Speciation and domestication in *Suiformes*: a genomic perspective

Laurent A. F. Frantz

Thesis committee

Promotor
Prof. Dr M.A.M. Groenen
Personal chair at the Animal Breeding and Genomics Centre
Wageningen University

Co-promotor
Dr O. Madsen
Research Associate at the Animal Breeding and Genomics Centre
Wageningen University

Dr H.-J. Megens
Research Associate at the Animal Breeding and Genomics Centre
Wageningen University

Other members
Prof. Dr D. Bradley, Trinity College, Dublin, Ireland
Prof. Dr M.E. Schranz, Wageningen University
Prof. Dr M. Schilthuizen, Naturalis Biodiversity Center and Leiden University
Prof. Dr ir J.L. van Leeuwen, Wageningen University

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Speciation and domestication in *Suiformes*: a genomic perspective

Laurent A. F. Frantz

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 26th January, 2015
at 1:30 p.m. in the Aula.

Laurent A.F. Frantz
Speciation and domestication in *Suiformes*: a genomic perspective
227 pages.

PhD thesis, Wageningen University, Wageningen, NL (2015)
With references, with summaries in English and Dutch

ISBN 978-94-6257-254-6

Abstract

Frantz, L.A.F. (2015). Speciation and Domestication in *Suiformes*: a genomic perspective. PhD thesis, Wageningen University, the Netherlands

The diversity of life on earth owes its existence to the process of speciation. The concept of speciation is primordial for evolutionary biologists because it provides a framework to understand how contemporary biodiversity came to be. Moreover, not only natural phenomena can result in the differentiation of life forms. Indeed, biodiversity can also be the result of direct and indirect human influence such as domestication. In this thesis, I investigate these evolutionary processes (speciation and domestication) in the *Suiformes* superfamily (pigs and related species). I use complete genome sequences to illuminate many specific aspects of the speciation and domestication in *Suiformes* as well as to draw general conclusions on these crucial processes. In chapter 2 I show how genomes provide an essential source of information to retrieve deep taxonomic relationships among *Suiformes*. This allows me to describe multiple novel aspects of their early evolutionary history such as the fact that *Suiformes* colonised North America at least twice. In this chapter, I further highlight and discuss novel methodological limitations that are inherent to phylogenomics. In chapters 3, 4 and 5 I use genome sequences to resolve the evolutionary history of the genus *Sus* (domestic pigs and wild boars species). More precisely, I show that, contrary to the expectation of simple models of speciation, the evolutionary history of these species involved alternating periods of gene-flow and genetic differentiation that are tightly linked to past climatic fluctuations that took place over the last 4 million years. In addition, these chapters also provide novel insights into the process of speciation by demonstrating that genetic differentiation between species can be achieved, even when gene-flow is strong. Lastly, in chapter 6 I tested multiple models of domestication for *S. scrofa*. In this chapter I show that models involving reproductive isolation between wild and domestic forms are incompatible with genomic data. Moreover, this chapter demonstrates that, while domestic pigs are morphologically homogenous, they are not genetically homogenous. Together, these findings have important implications for our understanding of the process of domestication because it shows that this process was not solely the result of captivity. Together, the results of this work not only provide a comprehensive evolutionary history for the *Suiformes*, but also novel insights into the complex processes (speciation and domestication) that are responsible for the diversity of life on earth.

Contents

1

General introduction

1.1 Introduction

A brief introduction to evolutionary genetics

The diversity of life on earth owes its existence to the process of speciation. Over millions of years, life forms have diversified as a result of genetic differentiation. The concept of speciation is key to evolutionary biology and provides a framework to understand how contemporary biodiversity came to be. The field of evolutionary genetics has allowed biologists to better understand speciation and has had a tremendous impact on evolutionary biology as a whole. Classical authors such as Wright, Fisher, Haldane and Kimura provided the basis of a field that would revolutionise our understanding of evolutionary biology. The theory of neutral evolution, introduced by Kimura in 1968 (Kimura 1968; Kimura 1983) provides a perfect example that illustrates how evolutionary genetics has revolutionised our understanding of evolution. This theory not only provided evolutionary geneticists with a null hypothesis to detect selection footprints but also allowed for the development of many other tools to study speciation. For example, the concept that DNA sequences may evolve without selective constraints provided geneticists with the necessary model to correlate genetic variations between populations or species with time and geography or to estimate demographic parameters. Indeed, the theory of neutrality allowed biologists to disentangle natural selection and molecular variation among and between species, leaving only random genetic drift, time, demography and geography as responsible for the observed variance. This model resulted in completely novel fields of studies such as phylogeography and molecular dating. Thus, the concept of neutrality directly allowed for a throughout investigation of the process of speciation through time, unravelling many aspects of the evolutionary history of life on earth, such as the effect that past climatic fluctuations have had on speciation and the geographical origin of diverse group of organisms (*e.g.* Hewitt, 2000; Hewitt, 2004; Meredith et al., 2011). It is important to recognise the importance that these early theories have had on work even carried today. It would be fair to say that without the idea of neutrality most of the work presented in this thesis would not have been possible.

In recent years, novel sequencing technologies have dramatically increased the amount of molecular data available to evolutionary biologists. This revolution of genomes has had a critical effect on our understanding of evolution. The critical boost in power afforded by genomics, compared with previous limited genetics studies, allows geneticists to test increasingly finer hypotheses. In this work I will provide concrete examples on how the genomic revolution impacted our understanding of speciation in general. In particular, I will concentrate on a superfamily, the *Suiformes* (pig and related species). This work will allow me to

draw general conclusions on the process of speciation as well as discuss the specificity of the process in *Suiformes*. In the first part (Chapter 2-4) I will provide a comprehensive evolutionary history of the superfamily from the Eocence (~40Ma) to the domestication of pigs. In a second part (Chapter 5-6), I will further provide many valuable insights into the process of domestication in pigs. Moreover, in these following introductory sections I will provide in depth definition of the concept of speciation and domestication and provide basic information on the methods used in this work.

A brief introduction to the concept of speciation

The word *speciation* was first coined by Cook in 1906 to define the process by which species differentiate. Cook described speciation as "the evolutionary process that leads to the origination or multiplication of species by subdivision, usually (if not always), as the result of environmental incidents". Thus, while Cook recognised the importance of natural selection in speciation he already realised that differentiation does not necessarily need to involve selection. This idea was later used to define one of the most commonly used models of speciation (allopatric speciation; Figure 1.1). Allopatric speciation provides the most basic model of species divergence. This model assumes the creation of a barrier to dispersal that divides a population into two sub populations. Such a barrier will have a direct impact on gene-flow and leads to divergence between the two populations. Allopatric speciation does not necessarily involve natural selection. Indeed, in the absence of gene-flow, random genetic drift alone can be sufficient to create, over time, large variations in allele frequency between sub populations eventually leading to their differentiation into two reproductively isolated species.

The idea of reproductive isolation between species led to the concept of biological species, in which two species cannot produce a fertile offspring (Mayr 1942). However, this model seems unrealistic in many cases (*e.g.* Schliewen, Tautz, & Pääbo, 1994; Dieckmann & Doebeli, 1999). For example, the well-known example of the Darwin Finches, in which subdivision is not the result of geographic isolation and natural selection clearly played a role in beak formation. This has led to the definition of other, more complex, models of speciation such as parapatric, peripatric and sympatric speciation (Figure 1.1). Sympatric speciation is the most extreme case. In this model two populations differentiate into two species with no physical barrier to gene-flow (*i.e.* complete range overlap). This means that genetic homogenisation of the two sub populations is not prevented by a physical barrier throughout the process of speciation. Such phenomena must involve disruptive natural selection and pre-zygotic reproductive isolation (as opposed to post-zygotic

isolation which means that hybrids are either infertile or cannot develop) as a mean to reduce gene-flow and induce genetic differentiation between the two conspecific, overlapping, populations (Kondrashov & Kondrashov, 1999). This has led to a reinterpretation of the concept of species. Indeed, the biological concept of species implies that no inter-specific gene-flow is possible due to post-zygotic reproductive isolation. The need for another concept, based on genetic differentiation, led to the idea of the Phylogenetic concept of species, which in my opinion, is tightly linked to complex speciation and provides a better, more general model. Speciation with gene-flow or complex speciation, in which natural selection plays a prominent role, is primordial to this work and is going to be refereed to many time in the following chapters.

