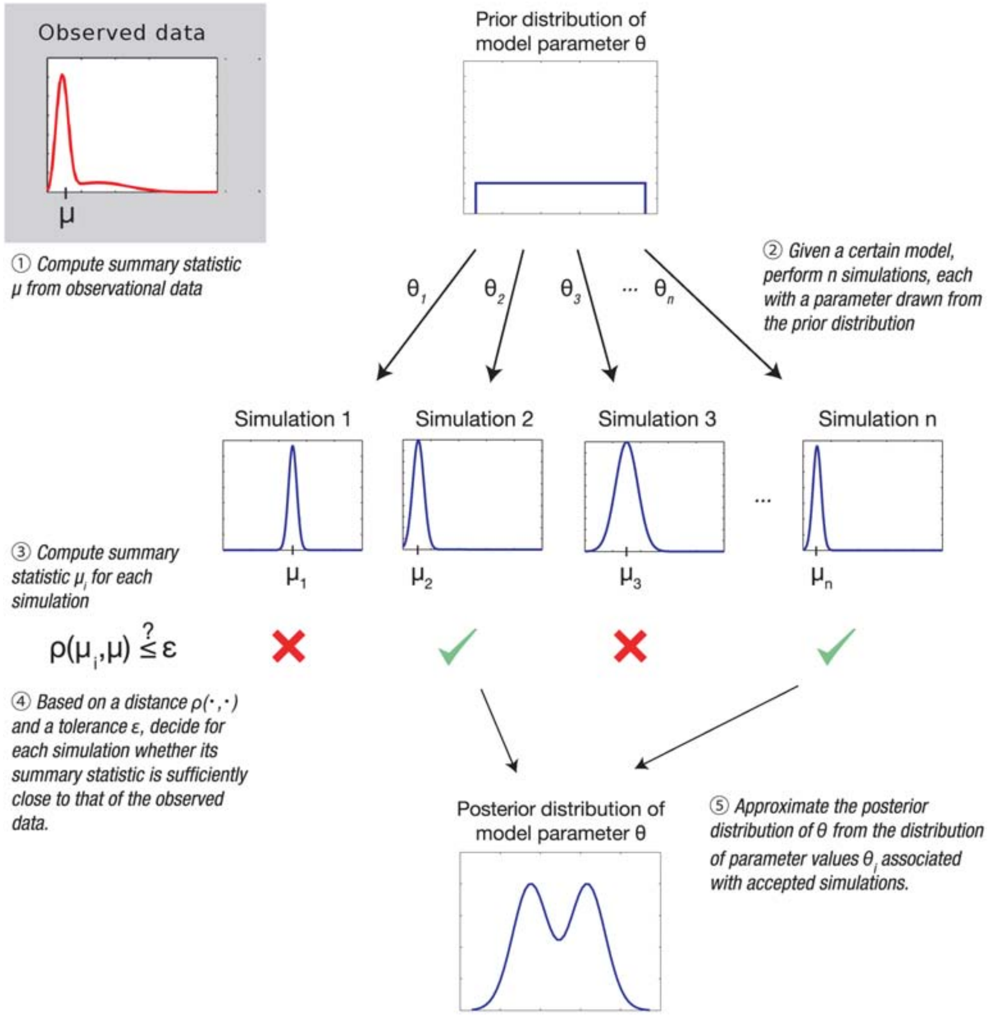
Observed data
μ
① Compute summary statistic μ from observational data
Prior distribution of model parameter θ
② Given a certain model, perform n simulations, each with a parameter drawn from the prior distribution
θ1
θ2
θ3
... θn
Simulation 1
Simulation 2
Simulation 3
Simulation n
...
μ1
μ2
μ3
μn
③ Compute summary statistic μi for each simulation
ρ(μi,μ) ≤? ε
✗
✓
✗
✓
④ Based on a distance ρ(·,·) and a tolerance ε, decide for each simulation whether its summary statistic is sufficiently close to that of the observed data.
Posterior distribution of model parameter θ
⑤ Approximate the posterior distribution of θ from the distribution of parameter values θi associated with accepted simulations.

***Figure 1.3**: ABC method for demographic inference. Adapted from Sunnåker* et al. *2013.*

Site Frequency Spectrum based methods

The last group of demographic inference method described here are methods based on the site frequency spectrum (SFS). SFS is the distribution of the count of alternative alleles in the population across biallelic variants (SNPs). The SFS depends on the past demographic events, so that changes in population size, bottlenecks and migrations leave a signature along the genome that modifies the SFS. Different population size fluctuations result in different shapes of the SFS, as show in Figure 1.4. If the population size is stationary, the number of sites is inversely proportional to its frequency in the population. A recent and strong increase of the population size will increase the number of singletons in the population, and decreasing the number of sites at medium and high frequency. On the other side, bottleneck results in a much more even distribution. The three scenarios represented in Figure 1.4 are the best case scenarios, more complex demographic history are not so clear and, worryingly, different demographic scenarios can result in the same SFS (Myers *et al.* 2008).

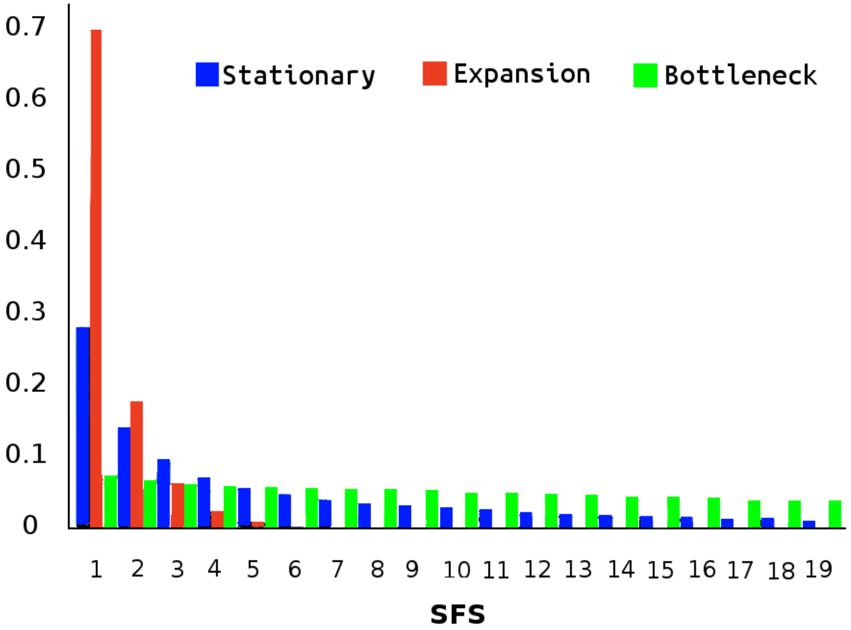

***Figure 1.4**: The effects of population expansions, contraction and no changes in population size on the site frequency spectrum.*

One advantage of inferring demography using the SFS is that increasing the number of variants analyzed or the number of individuals does not increase proportionally the computational time, but it increases the power of the analysis making the estimation more accurate (Gutenkunst *et al.* 2009; Excoffier *et al.* 2013). Mainly, SFS methods need the hypothetical demographic model as input, then the algorithm evaluates the model parameters that best fits the observed spectrum (Gutenkunst *et al.* 2009; Excoffier *et al.* 2013). Another advantage of SFS based methods is the inference of population interaction and of complex demographic scenarios. Using a joint SFS allows inferring migration events, the time since the split and possible fusion between two populations. A commonly SFS method used is *∂a∂*i (Gutenkunst *et al.* 2009) that uses a diffusion approximation method to infer the demographic history of up to 3 populations. *∂a∂*i assumes the spectrum is obtained from high quality independent sites (Gutenkunst *et al.* 2009): the variants must then be filtered to prune for linkage disequilibrium before SFS calculation and must be high coverage and high quality variants to distinguish between sequencing errors and singletons.

The input for *∂a∂*i is the observed SFS and the model for which to estimate the parameters. It then generates a SFS using the demographic model provided by the user, and compares it with the observed SFS to estimate the model's parameters in order to find the best fit estimates. Because the likelihood maximization has convergence problems, the estimation must be repeated various times to get maximum likelihood. For each model tested, the optimization step must be repeated, then, the best likelihood obtained are compared to evaluate which model fits better the data, for example using the Akaike Information Criterion, which measures the relative quality of a statistical model taking into account its likelihood and the number of parameters (Akaike 1974).

More recently Excoffier *et al.* (2013) developed another method based on SFS, fastsimcoal2, which uses conditional expectation maximization algorithm to find the best fit model. The main advantage of fastsimcoal2 is that it can estimate the demographic history of an arbitrary number of populations. Moreover, its convergence properties are better than those in *∂a∂*i.

Today, the SFS based methods are the fastest and less computational intensive methods to infer demography jointly for multiple individuals of multiple populations. Nevertheless, all methods results should be considered with caution. Using SNPs from whole genome sequencing to obtain the site frequency spectrum can be problematic, because different demographic history can result in the same SFS when variables are in LD (Myers *et al.* 2008).

CHAPTER II

Objectives

2. Objectives

The broad objective of this thesis was to study pig demography and the feralization process using polymorphisms obtained from high throughput whole genome resequencing data.

More specifically, the objectives of this work were:

- To evaluate and analyze pig genomewide variability using resequencing data from worldwide samples (Chapter 3);
- To infer the joint demographic history of Asian and European wild boars populations (Chapter 4);
- To analyze the genome and infer the origins of the feral population of *Isla del Coco* (Costa Rica), which remained isolated since 1793 (Chapter 5);
- To evaluate the effect of feralization on the hybrid (Asia and Europe) genomes of Cocos pig (Chapter 5).

CHAPTER III

A deep catalog of autosomal Single Nucleotide Variation in the pig

3. A deep catalog of autosomal Single Nucleotide Variation in the pig

Erica Bianco[1,2], Bruno Nevado[1,2*], Sebastián E. Ramos-Onsins[1], Miguel Pérez-Enciso[1,2,3 †]

[1] Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Spain. [2] Universitat Autònoma de Barcelona, Department of Animal Science, 08193 Bellaterra, Spain. [3] Institut Català de Recerca I Estudis Avançats (ICREA), Carrer de Lluís Companys 23, Barcelona, 08010, Spain. * Present address: Department of Plant Sciences - University of Oxford. South Parks Road Oxford, OX1 3RB, UK. † Corresponding author E-mail: miguel.perez@uab.es

PLoS ONE (2015) 10 (3) e0118867

Abstract

A comprehensive catalog of variability in a given species is useful for many important purposes, e.g., designing high density arrays or pinpointing potential mutations of economic or physiological interest. Here we provide a genomewide, worldwide catalog of single nucleotide variants by simultaneously analyzing the shotgun sequence of 128 pigs and five suid outgroups. Despite the high SNP missing rate of some individuals (up to 88%), we retrieved over 48 million high quality variants. Of them, we were able to assess the ancestral allele of more than 39M biallelic SNPs. We found SNPs in 21,455 out of the 25,322 annotated genes in pig assembly 10.2. The annotation showed that more than 40% of the variants were novel variants, not present in dbSNP. Surprisingly, we found a large variability in transition / transversion rate along the genome, which is very well explained (R2=0.79) primarily by genome differences in in CpG content and recombination rate. The number of SNPs per window also varied but was less dependent of known factors such as gene density, missing rate or recombination (R2=0.48). When we divided the samples in four groups, Asian wild boar (ASWB), Asian domestics (ASDM), European wild boar (EUWB) and European domestics (EUDM), we found a marked correlation in allele frequencies between domestics and wild boars within Asia and within Europe, but not across continents, due to the large evolutive distance between pigs of both continents (~1.2 MYA). In general, the porcine species showed a small percentage of SNPs exclusive of each population group. EUWB and EUDM were predicted to harbor a larger fraction of potentially deleterious mutations, according to the SIFT algorithm, than Asian samples, perhaps a result of background selection being less effective due to a lower effective population size in Europe.

Introduction

In this new era of sequencing, it is feasible to routinely obtain whole genome sequence data from an increasing number of individuals, making it possible the analysis of populations at the genomic level. The availability of this large amount of data allows us to study any species variability to an unprecedented detail. An intriguing observation from these studies is that, despite intensive selection and small effective sizes, animal domestic species harbor much more variability than anticipated [1–4]. In the particular case of the pig, this variability is especially remarkable and is likely caused by a complex demographic history and to the availability of a large amount of breeds [5–7].

The availability of a reference genome makes it feasible the large scale variant discovery with new sequencing technologies or 'next' generation sequencing (NGS). In pigs, the last assembly of porcine reference sequence (assembly 10.2) was released in November 2012 [8]. Although still incomplete, around 8% of the sequence is estimated to be missing from the assembly [8], it still constitutes the best resource to date in the genome of the pig. Currently, over a hundred pig sequences of about 20 different breeds and several countries have been published and are publicly available [8–12]. Despite these resources, so far, a comprehensive catalog of variants mined out these pig genomes is missing. Such a catalog is useful for many purposes: to design high density genotyping arrays, be it genome-wide or focused in specific genome regions or geographic origins of interest, to confirm SNPs from other experiments, to improve the reference genome, to identify variants of potentially large deleterious effect that can be followed up in functional studies, and to increase the general biological knowledge of a species. For instance, as we shall see, we discover a large mutational bias in the pig genome that is largely explained by the differential CpG content and recombination rate along the chromosomes.

Here, we report such a catalog (data have been submitted to dbSNP at the following URL: http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1062009 and they are also available at the website http://bioinformatics.cragenomica.es/numgenomics/), obtained from analyzing 128 genomes from wild boar and domestic pig samples worldwide distributed. In addition, we report annotation, allele frequencies in four major pig groups and we infer the ancestral allele for the majority of the SNPs. Knowledge of the ancestral allele is required for many statistical tests of selection so this is an additional value of the catalog here presented.

Materials and methods

Samples

We analyzed a total of 133 suid genomes (S1 Table), 128 pigs (*Sus scrofa*) and five outgroups (*S. barbatus, S. cebifrons*, *S. verrucosus, S. celebensis,* and an African warthog, *Phacochoerus africanus*). The 128 pig genomes were classified into four large groups: Asian Wild Boars (ASWB, n=41), that comprise wild boars from China, Japan and East Russia; Asian Domestics (ASDM, n=23), including 9 Chinese breeds like Meishan or Xian; European Wild Boars (EUWB, n=9) from Spain, France, Switzerland, and the Netherlands; and European Domestics (EUDM, n=55) from all major breeds Duroc, Large White, Landrace, Pietrain Hampshire and local breeds (Iberian, Tamworth). European domestics include as well American village pigs, which have a predominant European, although hybrid, origin [13].

All samples had been shotgun sequenced with Illumina's technology but at different depths, ranging from ~3× in a Tibetan wild boar [11] to 22× in a Wuzhishan pig, a miniature Chinese breed [9]. Here, we analyzed only two out of all available 25 lanes in the Wuzhishan pig so depths could be comparable across samples. The majority of genomes data were public [8,9,11,12] and 26 additional unpublished genomes were also used (E. Bianco *et al.*, submitted). Main sample details are in S1 Table. In all, over 28 x 10^9 reads, occupying around 2Tb of disk in bam format, were analyzed.

Alignment and variant calling

The detailed bioinformatics pipeline is in S1 Script. The samples from Groenen *et al.*, [8] (n=50) were downloaded as bam files mapped against assembly 10.2. For the rest of samples, raw reads were mapped against assembly 10.2 with BWA [14] allowing for 7 mismatches and using default options otherwise. Duplicate removal and sorting were done with samtools v 0.1.18-sl61, using rmdup and sort options, respectively [15]. For all bam files, both the downloaded bam files and those generated in-house, GATK v. 2.7 IndelRealigner [16] was ran to improve the alignment around indels, default options were used.

Genotypes were called for each individual separately using samtools (v 0.1.19+) mpileup option and filtered with vcfutils.pl varFilter [15] . We excluded indels in this analysis because of their low reliability for the range of depths in our samples [17]. For a SNP to be called, we set the minimum depth to 5× and the maximum depth of twice the average sample's depth plus one, minimum map quality and minimum base quality were both set to 20:

```
samtools mpileup -Q 20 -q 20 -m2 -Dugf Sus_scrofa10.2.fas
PIG_NAME.realigned.bam      |  bcftools  view  -vcg  -  |
vcfutils.pl  varFilter  -d  5  -D  (MEAN_DEPTH*2)+1  -Q  10  >
PIG_NAME.iSAM.flt.vcf
```

Individual vcf files were then merged using custom scripts. For each individual, missing variant positions were coded according to bcftools output without the "-v" flag to avoid variant calling; confident homozygous-reference calls were coded as '0/0' (homozygous for reference allele), and the position was marked as missing './.' otherwise.

VCFtools v0.1.12a [18] was used to filter the resulting multi individual vcf file, to extract outgroup genotypes, to analyze each of the four groups separately, and to filter out genotypes for which raw depth was ≥ 5 but where their high quality read depth was lower than 5 (--mindp). Allele frequency and allele count were calculated with the options --freq and --count, respectively, and transition /

transversion rate was calculated in windows of 1 Mb with the options --TsTv and –TsTv-summary. R version 3.0.2 was used to plot results [19].

Ancestral allele determination

The variant calling and the merging steps were done including the five outgroup samples. An *awk* script was used to ascertain the ancestral allele, applying the following criteria:

1. The SNP must be biallelic where at least one of the alleles is the reference allele.
2. A) The SNP must be present in at least two *Sus spp.* genomes, and be homozygous in all *Sus spp.* samples where the SNP is called;

   B) Otherwise, it must be present in *P. africanus* genome and be homozygous.
3. The ancestral allele must be either the reference or the alternative allele in *S. scrofa* (that is, a third allele must not be segregating).

In sites not complying with these conditions, the ancestral allele was considered as unknown.

Exclusive variants and diagnostic SNPs

Population allele frequencies were obtained with VCFtools [18]. We defined an exclusive segregating variant as a site in which the derived allele is segregating only in the target group and it is not present in any of the remaining groups. Only those biallelic RA sites (R refers to the reference allele and A, to the alternative), and where all four groups had at least 50% of individuals with genotypes called were used. Shared and private alleles were plotted with gplots R library with venn package [20].

Joint site frequency spectrum between groups was also calculated. For each group, we selected the modal group size (the number of samples *n* where the highest number of variants was called in exactly *n* individuals). The count of derived alleles was performed and plotted with R package lattice (levelplot option, [21]).

Genome context

We evaluated whether the number of SNPs per window and the transitions / transversion rate (Ts/Tv) correlates with genome features knowing to affect variability: GC content (%GC), CpG count, gene density and recombination rate. GC percentage and CpG count were calculated based on the *Sus scrofa* reference genome assembly 10.2 [8], and gene density was obtained as the percentage of the window sequence that is part of a gene according to *Sus scrofa* 2.75 GTF annotation. Genes overlapping two or more windows were discarded. We used the recombination rate (cM/Mb) from Tortereau *et al.*, [22] with the same genome partitioning as in that work, in windows of ~1 Mb long. In addition, we computed percentage of missing genotypes per individual per window, averaged over individuals. To quantify the effect of each variable, we fitted the following linear models using the R function lm [19]:

N_snps = β0 + β1 Ts_Tv + β2 log(rec_rate) + β3 GC_percentage + β4 CpG_count + β5 Gene_Density + β6 missing_rate + e, [*equation 1*]

and

Ts_Tv = β0 + β1 N_snps + β2 log(rec_rate) + β3 GC_percentage + β4 CpG_count + β5 Gene_Density + β6 missing_rate+ e, [*equation 2*]

We used the logarithm transformation of recombination rate because the raw values were highly skewed.

SNP annotation

All variants were annotated with Ensembl variant effect predictor (VeP) pipeline v76 [23] on Ensembl version 76 (using dbSNP *build* 140). Among the terms used in the annotation, we focused on stop gain and stop lost (sequence variants which cause a premature stop codon or the stop codon is changed resulting in an elongation of the protein), missense variants (non synonymous variants) and synonymous variants (a variant in a coding region that does not change the aminoacid). The definition of all terms used is available at http://www.ensembl.org/info/genome/variation/predicted_data.html. Variant annotation was performed both on the whole data set and by group. When there was more than one alternative allele, all possible alternatives were retained. The effect of the aminoacid changes was predicted using SIFT [24,25], a tool that tentatively predicts whether a missense variant affects protein function because of sequence homology and of the physical properties of amino acids. Default options were used.

Simulation of the bioinformatics pipeline

In this study, we used NGS data from different sources. These data are noisy and highly unbalanced, with highly variable depth across samples (S1 Table). Moreover, the pipeline applied is complex and the properties of the SNP calling procedure are not necessarily known. As a fundamental caution when analyzing such a complex data, it is advisable to evaluate, even if approximately, the performance of the pipeline applied. Here, we evaluated how reliable are the SNPs called and estimated how many SNPs were retrieved out of all those actually segregating in the samples by simulation. To do this, we employed Pipeliner [26], with small modifications. Pipeliner seamless integrates several steps and softwares:

1. Simulates, with the coalescence, genome data reflecting as much as possible the population analyzed.
2. Maps simulated SNPs into a reference DNA sequence, this is done by replacing the reference base by an alternative base in the SNP position for each haplotype and produces a fasta file for each sequence; next, each individual genome is created by randomly choosing two sequences.
3. Simulates the sequencing process producing reads that mimic Illumina's technology; we used ART (v. 1.5.1, [27]) to do so.
4. Runs BWA [14] to map the reads against the reference sequence.
5. Analyzes the output and reports several statistics of interest; among them:
    a) Recovery: percentage of original genotypes that are correctly identified.
    b) Sensitivity: percentage of callable genotypes, *i.e.* present in sites that pass the filters used, that are correctly identified.

c) False Discovery Rate (FDR): percentage of genotype calls performed that are incorrect

Pipeliner was fine tuned to duplicate as faithfully as possible the actual bioinformatics pipeline used to analyze the real data. First, to obtain the 'real' sequences in our sample, we simulated 256 sequences (128 diploid individuals) with MaCS [28], using the following structured population model:

```
NUMINDS=128
EUDM=55
ASDM=23;
EUWB=9
ASWB=41
macs NUMINDS*2 100000 -t 0.0005 -r 0.0005 -I 4 EUWB*2
EUDM*2 ASWB*2 ASDM*2 -n 1 0.2 -n 2 0.5 -n 3 2.5 -n 4
3 -em 0.049 2 4 5 -eM 0.06 0 -ej 0.07 2 1 -en 0.09 1
5 -en 0.2 1 7 -en 0.21 4 10 -ej 0.08 3 4 -ej 10 1 4
```

The command above simulates an older first split into two populations (Asia and Europe, -ej 10 1 4) and, a much more recent event, the split between domestic and wild populations in both continents (-ej 0.08 3 4 and -ej 0.07 2 1). This model is very similar to the model of [29]. Parameters were chosen such that estimates of nucleotide diversity were similar to those found in the real data. For the ART simulator [27], average depths were set for each individual as those empirically observed, ranging from 3x to 22× (S1 Table). As reference genome, we randomly chose one of the 'European' sequences, given that the assembly was derived from a Duroc specimen [8].

Finally, alignment, variant calling and merging were done following the pipeline used for real data, which resulted in a simulated multi individual vcf file. We used mstatspop v.0.998982beta [30] to evaluate the proportion of correctly called SNPs, the proportion of false SNPs and not identified SNPs. The whole process was repeated 100 times. Note that, despite we tried to faithfully represent the complexities of SNP calling for our specific set of samples, we ignored known difficulties in mapping due to structural variants or repetitive elements. The whole pipeline to do the simulation is in S1 Script.

Results

In silico evaluation of the bioinformatics pipeline

First, we evaluated our pipeline by simulation as described. We need to distinguish two issues. The first one is how many SNPs out of those segregating can be recovered. This is the main target in the real data analysis and, in this case, uniformly high depth and coverage may not be so critical because a SNP position that is not covered in one individual may have been covered in another one (provided is not a singleton). The second issue is how reliably called is each individual genotype. Accurate genotype calling is important for allele frequency estimates but not so much for SNP detection; for instance, suppose a heterozygous Reference/Alternative (RA) genotype is actually called as 'AA', the SNP will be equally identified, but frequency estimate will be strongly biased. With Pipeliner [26], we evaluated both issues as described in methods. Figure 3.1 illustrates the overall expected power and percentage of wrongly identified SNPs. As can be seen, we expect to have discovered about 95% of all SNPs that may have been segregating in the 128 samples analyzed; of those, less than 0.5% variants are expected to be false positives. By population group, the outcome varies according to depth, the Asian populations being slightly worse than European populations because of shallower depth (Figure 3.2). Even in those populations, power was 90% and FDR ~1%. The relatively high power of the pipeline, even at sallow depth, is due in part to the demographic model, which has very long branches between the Asian and European populations, followed by a bottleneck. This model, that reflects a plausible history of the pig genome, predicts an excess of non singletons compared to the neutral model; in turn, this means that a given SNP that is not called in one individual because of shallow depth can still be discovered in another sample. Singletons are unique and, therefore, this cannot happen in this case.

As for individual genotypes, we should expect according to the simulations, under the best case scenario, to recover ~70% of heterozygous genotypes, 73% of homozygotes for the alternative allele (AA) and 76% of the homozygotes for the reference allele (RR, S1a Figure). The average percentage of genotypes passing all quality filters that are correctly identified (sensitivity) is expected to be close to 1 for homozygote genotypes, and slightly lower for true RA genotypes (97.5%, S1b Figure). In all, the most likely reason for a SNP not to be correctly identified is that it was missed because of low depth or quality, rather than being incorrectly identified. FDR was very low, in the order of 1% for heterozygous genotypes (S1d,f Figures).

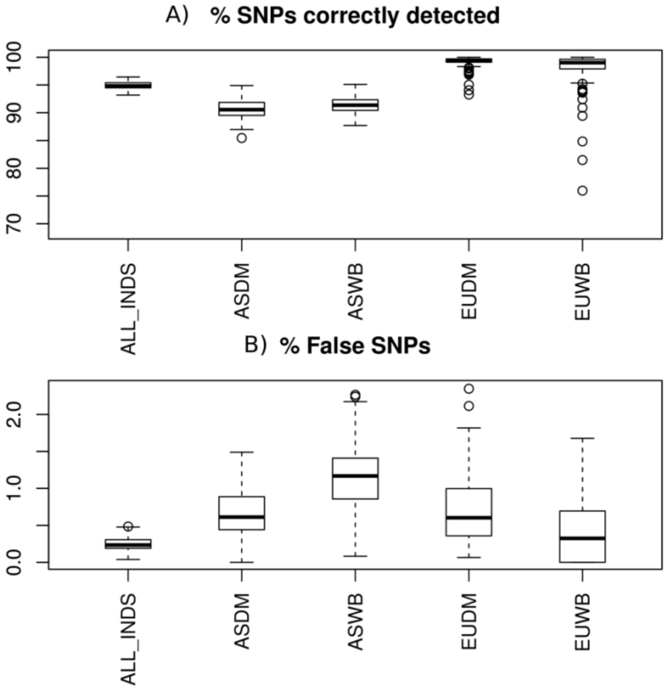

***Figure 3.1**: Using simulations, estimated percentage of segregating sites correctly detected (a) and percentage of false SNPs (b), according to the Pipeliner simulations. ALL_INDS: all samples; ASDM, Asian domestics; ASWB, Asian wild boar; EUDM, European domestics; EUWB, European wild boar.*

In all, Pipeliner simulations predict that we could expect our bioinformatics protocol to be highly reliable, allowing us to uncover about 95% of all SNPs in the samples with rather low false discovery rate. In practice, the real situation is likely to be worse than simulated because we are simulating the best case scenario, without considering the true complexities of the genome, *e.g.*, duplications, indels, repetitive sequences, unequal GC content and so on. It is nevertheless difficult to consider all these complicating factors in a simulation, and Pipeliner results should be taken as an upper limit, mainly valid for well aligned genome regions.