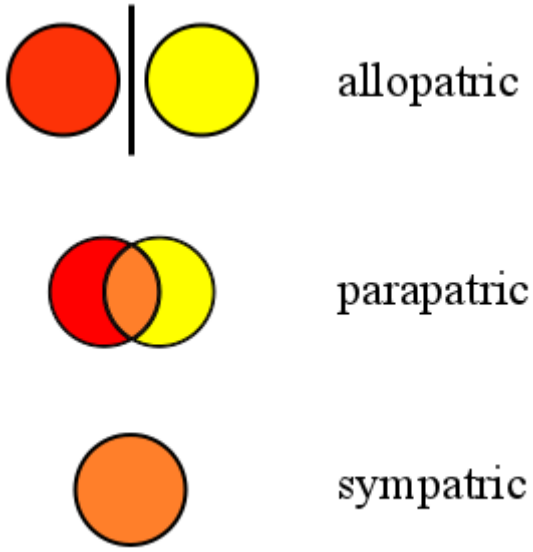


Figure 1.1: Models of Speciation. Each circle represents the geography of a species (red or yellow). Orange colour implies that the two species overlap in their range.

1.2 Mechanistic models of complex speciation

In the last decades, models of complex speciation have been used to explain patterns of inter-specific genetic variation. Many examples of complex speciation are available in the literature (Gourbière & Mallet, 2010; Seehausen, 2004). Thus, it seems that speciation with gene-flow is common (Nosil, 2008). In recent years, multiple studies have put forward models to explain how species differentiate in face of gene-flow (e.g. Basset, Yannic, Brünner, & Hausser, 2006; Noor & Bennett,

2009). These studies have had a tremendous effect on our understanding of this phenomenon. Most common, these models involve hybrid incompatibilities such as chromosomal rearrangement(s) and/or regions of high inter-specific divergence or "island of speciation". Hybrid incompatibilities predict poor pairing of chromosomes during hybrid meiosis or lower fitness of heterozygotes. Such a model is often referred to as the Dobzhansky-Muller model (Dobzhansky, 1937; Muller, 1942). However, this model seems unlikely to be able to explain sympatric or parapatric speciation as it seems implausible that incompatibilities arise while gene-flow is taking place. On the other hand, islands of speciation are regions in the genome that show a high degree of differentiation between diverging species, due to reduction of gene-flow induced by natural selection. Let's assume two species, A and B, and a single biallelic locus Z (a,b) that has an important effect on the phenotypic differences observed between these species. Introgression of the allele (a) more advantageous for species A into species B is expected to reduce the fitness of the F2 hybrid AB in species' B ecological niche. Under this model, we expect less homogenisation at locus Z compared to the rest of the genome due to a combination of natural selection and recombination. Such a phenomenon is likely to reduce global and localised (in the genome) inter-specific gene-flow (and inter-specific divergence) and to promote species differentiation. This concept provides a realistic and testable mechanistic model for complex speciation.

1.3 Methodology to study speciation

While non-exhaustive, these paragraphs should provide a list of methodologies and concepts that can be utilised to study speciation at different time scales. These first methods focus on retrieving taxonomy from DNA sequences. Thereafter, I will introduce available methods to test complex models of speciation using genome sequences.

In my opinion, the study of speciation finds its basis in systematic biology. Indeed, the taxonomic relationship of a group of species not only provides information about the process itself (*i.e.* on the chronology) but also provides the basis to design and test models of speciation. This often involves fitting a bifurcating model of evolution, also known as a phylogenetic tree, which depicts the relationship of a set of species. Phylogenetics is based on the concept of parsimony (Fitch 1971). Parsimony suggests that the simplest model is always the most likely (parsimonious) explanation for the data in hand. For example, it is more likely that vertebrae arose only once in evolution, thus it is more parsimonious to assume that all vertebrates share a more recent common ancestor before their common ancestor with invertebrates. Such powerful idea allows for the reconstruction of

the tree of life that provides the basis to understand and investigate speciation. Maximum parsimony can be applied to different types of data, such as morphological characteristics as well as DNA sequences. In recent years, the non-parametric approach of maximum parsimony has been replaced by more powerful parametric approaches, such as Maximum likelihood (ML; Felsenstein, 1981). While based on similar concepts, this method allows deriving the likelihood function of a tree given the data and an assumed model of DNA substitution. These models of substitution are based on the neutral theory and permit to compute the likelihood of a set of branch lengths (configuration of substitution on a tree). These models of DNA substitution allow biologists to accommodate for complex evolutionary processes such as taking into account unobserved DNA substitutions (Whelan & Goldman, 2001) and make ML a very powerful and statistically sound method for phylogenetics. For example a simple model, such as Kimura's K80 model that has two parameters to distinguish between α, the rate of transitions (A<->G or C<->T) and β, the rate of transversions (G<->C and A<-T>), provides a much more realistic way to model complex DNA substitutions through time. Such useful models of DNA substitutions have had a critical impact on our understanding of speciation. Indeed these provided a framework to infer the number of substitutions on a given lineage (on a known topology) and correlate these with time (Thorpe, 1982). This led to the development of molecular clocks. Molecular clocks allow evolutionary biologists to put phylogenetic trees into a geological context and to draw important conclusions upon the process of speciation. For example, such clocks have been used to test the hypothesis that the disappearance of the dinosaurs after the Cretaceous led to the diversification of mammals (*e.g.* Meredith et al., 2011; dos Reis et al., 2012). These clocks are often calibrated with known and dated fossils that can serve to translate substitution counts into years. There are still many issues with molecular clocks and their development is a very active area of evolutionary biology (Drummond, Ho, Phillips, & Rambaut, 2006). Some of these limitations include variable rate among (or even within) branches within a phylogenetic tree as well as fossil date uncertainty (used to translate substitution into years). However, these are beyond the scope of this introduction. Some further discussion will be provided in Chapter 2 and in the General Discussion.

Maximum likelihood reconstructions of phylogenetic trees from single DNA sequences is often "trivial" with current computing power. Indeed, in most cases ML estimation of phylogeny should converge toward the "correct" tree if models of DNA substitution are not strongly violated. This is not necessarily the case when trees are built from many loci. Incongruence is often hidden in the genome due to genealogical heterogeneity. Heterogeneity can arise from a genuine biological

signal (as well as analytical problems inherent to ML). Biologists often assume that the most common source of heterogeneity arises from lineage sorting. An incomplete lineage sorting (ILS) can be defined as a genealogy that does not match the underlying species phylogeny. These can arise due to ancestral polymorphisms that were present before speciation (or divergence). For example, the probability of a coalescence event between two speciation events T_1 and T_2 (T_1 older than T_2) depends on the effective population size N_e and $\Delta T = T_1 - T_2$. Thus, we can compute *P*, the probability of colaelescence in a random mating population of size N_e in an interval of time ΔT as $P = \left(\frac{1}{2N_e} \right)^{\Delta T}$. Therefore, it is unsurprising that such a phenomenon is often observed in real data from different populations or species as if N_e gets larger or ΔT gets smaller, the probability of coalescence, in between two speciation events, reduces as well as the number of genealogies in the genome that match the history of divergence (phylogenetic tree). This is why ILS can drastically reduce phylogenetic power. For example, the root of Eutherian mammals likely suffers from this problem (Figure 1.2). Many studies have attempted to resolve this node, however, it seems that each possible phylogenetic tree has been recovered by many different analyses with high support (McCormack et al., 2012; Meredith et al., 2011; Romiguier, Ranwez, Delsuc, Galtier, & Douzery, 2013; Teeling & Hedges, 2013). Researchers have applied different methodologies to tackle ILS. The first approach, also known as concatenation, works as a 'democratic vote'. In this method, all available loci are combined in a single analysis and the most likely tree for the whole data set can be retrieved. However, this method has been shown to lead to overconfidence and misleading results (Kubatko & Degnan, 2007). This can happen in the case of heterogeneity in coalescence patterns among loci and this violates the assumption of a single underlying genealogy. Indeed, because ML is a sum of likelihoods across sites, a very long DNA fragment can have an overwhelming effect in such analyses. The second approach uses a method at the interface of population genetics and phylogenetics and explicitly models incongruence among loci. In this method, incongruence is tackled by computing a tree for each available locus and ad-hoc species tree reconciliations by computing the likelihood of a species tree under the coalescence model (e.g. Liu, Yu, & Edwards, 2010). These two methods have allowed for the resolution of many taxonomic relationships within the tree of life.

Understanding the basic taxonomy of the group of organisms under study is the first step often required to characterize speciation. A simple phylogenetic tree can be seen as the backbone of the underlying process of speciation.

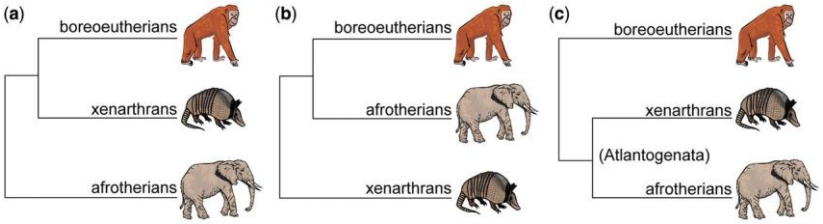


Figure 1.2: The root of the living mammals. This graph represents the three possible topologies of mammals (a) Afrotherian root, (b) Xenarthran root and (c) Atlantogenatan root. Adapted from Teeling and Hedges 2013.