As for individual genotypes, we should expect according to the simulations, under the best case scenario, to recover ~70% of heterozygous genotypes, 73% of homozygotes for the alternative allele (AA) and 76% of the homozygotes for the reference allele (RR, S1a Figure). The average percentage of genotypes passing all quality filters that are correctly identified (sensitivity) is expected to be close to 1 for homozygote genotypes, and slightly lower for true RA genotypes (97.5%, S1b Figure). In all, the most likely reason for a SNP not to be correctly identified is that it was missed because of low depth or quality, rather than being incorrectly identified. FDR was very low, in the order of 1% for heterozygous genotypes (S1d,f Figures).

In all, Pipeliner simulations predict that we could expect our bioinformatics protocol to be highly reliable, allowing us to uncover about 95% of all SNPs in the samples with rather low false discovery rate. In practice, the real situation is likely to be worse than simulated because we are simulating the best case scenario, without considering the true complexities of the genome, e.g., duplications, indels, repetitive sequences, unequal GC content and so on. It is nevertheless difficult to consider all these complicating factors in a simulation, and Pipeliner results should be taken as an upper limit, mainly valid for well aligned genome regions.

Individual missing and genotype rates

In the real data, we computed the number of identified SNPs in the whole sample that were not callable in each individual, as a percentage of all SNPs identified. The average individual SNP missing rate was 35%, ranging from 88% of an Asian Domestic (a Penzhou individual, [11]) to 4% of a European Domestic individual (an Iberian domestic, Bianco *et al*, submitted). Logically, missing rate was highly correlated with average depth: the lowest depth individuals, most of them ASDM and ASWB samples, had the highest missing rate (Figure 3.2a). This high variability in missing rate is reflected in the number of times a SNP was genotyped in the dataset: 61,665 variants were genotyped in only one individual and only 23 SNPs were genotyped in all individuals (Figure 3.2b).

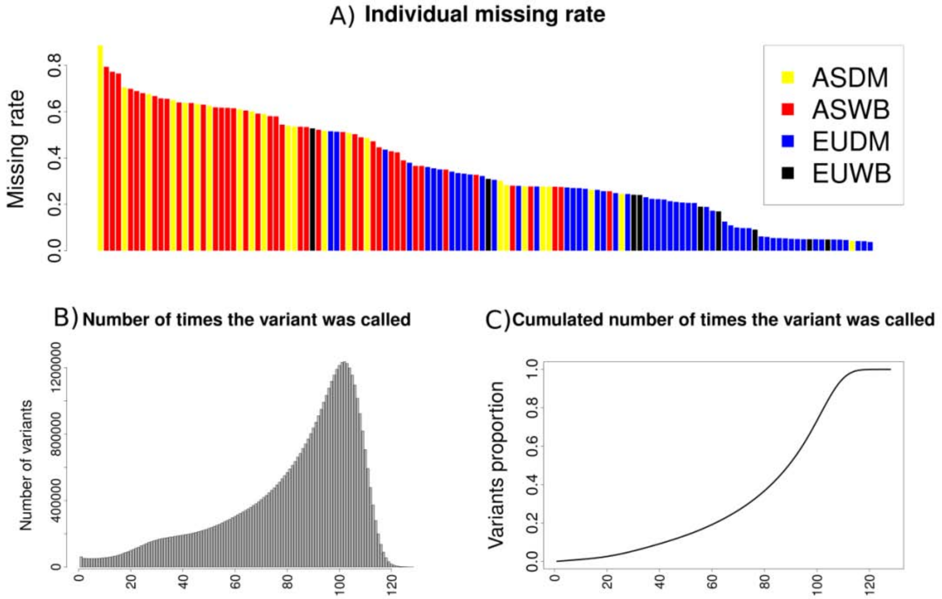

***Figure 3.2**: Individual missing rate (a); number of times a variant was called (b) and cumulative number of times a variant was called (c).*

The cumulative number of SNPs arranged by the times each SNP was called is in Figure 3.2c: 50% of SNPs were called in 90 individuals and the SNPs genotyped in more than 115 individuals were less than 1% of the total variants counts. In other words, we found basically the same number of SNPs in 115 samples than in the whole set (n=128).

General SNP statistics and genomic context

We found, among the 128 *S. scrofa* samples, a total of 48,119,476 SNPs that were called in at least one individual and passed all depth and quality filters. The majority (97.5%) were biallelic variants, whereas the rest presented 1, 3 or 4 alleles (Table 3.1). For 377,922 variants, only the alternative allele(s) were found; 12% of the 362,740 variants called as homozygote for the alternative allele were called in only one individual, but 71,081 (0.15% of total variants detected) were called in at least 64 individuals. These latter SNPs could reflect errors in the assembly. On average, we found ~19,000 ± 7,000 variants per Mb window (Figure 3.3). By chromosomes, chromosome 10 had the highest number of variants per bp, with 26.7 variants per kb. The lowest number of variants per kb was found in chromosome 1 (15.9 variants per kb) (S2 Table).

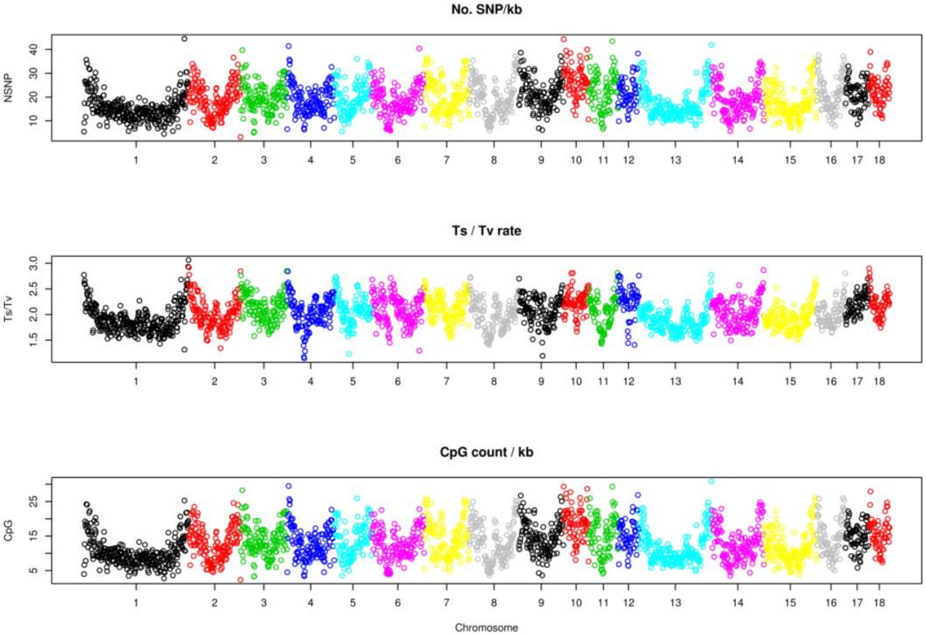

***Figure 3.3**: Number of SNPs per kb (top), average transition / transversion rate (middle) and CpG count per kb (bottom) per window. On the x axis, each dot represents a window of ~1 Mb long. Different colors correspond to different chromosomes, from SSC1 to SSC18.*

Average transition / transversion rate was Ts/Tv = 2.04 ± 0.28. This average rate is comparable to that found in other species [31–33]. A higher than one Ts/Tv is expected because of the molecular mechanisms behind transitions and transversion but there was, nonetheless, a genomewide large variability (Figure 3.3, middle); 125 windows showed Ts/Tv > 2.5 and 30 had Ts/Tv < 1.5, also see S2 Figure. Although mutational bias is known to vary widely along the genome, there was a striking apparent correlation between number of SNPs and transition / transversion rate, both increasing in telomere regions; CpG count followed also similar patterns (Figure 3.3).

We were intrigued by this observation, which seems not to have been reported previously. First, we noted that the rate of missing rate is correlated to the number of SNPs but this correlation was not too high, explaining only ~4% of variability for number of SNPs per window (Figure 3.4a). Therefore, contrary to what would have been expected, missing rate is not relevant to predict the number of SNPs in a window or, in other words, this means that mapping alignment quality and depth (the two most influential factors in calling SNPs from NGS data) are independently distributed of nucleotide variability, at least in our data (Figure 3.4a). In contrast, a much stronger relation was found between number of SNPs and Ts/Tv rate genomewide (R^2=0.36, Figure 3.4b).

In an attempt to explain how number of SNPs and Ts/Tv are interrelated, we fitted the linear models in equations 1 and 2. Table 3.2 shows the estimates of each independent variable, in increasing order of fit explained (R^2). For number of SNPs, Ts/Tv rate and recombination rate suffice to explain most of variability, whereas missing rate explains, marginally, only 3% increase in R^2. This indicates that the increase in number of SNPs per window is partly explained by an increase in the number of transitions. Gene density in turn is almost irrelevant, in agreement with our previous results [10]. All variables together explain 52% of variability, which means that there are still many more factors than those studied here that are relevant for explaining the number of SNPs per window (Figure 3.4c, Table 3.2).

The results for the Ts/Tv ratio are far more interesting. First, most of variability (61% out of 79%) is explained by CpG count (Table 3.2); this is likely due to the high instability of methylated CpG sites, which frequently mutate towards transitions [34]. Further, differences in recombination rate explain another 14% in R^2, whereas the rest of factors are only marginally relevant. Note that GC% per se, once corrected by the other factors, is not important, nor is gene density. In all, variability in Ts/Tv rates is fairly well explained (Figure 3.4d) by a differential composition in CpG in the genome and varying recombination rates. Our analyses also suggest, but do not prove, that the correlation of number of SNPs and Ts/Tv ratio that we observe is likely an indirect consequence of both variables being affected by the same genome features, *i.e.*, recombination rate and high mutability of CpG rich regions.

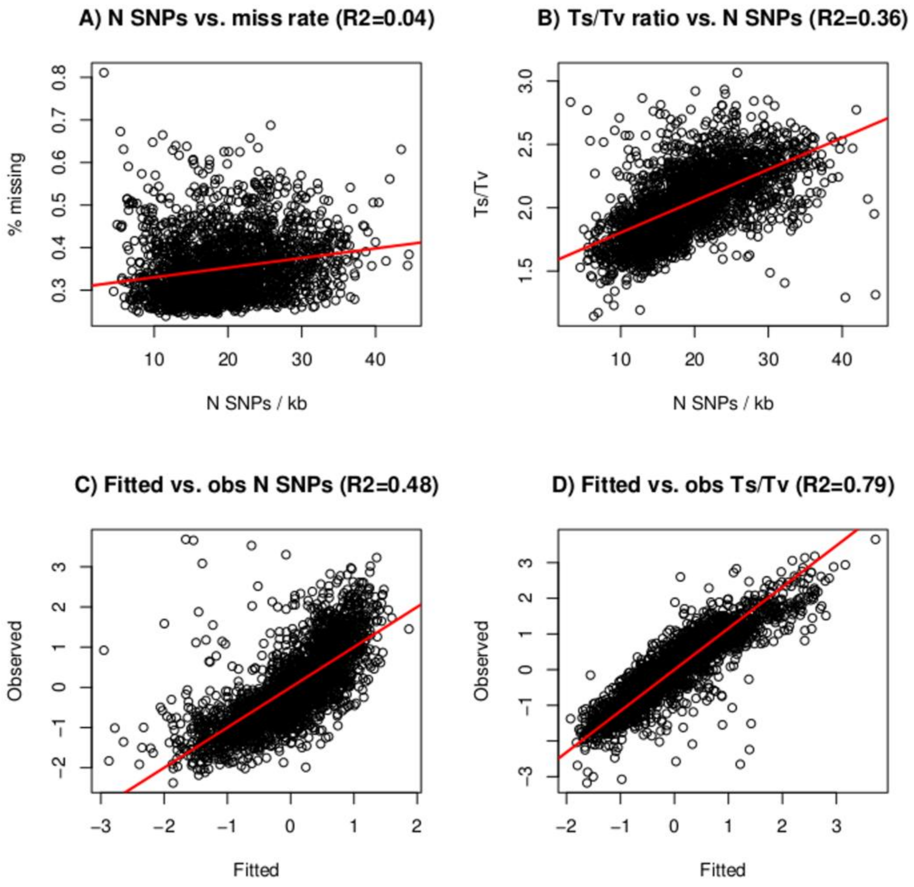

***Figure 3.4**: Number of SNPs vs. missing rate (a), and vs. Ts/Tv ratio (b); fitted vs. observed number of SNPs (c) and Ts/Tv (d) using equations 1 and 2, respectively.*

Ancestral allele

The ancestral allele was determined for biallelic (RA only) and monoallelic (AA) SNPs. Following the rules detailed in methods, it was possible to determine the ancestral allele for 39,017,375 out of 47,266,056 such SNPs (82%). Of them, 31,939,953 (82%) had the reference allele as ancestral, whereas the opposite occurred in the remaining SNPs. The number of times the reference allele is expected to be the ancestral allele can be approximated by the frequency of the ancestral allele across SNPs in a population of size 2N, where N is the number of individuals analyzed. This frequency q can be obtained from Ewen's sampling formula, $q_N = 1/(\sum_{i=1}^{2N-1} 1/i)$. Due to large variability in missing rate (Figure 3.2a), the number *N* to choose is not clear. Taking *N* = 100 (the modal sample size, Figure 3.2b), the expected frequency of the ancestral allele is 0.83 and for *N* = 128, *q* = 0.84, *i.e.*, very close to what was observed (82%).

Variants per population group

We calculated the number of variants per group (Table 3.3). The lowest number of variants was detected in European Wild Boars, which was also the group with fewest samples, whereas the highest number of variants was found in the Asian Wild Boars, also the most numerous group, although at lower average depth (S1 Table, Figure 3.2a). Note that the expected number of SNPs (S) to be detected is proportional to the number of samples sequenced, in a neutral model, E(S) = q_N θ L, where q_N is Ewen's sampling term, q is nucleotide diversity per base pair, and L, the length sequenced.

Next, we investigated for how many SNPs the derived allele was specific to one pig group or shared between two or more groups. To do so, we used the 34,500,122 variants where the ancestral allele was identified and called in at least 50% of the pigs in each group. Figure 3.5 shows the results in a Venn plot. A total of 4,052,639 (12%) variants was segregating in all four groups; in 39% of the variants (1,660,106 + 8,089,523 + 3,896,250), the derived allele was present only in Asian populations, whereas 18% (4,363,064 + 915,532 + 872,262) were exclusive of European populations. We found that ~ 9M SNPs were found exclusively in wild boars, whereas ~6.5 M variants were exclusive of domestics. Not unexpectedly, because of their higher variability compared to European wild boars, Asian wild boar had the highest number of unique variants (8,089,523 or 23.5% of the variants) and European wild boars, the lowest (only

915,532 or 2.65%). Note, however, that almost none of the SNPs had an exclusive allele fixed in any of the groups (only 14 in European wild boar, Table 3.3).

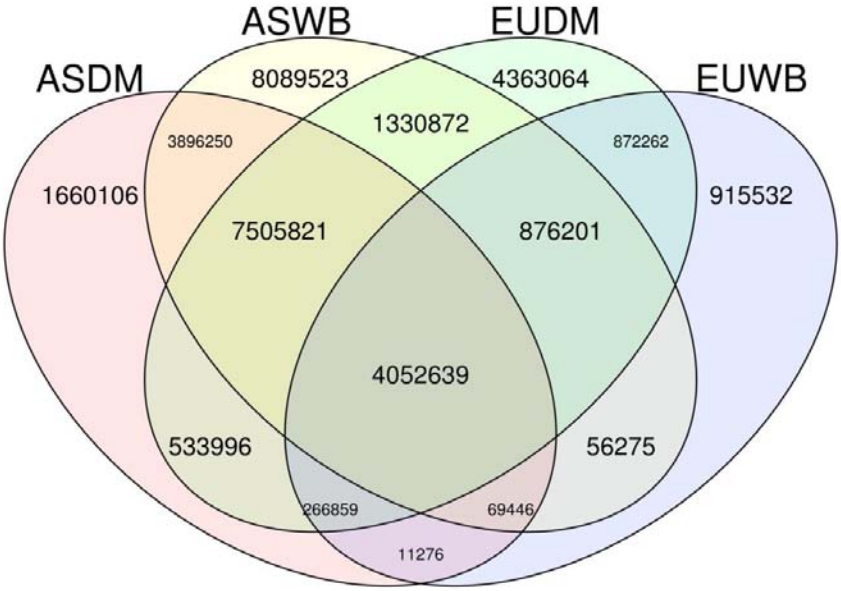

***Figure 3.5**: Number of exclusive and shared variants in the four groups. In each population, the variant must have been called for at least in 50% of the sample size.*

Joint site frequency spectra between populations, that is, how correlated are allele frequencies between populations, is a useful tool to infer demographic parameters (e.g., [35]). Here we computed the SNP joint site frequency spectrum between population groups: Domestics vs. Wild within continents and Asia vs. Europe within domestication status. Given that the number of individuals genotyped for each SNP varies, we only considered for these calculations the SNPs present in the modal group size, that is, for each group, the number of samples *n* that contained the maximum number of SNPs genotyped in exactly *n* samples (Figure 3.6). Note that the spectra are rectangular due to unequal number of samples per group. Comparisons of wild boar vs. domestics within continents show a diagonal pattern, that is, a positive correlation in allele frequencies between wild boar and domestics; this is the outcome of domestics being derived from local wild boars in each continent. There are some interesting differences between Asia and Europe though. In Europe, the pattern is somewhat less marked and with an increased density of markers at extreme frequencies (very low and very high allele counts).We interpret this as the result of low effective population size in Europe and the marked divergence of Asia and European groups. In contrast, the joint spectrum between continents was completely different, a result of the long evolutionary distance that separates Asian from European pigs (> 1 MYA), be it wild or domestics. In this case, the joint spectrum is dominated by alleles at extreme frequencies, particularly in Europe. For instance, consider the lowest and uppermost rows in ASWB vs. EUWB, they correspond to SNPs that may segregate at intermediate frequencies in Asia, but are singletons in Europe, and make most of the SNPs. This pattern is also observed when contrasting ASDM vs. EUDM although less marked, likely a result of Asian introgression in EUDM.

Variant annotation

The 48,119,476 variants called were analyzed by the VeP program (v-76,[23]). We found SNPs overlapping with 21,455 out of the 25,322 annotated genes and 25,166 transcripts. About half (22,336,270) of the variants were novel, *i.e.*, not present in dbSNP (*build* 140), variants. Most of the SNPs were annotated into intergenic (67.5%) or intronic (29.5%) regions, about 1% (463,030 SNPs) were annotated into coding regions. On average we found 1,013 SNPs per gene, including coding and non coding variants, as well as upstream, downstream regions and intronic variants. The detailed number of variants per category and per chromosome is in Supplementary material S3 Table. Among the 463,030 SNPs annotated into coding regions, 246,976 were synonymous. The most severe variant classes, according to their predicted functional consequences, are listed in Table 3.4 for all individuals and by group.

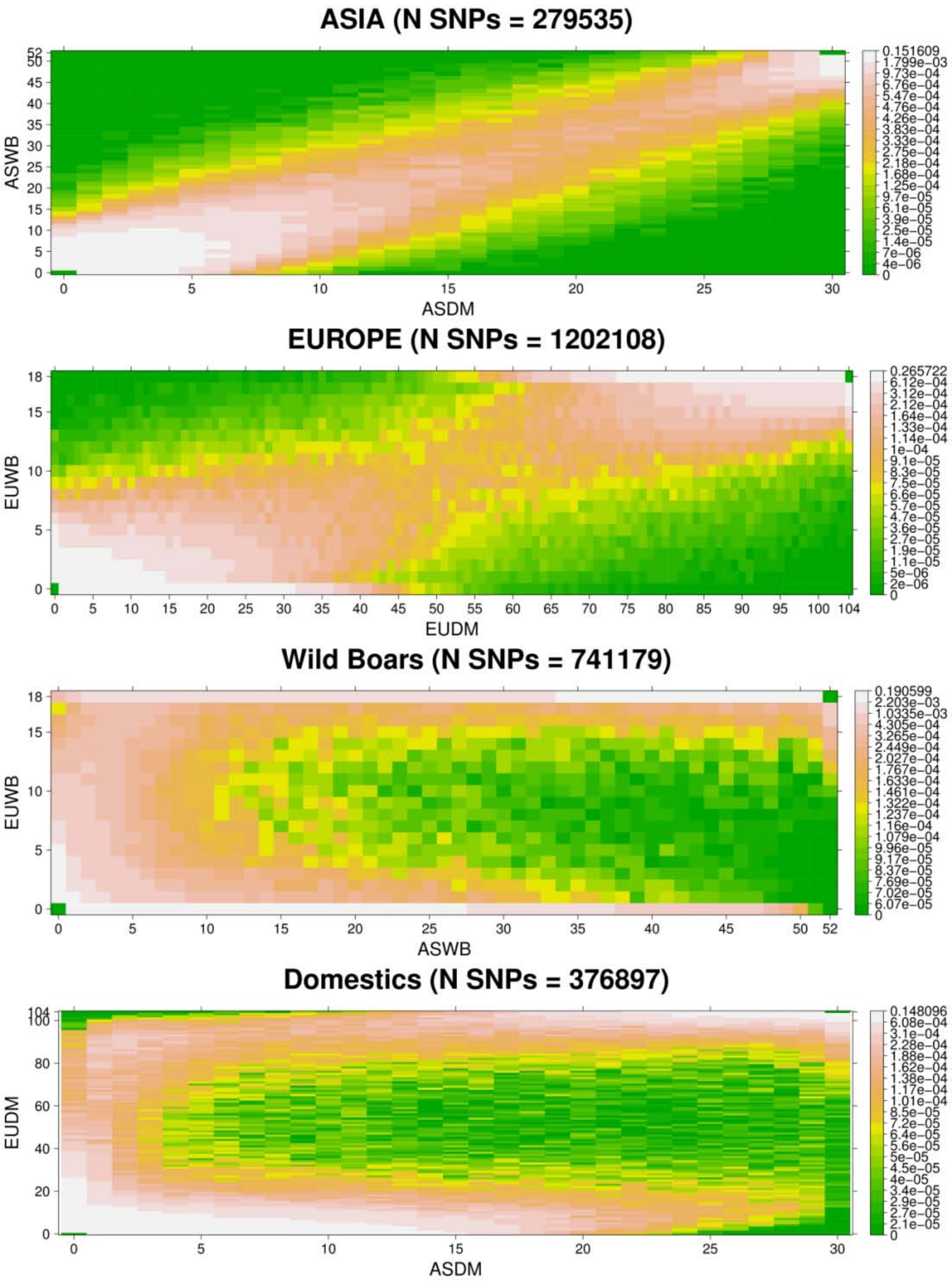

***Figure 3.6**: Joint site frequency spectra between population groups. Only SNPs found in the modal number of samples per groups were used. In each figure, x and y axis represent counts of the derived allele from 1 to 2N in each population, where N is the number of samples having the largest number of*

SNPs genotyped. Note that a count of 2N in say axis x means that the derived allele is fixed in that population but the same SNP can be segregating in the other population. The frequency of bivariate counts is represented in colors, with the log-scale as shown in the vertical bar. The more frequent a class is, the lighter the color, where dark green correspond to rare classes.

A total of 168,785 non-synonymous (missense) variants were found in 15,790 genes. SIFT predictions were obtained from 166,958 of these variants; 29% were predicted to have deleterious consequences (SIFT score < 0.05) on protein function (S3 Table). By population group, the percentage of predicted deleterious missense variants ranged from 12% (ASDM) to 28% (EUDM, S3 Table). We also identified how many SNPs with extreme frequency (>0.8 and <0.20) differences between wild boar and domestics were synonymous or non-synonymous. In contrast to Rubin *et al.*, [36], we did not find any over representation of non synonymous variants in domestics, neither in Europe nor in Asia (S5 Table). The most likely reason for this discrepancy is that Rubin *et al*. [36] pooled Asian and European wild boars; if we repeat the analyses with the same wild boar pool as these authors, we also retrieved an excess of non synonymous mutations in domestic pigs. A further complication for this analysis is that sample size is quite unbalanced, especially in Europe, so the presence of a new SNP in EWB can largely sift the population allele frequency.

Discussion

The pig is a highly variable and diverse species

To our knowledge, we present the most comprehensive SNP catalog of any livestock species to date. Using primarily published sequences, we identified over 48 million variants in the autosomal pig genome, which is more than the 28.3M SNPs recently reported in cattle [4]. Despite unequal and sometimes shallow coverage, the number of SNPs discovered per Mb was ~19,000 or one per 50 bp. This clearly shows how massive sequencing efforts have the ability to unfold vast amounts of hidden variability that could not have been detected until now. This work therefore expands dramatically the catalog of variants that are of potential interest in the pig breeding industry and beyond, given that the pig is also an important biomedical model. This effort, it should be noted, refers only to SNPs, similar works remain to be done for structural variants, mainly CNVs and indels.

Of the 48M identified SNPs, 46% were novel, indicating as well how incomplete are the porcine genomic resources available so far. These SNPs overlapped with 21,455 out of the 25,322 pig annotated genes, and we found an average of 1,013 variants per gene. Further, this catalog was obtained from worldwide samples, domestic and wild, making it an unbiased account of polymorphism in the species. Simulations suggest that the dataset generated should be highly reliable, at least for non complex regions where read mapping is not an issue. Simulation of the NGS pipeline with Pipeliner [26], using exactly the same options and comparable depth for each of the 128 pig samples, suggest that the SNPs reported are very likely to be real (FDR ~ 1%) and that we have uncovered a large percentage of the SNPs segregating in the populations sequenced. We estimate that, in the best case scenario, excluding NGS mapping problems in complex genome regions, about 90% of the SNPs segregating in the Asian samples sequenced and close to 100% for European samples sequenced may have been detected (Figure 3.1). At least for genome regions with good mapping properties, we have likely reported a large part of common SNPs in the pig species. Aside from genome complexities, It should be noted that, in current assembly, still about 8% of genome is estimated to be missing [8] so using an improved future assembly could even increase the amount of SNPs that can be retrieved from the same dataset.

Genomic context does matter

As in Drosophila and other species, including pigs [10,37,38], we found a significant correlation between recombination rate and number of polymorphisms, as predicted by models of hitchhiking and background selection [39,40]. In general, we also found an increased number of SNPs towards telomeric regions (Figure 3.3).

But perhaps the most surprising observation is that this increased number of SNPs is largely explained by a correlated change in mutational bias Ts/Tv, and not by the percentage of missing values caused by shallow depth (Figure 3.3, Figure 3.4b, Table 3.3). In turn, most of this Ts/Tv bias is explained by CpG content and recombination rate (Figure 3.4d). We are not aware of this phenomenon having been reported in other species, and whether this happens in other mammal or non mammal species should be investigated further. Our analyses suggest that an elevated CpG content subsequently increases the ratio of transition/transversion caused by methylation and affecting, indirectly, the number of SNPs.

The pig species has relatively few group exclusive SNPs

Ascertaining SNPs with extreme frequencies between groups is useful for traceability purposes, and to identify signatures of selection and of domestication. We looked for exclusive variants in all pairwise comparisons (ASDM, ASWB, EUDM, and EUWB), and also between domestics and wild boar between and within continents, setting the minimum sample size to half the group size per each group. About 4M SNPs were segregating in all groups, suggesting that these SNPs are very old, prior to divergence between the European and Asian clades that occurred ca 1 MYA or that they were introgressed more recently in European breeds from Asia [41]. Asian wild boar showed the highest number of private variants (> 8 million, Figure 3.5), in agreement with an Asian origin of the species [6,42], and the larger geographic span of sampling locations in Asia than in Europe. In contrast, less than 1M were exclusive of European wild boar. Interestingly, there were ~10 times more shared SNPs between domestics (ASDM vs. EUDM) than between wild boars (ASWB vs. EUWB). This could be due in part to the larger number of EUDM animals sequenced, but also to the introgression of Asiatic germplasm into European domestic breeds during late 18th century onwards, which likely has introduced alleles that had been lost in the European wild boar [43,44].

Higher frequency of potentially deleterious variants in Europe than in Asia

Annotation is one of the most critical and time consuming aspects of any genome, and that of the pig is still largely based on *in silico* automatized procedures. Therefore, the SNPs annotation provided here cannot be considered the definitive annotation; furthermore, about 8% of the pig genome is thought to be missing from current annotation [8] so these results should be taken with some caution. Similarly, a low SIFT [25] score cannot be taken as an infallible proof of damaging status because these algorithms are error prone and also, SIFT is based on the premise that function and protein evolution are correlated, and rely on protein conservation though species [24]. Nevertheless, they can serve as guide to prioritize variants that can be of interest for follow up studies.

With all these caveats in mind, it is nonetheless interesting to remark that we found a higher proportion of potentially deleterious variants in European Domestics (24%) and European wild boars (18%) than in Asian pigs (12%, Table 3.4). Although further works to verify this should be done, it could be due to the lower effective population size in European populations, as compared to Asia, which in turn results in natural selection being less effective in purging deleterious alleles. An alternative explanation would be that artificial selection in European breeds has resulted in an increase in alleles that are perceived as deleterious by current SIFT algorithms. However, this does not explain the increased frequency of potentially deleterious alleles in European wild boar.

Conclusions

We have carried out a large scale data mining effort of currently available pig genomes to uncover over 48M autosomal SNPs; a parallel simulation study suggests that false discovery rate should be very low, at least in genome regions with good 'mappability'. About 40% of the SNPs had not been reported, which shows how incomplete pig genome resources are. Intriguingly, we have found a large variability in mutational bias (transition / transversion rate) along the pig genome that is primarily explained by differences in CpG content and recombination rate. As for number of SNPs per kb, it is relatively insensitive to the rate of missing values and it depends mainly on Ts/Tv and recombination rates. The pig is a species with a very complex demographic history, where European and Asian branches isolated ~1 MYA only to be crossed in modern times to result in the most widely used pig breeds worldwide. As a result, there exists a relatively small percentage of SNPs that are exclusive of these European breeds compared to other populations. In contrast, the differences between Asian and European wild boars are much higher.

Ethics statement

DNA samples and genome analyses from all samples in this work have been published previously, and we refer to the original works for details [8–10,12,13,36,37]. DNA samples were obtained from blood samples collected according to national legislation, from tissue samples from animals obtained from the slaughterhouse, or from semen. For Spanish samples in particular, animal manipulations were performed according to the Spanish Policy for Animal Protection RD1201/05, which meets the European Union Directive 86/609 about the protection of animals used in experimentation.

Acknowledgments

We thank all groups and persons who provided data and / or samples.

Table 3.1: Total variants detected and number of variants per allele number at that locus.

Number of Allele(s) at position	Num. of positions	%
1 allele (AA)	362,740	0.75
2 alleles (all)	46,918,498	97.50
2 allele (R/A)	46,903,316	97.47
2 allele (A1/A2)	15,182	0.03
3 alleles (R/A1/A2)	828,854	1.72
4 alleles (R/A1/A2/A3)	9,384	0.02
Total number of variants	48,119,476	

R=Reference allele; A= Alternative allele; A1 = Alternative allele 1; A2 = Alternative allele 2; A3 = Alternative allele 3.

Table 3.2: Multivariate regression estimates for number of SNPs and Ts/Tv ratio.

Number of SNPs / kb			
Independent variable	Estimate ± SD	t-statistics	Increase in R2
Ts/Tv rate	0.56 ± 0.03	19.59***	0.36
Log(rec. rate)	0.34 ± 0.02	18.20***	0.12
Missing rate	-3.12 ± 0.39	-7.90***	0.03
Gene density	-0.12 ± 0.12	-7.69***	0.01
GC %	0.11 ± 0.03	3.82***	<0.01
CpG count / kb	-0.03 ± 0.04	-0.61	<0.01
Sum			0.52
Ts / Tv rate			
Independent variable	Estimate ± SD	t-statistics	Increase in R2
CpG count / kb	0.62 ± 0.02	24.18***	0.61
Log(rec. rate)	0.24 ± 0.01	20.19***	0.14
N SNPs / kb	0.25 ± 0.03	19.59***	0.04
Missing rate	0.67 ± 0.27	2.53*	<0.01
GC %	-0.02 ± 0.02	-1.21	<0.01
Gene density	0.01 ± 0.02	1.14	<0.01
Sum			0.79

All dependent and independent variables are standardized, except percentage of missing values; recombination rates are log-transformed; *, P<0.05; ***, P<10^{-3}. Note that variables tend to be significant even if its effect is small because of the large number of observations (windows). For that reason, the increase in R^2 due to each variable is a more useful assessment of its importance.

Table 3.3: Number of individuals and variants detected per group of populations.

Group	Number of individuals used	Number of Variants detected	Exclusive segregating variants	Exclusive fixed variants
ASDM	23	26,499,318	1,660,106	0
ASWB	41	35,719,205	8,089,523	0
EUDM	55	29,564,324	4,363,064	0
EUWB	9	12,562,569	915,532	14

Exclusive and fixed variants when filtering by SNPs called in at least 50% of the individuals in each group.

Table 3.4: Summary of the SNP annotation results for the most deleterious consequence obtained using VEP. Annotation term order is in decreasing order of severity according to Ensembl #.

	ALL	ASDM	ASWB	EUDM	EUWB
Splice donor variant	1,325	452	538	898	297
Splice acceptor variant	1,280	391	542	887	257
Stop gained	5,224	873	1,373	3,630	739
Stop lost	174	74	98	124	44
Initiator codon variant	359	155	209	237	98
Missense (non-synoynmous) variant	168,785	61,401	83,093	110,565	39,205
Splice Region variant	38,562	18,541	23,852	24,432	9,410
No. potentially deleterious (SIFT<0.05)	48,379	12,918	19,303	30,087	9,011
Total gene variants	21,741,159	10,732,921	14,632,536	12,645,639	4,941,666
Intergenic	33,988,809	17,284,550	23,626,744	21,137,970	7,829,133
Total variants*	48,119,476	25,427,907	34,471,527	29,073,261	11,602,230

*Note that the total of intergenic + genic variants is greater than the number of variants because it includes all those variants carrying more than one allele.

#***(http://www.ensembl.org/info/genome/variation/predicted_data.html)***

Supporting information legends:

Supporting information can be found at the following URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118867#sec024

S1 Script: Scripts used in the analyses of the real data and to simulate the performance of the pipeline. (doi:10.1371/journal.pone.0118867.s001)

S1 Figure: Simulated evaluation of expected genotype recovery (a) and sensitivity (b), error types for heterozygous genotypes (c), heterozygous genotype false discovery rate (d), error types in homozygous for the alternative allele (e), false discovery rate for homozygote alternatives alleles (f). RR=genotyped as homozygous for the reference; AA=genotyped as homozygous for the alternative; RA= genotyped as heterozygous. (doi:10.1371/journal.pone.0118867.s002)

S2 Figure: Genomewide distribution of standardized statistics by windows of ~1Mb. N SNPs, total number of SNPs per window; Ts/Tv, transition / transversion rate; CpG, number of CpG counts; log rec. rate, logarithm of recombination rate in cM/Mb from Tortereau *et al.* (doi:10.1371/journal.pone.0118867.s003)

S1 Table: Details of samples analyzed. ASDM, Asian Domestics; ASWB, Asian Wild Boar; EUDM, European Domestics; EUWB, European Wild Boar; N, number of samples; Average depth is calculated after filtering by base and map quality. (doi:10.1371/journal.pone.0118867.s004)

S2 Table: Number and kind of variants detected per chromosome. (doi:10.1371/journal.pone.0118867.s005)

S3 Table: Number of SNPs per annotation class and SIFT score of biallelic SNPs per population group.(doi:10.1371/journal.pone.0118867.s006)

S4 Table: The ten genes where the premature stop codon mutation was fixed, together with the SNP location. (doi:10.1371/journal.pone.0118867.s007)

S5 Table: Derived nucleotide substitutions showing marked allele frequency differences between wild boars and domestic pigs. (doi:10.1371/journal.pone.0118867.s008)

References

1. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464: 587–591. doi:10.1038/nature08832

2. Amaral AJ, Ferretti L, Megens H-J, Crooijmans RPM a, Nie H, Ramos-Onsins SE, et al. Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. PLoS One. 2011;6: e14782. doi:10.1371/journal.pone.0014782

3. Carneiro M, Rubin C-J, Di Palma F, Albert FW, Alfoldi J, Barrio a. M, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science (80). 2014;345: 1074–1079. doi:10.1126/science.1253714

4. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. Nature Publishing Group; 2014;46: 858–865. doi:10.1038/ng.3034

5. Porter V. PIGS - A handbook to the breeds of the world. HELM Information Ltd; 1993. p. 256.

6. Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, et al. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science. 2005;307: 1618–1621. doi:10.1126/science.1106927

7. Ramos-Onsins SE, Burgos-Paz W, Manunza a, Amills M. Mining the pig genome to investigate the domestication process. Heredity (Edinb). Nature Publishing Group; 2014; 1–14. doi:10.1038/hdy.2014.68

8. Groenen M a. M, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. Nature Publishing Group; 2012;491: 393–398. doi:10.1038/nature11622

9. Fang X, Mou Y, Huang Z, Li Y, Han L, Zhang Y, et al. The sequence and analysis of a Chinese pig genome. Gigascience. 2012;1: 16. doi:10.1186/2047-217X-1-16

10. Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silió L, et al. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics. 2013;14: 148. doi:10.1186/1471-2164-14-148

11. Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, et al. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet. 2013;45: 1431–1438. doi:10.1038/ng.2811

12. Ramírez O, Burgos-Paz W, Casas E, Ballester M, Bianco E, Olalde I, et al. Genome data from a sixteenth century pig illuminate modern breed relationships. Heredity (Edinb). 2014; 1–10. doi:10.1038/hdy.2014.81

13. Burgos-Paz W, Souza C a, Megens HJ, Ramayo-Caldas Y, Melo M, Lemús-Flores C, et al. Porcine colonization of the Americas: a 60k SNP story. Heredity (Edinb). 2013;110: 321–330. doi:10.1038/hdy.2012.109

14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

17. Neuman J a, Isakov O, Shomron N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. Brief Bioinform. 2013;14: 46–55. doi:10.1093/bib/bbs013

18. Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

19. Team R. R Development Core Team. R A Lang Environ Stat Comput. 2013; Available: http://www.r.project.org/

20. Warnes GR. gplots: Various R programming tools for plotting data. Journal of Phycology. 2012. pp. 569–575. doi:10.1111/j.0022-3646.1997.00569.x

21. Sarkar D. Lattice Graphics. R Doc. 2003;

22. Tortereau F, Servin B, Frantz L, Megens H, Milan D, Rohrer G, et al. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics. BMC Genomics; 2012;13: 586. doi:10.1186/1471-2164-13-586

23. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26: 2069–2070. doi:10.1093/bioinformatics/btq330

24. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11: 863–874. doi:10.1101/gr.176601

25. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4: 1073–1081. doi:10.1038/nprot.2009.86

26. Nevado B, Perez-Enciso M. Pipeliner: software to evaluate the performance of bioinformatics pipelines for Next Generation re-Sequencing. Mol Ecol Resour. 2014; doi:10.1111/1755-0998.12286

27. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28: 593–594. doi:10.1093/bioinformatics/btr708

28. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. Genome Res. 2009;19: 136–142. doi:10.1101/gr.083634.108

29. Pérez-Enciso M. Genomic relationships computed from either next-generation sequence or array SNP data. J Anim Breed Genet. 2014;131: 85–96. doi:10.1111/jbg.12074

30. Ramos-Onsins S, Ferretti L, Raineri E, Marmorini G, Burgos-Paz W, Vera G. mstatspop [Internet]. [cited 13 Oct 2014]. Available: http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html

31. Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, et al. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat Genet. 2000;24: 381–386. doi:10.1038/74215

32. Kijas JW, Lenstra J a, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, et al. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol. 2012;10: e1001258. doi:10.1371/journal.pbio.1001258

33. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science. 2012;193. doi:10.1126/science.1216872

34. Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci U S A. 1990;87: 4692–4696.

35. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: e1000695. doi:10.1371/journal.pgen.1000695

36. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109: 19529–19536. doi:10.1073/pnas.1217149109

37. Bosse M, Megens H-J, Madsen O, Paudel Y, Frantz L a F, Schook LB, et al. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genet. 2012;8: e1003100. doi:10.1371/journal.pgen.1003100

38. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic

Reference Panel lines. Genome Res. 2014;24: 1193–1208. doi:10.1101/gr.171546.113

39. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992;356: 519–520. doi:10.1038/356519a0

40. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993. pp. 1289–1303.

41. White S. From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. Environ Hist Durh N C. 2011;16: 94–120. doi:10.1093/envhis/emq143

42. Frantz L a F, Schraiber JG, Madsen O, Megens H-J, Bosse M, Paudel Y, et al. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biol. BioMed Central Ltd; 2013;14: R107. doi:10.1186/gb-2013-14-9-r107

43. Giuffra E, Kijas J, Amarger V. The origin of the domestic pig: independent domestication and subsequent introgression. Genetics. 2000;154: 1785–1791.

44. Bosse M, Megens H-J, Madsen O, Frantz L a F, Paudel Y, Crooijmans RPM a, et al. Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. Mol Ecol. 2014;23: 4089–4102. doi:10.1111/mec.12807

CHAPTER IV

Inferring the demographic history of European and Asian wild boar populations from the joint site frequency spectrum

4. Inferring the demographic history of European and Asian wild boar populations from the joint site frequency spectrum

Erica Bianco [1,2], Miguel Pérez-Enciso [1,2,3], Sebastián E. Ramos-Onsins[2]

[1] Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra, Spain; [2] Centre de Recerca en Agrigenómica (CRAG), 08193 Bellaterra, Spain; [3]Institut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

(Manuscript under development)

Abstract

In this study we analyzed the joint Site Frequency Spectrum (jSFS) of Eurasian wild boars and, using coalescence simulation and the diffusion approximation method of ∂a∂i, we try to infer the demographic model of Eurasian wild boars. A total of 1,968,814 genomewide polymorphisms from 9 European wild boars (EUWB) and 8 Asian wild boars (ASWB) were used to obtain the jSFS, which was dominated by extreme frequency SNPs classes, mostly in the EUWB population, and it presented a reduced number of SNPs in the classes at intermediate frequency in EUWB and at high frequency in ASWB. Trying to recover the shape of the observed spectrum, we simulated different scenarios with coalescence. Short and long evolutionary distance between the two populations did not recover the shape of the spectrum. Nevertheless, when migration was included in the demographic model, we found a joint spectrum coherent with the observed. With ∂a∂i, we tested 6 models that differed in number of bottlenecks and migration events. Despite reducing the bounds for the parameters to be estimated and taking into account for the severity of the bottleneck, only the simplest model converged. The more complex models failed to converge, but also suggest that migration between Asia and European wild boars after their split is needed to explain the results.

Keywords

Sus scrofa; Pig; NGS; Demographic Inference; joint Site Frequency Spectrum.

Introduction

Understanding the demographic history of a species is a key aspect in population genetics. Apart from completing the archeological evidence, knowing the demographic history of a species serves as null model in genome scans to detect regions under selection (Nielsen *et al.* 2007). The demographic history of *Sus scrofa* (wild boar) is a very complex one and is only partially known. *S. scrofa* originated in Island South East Asia during Plio-pleistocene (~ 5.3 - 3.5 MYA; Frantz *et al.* 2014). From there, it spread throughout the Eurasian continent. Studies on mitochondrial and nuclear DNA suggested the split between European and Asian wild boars occurred between 1.2 MYA and 0.8 MYA (Giuffra *et al.* 2000; Groenen *et al.* 2012; Frantz *et al.* 2014).

European wild boars demographic history have been studied using mitochondrial DNA (Scandura *et al.* 2008; Kusza *et al.* 2014), microsatellite data (Scandura *et al.* 2008) and whole genome sequence data (Groenen *et al.* 2012). After the split from the Asian population, European wild boars population size increased (Groenen *et al.* 2012), but later suffered a bottleneck, a contraction to southern *refugia* during last glacial maximum (~20,000 years ago), showed both with mtDNA (Scandura *et al.* 2008) and whole genome sequence data (Groenen *et al.* 2012). Nevertheless, a recent study on central Europe wild boars did not find this fluctuation in population size around the last glacial maximum (Kusza *et al.* 2014). More recent human mediated migrations made after World War II did not leave a detectable signature on European wild boar mtDNA (Scandura *et al.* 2008; Kusza *et al.* 2014).

Asian wild boars demographic history was recently studied using whole genome resequencing data, with PSMC algorithm (Groenen *et al.* 2012, Frantz *et al.* 2014). After the split with EUWB, Asian population increased. During the last glacial maximum, ASWB suffered a bottleneck, however, it was less pronounced than in European population (Groenen *et al.* 2012). After the split with EUWB, ASWB split into two populations, North and South China, around ~0.6 MYA (Frantz *et al.* 2014).

Up to date, wild boar demographic inferences did not include migration between populations, because the methods used implied only single individual or single population demographic inference. To evaluate migrations, Groenen *et al.* (2012) used the D-statistics (Patterson *et al.* 2012) and found that admixture occurred between European wild boars and North China wild boars throughout Pleistocene.

To our knowledge, only Groenen *et al.* (2012) made a comparison between the demographic histories of European and Asian wild boars, but there are no studies that consider jointly both populations demography, including migration events, and no estimation of the gene flow have been done using whole genome sequencing data. The aim of the current study is to infer a model for the joint site frequency spectrum of Eurasian wild boar, inferring the changes in population size, the time of the bottleneck and finally the amount of migration between Asian and European populations.

Materials and methods

Sample size and SNPs selection

High confident SNPs of Eurasian wild boars were obtained from Bianco *et al.* (2015). We included in the analysis all the European Wild boars of our previous study, but only Chinese and Russian wild boars of the Asian group. Tibetan wild boars (Li *et al.* 2013) were removed because of the low coverage and the unclear classification as wild boar (Frantz *et al.* 2015; Pérez-Enciso *et al.* 2015), whereas the Japanese wild boar was removed because it was an outlier in the Asian population structure (see Results, Groenen *et al.* 2012).

In total, 9 European wild boars (EUWB) and 8 Asian wild boars (ASWB) were used to obtain the joint Site Frequency Spectrum (jSFS). The complete list of individuals included in the analysis is shown in Table 4.1. Multi individual SNP (vcf) file was obtained from Bianco *et al.* (2015). The vcf file was then converted into plink using VCFtools v. 0.1.12a (Danecek *et al.* 2011; --plink option), retaining only biallelic SNPs. Silent SNPs (intergenic and synonymous) where ancestral allele is known were used (--extract). Pruning for linkage disequilibrium was done with plink (Purcell *et al.* 2007) --indep 50 5 2 parameters (default). After filtering for LD, a total of 6,135,186 SNPs were retained.

To calculate the jSFS, SNPs with missing data must be excluded. To increase the number of SNPs used, a strategy is to reduce the per SNP sample size (N) to the N at which the maximum number of SNPs have been genotyped in the population, and then subsampling those SNPs genotyped in more individuals. Missing rate was calculated with plink (--missing), separately for EUWB and ASWB, to calculate the N at which the maximum number of SNPs was genotyped: 9 and 6 respectively. Those polymorphisms that were genotyped in more than 6 individuals in ASWB population were randomly subsampled to 6 genotypes with a custom made tool. Per population derived allele frequency was obtained with plink (--freq) to calculate the unfolded jSFS.

Population geographic structure evaluation

In order to evaluate if a geographic structure was present between and / or within populations, we perform principal component analysis (PCA; Price *et al.* 2006; Patterson *et al.* 2006) on the SNPs used to calculate the jSFS. The PCA was then plotted with R (R Core Team 2013).

Simulated joint Site Frequency Spectrum

Different demographic scenarios were simulated using the coalescent simulator *ms* (Hudson 2002), trying to recover the appearance of the jSFS of EUWB and ASWB. In all scenarios we simulated 15 individuals divided in two populations, 9 and 6 respectively (to mime the sample size per population at which the maximum number of SNPs was sequenced), 1000 iteration and mutation parameter set to 500 (in order to have a high number of variant sites to calculate the jSFS):

```
ms 30 1000 -t 500 -I 2 18 12
```

We simulated different scenarios: starting from the simplest scenario of a split at different times in the past, we add complexity to the simulations. We tested if growth rate, for both growing and reducing the size of the two populations, recover the observed jSFS. Both the same and different growth rate per population was tested as well as changes in growth rate at a certain time in the past. All the changes in growth rate were tested at each of the time since the split tested before. Later we introduced migration event in the time closest to present, testing both equal and unequal gene flow at different proportions. The detailed simulation of split, growth rates and migrations values are reported in the Result section (Table 4.2). jSFS were drawn and simulated and observed jSFS were compared using *∂a∂*i (Gutenkunst *et al.* 2009).

Demographic inference

To perform demographic inference, we utilized the diffusion approximation method implemented in *∂a∂*i (Gutenkunst *et al.* 2009). *∂a∂*i calculates the jSFS of a given model and, through an optimization step, it adjusts the parameters of the model in order to increase the likelihood with the observed spectrum.

In total, we evaluated 6 demographic models:

- Model 1: exponential growth/reduction after split in two populations (Figure 4.1a);
- Model 2: model 1 with asymmetric migration between the two populations (Figure 4.1b);
- Model 3: exponential growth/reduction after the split and a change in growth rate after T1 have passed (two epochs model). (Figure 4.1c);
- Model 4: model 3 with asymmetric migration in the last epoch (Figure 4.1d);
- Model 5: three epoch model, growth rates change after T1 time and T2 times from the split in two populations (Figure 4.1e);
- Model 6: model 5 with asymmetric migration in the last epoch (Figure 4.1f).

For each model, effective population sizes (N) and time from split (T; in 2Ne generations) and migration (M; in 2Ne units, when present) were inferred by

∂a∂i and optimized to reach the best fit parameters, starting from random values chosen between boundaries (upper and lower) selected by the user (Gutenkunst *et al.* 2009).

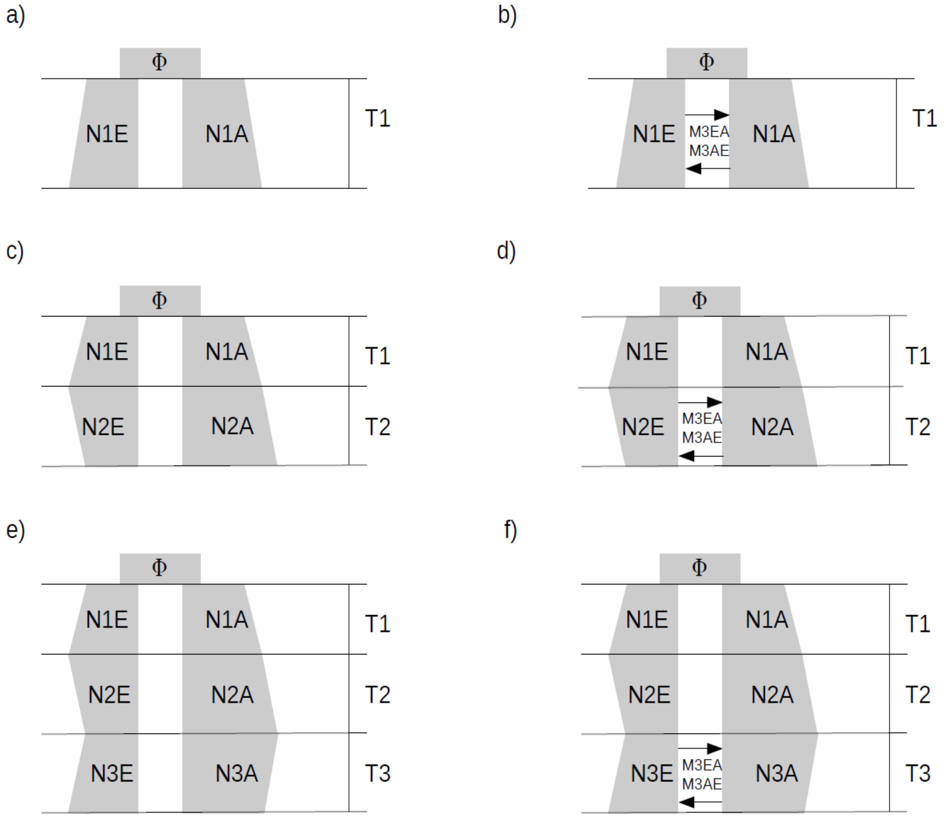

***Figure 4.1**: Demographic models tested with ∂a∂i. ɸ = common ancestor population. N§* indicates the effective size after the time T occurred, § = number of the epoch (1, 2, or 3); * = population (E = European Wild Boar, A = Asian Wild Boar). M3AE = migration from Asia to Europe; M3EA = migration from Europe to Asia. T§ = Epoch (1, 2 or 3).*

In mining the demographic history of a population, population size fluctuations are normally considered. When a population suffers a bottleneck, the same effects on variability can be produced by having a strong population size reduction during a short time period or by having a weaker reduction in size for a longer time (Fay & Wu 1999; Orengo & Aguadé 2004). In other words, the severity of a bottleneck is proportional to T/Nb, where T is the length of the period in which the bottleneck occurred and Nb is the population size during the bottleneck. In our model evaluations, these phenomena can be reflected by obtaining the same likelihood from different parameters when more than one epoch are included in the model (models from 3 to 6). In order to avoid this problem, we test additional models in which we set as fixed the time span of the first (models 3 and 4) epoch to 0.1, so that the severity of the bottleneck was only proportional to the population size fluctuation. In Table 4.3 there is a comprehensive list of the parameters used with upper and lower bounds indicated for each parameter.

The 6 models were tested using a single script in which we avoided the estimation of those parameters that were not present in the model to be tested (such as migration in models without migration). Basically, what we did was to set these parameters to a fixed value (the lower bound for time and migrations and 1 for population size), so that these parameters were not optimized. Each model was run from 60 to 80 different times, using different starting parameters to perform optimization.

Results

Observed and simulated joint Site Frequency Spectrum

A total of 1,968,814 SNPs were used to construct the jSFS, which were genotyped in exactly 9 EUWB and 6 out of 8 ASWB individuals. Principal component analysis was performed and no geographic pattern within EUWB was found (Figure 4.2). PC1 reflects the geographic separation between Asian and European wild boars, and explains ~22% of variance. PC2 reflects the distance between Japanese (WBJP) and the other Asian wild boars, because of this, WBJP sample was removed from the analysis.

A short evolutionary distance (0.01) between the two populations did not explain the jSFS we found (Figure 4.3b). The overall likelihood with observed data is low (likelihood ~ -1.8M). The classes along the diagonal were overrepresented, in contrast with the classes in which the two populations were fixed for different alleles. These classes were completely absent when two populations split close in the past.