These can represent simple null hypotheses, often implying a simple model of speciation. Departure from a strictly bifurcating tree can be detected and utilised to better understand speciation and detect complex patterns of gene-flow. Indeed, while incongruence can be problematic in a phylogenetic framework, it can also provide valuable information to test different models of speciation. It is possible, using coalescence theory and observed patterns of lineage sorting, to derive expectations of diverse parameters such as effective population size, divergence time and migration rate. This feature of incongruence has been applied in multiple methods used to analyse genome sequences (Li & Durbin, 2011; Mailund, Dutheil, Hobolth, Lunter, & Schierup, 2011; Yang & Rannala, 2010). So far I have only mentioned incongruence arising by chance due to stochastic lineage sorting. However, incongruence can also arise due to secondary contact between non mono-phylogenetic species (reticulation) and population substructure. In 2008 Slatkin and Pollack (Slatkin & Pollack, 2008) showed that under a simple model of divergence, there should be no asymmetric frequency of genealogies. Let's assume a species tree such as (A, (B, C)). Under a null model of no substructure or post divergence gene flow, the number of lineages (B,(A,C) and (C, (A,B) are expected to appear at roughly equal proportion in the genome. Alternatively, an excess of one or the other implies subdivision or gene-flow. This is a powerful idea, as it provides the basis to test complex models of speciation from a single genome sequence. Indeed, the near complete sampling of lineage in a genome sequence provides enough statistical power to test departure from a null hypothesis of symmetric genealogies. Such an idea was used to argue the possibility of secondary contact between Neanderthals and modern humans in Eurasia (Green et al., 2010) and other taxa (Eaton & Ree, 2013; Frantz et al., 2013; Prüfer et al., 2012). This method, also called Patterson's D-statistics (Durand et al., 2011), provides a direct way to challenge the bifurcating nature of a species tree. However, it does not allow for an

explicit statistical assessment of models of substructure versus admixture. Part of the work presented in this thesis will address this question and provides novel methods to distinguish between these hypotheses. The idea of asymmetry as a proxy to infer secondary gene-flow was then applied further to allele frequency, thus applied to multiple genomes of the same population (e.g. Pickrell & Pritchard, 2012). Instead of the pattern of lineage sorting these methods use drift asymmetry (change in allele frequency through time) to challenge population trees using allele frequency and also provide robust tools to infer complex speciation.

1.4 The Suiodae superfamily

The *Suidae* family (Order: Cetartiodactyla) (Meijaard et al. 2011), also known as boar, pigs, hog or suids form the superfamily *Suoidae* (also known as *Suiformes* or Suina), together with *Tayassudiae* (peccaries; America), share a common ancestor ~23-47 million years ago (Ma) (Gongora et al., 2011). Extant suids comprise six genera, *Sus* (domestic and wild boars) from Eurasia and Island Southeast Asia (ISEA), *Porcula* (pygmy hog) from India, *Babyrousa* (deer hog) from ISEA, *Potamochoerus* (bush pig and river hog), *Phacochoerus* (warthog) and *Hylochoerus* (forest hog) from sub-Saharan Africa. Besides, recent molecular studies (Gongora et al., 2011; Lucchini et al., 2005) the taxonomic relationship among Suidoae has typically been assessed using morphology (Orliac, 2013; Orliac, et al., 2010). However, many questions remain about the early evolution of the superfamily. Recent molecular studies lacked the power to confidently place the time of the most recent common ancestor (MRCA) of extent *Suiformes* in either the Eoecene (56-34 Ma) or the Oligocene (34-23 Ma) (Gongora et al., 2011). In addition, morphological analyses of fossils have inconclusively classified multiple Eocene fossils from Eurasia and North America as crown *Tayassuidae* or *Suidae* or as stem groups of *Suoidae* (see Orilac et al., 2010). These fossils include North American taxa (*e.g. Perchoerus*) and Eurasian taxa (Palaeochoeridae fossils such as *Doliochoerus* and *Palaeochoerus*). Thus, the monophyly of New World *Suiodae* and the possible multiple colonisation of America by suid-like species remains uncertain.

The speciation history of the genus *Sus* is also poorly known. Multiple studies have had little luck retrieving the phylogeny of these species and their mode of speciation remains fairly unknown (Larson et al., 2005; Larson et al., 2007; Lucchini et al., 2005; Randi et al., 1996). However, the genus is expected to provide an excellent model to study speciation. It comprises over 7 species (see Chapter 2) most of which live in the island of South East Asia. Already in the 19th century, Wallace recognised Island South East Asia as a natural laboratory for evolutionary

biology (Wallace 1855). Indeed the peculiar plaque tectonic (Hall, 1998) combined with large climatic fluctuations during the last million years (Zachos et al., 2001) most likely shaped the biogeography of the region (Lohman et al., 2011). Two major points need to be raised to understand the complexity of ISEA. Firstly, the region is an assemblage of multiple continental shelves that are sometime flooded by shallow seas themselves separated by deeper channels. Secondly, the great climatic fluctuations during the last few million years have greatly affected sea level through time. Indeed the sea level has risen to +30 m and decreased to -120 m (compared to contemporary sea-level) at many occasions, during the Plio-Pleistocene era (Elderfield et al., 2012). Taken together, these phenomena would have had an important impact on species-formation in this region by alternatively creating and erasing conditions for allopatric speciation. During cold periods, the sea level would reduce and led to the exposure of large continental shelves and resulted in connection between islands (sympatric or parapatric conditions), while warm period would lead to higher sea-level and disconnect islands by creating shallow seas on the low continental shelf (allopatric conditions). This process most likely resulted in the huge biodiversity that inhabits ISEA (Myers, Mittermeier, Mittermeier, Fonseca, & Kent, 2000). In the case of *Sus* these phenomena could have affected the power of previous analyses. In 2005, a team led by Greger Larson (2005) published a study showing that it is possible to geographically cluster *S. scrofa* populations (wild boar and domestic pigs) based on mtDNA. However, he showed that while these markers could be useful to characterise within *S. scrofa* phylogeography, they had almost no power to investigate inter-species relationship among ISEA species. Thus, many questions remain unanswered regarding the taxonomy and the mode of speciation of these species. These issues will be investigated at length in this work.

Another peculiarity of the genus *Sus* is its tight link to human evolution. This common evolution between the two species probably started before domestication. Indeed, pigs are large mammals that were likely hunted by early humans. This has probably resulted in many instances of human mediated translocation of pigs, especially throughout ISEA (Heinsohn, 2003). However, this phenomenon as well as its effect on the speciation of these species remains largely unknown.

1.5 A brief introduction to the concept of domestication

Before introducing the domestication history of pigs, it is necessary to define and discuss what domestication means and how it is achieved. Domestication is often associated with morphological and or behavioural changes induced by human

mediated involuntary or voluntary selection that results in direct control over breeding to improve traits that are beneficial for humans (the last step in the process). One extreme example is the reduction of brain size in domesticated animals compared to their wild ancestor, which most likely was the result of adapting the species to better control (breeding) by humans (Zeder 2012). Species that are solely captured cannot, in my opinion, be considered as domestic species.

Now that we defined domestication we can try to define the underlying process. The traditional paradigm of domestication is "**human induced selection**". However, I would like to first question how active humans were in this process. It is clear that humans played a direct role in the post-domestication selection of livestock and cultivated plants species. The question of an active participation of humans in the process of domestication itself is a key factor to obtain a comprehensive understanding of how domestication has taken place. This idea has led to the definition of two main models of domestication, the prey and commensal pathway (Vigne, 2011). The commensal pathway is a combination of first indirect human induced selection followed by directed breeding. This model has been put forward to explain the domestication of pigs and dogs (Ervynck et al., 2001; Vigne, 2011). In this model, the first step of domestication involves a habituation period. In other words, the early phase of the process does not necessarily involve an active participation by humans. Let's imagine a population of wolves that live nearby a human settlement. One could define the human settlement and the nearby surrounding as a human modified ecosystem (Vigne, 2011). Within such an ecosystem, human food wastes (*i.e.* bones) are being disposed. These can provide a source of food for nearby wild life such as our wolf population. In such a circumstance a division between wolf that are living in a human modified ecosystem (eating the scrap) and the wolf living in a natural ecosystem is to be expected. Selection may act, for example favouring wolfves that are less weary of humans, and as in complex speciation models could reduce gene-flow between the two ecotypes. However, such selection clearly does not involve direct human consent. Another model, the prey pathway, implies a constant active involvement by humans. This model has been suggested to fit herbivore domestication such as sheep, goat and cattle (Vigne, 2011). In this model humans first manage their wild herds for hunting purposes until this management involves a complete control over breeding (final step of domestication). These two models highlight the complexity of a process that remains elusive for many taxa.