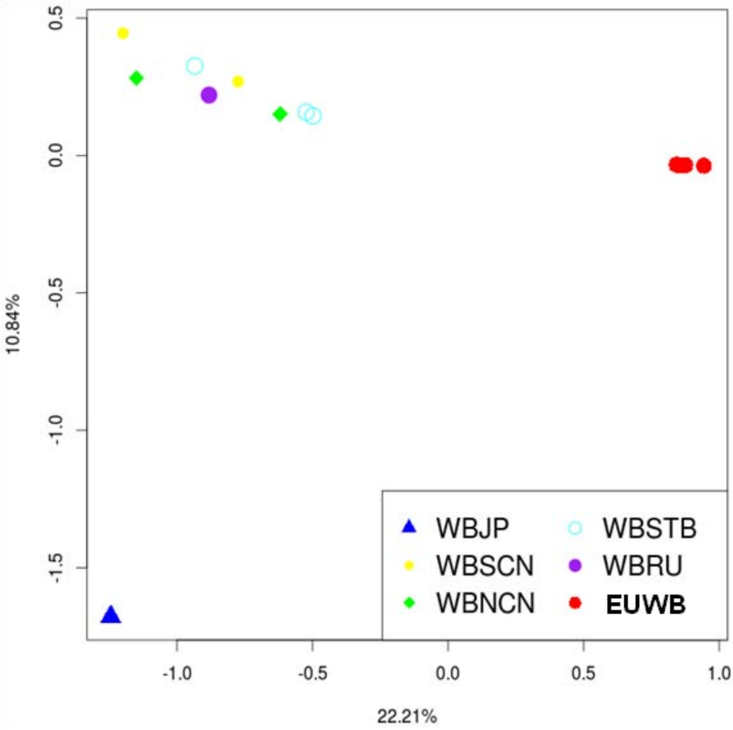

***Figure 4.2**: Principal component analysis of the wild boars based on the ~ 2M SNPs used in the study. EUWB = European wild boars; WBJP = Japanese wild boar; WBSCN = South China wild boars from Groenen* et al. *(2012); WBNCN = North China wild boars; WBRU = Russian wild boar; WBSTB = Tibetan wild boars from Li* et al. *(2013).*

When we simulated a long evolutionary distance (T=2), we also did not recover the observed shape of the jSFS (Figure 4.3c). In this case, the overall likelihood was higher than with short evolutionary distance, but still low (likelihood ~ -1.6M). In this case, in the simulated data all the classes with intermediate frequency in both populations are absent, due to the long distance between the populations, and all SNPs found were fixed in one of the populations. Changes in population size on one or both populations did not change the general jSFS shape.

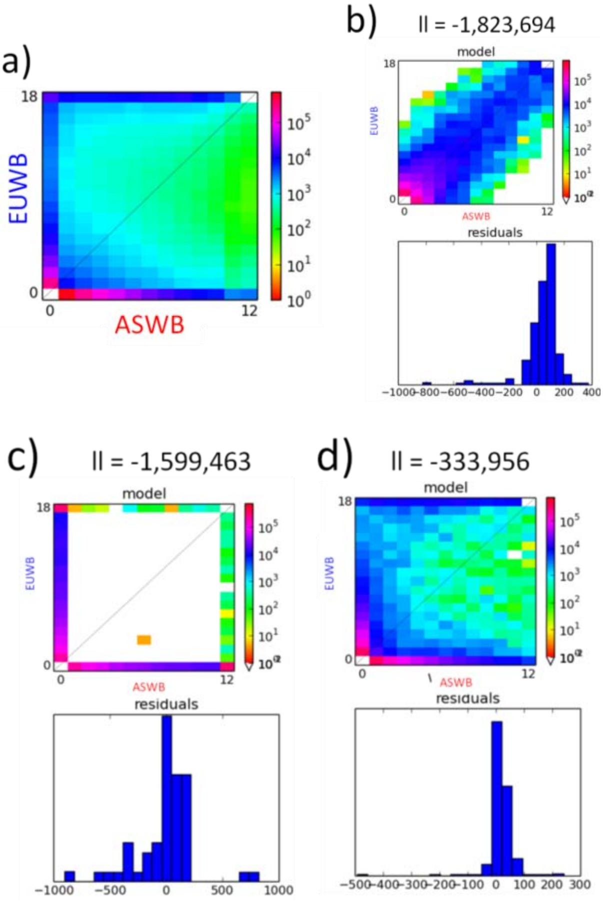

Figure 4.3*: Observed and simulated joint site frequency spectrum. The more frequent classes are in blue, magenta and red, less frequent classes in light blue and green. a) Eurasian wild boars observed joint site frequency spectra. b) jSFS obtained from simulated data: 0.01*Ne generations since split (upper); model fit to the data residuals (lower). c) jSFS obtained from simulated data: 2*Ne generations since the split (upper); model fit to the data residuals (lower). d) jSFS obtained from simulated data: Ne generations since the split and subsequent gene flow (upper); model fit to the data residuals (lower). Likelihoods on top of b) c) and d) indicate the overall fit of the simulated scenarios to observed data.*

The simulated scenarios that better recovered the "C" shape of the observed spectrum were those one which included migration. In Figure 4.3d we show an example of a demographic scenario with gene flow between the two populations in the last epoch. The overall likelihood was higher than scenarios without migration (likelihood ~ -300k), the "C" shape was recovered, the distribution of residuals was tighter around 0 than without migrations, and almost all classes were represented.

Parameters estimation, models and bounds selection

For each one of the six models, the parameters to be optimized can vary between the upper and the lower bounds defined by the user. We first run *∂a∂*i with the following upper and lower bounds (see also Table 4.3):

- T1, T2 and T3 :[10^{-2}, 10];
- N1E, N1A, N2E, N2A, N3E and N3A: [10^{-3}, 100];
- M3EA and M3AE: [10^{-3}, 10].

We focused on models 3 and 4 which had the best likelihoods in the first set of runs. Model 2 was excluded because of the low likelihood and models 5 and 6 because the complexity of a 3 epoch model made it difficult the calculation of best fit parameters and convergence.

Moreover, trying to improve parameters convergence, upper and lower bounds were reduced to:

- T1 andT2 :[10^{-2}, 3];
- N1E, N1A, N2E and N2A: [10^{-2}, 10];
- M3EA and M3AE: [10^{-2}, 10];

in order to reach more reasonable values and to improve the optimization process.

In this second set of run, we also take into account for severity (see Materials and methods). To do so, we tested two additional models 3b and 4b in which the time of the first epoch was fixed to 0.1 and only effective population sizes were optimized in the first epoch.

In Table 4.5 the top 10 likelihood models' parameters are shown. As found with the simulations done with *ms*, the best likelihood was given by model 4 which includes gene flow between the two populations. Nevertheless, even if the bounds were narrowed, the different runs did not converge to the same parameters, and taking into account severity did not improve convergence between runs: similar likelihoods were found with different estimated parameters (Table 4.5).

Discussion

Gene flow between Asian and European wild boars shaped actual diversity

In this study we try to infer the joint demographic history of Eurasian wild boars using the joint site frequency spectrum of the two populations: European and Asian wild boars. As shown in Bianco *et al.* (2015), the joint spectrum is dominated by SNPs at high frequency, with a little amount of SNPs at intermediate frequency in EUWB and fixed or nearly fixed in ASWB. Using simulated data, we found that short or long evolutionary distances (Figure 4.3b and c) between the two populations are not enough to explain the observed spectrum. A spectrum coherent with the observed was obtained simulating a demographic scenario that allows for gene flow subsequent from the split (Figure 4.3d). This hypothesis of a gene flow between European and Asian wild boars had already been suggested by Giuffra *et al.* (2000), who hypothesized Asian germplasm introgression into European wild boars in the last 200 years using mtDNA sequences. Groenen *et al.* (2012), using whole genome resequencing data, also suggested the gene flow occurred, but in late Pleistocene. In both cases, the authors suggested the gene flow occurred after the split and a period of isolation, as in the model we simulated.

Complex models fail to converge

Demographic models in which migration was included were compatible with the observed spectrum, as well as multiple epoch models. Wild boars population size fluctuations had already been found in previous studies (Scandura *et al.* 2008; Groenen *et al.* 2012; Kusza *et al.* 2014), as well as migration throughout Eurasia (Giuffra *et al.* 2000; Groenen *et al.* 2012). For that reason, we decided to test six different models with *∂a∂*i, trying to infer the demographic history of wild boars. Our idea was to evaluate first the simplest model of a split in the past with subsequent increasing or decreasing population size, and then add complexity including multiple population size fluctuations and migration(s). The first model converged to similar parameters (model 1: split in the past and change in populations sizes) in the different run performed (Table 4.3). Nevertheless, more complex models, which should explain more likely the demographic history of wild boars, did not converge to the same parameters in the different runs (more than 200 globally), although the best likelihood found per model was lower than model 1's likelihood.

One of the possible reasons two and three epochs models did not converge is the models tested were not realistic. We did not find a substructure in the PCA (Figure 4.2), but, as shown in Figure 4.4, there should be a separation between northern and southern samples (the red line), in concordance with Frantz *et al.* (2013) who dated the split between northern and southern Asian wild boar approximately 0.6 MYA. With this substructure in ASWB, a two populations' model will not explain wild boars demography, a second split within ASWB must be included in the model. In our study we only had 3 samples from northern location and 5 from southern location, which were not enough to recover this substructure in PCA. Increasing the number of samples will improve the power of detecting the split using principal component analysis.

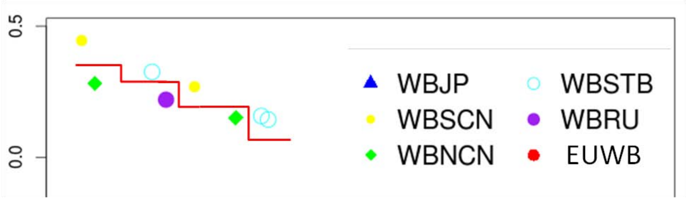

Figure 4.4*: Detail of PCA (Figure 4.2). The red line divides northern and southern ASWB individuals.*

Moreover, a previous study on SSCX found the presence of a large selective sweep that showed two distinct haplotypes between northern and southern wild boars in China (Ai *et al.* 2015). The authors suggest two possible explanations. First, the two haplotypes were maintained because of the low recombination of this region since their divergence. Despite Ai *et al.* (2015) found the two haplotypes diverged long before *Sus scrofa* diverged from the other *Sus*, 8.5 MYA, this time of divergence could had been increased by the strong selection on the two haplotypes (Ai *et al.* 2015), and northern and southern pigs subpopulations may have diverged later, as suggested by Frantz *et al.* (2013). The divergence of the two haplotypes on chromosome X and southern and northern wild boars confirm that a two populations' model is not correct to infer the demographic history of Eurasian wild boars. We must include a second split between southern and northern subpopulation after Asian wild boar diverged from European wild boars.

In addition, accordingly to Groenen *et al.* (2012), migration occurred between North China wild boars and European wild boars and we found that models including migration better explain the observed spectrum. Because of this, a model that includes migration between northern wild boars and EUWB after the split between north and south sub population should also be tested. Ai *et al.* (2015) also proposed that the haplotype found in northern individuals could have been introgressed from another extinct *Sus* species. If this is the case, it will be difficult to find a substructure within Asian wild boars. Nevertheless, European wild boars have the same haplotype than North China wild boars, so that again a two populations model might be too simple to explain Eurasian wild boar demography. In this case the model must include a first split between

South and North Asia wild boars, followed by a second one between European and North Asian wild boars. North China and South China wild boars have been found to cluster together in philogenetic trees (Groenen *et al.* 2012; Ai *et al.* 2015) and we did not found a population structure within Asian wild boars. If they have diverged previously than the divergence between EUWB and North China specimens, EUWB and North specimens have to be equally distant from South China wild boars. It is possible that a strong gene flow between North and South China groups have occurred after the split. To test this, an additional model that includes migration between North and South China wild boars after the split between European and North China samples should also be evaluated.

Another possible reason that difficulted the model to converge to the optimal parameters is that the likelihood was rather flat and present multiple local maxima, so that different demographic models result in joint site frequency spectra that fit equally better the observed spectrum (Myers *et al.* 2008). A deep and accurate analysis of likelihood behavior in the parametrical space and the comparison with other inference methods (such as Excoffier *et al.* 2013), will clarify this hypothesis and will allow to understand if a single demographic scenario can explain the observed spectrum.

In conclusion, here we show the first attempt to infer the joint demographic history of Eurasian wild boars. We found that the joint spectrum was explained with a demographic model that includes migration, but, when tested with *∂a∂*i, the more complex models did not converged to the same parameters. Future investigations are needed. Using multiple algorithms and testing for substructures within Asian population will help to recover Eurasian wild boars demographic history.

Acknowledgements

We thank all groups and persons who provided data and / or samples. EB is recipient of Ph. D fellowship BES-2011043530 from the Ministry of Science (Spain). This work was funded by AGL2010-14822 and AGL2013-41834-R (Ministry of Economy and Science, Spain) to MPE and SERO.

Bibliography

Ai H, Fang X, Yang B *et al.* (2015) Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, 1–11.

Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M (2015) A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig. *Plos One*, **10**, e0118867.

Danecek P, Auton A, Abecasis GR *et al.* (2011) The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, **27**, 2156–2158.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, **9**.

Fay JC, Wu C-I (1999) A Human Population Bottleneck Can Account for the Discordance Between Patterns of Mitochondrial Versus Nuclear DNA Variation. *Molecular Biology and Evolution*, **16**, 1003–1005.

Frantz LAF, Madsen O, Megens H-J *et al.* (2015) Evolution of Tibetan wild boars. *Nature Genetics*, **47**, 188–189.

Frantz LAF, Madsen O, Megens H-J, Groenen M a M, Lohse K (2014) Testing models of speciation from genome sequences: divergence and asymmetric admixture in Island Southeast Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Molecular ecology*, 1–9.

Frantz LAF, Schraiber JG, Madsen O *et al.* (2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome biology*, **14**, R107.

Giuffra E, Kijas J, Amarger V (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics*, **154**, 1785–1791.

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.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, **5**, e1000695.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, **18**, 337–338.

Kusza S, Podgórski T, Scandura M *et al.* (2014) Contemporary genetic structure, phylogeography and past demographic processes of wild boar Sus scrofa population in Central and Eastern Europe. *PLoS ONE*, **9**.

Li M, Tian S, Jin L *et al.* (2013) Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat Genet*, **45**, 1431–1438.

Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? *Theoretical Population Biology*, **73**, 342–348.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Andrew G (2007) Recent and ongoing selection in the human genome. *Nat Rev Genet.*, **8**, 857–868.

Orengo DJ, Aguadé M (2004) Detecting the footprint of positive selection in a European population of Drosophila melanogaster: Multilocus pattern of variation and distance to coding regions. *Genetics*, **167**, 1759–1766.

Patterson NJ, Moorjani P, Luo Y *et al.* (2012) Ancient Admixture in Human History. *Genetics*, **192**, 1065–1093.

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

Pérez-Enciso M, Burgos-Paz W, Ramos-Onsins SE (2015) On genetic differentiation between domestic pigs and Tibetan wild boars. *Nature Genetics*, **47**, 190–192.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**, 904–909.

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.

R Core Team (2013) R: A Language and Environment for Statistical Computing. *http://www.r-project.org/*.

Ramírez O, Burgos-Paz W, Casas E *et al.* (2014) Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity*, 1–10.

Scandura M, Iacolina L, Crestanello B *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? *Molecular ecology*, **17**, 1745–62.

Tables

Table 4.1: List of the individuals used to calculate the jSFS, with accession numbers.

Group	Country*	Sample ID	Accession num.	References
ASWB	CN	WB29U04_SChina	ERP001813	(Groenen *et al.* 2012)
ASWB	CN	WB29U12_SChina	ERP001813	(Groenen *et al.* 2012)
ASWB	CN	WB30U01_NChina	ERP001813	(Groenen *et al.* 2012)
ASWB	CN	WB30U08_NChina	ERP001813	(Groenen *et al.* 2012)
ASWB	CN	WBCN1851	SRA065458	(Li *et al.* 2013)
ASWB	CN	WBCN1852	SRA065458	(Li *et al.* 2013)
ASWB	CN	WBCN1853	SRA065458	(Li *et al.* 2013)
ASWB	RU	WBRU1064	-	Unpub.
EUWB	NL	WB21F05_Netherlands	ERP001813	(Groenen *et al.* 2012)
EUWB	NL	WB21M03_Netherlands	ERP001813	(Groenen *et al.* 2012)
EUWB	NL	WB22F01_NL	ERP001813	(Groenen *et al.* 2012)
EUWB	NL	WB22F02_NL	ERP001813	(Groenen *et al.* 2012)
EUWB	FR	WB25U11	ERP001813	(Groenen *et al.* 2012)
EUWB	SW	WB26M09_Malcantone	ERP001813	(Groenen *et al.* 2012)
EUWB	ES	WBES0494	SRP044261	(Ramírez *et al.* 2014)
EUWB	ES	WBES0717	-	Unpub.
EUWB	TN	WBTN0965	-	Unpub.

* ISO 3166 country code.

Table 4.2: List of the ms *simulations done to recover the observed jSFS. In bold the simulated data that were plotted in Figure 4.3b,c and d.*

Ts	T2	G1_1	G1_2	G2_1	G2_2	M12	M21
0.01*	**-**	**-**	**-**	**-**	**-**	**-**	**-**
2	**-**	**-**	**-**	**-**	**-**	**-**	**-**
0.25	-	-2	-3	-	-	-	-
0.01*	-	0.01	0.01	-	-	-	-
0.01*	-	0.02	6	-	-	-	-
0.01*	-	0.6	0.6	-	-	-	-
0.01*	-	6	6	-	-	-	-
0.1	0.09	-2	-6	-4	-3	-	-
0.01#	0.001	0.1	0.6	0.01	0.06	-	-
0.01#	0.001	0.1	0.6	0.01	0.01	-	-
0.01#	0.001	0.1	0.1	0.01	0.01	-	-
0.01#	0.001	0.01	0.06	0.1	0.6	-	-
0.01#	0.001	0.01	0.06	0.1	0.1	-	-
0.01#	0.001	0.01	0.01	0.1	0.1	-	-
0.01#	0.001	1	6	0.1	0.1	-	-
0.01#	0.001	1	1	0.1	0.1	-	-
0.01#	0.001	1	6	0.6	0.1	-	-
0.01#	0.001	0.1	0.6	1	6	-	-
0.01#	0.001	0.1	0.6	1	1	-	-
0.01#	0.001	0.1	0.1	0.1	0.1	0.01	0.01
0.01#	0.001	0.1	0.1	0.1	0.1	0.1	0.1
0.01#	0.001	0.1	0.1	0.1	0.1	0.5	0.5
0.01#	0.001	-2	6	6	3	0.5	2
1	**0.01**	**-2**	**6**	**6**	**3**	**0.5**	**2**
0.7	-	-	-	-	-	0.5	2

Ts=time from the split; T2=time at which growth rate changed; G1_1 = growth rate for population 1 at time 0; G1_2 = growth rate for population 2 at time 0; G2_1 = growth rate for population 1 at time T2; G2_2 = growth rate for population 2 at time T2; M12 = migration proportion from population 1 to population 2; M21 = migration proportion from population 2 to population 1.

* = the same growth rates were tested also using Ts=0.1 and Ts=1.

§ = the same growth rates and migration proportions (when tested) were also tested for T2=0.005 and T2=0.009 and for Ts=0.1 and T2=0.01, T2=0.09, T2=0.05; Ts=1 and T2=0.1, T2=0.9, T2=0.5.

Table 4.3: Number of parameters and the upper and lower bounds for each parameter and model tested.

#m	Np	T1	T2	T3	N1E	N1A
1	3	-	-	$[10^{-2}; 10]$	-	-
2	5	-	-	$[10^{-3}; 10]$	-	-
3	6	-	$[10^{-2}; 10]$ $[10^{-2}; 3]$	$[10^{-2}; 10]$ $[10^{-2}; 3]$	-	-
3b	5	-	0.1	$[10^{-2}; 3]$	-	-
4	8	-	$[10^{-2}; 10]$ $[10^{-2}; 3]$	$[10^{-2}; 10]$ $[10^{-2}; 3]$	-	-
4b	7	-	0.1	$[10^{-2}; 3]$	-	-
5	9	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-3}; 100]$	$[10^{-3}; 100]$
6	11	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-3}; 100]$	$[10^{-3}; 100]$

#m	N2E	N2A	N3E	N3A	M3AE	M3EA
1	-	-	$[10^{-3}; 100]$	$[10^{-3}; 100]$	-	-
2	-	-	$[10^{-3}; 100]$	$[10^{-3}; 100]$	$[10^{-3}; 10]$	$[10^{-3}; 10]$
3	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 100]$ $[10^{-2}; 10]$	-	-
3b	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	-	-
4	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 100]$ $[10^{-2}; 10]$	$[10^{-3}; 10]$ $[10^{-2}; 10]$	$[10^{-3}; 10]$ $[10^{-2}; 10]$
4b	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$	$[10^{-2}; 10]$
5	$[10^{-3}; 100]$	$[10^{-3}; 100]$	$[10^{-3}; 100]$	$[10^{-3}; 100]$	-	-
6	$[10^{-3}; 100]$	$[10^{-3}; 100]$	$[10^{-3}; 100]$	$[10^{-3}; 100]$	$[10^{-3}; 10]$	$[10^{-3}; 10]$

#= m: model; Np: number of parameters.

Table 4.4: Top 10 best likelihood per model in the first set of runs.

#m	T1*	T2*	T3	N1E	N1A	N2E	N2A	N3E	N3A	M3AE	M3EA	ll
1	0.01	0.01	0.6872	1	1	1	1	0.4790	12.6558	0	0	-161452
1	0.01	0.01	0.6951	1	1	1	1	0.5034	13.1570	0	0	-161491
1	0.01	0.01	0.6867	1	1	1	1	0.4918	13.1997	0	0	-161504
1	0.01	0.01	0.6756	1	1	1	1	0.4738	12.6325	0	0	-161551
1	0.01	0.01	0.6577	1	1	1	1	0.4776	12.2029	0	0	-162129
1	0.01	0.01	0.7538	1	1	1	1	0.5762	12.6388	0	0	-162692
1	0.01	0.01	0.6694	1	1	1	1	0.4581	9.9654	0	0	-162743
1	0.01	0.01	0.6636	1	1	1	1	0.4513	9.9843	0	0	-162818
1	0.01	0.01	0.6751	1	1	1	1	0.4969	9.9347	0	0	-163086
1	0.01	0.01	0.7163	1	1	1	1	0.5199	9.9804	0	0	-163134
2	0.001	0.001	9.9875	1	1	1	1	2.4652	17.6025	0.1095	0.0375	-111428
2	0.001	0.001	1.7860	1	1	1	1	0.8108	7.4660	0.2550	0.0659	-112522
2	0.001	0.001	1.2730	1	1	1	1	0.6631	7.3842	0.2075	0.0636	-113697
2	0.001	0.001	0.7955	1	1	1	1	0.4750	7.5321	0.1380	0.0041	-131705
2	0.001	0.001	1.2143	1	1	1	1	0.5584	3.8968	0.5552	0.0010	-140808
2	0.001	0.001	9.9308	1	1	1	1	0.2136	0.8377	2.1572	0.1879	-223291
2	0.001	0.001	3.8469	1	1	1	1	0.5319	2.1994	0.9194	0.0010	-238777
2	0.001	0.001	4.6968	1	1	1	1	0.0302	0.1738	9.8782	2.6847	-250876
2	0.001	0.001	1.5237	1	1	1	1	0.1284	5.8003	0.0267	1.4879	-251810
2	0.001	0.001	9.1775	1	1	1	1	2.9319	28.3547	0.0119	0.0203	-324327

#m	T1*	T2*	T3	N1E	N1A	N2E	N2A	N3E	N3A	M3AE	M3EA	ll
3	0.01	0.0462	0.3601	1	1	0.1003	9.3645	1.5870	3.3051	0	0	-77355
3	0.01	0.3231	0.1535	1	1	0.1374	9.9895	7.0000	2.2228	0	0	-78264
3	0.01	0.2588	0.1925	1	1	0.2017	6.1710	9.6571	2.0042	0	0	-145283
3	0.01	0.6444	0.2235	1	1	0.2933	9.8939	0.8259	6.6577	0	0	-173862
3	0.01	0.0146	0.7018	1	1	1.4303	3.4915	0.2846	4.1355	0	0	-177818
3	0.01	0.4215	0.6812	1	1	0.3290	9.8989	2.7299	2.7339	0	0	-182491
3	0.01	0.0415	0.5785	1	1	0.3049	2.9491	0.6164	1.5961	0	0	-194324
3	0.01	0.8561	0.0102	1	1	0.6560	5.2957	0.8996	9.8558	0	0	-196454
3	0.01	0.0229	0.2542	1	1	0.1197	2.9763	2.6657	2.9925	0	0	-199503
3	0.01	0.1478	0.3816	1	1	0.1633	0.5955	2.9700	2.9700	0	0	-296878
4	0.01	0.0266	0.6328	1	1	0.1000	5.9792	1.2367	3.2951	0.2171	0.0914	-60181
4	0.01	0.7780	2.6203	1	1	0.1000	3.4935	2.0464	8.9735	0.1403	0.0879	-64644
4	0.01	0.0273	1.9043	1	1	0.1568	2.4613	1.3484	7.1756	0.1379	0.1187	-65304
4	0.01	0.2025	1.4559	1	1	0.1102	1.0611	1.4481	7.3961	0.2771	0.1082	-73562
4	0.01	0.9069	0.3656	1	1	0.2341	9.8909	2.7356	1.9080	0.7740	0.0898	-85443
4	0.01	0.1666	1.0846	1	1	0.4202	2.9911	0.9712	9.8718	0.2139	0.0648	-90832
4	0.01	0.0141	0.7491	1	1	0.1236	0.9675	0.4595	9.3813	0.0282	0.1927	-121455
4	0.01	0.2786	0.2878	1	1	0.1061	9.7350	0.2736	2.1424	0.2489	0.6150	-133600
4	0.01	0.0100	2.6125	1	1	1.2911	0.6815	0.6699	4.8917	0.3943	0.0771	-136071
4	0.01	0.0669	0.6979	1	1	0.1869	2.9742	1.0703	1.9380	0.0478	0.0542	-138879
5	0.0840	0.3647	0.1356	0.6152	2.9398	0.5411	2.6466	0.1632	2.8863	0	0	-219585
5	0.1249	0.0102	0.4532	5.1586	0.8738	0.7689	0.5762	0.3202	5.4816	0	0	-250956

#m	T1*	T2*	T3	N1E	N1A	N2E	N2A	N3E	N3A	M3AE	M3EA	ll
5	0.0546	0.0100	0.2385	2.8863	0.2265	0.1021	2.5789	0.2022	0.7993	0	0	-351634
5	0.0472	0.0100	0.7793	0.1657	9.9464	1.1852	1.4215	0.4032	1.1651	0	0	-431958
5	2.0852	0.0236	0.1010	1.2353	8.1230	2.7177	9.9412	1.0359	1.5745	0	0	-471806
5	0.5320	0.1296	0.2686	0.1495	2.9687	0.1860	1.2168	1.1330	2.0331	0	0	-486874
5	0.7529	0.1126	0.1261	5.2263	9.3763	1.1767	5.2806	9.9365	0.3249	0	0	-613906
5	0.1162	1.5029	0.0277	7.6809	1.5506	0.4875	2.6505	9.4222	0.1049	0	0	-624156
5	0.0569	0.0342	0.0101	0.1010	0.7305	2.9700	2.9700	0.3098	0.1214	0	0	-796454
5	0.2155	0.0101	2.3744	0.1010	0.2935	2.7180	2.9700	1.4881	2.3063	0	0	-850649
6	0.0292	0.3636	0.2820	1.8595	9.4999	0.1014	4.2814	1.6978	2.2575	0.0174	0.3306	-61137
6	0.5687	0.0917	0.0525	0.2877	6.0307	9.3580	6.1212	2.6230	2.6267	0.0775	0.1998	-94416
6	0.9306	0.0181	0.0565	0.3702	9.8781	0.7011	5.4839	0.4010	1.5519	2.8382	0.6272	-134384
6	0.2479	0.0132	1.3507	0.4304	0.6739	0.4986	0.2905	0.1365	9.6597	0.3966	1.1402	-248048
6	1.2790	0.0101	4.9500	0.1994	1.1257	0.1701	2.9700	2.9700	2.9700	0.4955	0.0170	-296186
6	0.1356	0.0577	0.1937	0.1010	1.3597	0.6627	0.7352	0.2350	0.8165	0.7760	0.0101	-330021
6	1.4727	0.0668	0.0101	0.9906	2.9700	0.2670	2.9700	0.1010	2.9700	0.1964	1.9588	-397446
6	0.0101	0.8370	0.2923	0.1007	2.5102	5.4378	0.2026	0.3517	6.3866	2.4651	0.0102	-443607
6	0.1953	0.5225	1.2120	1.7695	2.8721	2.9234	0.1027	0.4717	2.2766	0.5899	0.0103	-451141
6	0.0137	0.3206	4.4146	0.1409	4.3125	8.1264	0.1377	0.5240	9.9891	1.0345	0.4204	-482417

*=model 2 was tested with a larger bound for times: $[10^{-3}, 10]$.