1.6 Domestication history and mechanisms in pigs.

S. scrofa is a widely distributed species with an extensive range covering most of Eurasia and part of North Africa (Meijaard et al., 2011) as well as some parts of ISEA. The range of this species is wider than any other wild ancestor of any domestic animal (including wolves). Since their domestication, pigs provided a crucial source of food in earliest cities found in the Levant, probably more than ovine and bovine (Zeder 1998). However, while pigs are one of the most consumed livestock worldwide even nowadays, we know very little about their domestication. Complete and independent domestication of pigs most likely took place at least twice, once in China and once in Anatolia (Larson 2005). Archaeologists have often looked at the domestication of pigs as if it was limited in time or in space (Albarella et al., 2007). This implies a strong distinction between wild and domestic pigs and the existence of hearth of domestication such as Anatolia (Jarman 1976; Zvelebil 1995). This view is supported by ancient DNA studies that showed that Anatolian farmers transported domestic pigs from the Levant into Europe as far as Paris (Larson et al., 2007). However, the domestication history of pigs is likely to be more complex. Indeed, shortly after their introduction, European domestic pig's mtDNA haplotypes from Anatolia were replaced by mtDNA haplotypes similar to those found in modern European wild boars (Ottoni et al., 2013). Such a finding implies that gene-flow between wild and domestic forms took place multiple times and contradict the classical dichotomy of wild versus domestic pigs. This raises questions regarding the number of domestication centres and the extent to which the whole process was repeated in different part of Eurasia. Is the spread of domestic pigs in Europe the result of a transfer of ideas or the result of a transfer of genetic material from Anatolia? If this is mainly the result of a transfer of ideas, are domestic pigs a defined or a loose genetic entity? Are common morphological characteristics of domestic pigs homoplasic? In the previous section I discussed the possibility of unintentional domestication of pigs. Such a model of domestication likely had an important impact on pig domestication as it implies that the process may have started, to some extent, in many places. Thus, to fully understand the process of pig domestication it is necessary to figure out how much of the process was repeated in different part of Eurasia.

1.6 Aims and outline

In the previous paragraphs I have raised many basic questions regarding the process of speciation and domestication in general and specifically for *Suiformes*. These questions include resolving the early evolution of the superfamily during the Eocene, understanding the process of speciation of *Sus* in ISEA, understanding

complex speciation and characterising the process of domestication in pigs. I propose to use modern evolutionary genetics techniques to investigate these questions. In addition, one of the major technical advances provided in this work is the genomic perspective. Indeed previous studies have focused on morphology and a few DNA markers to investigate these issues and often lacked the necessary resolution. The basic idea of this work is to use whole genome sequences to address these questions. The large genomic resources available at the Animal Breeding and Genomics group in Wageningen University (over 300 genomes of domestic and wild *Suiformes*) provide an ideal set-up. The structure of this work is arranged chronologically. I will start by presenting a genome-scale phylogenetic tree for the *Suiformes* and a time-scale for the early evolution of this superfamilly during the Eocene and Oligocene. Thereafter, I will evaluate complex models of speciation for *Sus* and the impact that Plio-Pleistocene climatic fluctuations have had on the biogeography of ISEA. Lastly I will test multiple models of domestication and provide clues upon this elusive process. Thus, the main aim of this thesis is to provide a comprehensive evolutionary history of *Suiformes* from speciation to domestication. Secondary aims include the development and testing of methods and the refinement of speciation and domestication theories.

References

Albarella U., Dobney K., Ervynck A. and Rowley-Conwy. 2007. *Pigs and Humans 10,000 years of interaction*. Oxford University Press, Oxford.

Basset, P., Yannic, G., Brünner, H., & Hausser, J. (2006). Restricted gene flow at specific parts of the shrew genome in chromosomal hybrid zones. *Evolution; international journal of organic evolution*, *60*(8), 1718-30. Cook O. F. 1906. Factors of species-formation. *Science* 23(120), pp.506–507.

Dieckmann, U., & Doebeli, M. (1999). On the origin of species by sympatric speciation. *Nature*, *400*(6742), 354-7. doi:10.1038/22521

Dobzhansky, T. 1937. *Genetics and the Origin of Species.* Columbia University Press, New York.

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C. J., & Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings. Biological sciences / The Royal Society*, *279*(1742), 3491-500. doi:10.1098/rspb.2012.0683

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. (D. Penny, Ed.)*PLoS biology*, *4*(5), e88. doi:10.1371/journal.pbio.0040088

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, *28*(8), 2239-52. doi:10.1093/molbev/msr048

Eaton, D. A. R., & Ree, R. H. (2013). Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic biology*, syt032-. doi:10.1093/sysbio/syt032

Elderfield, H., Ferretti, P., Greaves, M., Crowhurst, S., McCave, I. N., Hodell, D., & Piotrowski, A. M. (2012). Evolution of ocean temperature and ice volume through the mid-Pleistocene climate transition. *Science*, *337*(6095), 704-9. doi:10.1126/science.1221294

Ervynck, A., Hongo, H., Dobney, K., & Meadow, R. (2001). Born Free ? New Evidence for the Status of Sus scrofa at Neolithic Çayönü Tepesi (Southeastern Anatolia, Turkey). *Paléorient*, *27*(2), 47-73. doi:10.3406/paleo.2001.4731

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368-376. doi:10.1007/BF01734359

Fitch W.M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20, pp.406-416.

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H.-J., Bosse, M., Paudel, Y., Semiadi, G., et al. (2013). Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome biology*, *14*(9), R107. doi:10.1186/gb-2013-14-9-r107

Gongora, J., Cuddahee, R. E., Nascimento, F. F. D., Palgrave, C. J., Lowden, S., Ho, S. Y. W., Simond, D., et al. (2011). Rethinking the evolution of extant sub-Saharan African suids (*Suidae*, Artiodactyla). *Zoologica Scripta*, *40*(4), 327-335. doi:10.1111/j.1463-6409.2011.00480.x

Gourbière, S., & Mallet, J. (2010). Are species real? The shape of the species boundary with exponential failure, reinforcement, and the “missing snowball”. *Evolution; international journal of organic evolution*, *64*(1), 1-24. doi:10.1111/j.1558-5646.2009.00844.x

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., et al. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710-22. doi:10.1126/science.1188021

Hall, R. (1998). The plate tectonics of Cenozoic SE Asia and the distribution of land and sea. *Evolution*.

Heinsohn, T. (2003). Animal translocation: long-term human influences on the vertebrate zoogeography of Australasia (natural dispersal versus ethnophoresy). *Australian Zoologist*, *32*(3), 351-376.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, *405*(6789), 907-13. Macmillian Magazines Ltd. doi:10.1038/35016000

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *359*(1442), 183-95; discussion 195. doi:10.1098/rstb.2003.1388

Jarman M.R. 1976. Prehistoric economic development in sub-Alpine Italy, pp.375-399. In: Seiveking G.d.G, Longworth I.H. and Wilson K.E. (eds), *Problems in Economic and Social Archeology*, Duckworth, London.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217 (210), pp.624-626.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambrigde.

Kondrashov, A. S., & Kondrashov, F. A. (1999). Interactions among quantitative traits in the course of sympatric speciation. *Nature*, *400*(6742), 351-4. doi:10.1038/22514

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic biology*, *56*(1), 17-24. doi:10.1080/10635150601146041

Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Schibler, J., Tresset, A., Vigne, J.-D., et al. (2007). 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*(39), 15276-81. doi:10.1073/pnas.0703411104

Larson, G., Cucchi, T., Fujita, M., Matisoo-Smith, E., Robins, J., Anderson, A., Rolett, B., et al. (2007). 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*(12), 4834-9. doi:10.1073/pnas.0607753104

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., Lowden, S., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science*, *307*(5715), 1618-21. doi:10.1126/science.1106927

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493-6. doi:10.1038/nature10231

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, *10*(1), 302. doi:10.1186/1471-2148-10-302

Lohman, D. J., Bruyn, M. D., Page, T., Rintelen, K. V., Hall, R., Ng, P. K. L., Shih, H.-te, et al. (2011). Biogeography of the Indo-Australian Archipelago. *Annual Review of Ecology and Systematics*, *42*, 205-228. doi:10.1146/annurev-ecolsys-102710-145001

Lucchini, V., Meijaard, E., & Diong, C. (2005). New phylogenetic perspectives among species of South-east Asian wild pig (Sus sp.) based on mtDNA sequences and morphometric data. *Journal of Zooglogy*, (266), 25-35. doi:10.1017/S0952836905006588

Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G., & Schierup, M. H. (2011). Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS genetics*, *7*(3), e1001319. doi:10.1371/journal.pgen.1001319

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome research*, *22*(4), 746-54. doi:10.1101/gr.125864.111

Meijaard E., d'Huart J.P., Oliver W.L.R. 2011. Family *Suidae* (Pigs), pp248-291. In: Wilson DE, Mittermeier RA. (eds) *Mammals of the World. Volume 2*. Lynx Edicions, Barcelona, Spain; Lynx Edicions; 2011:248-291.

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. a, Fisher, C. a, Teeling, E. C., Goodbla, A., et al. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science*, *334*(6055), 521-4. doi:10.1126/science.1211028

Muller, H. J. 1942. Isolating mechanisms, evolution, and temperature. Biology Symposium 6m, pp.71–125

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities, *403*(February), 853-858.

Noor, M. A. F., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, *103*(6), 439-44. doi:10.1038/hdy.2009.151

Nosil, P. (2008). Speciation with gene flow could be common. *Molecular ecology*, *17*(9), 2103-6. doi:10.1111/j.1365-294X.2008.03715.x

Orliac, M. J. (2013). The petrosal bone of extinct Suoidea (Mammalia, Artiodactyla). *Journal of Systematic Palaeontology*, *11*(8), 925-945. doi:10.1080/14772019.2012.704409

Orliac, Maeva J., Pierre-Olivier, A., & Ducrocq, S. (2010). Phylogenetic relationships of the *Suidae* (Mammalia, Cetartiodactyla): new insights on the relationships within Suoidea. *Zoologica Scripta*, *39*(4), 315-330. doi:10.1111/j.1463-6409.2010.00431.x

Ottoni, C., Girdland Flink, L., Evin, A., Geörg, C., De Cupere, B., Van Neer, W., Bartosiewicz, L., et al. (2013). Pig Domestication and Human-Mediated Dispersal in Western Eurasia Revealed through Ancient DNA and Geometric Morphometrics. *Molecular biology and evolution*, mss261-. doi:10.1093/molbev/mss261

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. (H. Tang, Ed.)*PLoS genetics*, *8*(11), e1002967. Public Library of Science. doi:10.1371/journal.pgen.1002967

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., Koren, S., et al. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature*, *486*(7404), 527-31. doi:10.1038/nature11128

Randi, E., Lucchini, V., & Diong, C. H. (1996). Evolutionary genetics of the suiformes as reconstructed using mtDNA sequencing. *Journal of Mammalian Evolution*, *3*(2), 163-194. doi:10.1007/BF01454360

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., & Douzery, E. J. P. (2013). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular biology and evolution*, *30*(9), 2134-44. doi:10.1093/molbev/mst116

Schliewen, U. K., Tautz, D., & Pääbo, S. (1994). Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature*, *368*(6472), 629-32. doi:10.1038/368629a0

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in ecology & evolution*, *19*(4), 198-207. doi:10.1016/j.tree.2004.01.003

Slatkin, M., & Pollack, J. L. (2008). Subdivision in an ancestral species creates asymmetry in gene trees. *Molecular biology and evolution*, *25*(10), 2241-6. doi:10.1093/molbev/msn172

Teeling, E. C., & Hedges, S. B. (2013). Making the impossible possible: rooting the tree of placental mammals. *Molecular biology and evolution*, *30*(9), 1999-2000. doi:10.1093/molbev/mst118

Thorpe, J. P. (1982). The Molecular Clock Hypothesis: Biochemical Evolution, Genetic Differentiation and Systematics. *Annual Review of Ecology and Systematics*, *13*(1), 139-168. doi:10.1146/annurev.es.13.110182.001035

Vigne, J.-D. (2011). The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. *Comptes rendus biologies*, *334*(3), 171-81. doi:10.1016/j.crvi.2010.12.009

Wallace AR. 1855. On the law which has regulated the introduction of new species. *Ann Magazine Nature History*, 26, pp.184-196.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, *18*(5), 691-9.

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9264-9. doi:10.1073/pnas.0913022107

Zachos, J., Pagani, M., Sloan, L., Thomas, E., & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, *292*(5517), 686-93. doi:10.1126/science.1059412

Zeder M.A. 1998. Pigs and emergent complexity in the Near East, pp.109-122. In: Neslon S.M. (ed), *Ancestors for the pigs*, University of Pennsylvania Museum of Archeology and Anthropology, Philladelphia.

Zeder M.A. 2012. Pathways to Animal Domestication, pp.228-259 In: Damania A. and Gepts P. (eds) *Biodiversity in Agriculture: Domestication, Evolution and Sustainability*, Cambridge University Press, Cambridge.

Zvelebil M. 1995. Hunting, gathering or husbandry? Management of food resouces by late mesolithic communities in temperate Europe. MASCA Resarch Papers in Science and Archeology. Supplement 12, pp.79-104.

2

Using next-generation sequencing to infer phylogeny and divergence times: a Suiforme (*Suiodae*: Cetartiodactyla) case study

Laurent A.F. Frantz[1], Ole Madsen[1], Yogesh Paudel[1], Hendirk-Jan Megens[1], Jamie Gongora[2], Mirte Bosse[1], Richard P.M.A. Crooijmans[1], Greger Larson[3] and Martien A.M. Groenen[1].

[1]Animal Breeding and Genomics Centre, Wageningen University, The Netherlands. [2]Faculty of Veterinary Science, University of Sydney, Australia. [3]Durham Ancient DNA and Evolution, Department of Archeology, Durham University, UK.

Submitted

Abstract

Genome sequences provide the opportunity not only to resolve the tree of life, but also to understand and characterize genome-wide conflicting evolutionary histories. However, processing next-generation short-read genome sequences requires either a reference genome for alignment or a *de-novo* assembly, the latter of which is often prohibitively expensive for large genomes. In this study we sequenced and analysed the genomes of six species representing all the genera of the Suiodae superfamily for which only distant reference genomes are available. To do so we first evaluated the performance of multiple aligners to align reads to a distant reference genome. We then tested the effect of different variant calling methods. Our results show that while local aligners perform well over large hamming distances, different methods to call variants can have strong effects on nucleotide distance to the reference. However, we show that it is possible to overcome this issue using two reference genomes. Thereafter we simulated DNA sequences with sequencing errors under multiple phylogenetic tree shapes. We found that while errors have a strong effect on phylogenetic power, these are unlikely to positively bias phylogenetic analyses. We then investigated phylogenetic support across the genome by comparing the likelihood of different trees at multiple genomic scales (2, 5 and 10kb). We show that a concatenation approach leads to overly optimistic support values, whereas a supertree approach can lead to overly pessimistic support. We show that the latter is the result of incomplete lineage sorting and lack of phylogenetic signal in small genome segments. Thus, while we empirically demonstrate the presence of ILS at shorter inter-nodes our analysis also reveals that it is difficult to divide the genome in blocks small enough to detect ILS yet long enough to keep enough signal. We expect this phenomenon to be more problematic as inter-nodes get shorter and older. Lastly we perform a thorough molecular clock analysis to time the divergence of the two families *Suidae* and *Tayassuidae*. Our results support the view that New World *Suiodae* are paraphyletic and suggest two wave of colonization of America.

Keywords: phylogenetics, genomics, lineage sorting, molecular clock

2.1 Introduction

Genome sequences offer an unprecedented opportunity to resolve the tree of life (Rokas *et al.* 2003; Gatesy *et al.* 2007) and recent advances in next-generation sequencing (NGS) technology provide the means to sequence complete genomes at an affordable price. The analysis of genomic datasets is challenging however, and processing short-read sequences into an alignment requires either a closely related reference genome or a *de-novo* assembly. Assembling large genomes of animals and plants from short-reads *de-novo* is computationally demanding, and requires very deep sequencing of multiple kinds of libraries that are onerous to construct. In addition, mapping short-read sequences to a reference genome can also be problematic if only distantly related references are available (Prufer *et al.* 2010).

Because genomes are a mosaic of genealogical histories, reconstructing species trees is not trivial even with well-assembled genomes since genealogical heterogeneity can lead to phylogenetic incongruence (*i.e.* Lee *et al.* 2012; Salichos *et al.* 2013; Yoder *et al.* 2013). Incongruence can arise from analytical limitations (*i.e.* lack of phylogenetic signal in a gene or genome segment) or from biological factors including stochastic lineage sorting and lateral gene transfer (Degnan and Rosenberg 2006; Knowles 2009).

To address these issues, researchers have adopted two primary approaches. Firstly, multiple methods have been developed to tackle incomplete lineage sorting (ILS) by computing a single species tree based on incongruent gene trees (Madison and Knowles 2006; Liu 2008; Kubakto 2009; Liu *et al.* 2009; Liu *et al.* 2010). Other studies have taken a concatenation approach (hereafter referred as supermatrix). This approach requires the compilation of hundreds or thousands of genes/loci in a single data matrix (*i.e.* Rokas *et al.* 2003; Dunn *et al.* 2008; Smith *et al.* 2010). However, the assumption that all partitions of a supermatrix evolved according to the same genealogical history is often violated. This phenomenon can result in erroneously high support for an incorrect species tree (Kubakto and Degnan 2007). In addition, because recombination breakpoints are difficult to identify in an alignment, even methods that reconstruct species trees from gene trees potentially concatenate loci with different evolutionary histories (Gatesy and Springer 2013). These issues make it difficult to divide the genome into loci that possess just a single evolutionary history, yet simultaneously possess sufficient informative sites to resolve every node in a tree.