Table 4.5: Second set of runs, 10 best likelihood values. Model 3 is not shown because it fits worst the data than model 4 and 4b.

#model	T1	T2	N1E	N1A	N2E	N2A	M3AE	M3EA	ll
4	0.3314	0.3443	0.1212	9.9981	1.6591	2.4520	0.1697	0.1504	-41610
4b	0.1	0.4096	0.0762	9.9983	1.4204	2.9845	0.0344	0.1810	-46471
4	0.5190	0.2344	0.1726	6.3457	2.6811	5.7800	0.2693	0.1258	-46599
4b	0.1	0.3773	0.0752	9.9763	1.7598	3.0833	0.0100	0.1537	-47317
4b	0.1	0.4417	0.0654	5.7297	1.3624	4.0114	0.0206	0.2206	-50268
4	0.2365	0.5577	0.1439	8.5118	1.2975	4.3275	0.0252	0.1852	-55065
4	0.0608	2.5913	0.0100	2.0979	2.0984	8.7773	0.1296	0.1143	-56078
4b	0.1	0.3841	0.0921	2.2018	1.3871	10.0000	0.0100	0.1420	-56189
4	0.6600	0.1117	0.1989	9.9957	7.3322	1.3408	0.9668	0.1361	-57753
4b	0.1	0.5395	0.0354	6.9814	1.1709	3.9787	0.0267	0.3099	-58024

Time (T) is in 2Ne generation units; Migration (M) is in 2Ne units. Sample size N is in Ne units. Ne is the effective population size of the population at equilibrium ϕ.

CHAPTER V

The chimerical genome of *Isla del Coco* feral pigs (Costa Rica), an isolated population since 1793 but with remarkable levels of diversity

5. The chimerical genome of *Isla del Coco* feral pigs (Costa Rica), an isolated population since 1793 but with remarkable levels of diversity

Erica Bianco*†, Henry W. Soto‡, Lourdes Vargas§, Miguel Pérez-Enciso*†¶

* Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Spain. † Universitat Autònoma de Barcelona, Department of Animal Science, 08193 Bellaterra, Spain. ‡ Escuela de Zootecnia, Universidad de Costa Rica, 10501 San José, Costa Rica. § Sistema Nacional de Áreas de Conservación (SINAC), Ministerio de Ambiente y Energía (MINAE), San José, Costa Rica. ¶ Institut Català de Recerca I Estudis Avançats (ICREA), Carrer de Lluís Companys 23, Barcelona, 08010, Spain.

Molecular Ecology (2015) 24 (10): 2364-2378

Abstract

The history of domestic species and of their wild ancestors is a no simple one, and feral processes can clarify key aspects of this history, including the adaptive processes triggered by new environments. Here, we provide a comprehensive genomic study of *Isla del Coco* (Costa Rica) feral pigs, a unique population that was allegedly founded by two individuals and has remained isolated since 1793. Using SNP arrays and genome sequencing, we show that Cocos pigs are hybrids between Asian and European pigs, as are modern international pig breeds. This conclusively shows that, as early as the 18th century, British vessels were loading crossbred pigs in Great Britain and transporting them overseas. We find that the Y chromosome has Asian origin, which has not been reported in any international pig breed. Chinese haplotypes seem to have been transmitted independently between Cocos and other pig breeds, suggesting independent introgression events and a complex pattern of admixing. Although data are compatible with a founder population of N=2, variability levels are as high in Cocos pigs as in international pig breeds (~1.9 SNPs/kb) and higher than in European wild boars or local breeds (~1.7 SNPs/kb). Nevertheless, we also report a 10-Mb region with a marked decrease in variability across all samples that contains four genes (*CPE*, *H3F3C, SC4MOL* and *KHL2*) previously identified as highly differentiated between wild and domestic pigs. This work therefore illustrates how feral population genomic studies can help to resolve the history of domestic species and associated admixture events.

Introduction

Feralization, either by intentional releasing or accidental escape from human confinement, is an important and recurrent event from a genetics and ecological point of view. In general, feral animals are considered invasive species that disrupt the original ecosystem (Choquenot *et al.* 1996; Roemer *et al.* 2002; Cruz *et al.* 2005). Nevertheless, in the long run, feral animals usually adapt to their new territories (Oliver & Brisbin 1993), and are difficult to remove in practice. This makes feral animals unique evolutionary experiments, which tend to revert, in part, and recover some of their original characteristics present in their wild counterparts. Furthermore, a study of genetic differentiation across genome regions may point at genes whose selective coefficients have changed between captivity and the new wild environment. It is not rare that mutations with deleterious effects in the wild are selected during human captivity, a typical case is coat color mutations (Fang *et al.* 2009; Linderholm & Larson 2013). Finally, feral animals can also help to date and disentangle demographic events that occurred before feral and domestic animals diverged. This latter issue, which is the main topic of this work, is only possible if the history of the populations is known with sufficient detail. For instance, the genetic history of a specific feral population can be accurately inferred when the feral process is well-known and dated. This is, unfortunately, rarely the case because there can be a continuous, even if small, flux between wild and domestic specimens (e.g., between village pigs and local wild boars) or because historical documentation is lacking or unreliable. Here, we focus on the feral pigs from *Isla del Coco* ('Cocos Island'), in Pacific Costa Rica, and their relationships with other wild boar and domestic pig breeds.

The pig, *Sus scrofa*, originated in East Asia ca. 2-3 MYA and colonized most of temperate Eurasia and North Africa, but was absent from America before European colonization. The first recorded event of pig import into the New World, in the Caribbean, dates back to as early as the second Columbus trip in 1493 (Rodero *et al.* 1992; Crosby 2003; Zadik 2005). Pigs have always provided a cheap and reliable source of meat: they are easy to transport, as compared to cattle, much more prolific than sheep, and adapted to any terrain except for the driest. There is abundant historical evidence on this:

> ‘The pigs and cattle multiplied rapidly and quickly became feral, roving the islands and trampling the careful gardened landscape of Taíno cultivation. As early as 1507, the stock of cattle, pigs, and horses was so well established that breeding animals no longer needed to be imported’ (Higman 2010).

Two centuries later, starting in the early 18th century, English and Dutch ships were bringing Chinese and Siamese pigs in Europe, which had became fashionable due to their much better reproductive performance, docility and higher lard content than the autochthonous European pigs (Porter 1993). In the meantime, and ever since the beginning of long discovery trips, it was customary to release domestic animals in new territories, particularly islands, since these would serve as a source of meat in successive arrivals. Simultaneously, domestic animals were also traded, loaded in ships and released elsewhere. This is well documented, e.g., during the exploration of the Pacific by Captain Cook, who is credited for having released the first pigs in the New Zealand islands (Gascoigne 2007). It is interesting to remark that, according to Gascoigne (2007), Cook thought of delivering British agricultural advances to the Pacific through the export of farm animals suggesting that, at that time, improving British pigs by crossing them with Asian breeds was not that evident.

On July the 25th 1793, English captain James Colnett arrived at *Isla del Coco* commanding the whaling ship Rattler. After four days, they departed the island having left onshore one boar and one sow for the use of later visitors. In their own words:

> *‘We were much wearied, during the four days, we passed off this island, and prepared to quit. We therefore took on board, two thousand cocoa nuts; and, in return, left on shore, in the North*

> *bay, a boar and a sow, with a male and female goat.' (J. Colnett, p. 73)*

There is no further documented introgression of pigs in the island. While goats became extinct, the pigs have thrived and populated the island. Currently, the estimated census is of ~ 400 – 500 pigs, oscillating depending on food availability (Sierra 2001). During the late 17th and 18th centuries, the *Isla del Coco* was visited on numerous occasions by whaling ships, since the island is in the main area of whale transit in the East Pacific corridor. What makes the Cocos pigs unique is that there is one single documented pig release event and that, due to the protected and isolated area, this is likely to reflect the actual events (Arias-Sánchez 1993). Because of this, *Isla del Coco* pigs would be direct descendants of pigs living in the British islands more than two centuries ago, since at that time there was no, legally speaking, direct trade between Spanish and British colonies (Arias-Sánchez 1993).

The *Isla del Coco*, which became Costa Rican territory in 1869, is located in the Pacific Ocean, 500 km away from continental Costa Rica (05°31′N 87°04′O); it is only 24 km^2 large, but with difficult access due to a complex orography. The *Isla del Coco* National Park presents unique environmental conditions, extreme pluviometry/rainfall (~7000 mm/year), it is fully covered by cloud tropical forest and hosts numerous endemic species, including one of Darwin's finches (*P. inomata*) which genome was recently sequenced (Lamichhaney *et al.* 2015). It was designated a National Park and Biological Reserve by the government of Costa Rica in 1978 and World Heritage Site by UNESCO in 1997 (Government of Costa Rica 1996).

There are no genetic studies on Cocos feral pigs yet, despite their unique history and interest. The only genetic study so far was an analysis of three microsatellites by Sierra (2001), who found that, despite strong bottlenecks, these pigs were remarkably heterozygous. It is certainly of great historical and practical interest to disclose the relationship between current Cocos pigs and modern pig breeds, especially with those of English origin, since Cocos pigs can provide a yardstick against which to compare the changes brought about by artificial selection during the last century. Studying the genomes of Cocos pigs should reveal, as well, traces of the bottleneck and possible adaptive signals to the new environment. In this work, we present a comprehensive genomic analysis of Cocos feral pigs with these purposes in mind, framed in a species-wide context.

Materials and methods

Tissue (tail) samples from *Isla del Coco* pigs were collected by specialized personnel in the *Isla del Coco* National Park, during normal management practices of the population, and were preserved in ethanol until further processing. For this work, a diversity of genomic data was obtained from several porcine datasets. Since different analyses were carried out in each dataset, this section is arranged by dataset and analysis to facilitate reading.

Array Genotyping and population structure

Twelve Cocos pigs were genotyped with the 60k SNP array from Illumina (Ramos *et al.* 2009), and was performed by GeneSeek (Lincoln, NE, USA). These genotypes were combined with a wide pig biodiversity panel, fully described in Burgos-Paz *et al.* (2013), that comprises feral and village pigs from several American countries, European and Chinese domestic breeds, and European wild boar. These pigs had also been genotyped with the 60k array, and a total of 411 genotyped samples were available in total. Merged data were filtered using PLINK (Purcell *et al.* 2007) as described in Burgos-Paz *et al.* (2013). In short, SNPs were pruned if they were monomorphic, had a minor allele frequency below 0.05, were located on the sex chromosomes, had more than 5% missing genotypes, were not mapped on the Sscrofa10.2 assembly or the position was ambiguous. Among the 62,163 SNPs initially present in the chip, 46,211 were finally retained.

The 60k SNP array data were used to visualize genetic distances between worldwide porcine populations. Principal Component Analysis (PCA) was carried out as implemented in Eigenstrat (Price *et al.* 2006). A complete relationship between individuals was drawn via a Neighbour Joining (NJ) algorithm and visualized with DENDROSCOPE software (Huson & Scornavacca 2012) using pairwise identity-by-state genetic matrix distance (1-IBS) as obtained with PLINK (Purcell *et al.* 2007). To examine potential origins of each population, the Maximum Likelihood approach implemented in ADMIXTURE v. 1.22 (Alexander *et al.* 2009) was employed in the 60k array data. First, ADMIXTURE was run in a semi-supervised manner to estimate the percentage of the Asian component in each population; this was done by assigning Chinese pigs to a first cluster and European wild boars and Iberian pigs to a second cluster (K=2), and letting ADMIXTURE assign each cluster percentage to the rest of individuals genotyped with the SNP array. ADMIXTURE was also run in a completely unsupervised manner with a variable number of clusters, using cross-validation to choose an optimum K-value, as suggested by the authors. Both PCA and ADMIXTURE were run by pruning markers in high-linkage disequilibrium using the option --indep 50 5 2 in PLINK.

Mitochondrial control region sequencing

About 694 bp from the mitochondrial control region was sequenced in the diversity panel described above, including the 12 Cocos pigs, using primers and conditions as in (Alves *et al.* 2003). Sequences were edited and aligned with SeqScape® Software v2.7 against the *Sus scrofa* mitochondrion complete genome (NCBI ref. AF034253). Additional sequences were downloaded from GenBank. Multiple sequence alignment was performed using Muscle v3.8.31 (-diags -maxiter 4; Edgar 2004). To represent the sequences, networkv.4.6.1.2 (Bandelt *et al.* 2000) was employed.

Microsatellite genotyping

In order to investigate whether the number of alleles in Cocos pigs was compatible with a founder bottleneck of two individuals, we genotyped 14 Cocos pigs, 10 Iberian and 15 Large White pigs with a 12-microsatellite panel. The twelve microsatellites are among those recommended by ISAG for traceability purposes (http://www.isag.us/committees.asp?autotry=true&ULnotkn=true) and are therefore known to be highly polymorphic: SW240, SW857, SW911, S0090, S0090, SW936, SW936, SW72, S0155, SW951, S0386, S0101, S0355 and SW24. Genotyping was carried out by the *Servei Veterinari de Genètica Molecular* (http://www.svgm.es/eng, Barcelona, Spain).

Shotgun sequencing, alignment and variant calling

Shotgun sequence (NGS) from 16 pigs was analyzed: two Cocos pigs (IC, one boar and one sow), Iberian from Spain (IB, n=2), European wild boar from Spain and Switzerland (WB, n=2), Large White (LW, n=2), Duroc (DU, n=2), Meishan (MS, n=2), one Wuzhishan (WU), from South China, one Tamworth, a local endangered British breed (TW), one Yucatan, a minipig strain developed in the USA from local Mexican pigs (YU), and one Guatemalan Creole, a village pig (CR). The two Cocos samples, one Iberian, the Tamworth and the Yucatan were sequenced for this work, whereas the rest of the sequences were downloaded from the SRA archive (Table 5.1).

Shotgun genome sequencing was performed in the *Centro Nacional de Análisis Genómico* (CNAG, www.cnag.cat, Barcelona, Spain) using the HiSeq2000 Illumina platform. The library preparation was performed according to the Illumina paired-end sequencing protocol, with minor modifications. For seven samples (DU23M01, DU23M02, LW36F04, LW36F05, MS20U10, MS21M14 and WB26M09), aligned reads in bam format were downloaded from the SRA archive(ERP001813); for the rest of the samples, raw reads were mapped against assembly 10.2, which is from a Duroc pig (Groenen *et al.* 2012), with BWA (Li & Durbin 2009) allowing for seven mismatches and using default options otherwise. Samtools v. 0.1.18-sl61 was employed for duplicate removal and sorting using rmdup and sort options, respectively (Li *et al.* 2009). For both the downloaded bam files and those generated in-house, GATK v. 2.7 IndelRealigner (McKenna *et al.* 2010) was run to improve the alignment around indels, using default options. Genotypes were called for each individual separately using the samtools (v. 0.1.19) mpileup option and filtered with vcfutils.pl varFilter (Li *et al.* 2009). Indels were excluded in this work. For a SNP to be called, we set the minimum depth to 5× and the maximum depth at twice the average sample's depth plus one; minimum map quality and minimum base quality were each set to 20. Individual vcf files containing the genotypes were then merged using custom scripts in a multi-individual file. In short, for each individual, missing variant positions were coded according to bcftools output without the '-v' flag to avoid variant calling; confident homozygous-reference calls were coded as '0/0' (homozygous for the reference allele), and the position was marked as missing './.' otherwise. The general bioinformatics pipeline is fully described in Bianco *et al.* (in press) Although large genetic divergence between the sample and the assembly may cause biases in the SNP calling at low depth (e.g., Nevado *et al.* 2014), in Bianco *et al.* (in press) we showed that the bioinformatics pipeline employed here is expected to have a low false discovery rate, in the order of 1% for heterozygous genotypes, and a power of ~95%.

In the case of chromosome Y, we mapped the vcf file SNPs into the assembly using Pipeliner vcf2fas tool (Nevado & Perez-Enciso 2014), replacing unaligned positions by N's. Neighbor Joining (NJ) tree was then obtained with MEGA5 using pairwise deletion and otherwise default options (Tamura *et al.* 2011).

Sequence-based diversity estimates

Despite being massive, NGS data result in highly unbalanced datasets due to unequal coverage, insufficient map and/or base qualities across regions and samples. As a result, genotype files normally contain a large number of missing data, requiring methods that specifically account for this. Here, we employed the methods developed by Ferretti *et al.* (2012) and implemented in mstatpop software (Ramos-Onsins, unpublished, available at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/) to infer nucleotide diversity and Fst from the shotgun-sequence NGS data. Watterson’s theta estimators of diversity (θ) were calculated in each population. Fst and diversity were computed in 100 kb non-overlapping windows, setting the restriction of a minimum of 20 kb aligned per window and per individual.

Identity by descent (IBD) tracts and runs of homozygosity (ROH)

Sequence-based genotypes were phased from the 16 samples and Identity by Descent (IBD) tracts were inferred with Beagle4 (Browning & Browning 2013a). Data were previously filtered to remove all SNP with a missing rate greater than 20%.To identify long runs of homozygosity (ROHs), we combined the two approaches proposed by Browning and Browning: IBDseq (Browning & Browning 2013b) and Beagle4(Browning & Browning 2013a). The former allows for potential genotyping errors from NGS data analysis, whereas the latter allows for better control, as it reconstructs both haplotypes. Once IBD segments had been inferred, we analyzed if there was a co-occurrence of Asian origin in the same genome segments between different pairs of breeds. In particular, we were interested in studying whether tracts IBD between Chinese and Cocos pigs were also IBD between Chinese and other breeds. The goal was to identify a potential common Asian signature that was shared between the Cocos and modern European breeds. To do so, we computed the probability of a given breed 'A' having an Asian haplotype when the Cocos pigs also harbor an Asian haplotype in the same genomic region using:

$$P(A{\equiv}MS \mid IC{\equiv}MS) = L(A{\equiv}MS \cap IC{\equiv}MS) / L(IC{\equiv}MS),$$

where $P(A{\equiv}MS \mid IC{\equiv}MS)$ is the probability of any base-pair in breed A being IBD with Meishan (MS), given that this position is also IBD between Cocos (IC) and Meishan, and $L(A{\equiv}B)$ is the genome-wide sum of lengths of the IBD segments between A and B breeds obtained from Beagle4. Length $L(A{\equiv}MS{\cap}B{\equiv}MS)$ was obtained by intersecting $L(A{\equiv}MS)$ and $L(B{\equiv}MS)$ using BEDtools (Quinlan & Hall 2010). For comparison, the probabilities, with respect to Iberian (IB), were also computed:

$$P(A{\equiv}IB \mid IC{\equiv}IB) = L(A{\equiv}IB{\cap}IC{\equiv}IB) / L(IC{\equiv}IB).$$

Since there is ample evidence that Iberian pigs have not been crossed with Asian pigs (Fernández *et al.* 2011; Ramírez *et al.* 2015), they provide a baseline against which to compare Asian introgression. The values reported are averages over both haplotypes and the two individuals of each population, except for Tamworth, where a single individual was sequenced.

Bayescan (Foll & Gaggiotti 2008)was employed to test whether selection could be responsible for observed differentiation patterns. Bayescan models Fst coefficients in a hierarchical manner, a population-specific effect shared by all loci and a locus (potentially selective) component, and computes the Bayes factor of the model with selection vs. without selection. Bayescan was employed using the sequence-based SNPs in the largest ROH identified.

Demographic model simulation

To study whether two founder pigs could account for the observed nucleotide variability, combined coalescence - forward simulations were performed. First, Asian (100 diploid individuals) and European (25 individuals) populations were simulated with coalescence using MaCS (Chen et al. 2009) with command

macs 250 100000 -t THETA -r 0.001 -I 2 50 200 -n 2 DIFF -ej 3 2 1, (1)

where DIFF is the ratio of Asian/European effective sizes (Ne), and THETA, the variability of the European population before introgression. Next, we employed forward simulation using the output of the coalescence as starting sequences with Slim (Messer 2013). In this part, we simulated an F1 that resulted from crossing Asian and European pigs followed by a backcross, then 1, 4 or 6 generations of random mating to mimic the stage after the Chinese pig import into the UK, and by a bottleneck of N=2 individuals lasting one generation to represent the founding of the Cocos population; finally, the population grew to Ne =50 and continued with constant Ne for 100 generations, the time from when the feral population was founded until sampling (we assumed an average generation interval of two years). Variability (Watterson's θ) in the admixed population was estimated from four randomly sampled sequences, i.e., two diploid individuals. The whole process was run 2,000 times, the length simulated was 100 kb in each replicate.

We were primarily interested in assessing whether Cocos variability was compatible with a bottleneck N=2, conditional on observed variability for putative European and Asian ancestors. In the absence of such information, we used Iberian and Meishan as proxies. To do so, we chose 2,000 random 100kb windows (one for each replicate) from the Iberian and Meishan genomes, and we used as input in equation (1) for DIFF the observed values from the real Meishan/Iberian data, and for THETA the variability of the Iberian pig population.

Results and discussion

Table 5.1 and S1 (Supporting Information) show the main statistics of the genotyped and shotgun sequenced samples, respectively. After filtering by quality, average sequencing depths were 8.7 and 12.7 for the sow and boar Cocos pigs, respectively.

Cocos feral pigs make a highly differentiated population and are the result of admixing Chinese and European germplasms.

The PCA obtained with the 60k-array autosomal genotypes in the diversity panel shows that the *Isla del Coco* pigs are of European origin (Figure 5.1a); otherwise, they would be positioned closer to Asian pigs than the rest of the European pigs analyzed, e.g., Large White or Duroc. Note that the Cocos feral pigs are tightly grouped, and make a cluster of their own, likely the result of having evolved under isolation with a low effective population size. The PCA suggests that Cocos pigs can be considered a distinct population today, separate from the rest of the pig breeds studied. Interestingly, the closest populations, in terms of the 60k SNP-based metrics, are Guadeloupe, Colombian Creoles and Ossabaw pigs (Figure S1, Supporting Information), all of which are village or feral pigs from Central America. It can be hypothesized, therefore, that these Central American pigs are descendants from a common wave of colonizing events that started in Europe from the 16th century onwards. In agreement with the 60k panel data, the maternally inherited mitochondrial data also indicate a European origin (Figure 5.1b). We found a single haplotype in the mitochondrial control region of all 12 Cocos individuals, consistent with a tight clustering in the 60k array PCA (Figure 5.1a). This haplotype is quite frequent in European pigs, and has also been reported in Duroc, Large White, Iberian, wild boar and American village pigs.

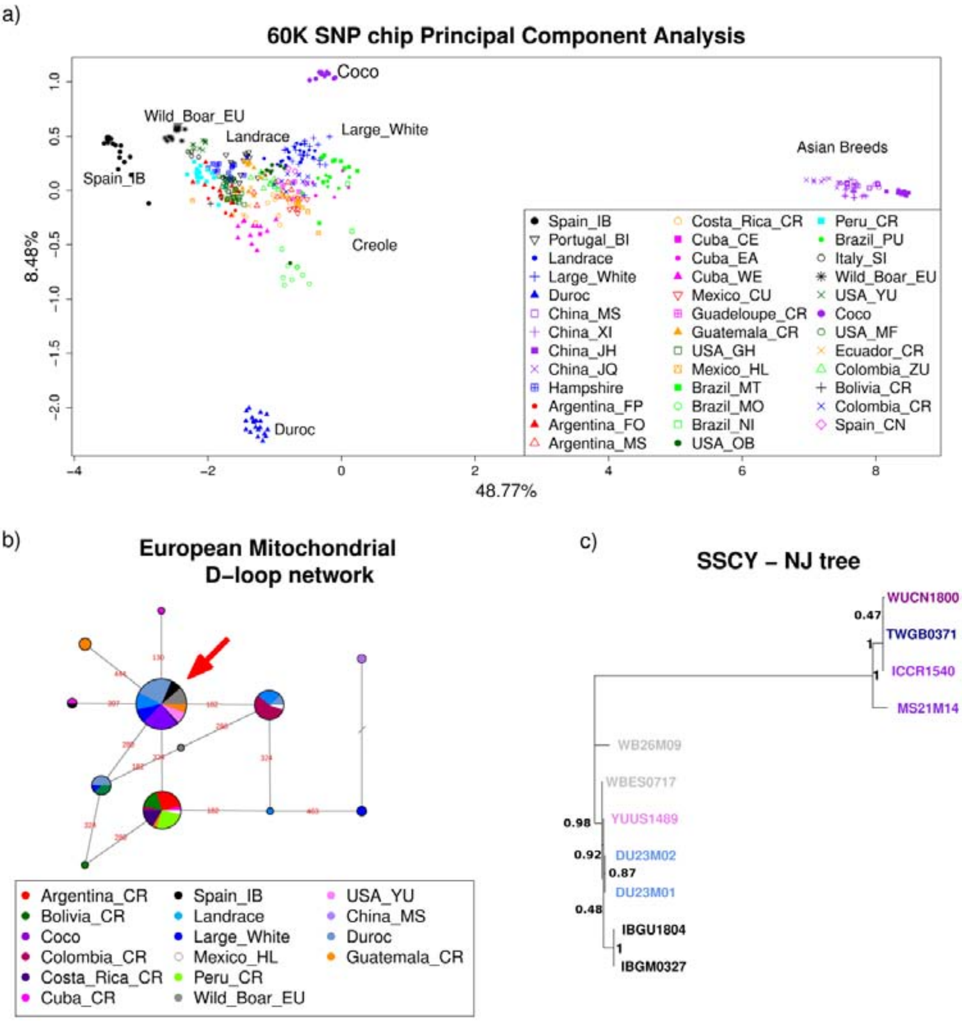

***Figure 5.1**: Representation of genetic distances according to differently inherited loci. (a) Autosomal loci: Principal Component Analysis of the 60k SNP array genotypes from the biodiversity panel. (b) Maternally inherited locus: mitochondrial control-region network of a sample of European haplotypes; all Cocos pigs shared the same haplotype. The arrow indicates the haplotype found in Cocos pigs. For clarity, only European haplotypes are shown, the closest Asian sequence differed by 13 mutations. (c) Paternally inherited locus: Neighbor-Joining tree of the chromosome Y SNPs from the sequenced boars.*

The phylogeography of the porcine Y chromosome is poorly studied. Unfortunately, the 60k panel does not contain any SNP from the Y chromosome. To remedy this, the shotgun sequence data from the 16 individuals were used, which are limited by having considerably fewer samples than those in the 60k biodiversity panel: only 11 boars were sequenced (Table 5.1). Moreover, the current porcine assembly of the Y chromosome is currently limited to unordered BACs that were derived from a Meishan boar (Groenen *et al.* 2012) and cannot be considered as a reliably assembled chromosome; nevertheless, it is useful for traceability purposes. Once heterozygous SNPs were filtered out, we were able to recover 2,491 SNPs in the Y chromosome sequence data. Importantly, and in contrast to the mitochondrial control region, Cocos Y chromosome is clearly of Asian origin (Figure 5.1c). Furthermore, the Cocos SSCY haplotype is close to that of Tamworth, differing in 20 positions, and in 91 and 120 positions with Wuzhishan and Meishan, respectively. In contrast, the average number of differences between Cocos and European Y chromosomes was 955. It should be noted that the Asian signature is remarkably absent from the Y chromosome in all international pig breeds that have been introgressed with Asian germplasm except, precisely, in the Tamworth breed (Ramírez *et al.* 2009). Here, we confirm this observation by using a much larger number of SNPs. The absence of the Y chromosome signature in European breeds had been interpreted as an asymmetrical flow between males and females from Asia into Europe. These new data confirm that both sows and boars were imported from Asia, perhaps in unequal numbers, and that the Asian Y chromosome was lost in the process of breed development later, in the 19th century and onwards. The fact that the Tamworth and Cocos pigs share a very similar Y chromosome may lead to the hypothesis that these two populations are closely related but, as we shall see, this is not the case.