To investigate these challenges, we sequenced and analysed the genomes of seven species of Suoidae. The *Suidae* family (Order: Cetartiodactyla) (Meijaard et al. 2011), also known as boar, pigs, hog or suids form the superfamily Suoidae (also known as Suiformes or Suina), and together with Tayassudiae (peccaries), share a

common ancestor ~23-47 million years ago (Ma) (Gongora *et al.* 2011). Extant suids comprise six genera, *Sus* (domestic and wild boars) from Eurasia and Island Southeast Asia (ISEA), *Porcula* (pygmy hog) from India, *Babyrousa* (deer hog) from ISEA, *Potamochoerus* (bush pig and river hog), *Phacochoerus* (warthog) and *Hylochoerus* (forest hog) from sub-Saharan Africa.

The taxonomic relationships among Suoidae have typically been assessed using morphological characters from fossils (*e.g.* Orilac *et al.* 2010a; Orilac 2013); though molecular studies focusing on few nuclear and mitochondrial genes (Randi *et al.* 1996; Gongorra *et al.* 2011) have also been employed. Both of these approaches have shortcomings and our understanding of the early evolution of the superfamily remains very limited. For instance, a recent molecular study lacked the power to confidently place the time of the most recent common ancestor (MRCA) of extent Suiforme in either the Eoecene (56-34 Ma) or the Oligocene (34-23 Ma) (Gongora *et al.* 2011). In addition, morphological analyses of fossils have inconclusively classified multiple Eocene fossils from Eurasia and North America as crown *Tayassuidae or Suidae* or as stem groups of *Suoidae* (*e.g.* Orilac *et al.* 2010a). These fossils include North American taxa (*e.g. Perchoerus*) and Eurasian taxa (Palaeochoeridae fossils such as *Doliochoerus* and *Palaeochoerus*). Thus, the monophyly of New World Suiodae and the possible multiple colonization of America by suid-like species remains uncertain. Narrowing down the confidence interval around the time of the MRCA of the superfamily could shed light on the evolutionary time scale of New World Suoidae (*Tayassuidae*), the status of the early fossils, provide valuable information related to the early radiation of even-toe ungulates (*Cetartiodactyla*) (Orilac et al. 2010b) and the possible multiple colonization of America by Suiodae.

Here, our first aim is to investigate how different sources of potential biases, arising from NGS can affect phylogenomic analyses. We then investigate how phylogenetic support varies across the genome and the tree by comparing results from supertree and supermatrix approaches. This allowed us to retrieve a well-supported tree for Suiodae and supplied novel insight into the evolutionary history of the superfamilly. Lastly, we performed a molecular clock analysis that allowed us to not only establish an evolutionary time scale for this superfamily, but also to test whether extant *Tayassuidae*/*Suidae* originated in the Eocene or the Oligocene.

2.2 Material and Methods

Whole genome alignment of reference genomes

In order to use both *S. scrofa* (Ssc10.2) and *B. taurus* (UMD3.1) reference genome assemblies for comparison of short-read alignments, we first conducted a whole genome alignment (WGA) between the two reference assemblies using a combination of Mercator and Mavid (Dewey 2007). Both reference genomes were downloaded from Ensembl (release 70). *Bos taurus* was chosen because it is the most closely related species to Suiodae for which a high quality draft reference genome is available. First we used Mercator, which automates BLAT (Kent 2002), to identify exon sequence similarity and builds a one-to-one orthology map between the two genomes. To do so, we used the annotation provided by Ensembl (release 70). Thereafter, we used Mavid (Bray *et al.* 2003) to align large one-to-one orthologous blocks at the nucleotide level.

DNA extraction and sequencing

DNA was extracted from blood or tissue using DNeasy blood & tissue kits (Qiagen, Venlo, NL) for seven species, *P. tajacu* (peccary), *B. babyrussa* (deer hog), *P. africanus* (warthog), *P. larvatus* (river hog), *P. porcus* (bushpig) and *S. celebensis* (Sulawesi warty pig). Quality and quantity was measured with the Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA). Libraries of ~300 bp fragments were prepared using Illumina paired-end kits (Illumina, San Diego, CA) and 100bp paired-end sequenced with Illumina HiSeq.

Short-read alignment and Genotype calling

Short-read sequences obtained from Illumina Hi-Seq were first trimmed using sickle (https://github.com/najoshi/sickle) with a minimum base quality (BQ) of 13. Reads were then aligned separately to Ssc10.2 and UMD3.1. We tried multiple aligners to test their speed and sensitivity to align short-read sequences to divergent reference genomes (Supplementary Table 2.1). We randomly selected 2,000,000 read-pairs for each species as a test dataset to explore the efficiency of different aligners. We first aligned the subsets of reads using BWA (Li and Durbin 2009). We also tested Stampy (Lunter and Goodson 2010), Bowtie2 (local / very sensible option; Langmead and Salzberg 2012) and SMALT (Ponstigl 2010). Based on this comparison we choose SMALT (k=13, s=3) to align the full data set to both reference genomes (see Supplementary Material). Local re-alignment was also performed using GATK localRealigner (McKenna *et al.* 2010).

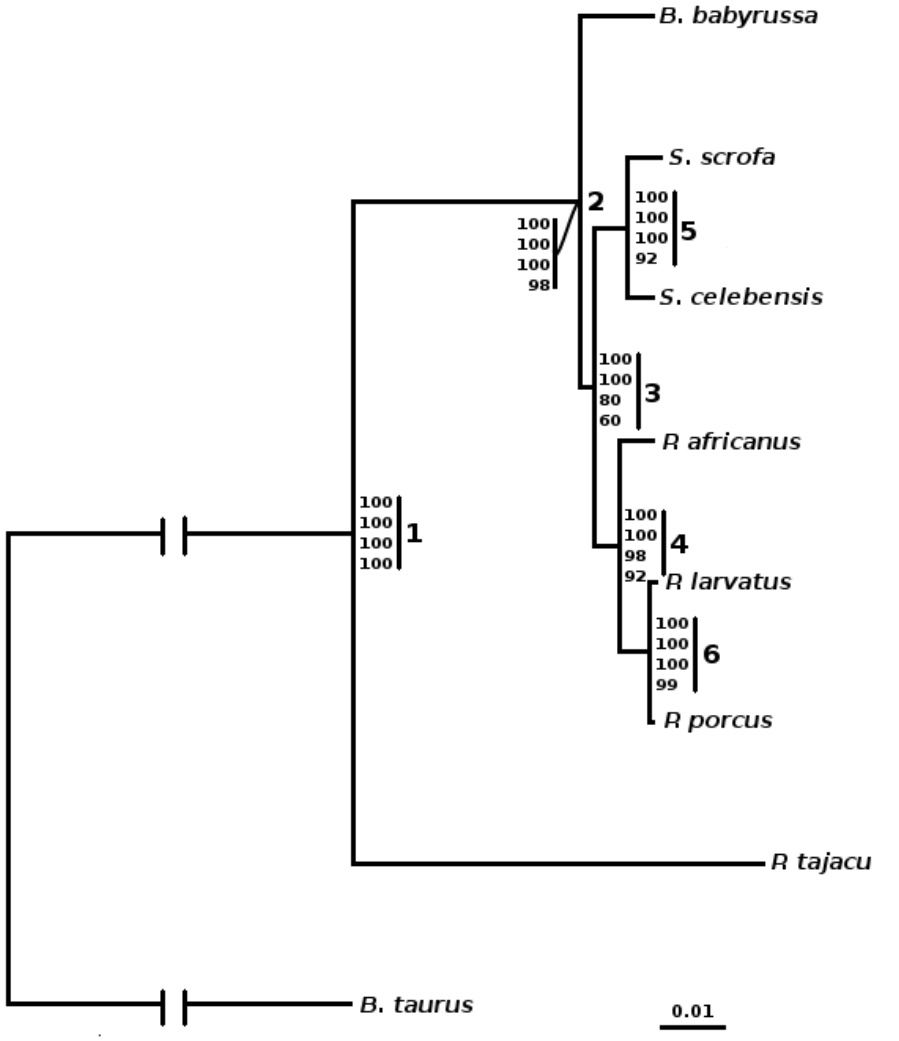


Figure 2.1: Species tree with support values from diverse analysis. The branch lengths were obtained by choosing a single random supermatrix (see Methods). Single digits (1-6) represent node numbering. The four values at each node represent the support from various analyses, supermatrix support and eMRC support for 10kb, 5kb and 2kb bins, respectively.