We conclude that the *Isla del Coco* ancestors living in the late 18th century were already hybrids between Chinese and European pigs. Since trading between Spanish and British colonies at that time was greatly limited, and there is no evidence of Spaniards importing Chinese pigs, it is more than likely that the pigs to be released in *Isla del Coco* were onloaded in England, where crosses with Chinese pigs had become quite popular at that time (Arias-Sánchez 1993; Porter 1993; White 2011). Moreover, since Cocos pigs cluster closer to Europe than to Asia (Figure 5.1a), it can also be inferred that they were not an F1 intercross between China and Europe; if this were the case, they would have been positioned equidistantly in the PCA graph of Chinese and European pigs (McVean 2009). Instead, they must be the result of additional generations of backcrossing with English pigs of 'pure' European ancestry. To estimate, even if grossly, the percentage of Asian germplasm in Cocos pigs, we ran a partially supervised ADMIXTURE analysis with K=2 using the 60k SNP data (see Methods). The estimated percentage of the Asian component is ~24% for Cocos feral pigs (Figure 5.2). Note that this percentage can be attained with only one generation of backcross following the F1 between local English and Chinese pigs. We find a similar, albeit somewhat lower, Asian component in Landrace and Large White breeds (20%-22%). This percentage is similar to the estimate obtained by Bosse *et al.* (2014a) using genome shotgun-sequence data (see their Fig. 2). In a previous study using Approximate Bayesian Computation (ABC), Ojeda *et al.* (2011) had estimated the Asian component of international pig breeds was ~30% for a 2-Mb region of chromosome 4, and Fang & Andersson (2006) also found that ~30% of mitochondrial lineages in European pigs were of Asian descent. As for the rest of the pig populations, the Asian component ranged from 10% in several village pigs with strong Iberian influence (Peru, Yucatan) to almost 30% in some Brazilian pigs (Nilo and Monteiro breeds) that may have direct Asian influence (Burgos-Paz *et al.* 2013). The Duroc and Hampshire international pig breeds have a more modest percentage of the Asian component (10%-14%) than have the Large White, equivalent to two generations of backcross with European pigs after the F1, approximately.

The fact that Cocos, Large White and Landrace modern pig breeds show a similar amount of Chinese introgression suggests that either most of the Chinese introgression occurred during the 18th century, i.e., before the Cocos pig feral population was founded, or that management practices remained constant, i.e., an F1 followed by one generation of backcross. Otherwise, the Chinese component would have steadily increased over the centuries and we should have observed a larger Asian component in modern European breeds than in Cocos pigs. Since it is documented that Chinese inflow continued (Porter 1993; White 2011), we argue that the second hypothesis is far more likely than the former, and that management practices resulted in a steady Asian component of British pig breeds.

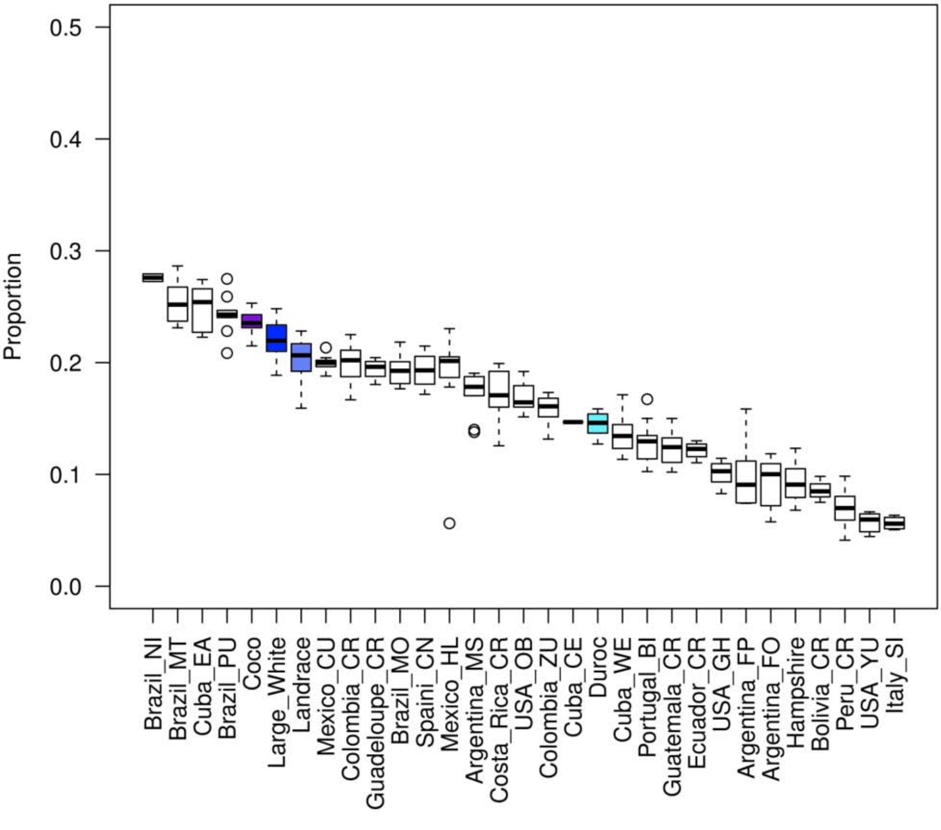

***Figure 5.2**: Percentage of the Asian component as estimated from a partially supervised ADMIXTURE analysis with K=2 (Europe and Asia) from the 60k array genotypes in the worldwide diversity panel. Iberian and European wild boars were assigned cluster K=1, whereas Chinese breeds were assigned cluster K=2. Breed codes are as in Table S1, Supporting Information.*

Cocos pigs show remarkable levels of diversity, yet is compatible with a founder size of N=2.

Next, we employed sequence data to estimate nucleotide diversity, since the 60k data are biased by the SNP ascertainment process. After filtering for coverage and quality, we found 26,391,661 variant sites in the 16 samples analyzed. Of them, 4,909,777 were segregating in Cocos pigs; a similar amount of SNPs was found in Duroc and Large White individuals: 4,946,086 and 5,098,709, respectively. Note, nevertheless, that the raw number of SNPs is a biased estimator of variability because it is influenced by the number of bases aligned and depth, so we estimated nucleotide diversity using the methods in (Ferretti *et al.* 2012). Table 5.1 presents the Watterson's estimates of nucleotide diversity (θ) by individual and by breed. Remarkably, Cocos pigs were as variable (θ_{IC}=1.9 SNPs / kb) as were International domestic pigs, and were more variable than were wild boar (θ_{WB}=1.7) or local European breeds (θ_{IB} = 1.4 and θ_{TW} = 1.8); Large White and Guatemalan Creole were the only non-Asian populations with nucleotide diversities larger than those of Cocos pigs. In contrast to autosomal diversity, Non-Pseudoautosomal Region (NPAR) nucleotide variability was remarkably low in all sows sequenced (Table 5.1), irrespective of their origin. In agreement with previous studies (Esteve-Codina *et al.* 2013; Ma *et al.* 2013; Fernández *et al.* 2014), our data confirm that NPAR exhibits a much lower variability than that expected under neutrality, i.e., 3/4 of the autosomal diversity (θ_A). Here, $\theta_{NPAR} \leq 1/4\ \theta_A$ or about one-third of the expected diversity was observed, either in Asia or in Europe.

Given that the levels of variability in Cocos pigs are as high as in international pig breeds, and were higher than in local European breeds or wild boar, it is pertinent to ask whether just two founder pigs suffice to explain this observation. To respond to this question, a combined set of coalescence and forward simulations was carried out (see Methods). Among the scenarios simulated, that differed in the number of generations of random mating after the cross between Chinese and European pigs, the one fitting the real data best in terms of θ was the model in which an F1 between Asian and 'primigenius' European pigs was followed by a backcross and only one generation of *panmixia* before the bottleneck (Table S2, Figure S2, Supporting Information). Increasing the number of generations between backcross and bottleneck reduced the variability, as did decreasing the percentage of Asian germplasm.

The microsatellite data were also compatible with the hypothesis of two founders for the Cocos population. Out of the 12 microsatellites genotyped, four monomorphic markers (S0090, SW951, S0386 and S0355), four biallelic markers (SW240, S0155, S0101 and SW24) and only one marker (SW936) with four alleles were found. In contrast, mean allele numbers per microsatellite were 4 and 5 in Iberian and Large White, respectively.

Heterogeneous haplotype sharing suggests multiple events of Asian introgression.

It is tempting to hypothesize that Cocos pigs share direct ancestors with modern English breeds, e.g., Large White or Tamworth. The fact that the estimated percentage of the Asian component in Cocos pigs is comparable to that of international pig breeds, and that Tamworth and IC pigs share the Y chromosome would lend initial support to this hypothesis. However, the presence of several Chinese and Siamese pig strains in England by the early 19th century is historically well-accredited (Parkinson 1910, quoted by Porter 1993). Further, an unsupervised ADMIXTURE analysis of the 60k SNP data (Figure S3, Supporting Information) assigned its own cluster to Cocos pigs, distinct from other modern pig breeds such as Landrace, Large White or Hampshire, in agreement with the PCA (Figure 5.1a). Interestingly, part of the Asian component seems to be shared between Landrace and Large White, but not with Cocos pigs. Unfortunately, no Tamworth was genotyped with the 60k array, so we cannot quantify its Asian component. From the analysis of the 60k array data, it is not evident that IC and modern British breeds share the same ancestors.

We wished to further investigate whether there existed a common Asian footprint, shared between the Cocos pigs and modern breeds. To do so, we quantified how many genome segments were shared across breeds from the haplotypes as reconstructed with Beagle4. In this case, Asian introgression homogeneity between Cocos and another breed 'A' can be analyzed by computing the probabilities of a segment tract of Breed A being IBD with Meishan, given that the corresponding segment is also IBD with a Cocos pig (see Methods). These probabilities therefore measure the similarity of genetic histories across breeds. First, note that $P(MS1 \equiv MS2 \mid IC2 \equiv MS2)$, where MS1 and MS2 are the two Meishan pigs sequenced, and $P(IB1 \equiv IB2 \mid IC2 \equiv IB2)$, IB1 and IB2 being the two Iberian pigs sequenced, are inversely proportional to variability: they quantify how likely is that two MS or IB haplotypes are IBD. Unsurprisingly, this probability is much higher in Iberian than in Meishan breeds, 0.90 and 0.51, respectively (Table 5.2). Equivalently, for A=IC, these probabilities measure how similar are the Cocos haplotypes of either Meishan or Iberian origin. In this case, these probabilities are a function of both the diversity within IB or MS and of Cocos recent demographic history. This probability was very high for Iberian haplotypes, $P(IC1 \equiv IB \mid IC2 \equiv IB) = 0.87$, showing that the Iberian component of Cocos pigs is more homogeneous than that of the Meishan component $P(IC1 \equiv MS \mid IC2 \equiv MS)=0.61$. Conditional IBD sharing between MS and breeds that underwent Asian introgression (LW, TW and DU) is similar to that in breeds without an Asian component, i.e., Iberian, $P(LW \equiv MS \mid IC \equiv MS) \sim P(DU \equiv MS \mid IC \equiv MS) \sim P(TW \equiv MS \mid IC \equiv MS) \sim P(IB \equiv MS \mid IC \equiv MS) = 0.15$. Note that $P(IB \equiv MS \mid IC \equiv MS)$ can be interpreted as a baseline IBD sharing due to haplotypes that have remained IBD between distant breeds such as IB and MS, perhaps because of their very low polymorphism, and not because of direct Chinese introgression (since Iberian pigs were not crossed to Asian pigs). Therefore, these data suggest that, within a given region, there is not an increased probability of Asian origin in Large White, Duroc or Tamworth when the genome is of Meishan origin in Cocos pigs. In other words, the Asian haplotypes seem to have segregated independently in each of these breeds. These results, in agreement with historical records (Porter 1993), suggest that several independent introgression events from Asia made up the genomes of modern pig breeds, events that are uncorrelated with those intervening in the founding of the *Isla del Coco* feral population.

It is also interesting to investigate the patterns of IBD sharing between the Cocos pigs and the rest of the sequenced individuals (Figure 5.3). Not unexpectedly, maximum IBD sharing was between the two Cocos pigs, but this is due to a larger number of IBD segments rather than to an increased length in each IBD tract. Except for Asia, the rest of pigs analyzed shared a similar total IBD length with Cocos pigs, be it wild boar, Iberian or other European breeds.

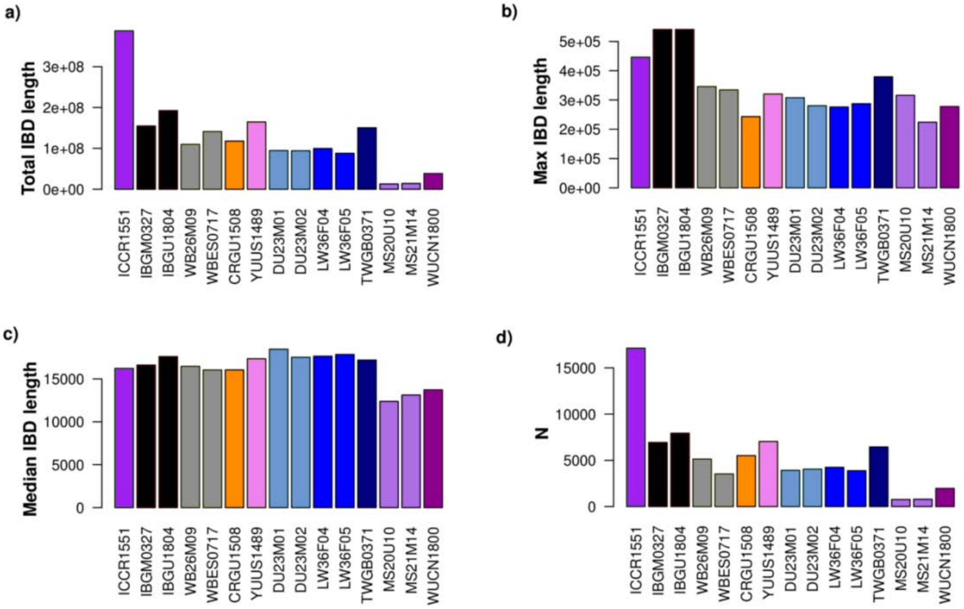

Figure 5.3*: Distribution features of IBD patterns between the Cocos boar (ICCR1540) and the rest of the samples sequenced. (a) Total length of IBD tracts shared between samples. (b) Length of longest IBD segment. (c) Median length of IBD segments. (d) Number of shared IBD segments. Lengths are in base-pair units, sample names are as in Table 5.1.*

The median haplotype length was quite similar across samples (~ 17 kb), except with Asian pigs. Fewer and shorter IBD tracts between Asian and European breeds than between European breeds were also found by Bosse *et al.*(2014b),and is the expected outcome of the large number of recombinations that have occurred since European and Asian clades diverged, ca. 1.2 MYA. Although average IBD sharing between Cocos and other European populations was similar, it was greater with Iberian pigs (173 Mb) than with Large White (94

Mb) and Tamworth (150 Mb) British pig breeds. This correlates with a smaller Fst between Cocos and Iberian pigs ($Fst_{IC\text{-}IB}$= 0.28) than with Large White ($Fst_{IC\text{-}LW}$= 0.32) or with Tamworth ($Fst_{IC\text{-}TW}$= 0.34). The observation that Cocos pigs are actually closer genetically to Iberian than to modern British breeds could be explained if the ancestors of Cocos pigs were crossed with Iberian pigs after departing from England, or if English ancestors of Cocos pigs were closer genetically to Iberian pigs than are, say, to modern Large White. Given historical accounts (Arias-Sánchez 1993), the former hypothesis is unlikely, whereas the second possibility may have happened if multiple, distinct events of Asian introgression occurred in the ancestors of Cocos pigs and of Large White. As we have argued, this latter hypothesis is supported by the data at hand. Nevertheless, more sequence data, especially from Chinese and South East Asian pigs, are needed to fully characterize the full history.

Extreme ROH

Runs of homozygosity (ROHs) provide insight on demographic history and correlate with genome features such as recombination rate (Bosse *et al.* 2012). The longest ROH identified in Cocos pigs was in SSC8: 44,705,282-72,057,361, i.e., a 28-Mb region. This long region is also identifiable from the 60k genotyping data (marked with an arrow in Figure S4, Supporting Information). Its sequence based nucleotide diversity was θ =0.43 per kb, or about five times lower than the genome-wide autosomal level, θ =1.9 (Table 5.1). Although the region contains or is near the centromere, the recombination rate is not completely suppressed (0.17 cM/Mb compared to a genome-wide rate of 0.69 cM/Mb, Tortereau *et al.* 2012). Interestingly, Cocos pigs seem to be enriched in the Asian component for this region, opposite to what is observed for the international pig breeds Landrace and Large White (Figure 5.4). Some Duroc pigs share a similar haplotype to Cocos, but they are scattered in the PCA. Nevertheless, recombination events have occurred since Asia – Europe admixing. This can be inferred from the erratic pattern in 1-IBS distances between Cocos and either Iberian or Wuzhishan pigs (Figure 5.4b).

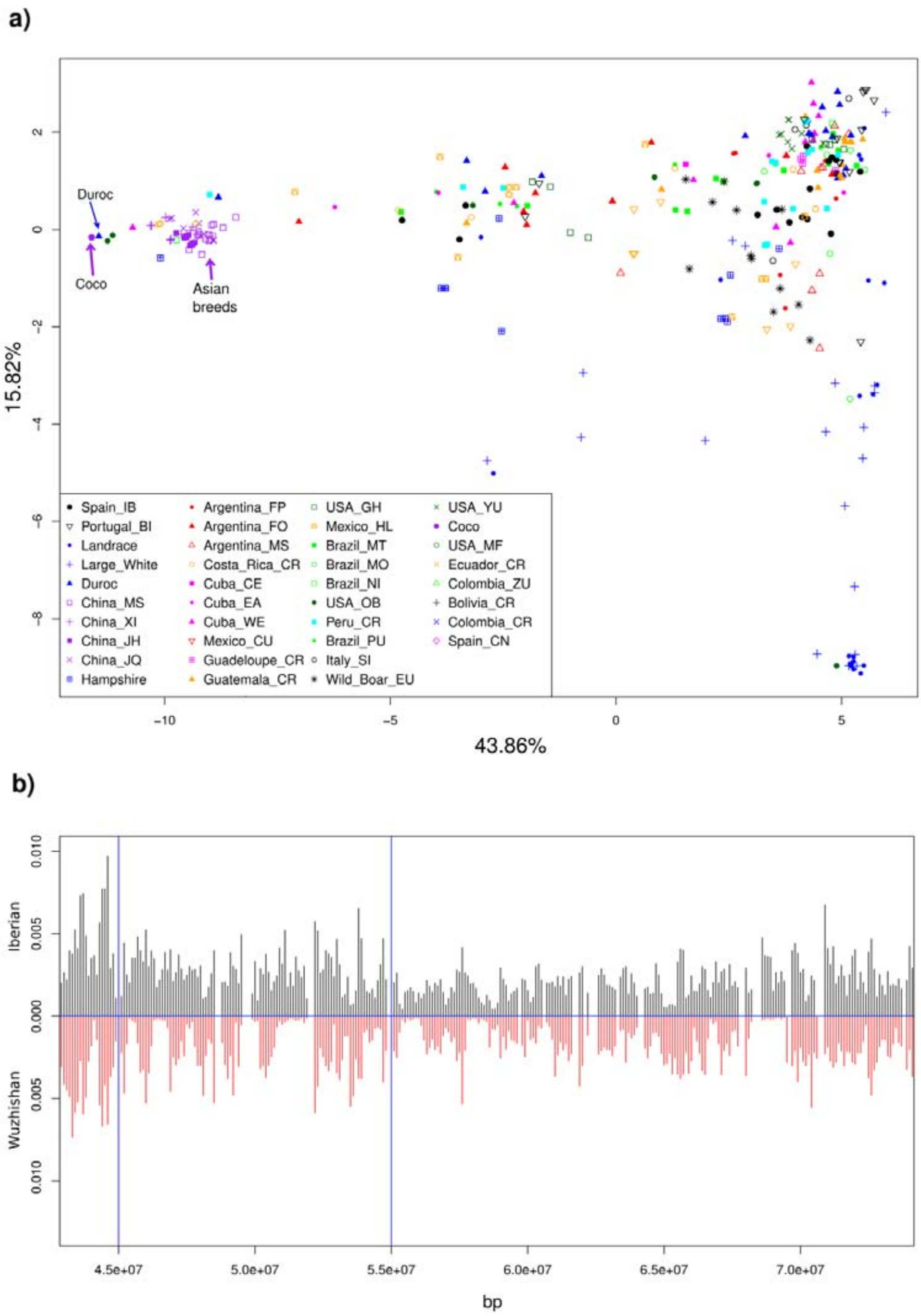

***Figure 5.4**: Features of the 28-Mb ROH. (a) Principal Component Analysis of the region using the genotypes of the 60k array from the biodiversity panel.*

(b) Distribution of allele-sharing distances (1-IBS) between Cocos and Iberian (top) and between Cocos and Wuzhishan (bottom) pigs in 100-kb windows along the region. The shorter the bars, the shorter the genetic distance. Note that, for most windows, Cocos pigs are genetically closer to Asian pigs than to European pigs, but that the erratic pattern suggests the presence of numerous recombination events. The blue vertical bars limit the lowest-diversity 10-Mb window (SSC8:45,395,481-55,290,347).

Further inspection showed that this 28-Mb region also showed a decreased variability in the rest of the samples analyzed, and that variability was lowest in a 10-Mb segment (SSC8:45,395,481-55,290,347) across most samples (Table 5.1). This suggests that a putative selective sweep, if responsible for the decreased diversity, should predate breed divergence or has repeatedly occurred during pig evolution. The 10-Mb region contains 47 annotated genes (Table S3, Supporting Information). Figure 5.5 presents standardized nucleotide diversities "z" for each gene and population, where the dashed lines show the lowest possible diversity (θ=0, i.e., equivalent to z=-μ/σ) for Meishan and Iberian, the most and least diverse populations, respectively. Note that the two Duroc and Large White international pig breeds are mostly devoid of any SNP throughout all genes in the region (Table S3, Supporting Information). The first four genes (*CPE, H3F3C, SC4MOL and KHL2*) are located within one of the regions identified by Rubin *et al.* (2012, their Table S1, Supporting Information) as extremely differentiated between wild boar and domestic pigs, and being putative domestication sweeps. In agreement with their results, we also observed that variability for those genes is not below the mean in wild boar (Figure 5.5a). In contrast, the region containing genes *PDGFC, U12, GLRB, SNORA11, GRIA2* and ENSSSCG00000023923 (orthologous to a miRNA) showed almost no variability in any sample, including wild boar, with only 218 SNPs in genic regions among all individuals. Considering as outliers those SNPs with a Bayes factor greater than 10, Bayescan detected three distinct outlier clusters with a negative selective coefficient α (b); the clusters span approximately 46,130-46,132 kb, 48,635-48,645 kb and 50,655-50,755 kb.

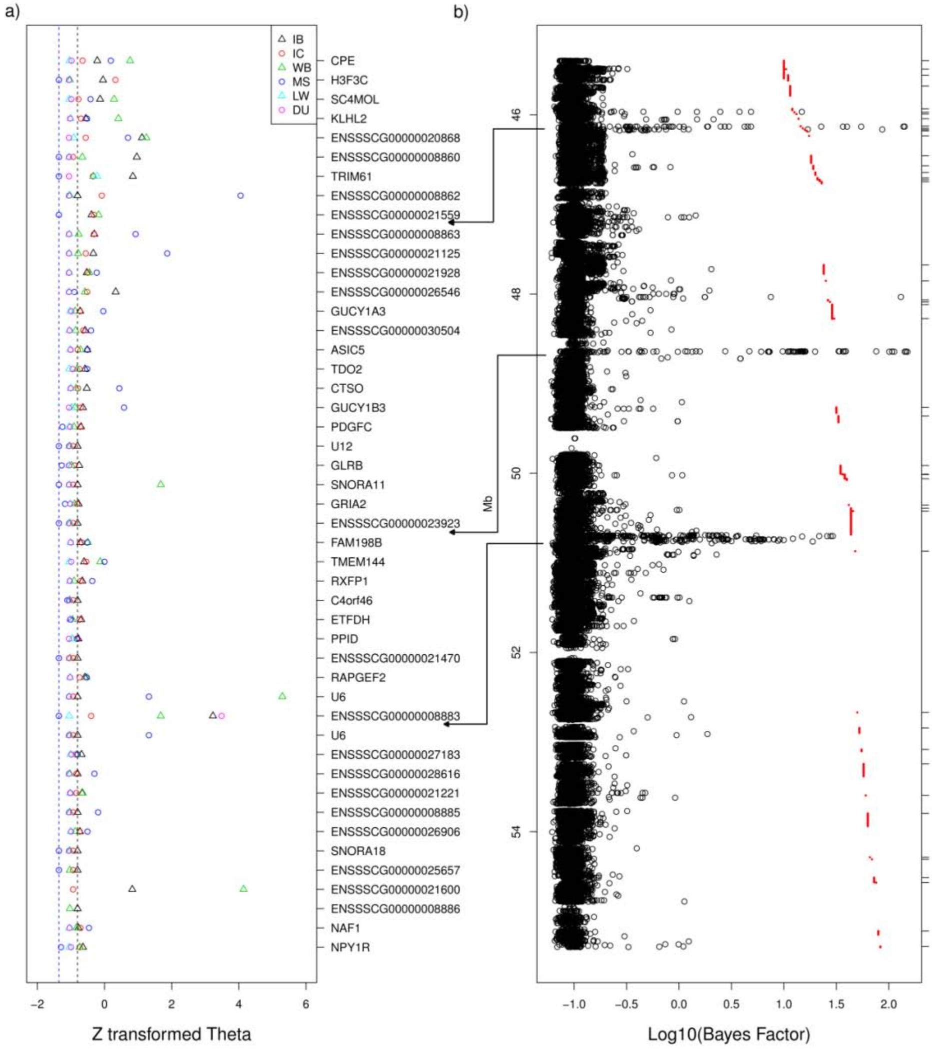

***Figure 5.5**: (a) Scaled variability per gene and breed; each dot corresponds to the scaled diversity in a gene and breed. The dashed lines correspond to the minimum possible diversity (θ=0) in the Iberian (lower line) and Meishan (upper line) breeds. Breed codes are as in Table 5.1. Arrows connect positions of putative selective events with nearest genes. (b) Bayescan results of the 10-Mb low-variability region (SSC8:45,395,481-55,290,347), red bars represent the 47 annotated genes; each dot corresponds to a single SNP.*

A negative α can be caused by balancing or purifying selection, but strong purifying selection is a plausible explanation here, given the low overall variability of the region. Although no genes are annotated within these three regions, interestingly, the second cluster is next to the region described above (encompassing genes *PGDFC* to ENSSSCG00000023923), which exhibits a negligible genic variability.