We extensively tested the effect of different methods to infer genotypes from short-read alignments. We inferred genotypes using a model free approach (custom; Supplementary Material) and a popular Bayesian approach as implemented in GATK (McKenna *et al.* 2010). To evaluate these methods we used a triangulation approach. More precisely, we computed the proportion of genotypes

that were identically inferred at orthologous bases (congruence) in our short read alignments to *B. taurus* and *S. scrofa* reference genomes. We also evaluated the effect of prior specification in GATK. A detailed account of these analyses is provided in the Supplementary Material.

Phylogenetic analysis

We first divided the large one-to-one orthologous regions obtained from the Mercator/Mavid alignment into consecutive bins, each spanning 10kb. In order to limit any bias arising from missing data, we required that each bin had at least 9kb (90%) covered sequence in all species. In order to investigate the effect of bin size on tree inference, we divided each 10kb interval into 2 bins of 5kb and 5 bins of 2kb. We then inferred a ML tree with 100 bootstrap replicates for each bin separately under the GTR+Γ4 model of substitution as implemented in RAxML v7.2.8 (Stamatakis 2006). We then constructed a consensus tree using the extended majority rule consensus (eMRC) as implemented in PHYLIP 3.69 (Felsenstein 1989) for each size category (10kb, 5kb and 2kb). We also used STELLS (Wu 2011) to compute a species tree using ML gene-trees obtained from RAxML. STELLS was run separately for the 10kb, 5kb and 2kb bins dataset

We then tested the supermatrix approach. We inferred a ML tree for 100 supermatrices of 1Mbp using RAxML (GTR+Γ4). Each supermatrix was generated by randomly selecting 100 of the 10kb bins, with each 10kb interval treated as a separate partition (sharing the same evolutionary history as the other partitions, but with its own model parameters).

In order to explore the relationship between bin size and phylogenetic support across the genome, we compared the likelihood of the species tree in Figure 2.1 to the likelihood of alternative topologies in different bin sizes in a subset of samples. To do so, we extracted the sequence of four taxa: *S. scrofa, P. larvatus, B. babyrussa* and *P. tajacu* from our previous 10kb bins. Each bin was then divided again into bins of 5kb and 2kb. Thereafter we computed the log likelihood (lnL) of the three possible rooted topologies in each bin: T0=(*P .tajacu*, (*B. babyrussa*, (*S. scrofa*, *P. larvatus*)), T1=(*P.tajacu*, (*S. scrofa*, (*B. babyrussa*, *P. larvatus*)) and T2=(*P. tajacu*, (*P. larvatus*, (*B. babyrussa*, *S. scrofa*)) under the GTR+Γ4 model as implemented in RAxML v7.2.8 (Figure 2.2a; thereafter referred as ILS1). For each bin we then computed the difference of the log likelihood of T0 (species tree) and T1 or T2 (alternative topologies) as:

$$(1) \quad \Delta ln = -lnT_0 + lnT_i$$

We repeated the same analysis (ILS2) on another subset of the tree, with T0: (*P. tajacu*, (*P. africanus*, (*P. larvatus*, *P. porcus*)), T1: (*P. tajacu*, (*P. porcus*, (*P. africanus*, *P. larvatus*)) and T2: (*P. tajacu*, (*P. larvatus*, (*P. africanus*, *P. porcus*)) (Figure 2.2b). Lastly, we explored the effect of sequencing and genotyping error on phylogenetic inference using the sequence simulation software package Seq-Gen (Rambaut and Grass 1997) (See Supplementary Material for details).

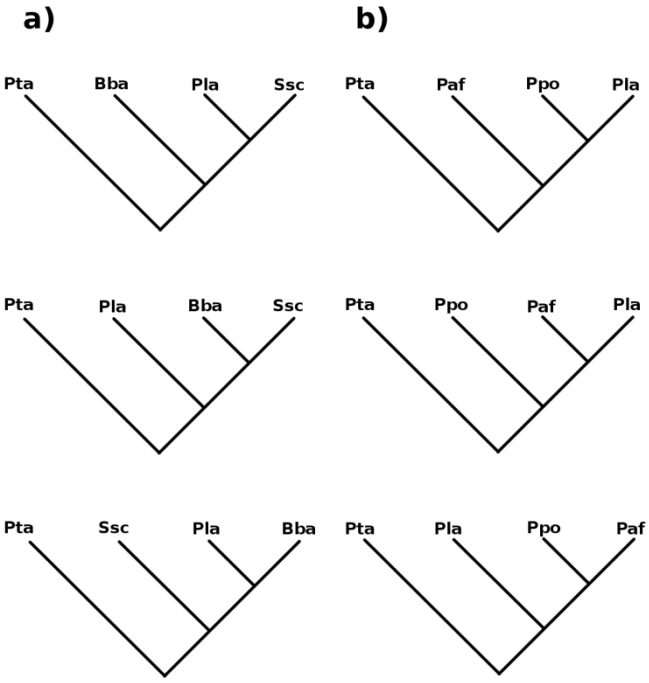


Figure 2.2: Topologies tested in lnL comparison. A) Three topologies tested in the ILS1 analysis. Species tree from Figure 2.1 (first row) and two alternative topologies (row 2 and 3). B) Three topologies tested in the ILS2 analysis. Ssc=*S. scrofa*; Pla=*P. larvatus*; Bba= *B. babyrussa*; Pta=*P. Tajacu*. Ppo= *P. porcus*; Paf= *P. africanus*.

Molecular clock analysis

Molecular clock analyses are often computationally demanding and are difficult to perform with whole genome data. Moreover, this type of analysis could be very sensitive to alignment and genotyping errors. In order to eliminate any possible bias stemming from this kind of error we only used coding sequences (CDS) since the alignment cow/pig is more reliable in these regions (see Results). We extracted the genotype of each species from our custom filtering approach (see

Supplementary Material) in the CDS of one to one orthologous genes. For each species, we masked (using Ns) any position at which the genotype call did not match between our short-read alignments to Ssc10.2 and UMD3.1. As a result, we were able to eliminate every CDS that had >10% missing data (ambiguous base calls such as N) in every species.

We used an approximate likelihood method to estimate divergence time as implemented in MCMCTREE (Yang 2007) with an auto-correlated rates clock model (clock=3 in MCMCTREE) and a HKY+Γ4 substitution model. We excluded any sites in the alignments that had an ambiguous base call in any species and used four fossil calibration points with soft bounds. For the root (Cetartiodactyla) we used a uniform distribution with minimum soft bound at 48.5 Ma and a maximum soft bound at 65.8 Ma (Meredith *et al.* 2011; do Reis *et al.* 2012).

MCMCTREE allows users to specify calibration using a heavy tailed Cauchy distributions, with a minimum (or maximum) bound age (*tL* or *tU*) and two parameters, offset (*p*) and scale (*s*=*ctL*) to model uncertainty in fossil age. For node 1 (root of Suiodae) we used a minimum soft bound to represent the earliest *Tayasuidae* fossil (peccaries). Peccary fossils are divided in three groups, 'New world' peccaries (*i.e. Cynorca*), that unequivocally appear in the fossil record in the early Miocene of North America (~20 Ma; Harris and Lui 2007), Eocene (~35Ma) North American Suiodae fossils (*i.e. Perchoerus*; Prothero 2009) and 'Old World' peccaries (*i.e. Palaeochorids: Doliochoerus and Palaeochoerus)* that appear in the fossil record 40Ma (Ducroq 1994). However, the monophyly of both New World (extent *Tayassuidae* and *Perchoerus*) and Old World peccaries (extent *Tayassuidae* and Palaeochorids) remains a source of debate (Ducrocq 1994; Van der Made 1997; Ducrocq et al. 1998; Liu 2001; Geisler and Uhen 2003, 2005; Theodor and Foss 2005; Harris and Lui 2007; O'Leary and Gatesy 2008; Prothero 2009; Spaulding et al. 2009; Orliac *et al.* 2010a; Orliac 2013). Given these uncertainties, we used a calibration with minimum age (*tL*) at 20 Ma, to represent the first unequivocal appearance of the 'New World' peccaries (*Tayassuidae*) in the fossil record. However, because of the possible earlier occurrence of peccaries, as early as 40-35Ma in Eurasia and North America (Ducroq *et al.* 1994; Prothero 2009) we used a flat prior, with a scale parameter of c=2, allowing the MCMC to explore a wide range of time for the divergence between *Tayassuidae* and *Suidae* (*tL*=2 [20Ma], *p*=0.1, *c*=2).

For node 3 (MRCA of sub-Saharan African *Suidae* and *Sus*) we used the same fossil calibration as in Frantz *et al.* 2013 and Gongora *et al.* 2011 (*tL*=0.55 [5.5Ma], *p*=0.9, *c*=0.5) (Brunet and White 2001). For node 5 (MRCA of *Sus*), we used a minimum bound at 2 Ma [*tL*=0.2 [2Ma], *p*=0.1, *c*=0.5] to represents the earliest appearance of

Sus in the fossil record of Island Southeast Asia (for detailed information about this fossil calibration please refer to Frantz et al. 2013, additional file 6). We modelled the age of non-calibrated nodes as a uniform distribution.