General Discussion

Herein we provide a comprehensive genomic analysis of *Isla del Coco* feral pigs, a unique population that has remained isolated for over 200 years and was likely founded by a single sow and boar. Our work conclusively shows that, as early as the late 18th century, British vessels were loading crossbred pigs in Great Britain and transporting them overseas. Cocos pigs are biological relics of these events and can serve as reference points to study the evolution of highly selected, modern pig populations versus their putative ancestors, when Asian admixture was beginning. Asian introgression left a dramatic and enduring footprint in the genomic makeup of local European pigs (Bosse *et al.* 2014b). We have shown that this flow was not incremental despite the continuous importing of foreign pigs; instead, we find a rather constant percentage of Asian germplasm in either Cocos pigs or other admixed breeds like Large White or Tamworth. A search of an ancestral, homogeneous Asian signature shared across populations was unsuccessful though. This agrees with previous works showing that Fst differentiation signals between modern pig breeds and wild boars are mainly breed-specific (e.g., Amaral *et al.* 2011). Our observations, instead, point to several independent waves of Asian pig import followed by standardized management practices that essentially wiped out the non-crossbred local pigs that originally lived in the UK (and possibly other European countries).

Each genome region in a living individual tells different demographic stories about the past of its population. Since Cocos pigs are hybrids, some of these stories can be dramatically different, e.g., Cocos pigs harbor a mitochondrial genome of European origin and an Asian Y chromosome. So far, the only pig population with a reported Asian Y chromosome outside Asia was Tamworth (Ramírez *et al.* 2009). Despite sharing the Y chromosome with Tamworth, however, Cocos pigs were not more genetically related to Tamworth pigs than they were to Large White. This is additional, indirect, evidence that Asian haplotypes are uncorrelated across breeds. In fact, the closest population to Cocos pigs was the Iberian breed, a local Spanish breed that has not been introgressed with Chinese pigs. Ruling out the crossing of Spanish and British pigs in the 17th century, a likely explanation is that native British pigs, before introgression, were more closely related to modern Iberian pigs than they are to current international pig breeds. Ancient DNA studies will settle this matter, but some facts support this hypothesis, among them the low variability that has been found in European wild boar and Iberian pigs and the low differentiation between 16th century pigs, Iberian pigs and European wild boar (Ramírez *et al.* 2015).

Feral animals, despite strong founder effects, can still be important reservoirs of DNA diversity. The levels of variability observed in Cocos pigs were similar to those in international pig breeds, and higher than in European wild boar or local Iberian pigs. The reason for this is their admixed nature and likely short duration of the bottleneck that followed the cross between Asian and European pigs. Symmetrically, it can be argued that current international pig breeds harbor a nucleotide diversity that could be explained by just two hypothetical founders two centuries ago. Interestingly, simulations suggest two founders suffice to explain the observed variability, provided the Cocos population was founded immediately after introgression from Asia into Europe, i.e., that not many generations of recombination occurred in hybrid pigs before the founding bottleneck.

Finally, Cocos pigs are also excellent models to study the dynamics of feral events in an admixed genome. The differential increase or decrease of Asian haplotype frequencies vs. those in international pig breeds, and the detection of long runs of homozygosity, can be promising approaches to unravel adaptation signals to Coco Island extreme environmental conditions. A 28-Mb segment located in the centromere region of SSC8 was pinpointed as being the longest ROH in Cocos pigs; overall, this region showed about 20% the variability observed in the average genome. Interestingly, a nested 10-Mb region (Table 5.1) exhibited a marked decrease in variability across all samples sequenced, a two-fold reduction in wild boar, four-fold in Cocos pigs and over ten-fold in Large White or Tamworth. This region contains 47 annotated genes, including four genes (*CPE, H3F3C, SC4MOL and KHL2)* previously identified as highly differentiated between wild boar and domestic pigs (Rubin *et al.* 2012).

Acknowledgements

We would like to thank the National Parks System (Sistema National, Areas de Conservación, MINAE) in Costa Rica for providing the samples, A. Arias for useful comments on the history of *Isla del Coco*, and C. Simmons, native English-speaking University Instructor of English, for proofreading the manuscript. EB is the recipient of a PhD fellowship from the Ministry of Science (Spain). This work was funded by AGL2010-14822 and AGL2013-41834-R grants (Ministry of Economy and Science, Spain) to MPE.

References

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

Alves E, Ovilo C, Rodriguez MC, Silio L (2003) Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Animal Genetics*, **34**, 319–324.

Amaral AJ, Ferretti L, Megens H-J *et al.* (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PloS One*, **6**, e14782.

Arias-Sánchez R (1993) La isla del Coco: perspectiva histórica y análisis de una leyenda. Universidad De Costa Rica.

Bandelt HJ, Macaulay V, Richards M (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Molecular phylogenetics and evolution*, **16**, 8–28.

Bianco E, Nevado B, Ramos-Onsins SE, Perez-Enciso M (2015) A deep catalog of autosomal single nucleotide variation in the pig. PLoS One (in press).

Bosse M, Megens H-J, Frantz LF *et al.* (2014a) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature communications*, **5**, 4392.

Bosse M, Megens H-J, Madsen O *et al.* (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS genetics*, 8, e1003100.

Bosse M, Megens H-J, Madsen O *et al.* (2014b) Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. *Molecular ecology*, **23**, 4089–4102.

Browning BL, Browning SR (2013a) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**, 459–471.

Browning BL, Browning SR (2013b) Detecting identity by descent and estimating genotype error rates in sequence data. *American journal of human genetics*, **93**, 840–851.

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

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. *Genome research*, **19**, 136–142.

Choquenot D, Mcilr J, Korn T (1996) *Managing Vertebrate Pests: Feral Pigs*. Australian Government Publishing Service, Canberra.

Crosby AW (2003) *The Columbian Exchange*. Greenwood Publishing Group, London.

Cruz F, Josh Donlan C, Campbell K, Carrion V (2005) Conservation action in the Galápagos: feral pig (Sus scrofa) eradication from Santiago Island. *Biological Conservation*, **121**, 473–478.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**, 1792–1797.

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.

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 Series B*, **273**, 1803–1810.

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 genetics*, **5**, e1000341.

Fernández AI, Alves E, Óvilo C, Rodríguez MC, Silió L (2011) Divergence time estimates of East Asian and European pigs based on multiple near complete mitochondrial DNA sequences. *Animal genetics*, **42**, 86–88.

Fernández AI, Muñoz M, Alves E (2014) Recombination of the porcine X chromosome: a high density linkage map. *BMC Genetics*, **15**.

Ferretti L, Raineri E, Ramos-Onsins S (2012) Neutrality tests for sequences with missing data. *Genetics*, **191**, 1397–1401.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Gascoigne J (2007) *Captain Cook*. Hambledon Continuum, London.

Government of Costa Rica (1996) *Coco's Island marine and terrestrial conservation area: Nomination for inclusion in the World Heritage List of natural properties.* Government of Costa Rica, San José, Costa Rica.

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

Higman BW (2010) *A Concise History of the Caribbean*. Cambridge University Press, Cambridge.

Huson DH, Scornavacca C (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, **61**, 1061–1067.

Lamichhaney S, Berglund J, Almén MS *et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, **518**, 371–375.

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 *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Linderholm A, Larson G (2013) The role of humans in facilitating and sustaining coat colour variation in domestic animals. *Seminars in cell & developmental biology*, **24**, 587–593.

Ma J, Gilbert H, Iannuccelli N *et al.* (2013) Fine mapping of fatness QTL on porcine chromosome X and analyses of three positional candidate genes. *BMC genetics*, 14, 46.

McKenna A, Hanna M, Banks E et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303.

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

Messer PW (2013) SLiM: simulating evolution with selection and linkage. *Genetics*, **194**, 1037–1039.

Nevado B, Perez-Enciso M (2014) Pipeliner: software to evaluate the performance of bioinformatics pipelines for Next Generation re-Sequencing. *Molecular ecology resources*, **15**, 99–106.

Nevado B, Ramos-Onsins S, Perez-Enciso M (2014) Re-sequencing studies of non-model organisms using closely-related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, **23**, 1764–1779.

Ojeda A, Ramos-Onsins SE, Marletta D *et al.* (2011) Evolutionary study of a potential selection target region in the pig. *Heredity*, **106**, 330–338.

Oliver WLR, Brisbin IL (1993) Introduced and Feral Pigs, Problems, Policy and Priority. In: *Pigs, Peccaries and Hippos* , pp. 179–191. IUCN, Gland, Switzerland.

Porter V (1993) *PIGS - A handbook to the breeds of the world.* HELM Information Ltd.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**, 904–909.

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.

Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Ramírez O, Burgos-Paz W, Casas E *et al.* (2015) Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity*, **114**, 175-184.

Ramírez O, Ojeda A, Tomàs 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.

Ramos AM, Crooijmans RPM, Affara NA *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.

Rodero A, Delgado JV, Rodero E (1992) Primitive andalusian livestock and their implications in the discovery of america. *Archivos de Zootecnia*, **41**, 383–400.

Roemer GW, Donlan CJ, Courchamp F (2002) Golden eagles, feral pigs, and insular carnivores: how exotic species turn native predators into prey. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 791–796.

Rubin C-J, Megens H-J, Martinez Barrio A *et al.* (2012) Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19529–19536.

Sierra C (2001) El cerdo cimarrón (Sus scrofa, Suidae) en la Isla del Coco, Costa Rica: Escarbaduras, alteraciones al suelo y erosión. *Revista de Biología Tropical*, **49**, 1158–1170.

Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

Tortereau F, Servin B, Frantz L *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, **13**, 586.

White S (2011) From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. *Environmental History*, **16**, 94–120.

Zadik BJ (2005) *The Iberian Pig in Spain and the Americas at the Time of Columbus*. University of California, Berkley.

Data Accessibility

DNA sequences: SRS accessions SRS869969, SRS869970, SRS869971, SRS869976, SRS875335.

High density genotyping data, control region sequences, microsatellite genotypes: Dryad doi:10.5061/dryad.p610q.

Author contributions

EB wrote software and analyzed data; MPE, HWS and LV contributed new reagents and analytical tools; MPE designed research and wrote the paper with contributions from the rest of the authors.

Supporting information

Supporting information can be find at
http://onlinelibrary.wiley.com/doi/10.1111/mec.13182/suppinfo

Table 5.1: Samples sequenced and main statistics.

Breed	Breed code	Sample code	Sex	Country	Average depth	Individual autosomal θ (SD)[a]	Individual NPAR θ (SD)[b]	Breed θ[c]	SSC8: 44-72Mb θ[d]	SSC8: 45-55 Mb θ[e]	SRA Accession
Cocos	IC	ICCR1540	M	CR	12.7	1.82 (2.57)	NA	1.91	0.43	0.43	SRR1873279
		ICCR1551	F	CR	8.7	1.82 (2.45)	0.41 (1.20)	-	-	-	SRR1873280
Creole	CR	CRGU1508	F	GU	12.0	2.23 (2.70)	0.36 (1.32)	2.21	0.34	0.13	SRR1513309
Yucatan	YU	YUUS1489	M	USA	14.1	1.78 (2.66)	NA	1.78	0.21	0.37	SRR1873293
Wild Boar	WB	WB26M09	M	CH	14.4	1.09 (1.65)	NA	1.70	0.72	0.91	ERX149181
		WBES0717	M	ES	13.0	1.48 (2.15)	NA	-	-	-	SRR1513306
Iberian	IB	IBGM0327	M	ES	13.0	1.16 (1.98)	NA	1.38	0.56	0.44	SRR1513307
		IBGU1804	M	ES	14.5	1.02 (1.98)	NA	-	-	-	SRR1917381
Duroc	DU	DU23M01	M	Intl.	10.7	1.83 (2.24)	NA	1.90	1.28	1.24	ERX149133
		DU23M02	M	Intl.	11.6	1.84 (2.24)	NA	-	-	-	ERX149134
Large White	LW	LW36F04	F	Intl.	9.5	1.92 (2.38)	0.36 (0.78)	1.95	0.37	0.13	ERX149159
		LW36F05	F	Intl.	8.6	1.77 (2.34)	0.36 (0.82)	-	-	-	ERX149160
Tamworth	TW	TWGB0371	M	UK	13.3	1.76 (2.66)	NA	1.76	0.11	0.08	SRR1873281 SRR1873282
Meishan	MS	MS21M14	M	CN	10.1	2.36 (2.39)	NA	2.61	1.04	1.33	ERX149165
		MS20U10	F	CN	9.2	2.55 (2.34)	0.33 (0.78)	-	-	-	ERX149162
Wuzhishan	WU	WUCN1800	M	CN	22.2	1.82 (2.78)	NA	1.82	0.15	0.12	SRA051254

[a] Autosomal genome-wide nucleotide diversity per kb and standard deviation (SD) over windows of 100-kb.
[b] NPAR: Non pseudoautosomal region.
[c] Autosomal breed diversity per kb.
[d] Breed diversity of chromosome 8 region 44.7-72 Mb.
[e] Breed diversity of chromosome 8 region 45-55 Mb.

Table 5.2: Conditional bivariate probabilities of segment IBD as inferred with Beagle4.

Population A	**P(A≡MS | IC≡MS)a**	**P(A≡IB | IC≡IB)a**
Meishan	0.51b	0.01
Large White	0.14	0.37
Duroc	0.13	0.41
Tamworth	0.13	0.46
Iberian	0.15	0.90c
Cocos	0.61d	0.87d

[a] The probabilities P(A≡B | IC≡B) were computed from assessing the length of haplotypes identified IBD by Beagle4 that were shared simultaneously between individuals from breed A and B, and between Cocos and Meishan, divided the length of haplotypes IBD shared between Cocos and breed B, either Meishan or Iberian.
[b] P(MS1≡MS2 | IC2≡MS2) refers to the probability of being IBD between the two
Meishan samples MS1 and MS2.
[c] P(IB1≡IB2 | IC2≡IB2) refers to the probability of being IBD between the two Iberian samples IB1 and IB2.
[d] P(IC1≡MS | IC2≡MS) and P(IC1≡IB | IC2≡IB) refer to the probability of being IBD for each of two Cocos samples IC1 and IC2.

CHAPTER VI

General discussion

6. General discussion

Since 1999, when the first complete mitochondrial genome of a pig was sequenced (Lin *et al.* 1999), dramatic advances in sequencing technologies and in pig genomics knowledge have occurred. The sources available to study pig genetics and genomics has vastly increased in the last 15 years. The 60k Porcine SNP chip of Illumina was published in 2009 (Ramos *et al.*), and a new Affimetrix 650k high density SNP chip will be soon available (1). Meanwhile, pig genomic knowledge increased due to NGS technologies and in 2012, Groenen *et al.* published the pig reference sequence (Sscrofa 10.2). In this thesis we used NGS methods to explore worldwide and genomewide variability in *Sus scrofa* (Chapter 3, Bianco *et al.* 2015a). We then used the information we obtained from the catalog of variants to explore the joint demography of Eurasian wild boars (Chapter 4, Bianco *et al.*, in prep.). In parallel, we used NGS data to evaluate the effects of feralization on the genome in a domestic breed, which is isolated since 1793 in Cocos Island (Chapter 5, Bianco *et al.* 2015b).

6.1. A database of pig variants across the globe: a resource for future studies

Next generation sequencing methods and the reduced cost of genome resequencing in the last decade, together with the availability of a reference genome, increased the number of studies based on individual resequencing. In the last 3 years, since the new version of pig reference genome Sscrofa10.2 is available (Groenen *et al.* 2012), more than 270 whole genome pig sequences have been published (Fang *et al.* 2012; Groenen *et al.* 2012; Li *et al.* 2013, 2014; Esteve-Codina *et al.* 2013; Ramírez *et al.* 2014; Molnár *et al.* 2014; Ai *et al.* 2015; Bianco *et al.* 2015b; Kim *et al.* 2015; Moon *et al.* 2015)*. At the beginning of 2014, the sequence of 100 domestic pigs and wild boars were available, either mapped against the reference genome (Groenen *et al.* 2012; Ramírez *et al.* 2014) or used for a *de novo* assembly (Fang *et al.* 2012; Li *et al.*

(1) http://www.animalgenome.org/pig/newsletter/No.121.html

* last update: August, the 10th 2015.

2013). All these sequences were publicly available, but studied and analyzed separately because of the different samples origins and different studies. These individuals belonged to 20 different domestic breeds and wild boars of both Asian and European origin and, for the first time, it was then possible to have an overview of the genomewide variability across the globe of the pig species. In addition, five individuals of other species had also been sequenced in the pig reference genome project (Groenen *et al.* 2012), but no consensus ancestral allele was determined. In Chapter 3 (Bianco *et al.* 2015a) we generated a database of variants using the more than 120 whole genome sequences available at the time. We generated a genomewide worldwide comprehensive catalog of SNPs to evaluate the number of SNPs present worldwide and genomewide, and the number of SNPs exclusive per groups of populations (divided by continent and domestication state).

This catalog is a useful tool for many purposes. Even with the shallow coverage of some of the animals used in our study, we managed to recover ~48M SNPs. The number of variants we found is higher than the ~30M SNPs found by Kim *et al.* (2015) analyzing 70 individuals at high coverage and comparable with the ~40M SNPs found by Ai et al. (2015) who also used 70 individuals at a high coverage, even if these individuals were mapped against Wuzhishan reference genome (Fang *et al.* 2012) instead of Sscrofa10.2 (Groenen *et al.* 2012). We did not recover all the variability present in the pig species, because of the low coverage of some individuals and the absence of many breeds, but still we increased of ~40% the amount of variants present in public database (compared with dbSNP *build* 140) and, together with the following up works, we are confirming that pig is a very variable species. The availability of an increasing number of individuals of different breeds and population sequenced at high coverage will allow the discovery of more SNPs: a new analysis of 288 of the individual resequenced resulted in ~75M SNPs (J. Leno-Colorado and M. Pérez-Enciso pers. Comm.). The number of SNPs of pig species is a resource in continuous development and it provides raw material to facilitate further applications.

The annotation features we described and report in chapter 3 are another tool for future investigations. Pig annotation is still improving. In chapter 3, all variants found were annotated against the version 76 of the Variant effect Predictor pipeline (McLaren *et al.* 2010), and in July 2015 the version 81 was published with no changes reported in *S. scrofa* gene annotation, so that our annotation statistics are still the most updated and useful in a genomewide and worldwide frame.

Finally, chapter 3 also provides the list of ancestral alleles for ~39M SNPs. Determine which allele is ancestral, and which is derived, is fundamental and required for many statistical test for selection in evolutionary studies, in order to understand the direction of the mutation. In summary, the catalog provided in chapter 3 will be a useful tool in future studies on pig selection and evolution.

Mutational bias and recombination rate explain the variation in the number of polymorphism across the genome

One of the applications of a catalog of variants is the possibility to extrapolate genomic statistics and to evaluate genomewide, in a domestic species, what is known in other species or from theoretical biology. For instance, transition/transversion rate (Ts/Tv) was not known for *Sus scrofa*. In general, Ts/Tv is about 2 because of the molecular mechanisms that regulate the two types of mutation. In our study we confirm the average of pig genome is ~2, but we also found a correlation with the number of variants, which are more abundant towards telomeres (Bianco *et al.* 2015a). A similar pattern was found in the count of CpG along the chromosomes (Figure 3.3). The correlation between missing data and the number of SNPs was very small, only 4% of SNPs variability along the genome was explained, which support the good quality of our dataset, also evaluated through simulations (false discovery rate ~1%). The most striking correlation we found was between Ts/Tv and CpG count. The number of SNPs was highly explained by Ts/Tv and both the number of SNPs and Ts/Tv were explained by the variance in recombination rate. Our analysis suggests genome variability is the consequence of recombination and the higher mutability of CpG sites. The mutational effect of recombination has not yet been proved and further analyses are needed to demonstrate recombination cause mutation, but still, we found that regions with higher recombination rate are also those one with a higher number of variants along the genome.

6.2. Using polymorphisms to study the complex wild boar demography

Humans shaped domestic breeds' variability capitalizing on variability already present in wild boars, a result in turn of wild boar own demographic history. The demography of wild boars is complex. *Sus scrofa* originated in South East Asia 5 - 3 MYA and from there it spread throughout the Eurasian continent (Frantz *et al.* 2014). The following split between European and Asian wild boars occurred 1.2 - 0.8 MYA (Giuffra *et al.* 2000; Groenen *et al.* 2012; Frantz *et al.* 2014). A second split within Asian wild boar occurred ~0.6 MYA, between North and South China individuals (Frantz *et al.* 2014). With the publication of pig reference sequence, Groenen *et al.* (2012) evaluated the demographic history of European and Asian wild boars with the PSMC method. This is up to date the most comprehensive study of wild boar demographic history using whole genome sequence data. One of the main pitfalls of the PSMC method is that each population is analyzed separately and it uses a single sequence to recover the history of a population.

In chapter 4 we used the data obtained from the analysis of chapter 3 to try disentangling the joint demography of European and Asian wild boars, in order to evaluate not only fluctuation in population effective size but also migrations within Eurasia. Not all the wild boars samples used to create the catalog of SNPs were included in the demography analysis. All the Tibetan wild boars from Li *et al.* (2013), because of the shallow coverage and the unclear state as wild or domestic, and the Japanese wild boar from Groenen *et al.* (2012), because of the distance between Japanese wild boars from the other Asian individuals, were excluded. Despite the reduced number of samples, the general shape of the spectrum did not change (Figure 6.1). In both cases the spectrum was dominated by extreme frequency SNPs classes, mostly in the

European population, and had a reduced number of SNPs in the classes at intermediate frequency in European individuals and at high frequency in Asian population. This means that the reduced number of samples approximately recovered the joint variability of Eurasian wild boars.

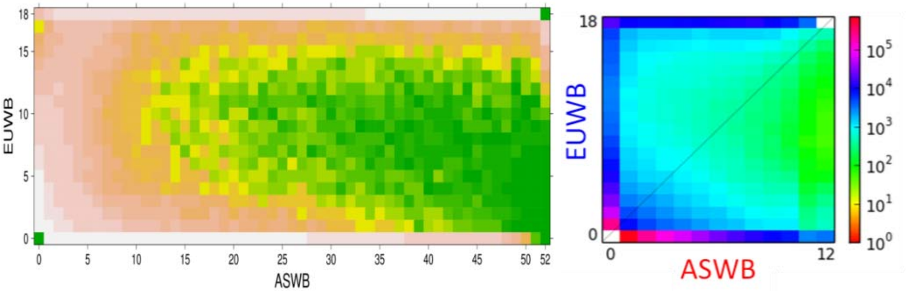

***Figure 6.1**: joint Site Frequency Spectrum of Eurasian wild boars. On the right: spectrum from chapter 3. On the left: spectrum from chapter 4. Reducing the number of Asian individuals did not change the general aspect of the spectrum.*

Drawbacks of using a reduced number of samples at medium-low coverage

Even if the spectrum was approximately recovered by removing outliers and samples of uncertain classification (*e.g.*, Tibetan wild boars), we did not resolve the convergence issue when try to infer the parameters of possible demographic models. The possible causes of this convergence issue can be that the likelihood was flat and the algorithm fails to converge because of multiple local maxima, that means different demographic history resulted in the same spectrum (Myers *et al.* 2008); or that the models tested are not realistic.

∂a∂i's algorithm allows to summarize genomewide SNPs information from up to 3 populations into the site frequency spectrum and uses it to infer demography in a reasonable amount of CPU time, and small memory usage (Gutenkunst *et al.* 2009). A side issue is that this algorithm has convergence problems so that the inference must be repeated various times in order to find the best fit estimates for a given demographic model (Gutenkunst *et al.* 2009; Excoffier *et al.* 2013). In other studies, the authors considered as plausible those parameters to which the software converged at least 3 times in 200 runs,

increasing the number of runs if the convergence did not occurred at least 3 times (Huber *et al.* 2014), whereas Excoffier *et al.* (2013) when trying to compare the performances of their own software to *∂a∂*i, they have to exclude some runs because the resulted demographic parameters were outliers and biased the comparison. This convergence problem can be due to the flatness and multimodality of the likelihood. A detailed analysis of likelihood behavior while varying one by one all the tested parameters will clarify if likelihood is in fact flat enough so that different local maxima are reached in different runs of estimation. Moreover, the spectrum can be the result of different demographic histories (Myers *et al.* 2008), and, joint with a flat likelihood, different demographic models can nearly equally fit the observed spectrum.

Another issue is that the parameter estimates can be sometimes not coherent with known wild boar demographic history. For example, in the only model that converged (model 1, Figure 4.1a) the split time was ~0.65 and the Asian effective population size at present was > 10 times the ancestral population size. In *∂a∂*i, time is in 2Ne generations units, so using a generational time 2 years and an ancestral effective population size of ~2x10^4 (Groenen *et al.* 2012), the time from the split results to be ~ 24,000 years, way less the split time between European and Asian wild boars, that is 1.2 - 0.8 MYA (Giuffra *et al.* 2000; Groenen *et al.* 2012; Frantz *et al.* 2014). In the same model, the actual population size of Asian wild boars was more than 10 times the ancestral effective population size (~2x10^5), again in contrast with what found by Groenen *et al.* (2012) which found the actual population size is half the ancestral for Asian wild boars (~10^4). One possible explanation is that the model is not adequate to explain the observed spectrum. *∂a∂*i needs the model as input, than it is forced to find the best fit parameters. When a model is not realistic, the parameters *∂a∂*i found are logically not coherent with what is already known about the demography of the species. A deeper analysis of the internal structure of both European and Asian populations, with an increased number of individuals from different locations, will help disentangle which could be a coherent model to test and clarify the demographic history of wild boars.

Demographic history complexity increased by hybridization

Even though there were convergence issues, both the analysis with coalescence (*ms*) and with *∂a∂*i pointed to admixture or migration events after the split between European and Asian wild boars. This increases demographic inference difficulty: in the case of wild boars, signatures of admixture between European and Asian populations were already been found in previous studies (Giuffra *et al.* 2000; Groenen *et al.* 2012). Wild boars admixture is unlikely to be due to human translocation of wild boars (Scandura *et al.* 2008; Kusza *et al.* 2014), but can have been mediated by feral domestic admixed individuals.

Admixture between Asia and Europe is more common in domestic individuals than in wild boars. Around 20% of a commercial breeds individual's genome has Asian origin (Bosse *et al.* 2014a, 2015; Bianco *et al.* 2015b). Asian germplasm introgression to improve domestic pigs started in the late 18th century, because breeders recognized Asian pigs had bigger litter size and reach sexual maturity earlier. In chapter 5 we analyzed the genome of Cocos Island feral pigs, which can be seen as a screenshot of a domestic breed genome soon after the introgression event. This population originated in 1793, when a couple of pigs was left on shore by an English whaler. Since then, no introgression events have been recorded, so that the Asian haplotypes of the population came from this single couple of individuals. Cocos pigs have ~24% of Asian origin genome. We found that, despite having a similar average of Asian germplasm, Cocos, Landrace and Large White breeds did not share a single Asian ancestor, because we did not found correlation between Asian haplotypes of different breeds. Moreover, the Y chromosome haplotype is almost identical of a Tamworth individual, but autosomal IBD between the two populations was not higher that IBD between Cocos and other commercial breeds. This means that Asian germplasm introgression into English breeds was not a single event, more likely, various events of admixture contributed to actual commercial breeds' genome composition. Further analyses are needed to characterize the particular Asian ancestors in each European breed.

European commercial breeds are hybrids. Moreover, because of their origin and domestic and wild pigs interbred, the genetic variability of European wild boars is still present in domestic breeds, jointly with the Asian origin variability introduced by hybridization. Studying the demography of this mosaic of genomes in one individual will result in very different demographic histories, according to the different region of the genome. Cocos pigs are a living proof of the short time effect of hybridization on a domestic breed's genome, and can be used as a model to discover, for example, the different haplotypes that have been selected in the last 200 years of selection and breeding strategies.

6.3. Using polymorphisms to study feralization and better understand domestication

Pigs easily adapt to new environments, and often animals escape from herds, disrupting the local ecosystem (Choquenot et al. 1996; Sierra 2001; Roemer et al. 2002; Cruz et al. 2005). A feral animal can be defined as a domestic individual that was released or escaped or its descendants (Daniels & Bekoff 1989). Feral animals recover, at least in part, some of the phenotypic and genotypic traits of their wild counterparts. Studying the process of feralization can help understanding the domestication process because it can be seen as reverse domestication. In chapter 5 we evaluated the effect of feralization on a domestic population genome. This population was founded by a single couple of hybrid individuals and they were left free to reproduce, since no human was permanently living on the island, having been isolated for over 200 years.

The introgression of Asian germplasm before the event of feralization complicates the study of the mechanisms of the "reverse" process of domestication, because the different regions of the genome tell different domestication stories. This population allowed us to analyze the trajectory of a hybrid genome that became feral and remained isolated for over 200 years. Intriguingly, the levels of variability were comparable to those of extant modern pig breeds. This can be explained if the duration of the bottleneck was very short. Nevertheless, there were also some long homozygous regions along the genome. The longest one is a 28-Mb region in SSC8. This region had also been detected with the 60k SNP chip data, and was confirmed with the analysis of whole genome sequence. A detailed analysis of this region in several populations showed that this long ROH was not exclusive of Cocos pigs, and that it was present in other European commercial breeds as well, but not in European wild boars. A shorter sub-region at the beginning of this long ROH was already recognized by Rubin *et al.* (2012) as a candidate region for selection during the domestication process.

6.4. Perspectives

The three studies that make up this thesis have increased our knowledge about genomewide variability in the pig genome, but also pose some unsolved questions and suggest new avenues for future research.

First of all, the increased number of publicly available sequences, most of them of Asian origin (Ai *et al.* 2015; Kim *et al.* 2015; Moon *et al.* 2015), should fill the gap in the database we created in chapter 3 (Bianco *et al.* 2015a). The higher coverage of the newly published sequences will allow reducing SNP calling error and will reduce the number of variants we found only in few individuals because of quality and coverage. Therefore, a more specific and in-depth analysis of groups and breeds exclusive variants can now be performed.

The availability of new Asian sequences will also allow a better inference of wild boar and domestic pig demographic history. The increased number of wild boars sequenced is fundamental to recover the substructure within the Asian continent. Moreover, samples from Russia and near East Asia should help clarifying the pattern of migration between East Asian and European populations. The next step will be to include domestication in demographic inference. As more breeds are resequenced, it will be possible to trace the joint demographic history of European and Asian breeds, both at genomic and haplotypic level.

Finally, *Isla del Coco* pig population gave us a first insight on the genetic consequences of feralization in a hybrid domestic population, but it only scratches the surface of the iceberg in what concerns village and creole pigs.

The comparison of Cocos haplotypes with commercial breeds can help to identify which haplotypes may have been selected due to modern breeding practices. Moreover, an analysis of the phenotypes that have been modified in Cocos pigs during the last 200 years and their genotypes will increase our knowledge of adaptation to extreme climates.

CHAPTER VII

Conclusions

7. Conclusions

1) Here we provide the first genomewide and worldwide catalog of variants in the pig species. Among the ~48 M variant sites found, we were able to assess the ancestral allele of ~39M. As expected from the Asian origin of *Sus scrofa*, Asian populations exhibited the highest number of exclusive variants. Within Europe, the group of domestic breeds had a higher number of variants than wild boars, likely because of Asian introgression in the last 200 years of breeding. Moreover, we found a strong genomewide correlation between number of SNPs and transition/transversion rate.

2) The analysis of wild boar joint site frequency spectrum suggests that migration events after the split and a period of isolation are necessary to explain wild boar demographic history. Additional samples from Asian regions are needed, nevertheless, for a better assessment of the model. The algorithm to infer demography is also critical for obtaining reliable estimates, since we found important convergence problems with *∂a∂*i.

3) Cocos island pigs are descendants of an English population that was Asian – European hybrid, as are modern international pig breeds. Nevertheless, our analyses suggest multiple introgression events from Asia occurred in the different modern pig breeds and in Cocos population. Further, the Asian component did not increase throughout the years despite the continuous documented import of Asian animals.

4) Despite being isolated for over 200 years, the Cocos feral population is still as variable as modern commercial breeds. This variability pattern is the result of admixture followed by a likely very short bottleneck before the expansion in the island. Nevertheless, we found long runs of homozygosity (ROH) in Cocos individuals. The longest ROH spanned 28-Mb in SSC8, which includes a 10-Mb candidate region for a signature of selection in domestication process.

Bibliography

Ai H, Fang X, Yang B *et al.* (2015) Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, 1–11.

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**, e56001.

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Altmann A, Weber P, Bader D *et al.* (2012) A beginners guide to SNP calling from high-Throughput DNA-sequencing data. *Human Genetics*, **131**, 1541–1554.

Alves E, Ovilo C, Rodriguez MC, Silio L (2003) Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Animal Genetics*, **34**, 319–324.

Amaral AJ, Ferretti L, Megens H-J *et al.* (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PloS one*, **6**, e14782.

Beaumont M a. (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Beaumont M a., Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Bertorelle G, Benazzo a., Mona S (2010) ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology*, **19**, 2609–2625.

Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M (2015a) A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig. *Plos One*, **10**, e0118867.

Bianco E, Soto HW, Vargas L, Pérez-Enciso M (2015b) The chimerical genome of Isla del Coco feral pigs (Costa Rica), an isolated population since 1793 but with remarkable levels of diversity. *Molecular Ecology*.

Bosse M, Madsen O, Megens H-J *et al.* (2015) Hybrid origin of European commercial pigs examined by an in-depth haplotype analysis on chromosome 1. *frontiers in Genetics*, **5**, 1–9.

Bosse M, Megens H-J, Frantz LAF *et al.* (2014a) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature communications*, **5**, 4392.

Bosse M, Megens H-J, Madsen O *et al.* (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS genetics*, **8**, e1003100.

Bosse M, Megens H-J, Madsen O *et al.* (2014b) Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. *Molecular ecology*, **23**, 4089–4102.

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

Choquenot D, Mcilr J, Korn T (1996) *Managing Vertebrate Pests: Feral Pigs*. Australian Government Publishing Service, Canberra.

Cock PJ a, Fields CJ, Goto N, Heuer ML, Rice PM (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, **38**, 1767–1771.

Crosby AW (2003) *The Columbian Exchange*. Greenwood Publishing Group, London.

Cruz F, Josh Donlan C, Campbell K, Carrion V (2005) Conservation action in the Galàpagos: feral pig (Sus scrofa) eradication from Santiago Island. *Biological Conservation*, **121**, 473–478.

Daniels TJ, Bekoff M (1989) Feralization: The making of wild domestic animals. *Behavioural Processes*, **19**, 79–94.

Darwin C (1868) *The Variation of Animals and Plants under Domestication.* John Murray, London.

Elliott JH (2006) *Empires of the Atlantic World: Britain and Spain in America 1492 - 1890*. Yale University Press.

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.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, **9**.

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

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

Fernández a I, Alves E, Óvilo C, Rodríguez MC, Silió L (2011) Divergence time estimates of East Asian and European pigs based on multiple near complete mitochondrial DNA sequences. *Animal genetics*, **42**, 86–88.

Fernández A, Muñoz M, Alves E (2014) Recombination of the porcine X chromosome: a high density linkage map. *BMC Genetics*, **15**.

Frantz LAF, Madsen O, Megens H-J, Groenen M a M, Lohse K (2014) Testing models of speciation from genome sequences: divergence and asymmetric admixture in Island Southeast Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Molecular ecology*, 1–9.

Frantz LAF, Schraiber JG, Madsen O *et al.* (2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome biology*, **14**, R107.

Giuffra E, Kijas J, Amarger V (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics*, **154**, 1785–1791.

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.

Gusev A, Palamara PF, Aponte G *et al.* (2012) The architecture of long-range haplotypes shared within and across populations. *Molecular Biology and Evolution*, **29**, 473–486.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, **5**, e1000695.

Herrero-Medrano JM, Megens H-J, Groenen M a M *et al.* (2013) Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC genetics*, **14**, 106.

Howrigan DP, Simonson M a, Keller MC (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC genomics*, **12**, 460.

Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping It Local: Evidence for Positive Selection in Swedish Arabidopsis thaliana. *Molecular Biology and Evolution*, **31**, 3026–3039.

Kim H, Song KD, Kim HJ *et al.* (2015) Exploring the Genetic Signature of Body Size in Yucatan Miniature Pig. *Plos One*, **10**, e0121732.

Kirin M, McQuillan R, Franklin CS *et al.* (2010) Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE*, **5**, 1–7.

Kusza S, Podgórski T, Scandura M *et al.* (2014) Contemporary genetic structure, phylogeography and past demographic processes of wild boar Sus scrofa population in Central and Eastern Europe. *PLoS ONE*, **9**.

Larson G, Albarella U, Dobney K *et al.* (2007a) Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15276–81.

Larson G, Cucchi T, Fujita M *et al.* (2007b) Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania.

Proceedings of the National Academy of Sciences of the United States of America, **104**, 4834–9.

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, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–6.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li M, Tian S, Jin L *et al.* (2013) Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat Genet*, **45**, 1431–1438.

Li M, Tian S, Yeung CKL *et al.* (2014) Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Scientific reports*, **4**, 1:7.

Lin CS, Sun YL, Liu CY *et al.* (1999) Complete nucleotide sequence of pig (Sus scrofa) mitochondrial genome and dating evolutionary divergence within Artiodactyla. *Gene*, **236**, 107–114.

Lohmueller KE, Bustamante CD, Clark AG (2009) Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics*, **182**, 217–231.

Van der Made J (1999) Ungulates from Atapuerca TD6. *Journal of human evolution*, **37**, 389–413.

McLaren W, Pritchard B, Rios D *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, **26**, 2069–2070.

Molnár J, Nagy T, Stéger V *et al.* (2014) Genome sequencing and analysis of Mangalica, a fatty local pig of Hungary. *BMC genomics*, **15**, 761.

Moon S, Kim T-H, Lee K-T *et al.* (2015) A genome-wide scan for signatures of directional selection in domesticated pigs. *BMC Genomics*, **16**, 1–12.

Muñoz M, Alves E, Ramayo-Caldas Y *et al.* (2012) Recombination rates across porcine autosomes inferred from high-density linkage maps. *Animal genetics*, **43**, 620–3.

Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? *Theoretical Population Biology*, **73**, 342–348.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Andrew G (2007) Recent and ongoing selection in the human genome. *Nat Rev Genet.*, **8**, 857–868.

Noce a., Amills M, Manunza A *et al.* (2015) East African pigs have a complex Indian, Far Eastern and Western ancestry. *Animal Genetics*.

Ojeda A, Ramos-Onsins SE, Marletta D *et al.* (2011) Evolutionary study of a potential selection target region in the pig. *Heredity*, **106**, 330–338.

Ottoni C, Flink LG, Evin A *et al.* (2013) Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics. *Molecular biology and evolution*, **30**, 824–32.

Porter V (1993) *PIGS - A handbook to the breeds of the world.* HELM Information Ltd.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American journal of human genetics*, **69**, 1–14.

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

Ramayo-Caldas Y, Mercade A, Castello A *et al.* (2012) Genome-wide association study for intramuscular fatty acid composition in an Iberian x Landrace cross. *Journal of Animal Science*, **90**, 2883–2893.

Ramírez O, Burgos-Paz W, Casas E *et al.* (2014) Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity*, 1–10.

Ramírez O, Ojeda A, Tomàs 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.

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.

Ramos-Onsins SE, Burgos-Paz W, Manunza a, Amills M (2014) Mining the pig genome to investigate the domestication process. *Heredity*, 1–14.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ (2014) ABC inference of multi-population divergence with admixture from un-phased population genomic data. *Molecular Ecology*, **23**, 4458–4471.

Rodero A, Delgado J V., Rodero E (1992) Primitive andalusian livestock and their implications in the discovery of america. *Archivos de Zootecnia*, **41**, 383–400.

Roemer GW, Donlan CJ, Courchamp F (2002) Golden eagles, feral pigs, and insular carnivores: how exotic species turn native predators into prey. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 791–796.

Rubin C-JC-J, Megens H-JH-J, Martinez Barrio A *et al.* (2012) Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19529–19536.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Scandura M, Iacolina L, Crestanello B *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? *Molecular ecology*, **17**, 1745–62.

Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, **46**, 919–925.

Sierra C (2001) El cerdo cimarrón (Sus scrofa, Suidae) en la Isla del Coco, Costa Rica: Escarbaduras, alteraciones al suelo y erosión. *Revista de Biología Tropical*, **49**, 1158–1170.

Sunnåker M, Busetto AG, Numminen E *et al.* (2013) Approximate Bayesian Computation. *PLoS Computational Biology*, **9**, e1002803.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Tortereau F, Servin B, Frantz LAF *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC genomics*, **13**, 586.

Veroneze R, Lopes PS, Guimarães SEF *et al.* (2013) Linkage disequilibrium and haplotype block structure in six commercial pig lines. *Journal of Animal Science*, **91**, 3493–3501.

Vilaça ST, Biosa D, Zachos F *et al.* (2014) Mitochondrial phylogeography of the European wild boar: The effect of climate on genetic diversity and spatial lineage sorting across Europe. *Journal of Biogeography*, **41**, 987–998.

Wallace AR (1855) On the Law which has regulated the Introduction of New Species. *The Annals and Magazine of Natural History*, **XVI**, 183–196.

Wang Y, Tang Z, Sun Y *et al.* (2014) Analysis of genome-wide copy number variations in chinese indigenous and Western pig breeds by 60 k SNP genotyping arrays. *PloS one*, **9**, e106780.

Wetterstrand KA (2014) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *www.genome.gov/sequencingcosts*.

White S (2011) From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. *Environmental History*, **16**, 94–120.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Zadik BJ (2005) *The Iberian Pig in Spain and the Americas at the Time of Columbus*. University of California, Berkley.

Resumen

Las nuevas tecnologías de ultrasecuenciación (NGS) han alterado espectacularmente la investigación en la genómica de las especies domesticas, entre ellas la del cerdo. Usando datos de genómica es posible, por ejemplo, comprender mejor la demografía de los jabalíes y su impacto en el proceso de domesticación. Además, el estudio de los cerdos ferales mejoran el conocimiento de las dinámicas de feralización, y sirven de comparación con las razas domesticas actuales. Este trabajo es un estudio sobre la demografía y los procesos de feralización en la especie porcina, a través del uso de polimorfismos genómicos.

En la primera parte, se ha generado el primer catálogo de variantes SNPs a nivel genómico y mundial, analizando el genoma de 128 cerdos y 5 "outgroups". Entre las ~48 millones de variantes que encontramos, pudimos inferir el alelo ancestral de ~39 millones. El numero de variantes derivadas exclusiva de razas europeas (~6 millones) es menor que de las asiáticas (>13 millones), tal como se espera por el origen asiático de *S. scrofa*. También encontramos una fuerte correlación en la frecuencia alelica entre cerdos domestico y jabalíes dentro de Asia y dentro de Europa. Esta correlación no se encontró entre continentes, debido a la gran distancia evolutiva entre cerdos de ambos continentes (~1 millón de años).

En la segunda parte de la tesis, intentamos aclarar la historia demográfica de los jabalíes. Analizamos el espectro de frecuencia conjunto de unos 2 millones de SNPs, encontrados en el trabajo anterior, de 9 jabalíes europeos y 8 jabalíes asiáticos usando coalescencia y la inferencia analítica de ∂a∂i. Con coalescencia evaluamos si la separación de las dos poblaciones es suficiente para explicar el espectro observado, pero incluyendo migración en los modelos, el espectro conjunto es coherente con el observado. Con ∂a∂i, comparamos 6 modelos que difieren en el número de cuellos de botellas y eventos migratorios. En las diferentes iteraciones, los parámetros demográficos convergieron en los mismos valores solo con el modelo más sencillo. A pesar de este problema de

convergencia con los modelos más complejos, ambos métodos muestran que la migración es necesaria por explicar la historia demográfica de los jabalíes.

En la tercera parte de este trabajo estudiamos las dinámicas de la feralización. Analizamos el genoma de los cerdos ferales de la Isla del Coco (Costa Rica), que ha estado aislada desde su fundación en 1793 y es un excelente modelo para estudiar de las dinámicas de la feralización. En este estudio confirmamos que los cerdos domésticos ingleses ya eran híbridos entre razas europeas y asiáticas al final del siglo XVII. Sorprendentemente, a pesar del cuello de botella, la variabilidad promedio de la población de la Isla del Coco es similar a la variabilidad de las actuales razas comerciales, tales como Large White o Duroc. Además, encontramos una región de unas 10-Mb con un marcado descenso de la variabilidad en todas las muestras analizadas, previamente identificada como altamente diferenciada entre jabalíes y razas domesticas.

La domesticación y la feralización son eventos simétricos de la historia del cerdo. El análisis de la demografía de los jabalíes sirve como hipótesis nula para el estudio de las dinámicas selectivas previas a la domesticación. Por otro lado, el análisis de los cerdos ferales de la Isla del Coco permite reconstruir los genomas de los cerdos previos a la selección moderna pero posterior a la hibridación con Asia. Además ayudan al estudio de los efectos de la feralización en un animal híbrido. Este trabajo ha sido posible solo gracias a la evolución de las técnicas de secuenciación que permitieron la publicación de un número creciente de secuencias de genomas completos.

Riassunto

Le nuove tecniche di sequenziamento (NGS) hanno alterato in modo spettacolare la ricerca in campo genomico nelle specie domestice, tra le quali il maiale. Attraverso l'uso di dati genomici è possibile avere una maggior comprensione della demografia del cinghiale e del suo impatto sul processo dell'addomesticamento. Inoltre, lo studio dei maiali inselvatichiti incrementa la conoscenza del processo di inselvatichimento e serve come confronto con le razze domestiche. Attraverso l'analisi di SNPs genomici, in questa tesi si analizzano la demografia e il processo di feralizzazione nella specie porcina.

Nella prima parte, utilizzando SNPs ottenuti dal genoma di 128 maiali, domestici e selvatici, e di 5 outgroup, è stato elaborato il primo catalogo di SNPs a livello genomico e mondiale della specie *Sus scrofa*. Dei ~ 48 milioni di SNPs genotipati è stato calcolato quale è l'allele ancestrale e quale il derivato di ~39 milioni. Coerentemente con l'origine asiatica della specie, il numero di SNPs esclusivi delle razze europee (>6 Milioni) è inferiore rispetto a quello delle razze asiatiche (>13 milioni). Una forte correlazione della frequenza allelica tra il gruppo di animali domestici e cinghiali è stata trovata. Questa correlazione è assente tra continenti, a causa della lunga distanza evolutiva tra le popolazioni asiatiche e europee (~1 milione di anni).

Usando gli SNPs dello studio precedente, abbiamo analizzato la storia demografica del cinghiale eurasiatico. Abbiamo analizzato lo spettro di frequenza congiunto di circa 2 milioni di SNPs sequenziati in 9 cinghiali europei e 8 asiatici, usando coalescenza ed il metodo analitico di *∂a∂*i. Con coalescenza sono stati simulati differenti scenari demografici per capire se uno *split* era sufficiente per spiegare lo spettro osservato, però, solo con la presenza di eventi migratori nello scenario demografico simulato è stato possibile ottenere uno spettro di frequenza coerente con l'osservato. Con *∂a∂*i, abbiamo confrontato 6 modelli demografici, che si differenziano nel numero di *bottleneck* e di eventi migratori. Nelle differenti iterazioni effettuate, si è ottenuta la convergenza dei parametri sugli stessi valori solo con il modello più semplice.

Nonostante i problemi di convergenza, entrambi i metodi hanno evidenziato che la migrazione ha un ruolo fondamentale per spiegare la storia demografica del cinghiale.

Nella terza parte sono state studiate le dinamiche della feralizzazione. Abbiamo effettuato l'analisi del genoma della popolazione di maiali dell'Isola del Cocco (Costa Rica), fondata nel 1793 e isolata da allora, un eccellente modello per lo studio delle dinamiche della feralizzazione. In questo studio si conferma che alla fine del XVII secolo le razze domestiche di origine inglese erano ibridi tra razze europee e asiatiche. Nonostante il forte effetto fondatore della popolazione del Cocco, la variabilità media del genoma è similare a quella delle attuali razze commerciali, come Large Whithe o Duroc. Inoltre, nonostante l'alta variabilità media, è stata identificata una regione di 10-Mb che presenta una bassa variabilità in tutti i campioni analizzati, precedentemente identificata come altamente differenziata tra cinghiale e maiale.

La domesticazione e la feralizzazione sono eventi simmetrici nella storia di *Sus scrofa*. L'analisi della demografia del cinghiale serve come ipotesi nulla nello studio delle dinamiche selettive precedenti la domesticazione. L'analisi dei maiali inselvatichiti dell'isola del Cocco permette di ricostruire il genoma dei maiali presente prima degli attuali processi di selezione, ma posteriori all'ibridazione tra Asia e Europa. Inoltre, permettono lo studio degli effetti della feralizzazione su un animale ibrido. Questo studio è stato possibile solo grazie all'evoluzione delle tecniche di sequenziazione, che hanno permesso la pubblicazione di un numero sempre crescente di sequenze di genoma completo.

Curriculum vitae

Short biography

Erica Bianco was born in Carmagnola (Turin, Italy) in 1986, she graduated in Biological Sciences in 2008 at the Università degli Studi di Torino. In 2008, she worked at the Laboratorio di Biologia Marina Torino, where she performed her bachelor thesis research project. In 2009 she earned the Erasmus fellowship and the Vinci fellowship to obtain a binational master between the Università degli Studi di Torino and the Université Joseph Fourier of Grenoble. Between 2009 and 2010 she worked in the Laboratoire d'Ecologie Apline at Grenoble, where she performed her master thesis research project. In 2011 she obtained the master in Conservation and Animal biodiversity at the Università degli Studi di Torino and the master in Biology, Ecology and Biodiversity (research) at the Université Joseph Fourier of Grenoble. In 2011 she obtained a FPI fellowship from the Spain Science ministry to perform a Ph. D in Animal Production at the Animal Genetics group of Veterinary faculty in the Universitat Autònoma de Barcelona. In 2013 she got an EEBB fellowship to perform a stay abroad at the Department of Medical Biochemistry and Microbiology at Uppsala Universitet (Sweden).

List of publications

Ramírez O, Burgos-Paz W, Casas E, Ballester M, **Bianco E**, Olalde I, Santpere G, Novella V, Gut M, Lalueza-Fox C, Saña M, Pérez-Enciso M. (2014) ***Genome data from a sixteenth century pig illuminate modern breed relationships.*** Heredity, 1–10. doi: 10.1038/hdy.2014.81

Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M (2015) ***A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig***. Plos One, 10, e0118867. doi: 10.1371/journal.pone.0118867

Bianco E, Soto HW, Vargas L, Pérez-Enciso M (2015) *The chimerical genome of Isla del Coco feral pigs (Costa Rica), an isolated population since 1793 but with remarkable levels of diversity.* Molecular Ecology. doi: 10.1111/mec.13182

Bonin A, Paris M, Frérot H, **Bianco E**, Tetreau G, Després L (2015) ***The genetic architecture of a complex trait: resistance to multiple toxins produced by*** **Bacillus thuringiensis israelensis** ***in the dengue and yellow fever vector, the mosquito*** **Aedes aegypti.** Infection, Genetics and Evolution. doi: http://dx.doi.org/10.1016/j.meegid.2015.07.034

Conference participations:

Oral communications:

Bianco E, Soto HW, Vargas L, Gut M, Nevado B, Ramos-Onsins SE, Pérez-Enciso M. ***El genoma de los cerdos ferales de la Isla del Coco (Costa Rica).*** XXXIX Congreso de la Sociedad Española de Genética – Gerona, September 2013

Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M. ***A comprehensive catalogue of pig genome polymorphism from massive resequencing data- the chromosome 18.*** XVII Reunión Nacional de Mejora Genética Animal – Bellaterra, June 2014

Bianco E, Soto HW, Vargas L, Pérez-Enciso M. ***Cocos Island feral pigs and what happens when an admixed population stays isolated for 200 years.*** 2nd Annual Congress of Young Researchers (ACYR); CRAG - Bellaterra, June 2015

Poster:

Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M. ***The chimerical genome of the isolated feral pig population of*** **'Isla del Coco'** ***national park.*** Next Generation Sequencing Conference (NGS) – Barcelona, June 2014

Acknowledgements

I would like first to thank my thesis director Miguel Pérez-Enciso for giving me the opportunity to learn about bioinformatics and pig genomics, and for allowing me to perform this exiting research in these very interesting topics and with the newest technologies.

A big thank to all the members of the Animal Genetics department at CRAG, in particular to all those people who shared with me the office 325: Kuin, William, Arianna, Ali, Maria, Sarai, Bete, Bruno, Juan Carlos, Antonia, Taina, Angela, Jordi, Fabiana, Marta and Sara, or the lunches and coffees: Rayner, Anna Puig, Manu, Dani, Ediane, Anna Castelló, Anna Mercader, Betlem, Tania, Joanna, Samir, Ligia, Yuliaxis, Jordi Corominas, Xavi, Oriol, Ingrid, Veronica, Sebas, Marcel, Alex, JM, Quim, Júlia… thanks for all the moments, all the good words and all the fun together.

Thanks to all the people of the IMBIM at BMC, who hosted me in my stay in Sweden, in particular to all the Ph. Ds, to Calle, Freyja, Nima, Marco and Rakan. Also a big thanks to Malin, Micke, Alfred and Klara, thanks for your hospitality and your friendship.

Thank to the facebook group "Gli Italiani a Barcellona", without these people I would have never survived the first period here in BCN and I would not have known some important people for my life, thanks Silvia, Martina and Dani, I love all of you.

A special thanks to Clara and Klayna, my very best flat mates and friends, that year in our apartment was amazing, and made friendships that will never end. I love you girls.

To my family and my new extended family, mamma, Adriana, Valeria, papà, nonni, Gianni, Franca, Maurizio, Marta, Simona, Dolores, Daniele, Dado, Marianna and Federico thank you for being there when I came back to Italy and for supporting me, host me, to look for me at the airport and tolerate the "tramp" side of me.

Thank you Marti, always…

And last, but not least, thank you Ivan, for everything.

Cover picture has been drawn by Adriana and Valeria Bianco.

Colophon

This work was funded by AGL2010-14822 and AGL2013-41834-R (Ministry of Economy and Science, Spain). Erica Bianco is recipient of a FPI fellowship BES-2011043530 from the Ministry of Economy and Science (Spain).