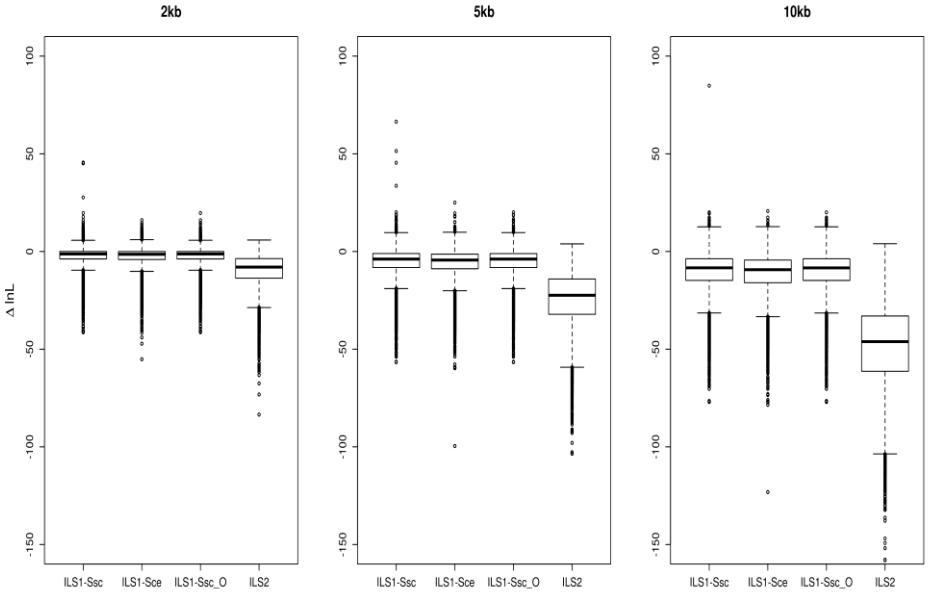


Figure 2.3: Distribution of ΔlnL for different bin sizes. Box-plots representing the distribution of ΔlnL values for 10, 5 and 2kb bin size. Labels on the X-axis specify the data-set (Figure 2.2) and ingroup taxa used (*S. scrofa* or *S. celebensis*). Each panel corresponds to a bin size. Ssc=*S.scrofa* as ingroup. Ssc_O= *S. scrofa* as ingroup with high-depth regions removed (see results); Sce=*S.celebensis*.

MCMCTREE requires further prior specification such as prior on substitution rate (rgene) and rate-drift parameter (σ2) which are modelled via a gamma distribution with two parameters, G(α [shape], β [scale]). We first estimated a rough mean for the substitution rate per unit of time over all loci, by fixing the root at 60Ma and fitting a strict clock to each locus using BASEML (Yang 2007). The mean substitution per unit of time was then used to assign a diffuse prior (α =1). The rate drift parameter was set to G(1,6). We further investigated the impact of the prior on σ2 by multiplying the parameter α by 0.1 or 10, so that both variance and mean of the distribution was reduced or increased by 10 (Inoue *et al.* 2010).
In order to increase the computational efficiency of MCMTREE, we partitioned our dataset into 10 partitions according to their substitution rate. To do so we first

estimated branch length using RAxML with a GTR+Γ4 model of substitution for each CDS and computed a relative branch length matrix. This matrix was used to perform a PCA using R. We clustered CDS using the PAM algorithm and the first two axes of the PCA (dos Reis *et al.* 2012). For each prior combination, posterior distributions were approximated using two independent MCMC runs of 10^6 samples (25% burn-in). Each MCMC analysis was also run without sequence data to ensure a reasonable prior age distribution at each node. Lastly, each run was manually inspected using TRACER (Rambaut 2009) to ensure convergence.

Previous studies have shown that the prior on rate-drift parameters can have a strong influence on the posterior age distribution (Inoue *et al.* 2010). This prior represents how substitution rates vary through time and can be difficult to specify. A recent study proposed modelling these changes as a Dirichlet process prior (DPP) (Heath *et al.* 2012). This allows each branch in the tree to be assigned a rate class regardless of the position of the branch in the tree. Uncertainty on this hyperparameter (number of rate classes) can be modeled via hyperprior (gamma distribution) allowing for a significant flexibility. A DPP can also be used as a hyperprior on the parameters for calibrating prior age density of fossils in the tree (Heath 2012). Approximation of the posterior distribution of the various parameter and hyperparameters (*i.e.* age, number of rate etc.) can be achieved via a Markov Chain Monte Carlo (MCMC) sampling.

Here, we used the program DPPDIV (Heath *et al.* 2012) that implements this approach. We used a prior mean number of rate categories of 2 and 3 for the prior on calibration clusters. The age distribution of uncalibrated nodes was modeled as a birth-death process. We used the same calibration as used in the MCMCTREE analysis. However, prior age distributions on calibrated nodes were specified as exponential distribution at nodes 3 and 5. Uniform distributions for the root (U[48, 65.8]) and node 1 (U[20, 50]) were also used. We ran two independent MCMC chains of 10^6 samples and combined the two runs using DendroPy 3.2.0 (Sukumaran & Holder 2010). Convergence was assessed using TRACER (Rambaut 2009).

2.3 Results

Alignment and genotype calling

We tested the performance of multiple aligners (BWA, SMALT, Bowtie2 and Stampy) to align short-reads to a 'foreign' reference genome using a subset of our data. Our results show that SMALT, Bowtie2 and Stampy performed well over a wide range of hamming distance (Supplementary Material). However, while Stampy provided slightly better alignment statistics than SMALT and Bowtie2, its

running time was almost 10 fold longer (Supplementary Table 2.1&2.2) making it too computationally expensive for large mammalian genomes. We therefore chose SMALT to align the full data set (Supplementary Material). Table 2.1 displays the empirical average depth of coverage for the full alignment using SMALT. Overall, the depth of coverage was higher on Ssc10.2 than UMD3.1 as expected.

Table 2.1. Mean depth of coverage from SMALT alignment. Ssc10.2 = *S. scrofa* reference genome. UMD3.1 = *B. taurus* reference genome.

	Ssc10.2	UMD3.1
S. celebensis	28	17
P. africanus	15	10
P. larvatus	11	7
P. porcus	10	7
B. babyrussa	12	9
P. tajacu	12	9

We compared different approaches to infer genotypes (Material and Methods; Supplementary Material). Our results demonstrate that the choice of prior on the probability of heterozygous calls in the Bayesian approach implemented in GATK can have a strong impact on the mismatch proportion (Supplementary Table 2.3). This phenomenon increased with distance to the reference (Supplementary Material).

We also compared the congruence of genotype calls. We defined congruence as the proportion of genotype calls that are identically called from the alignment to *B. taurus* and *S. scrofa* reference genomes. Congruence was computed for each species separately (Supplementary Material). We show that congruence was high (>97%; Supplementary Table 2.4) for both GATK (with reasonable priors) and our custom model-free approach. The congruence was higher in CDS (>99%; Supplementary Table 2.5). This suggests that some of the 3% could be attributed to miss-alignment of the two reference genomes. In a latter section we investigate how that degree of error could influence phylogenetic analyses.

Our results suggest that little reference bias is expected, even when aligning to a distant reference genome (Supplementary Material; Supplementary Table 2.4&5, 0< D <=0.002). The following phylogenetic analysis was performed using genotype calls obtained from our custom method.

Phylogenetic analysis

The data matrix used for the full phylogenetic analysis contained approximately 238 million sites and over 28 million polymorphic sites (polymorphic in the ingroup,

excluding *B. taurus*). Thus, our simple binning approach of 10kb, 5kb and 2kb resulted in approximately 1100, 550 and 220 polymorphic sites per bin, respectively. The same bins were used in our four-taxon phylogenetic analyses (ILS1 and 2, see methods for details). However, the ingroups (*S. scrofa*, *P. larvatus* and *B. babyrussa*) were more closely related. As a result, the number of polymorphic sites was significantly reduced compared to the full data set because of the exclusion of *P. tajacu* (used as an outgroup). Thus, the expected number of polymorphic sites per bin was approximately 340, 170 and 70 for bins size of 10kb, 5kb and 2kb, respectively, in the ILS1 data set (Figure 2.2a) and 175, 90 and 35 in the ILS2 data set (Figure 2.2b).

Our phylogenomic approach resulted in a well-resolved tree. The supermatrix, eMRC and STELLS approaches all supported the same tree. However, the support values were different between supermatrix and eMRC. The supermatrix analysis gave 100% support to the topology shown in Figure 2.1, while eMRC resulted in overall lower support. Moreover, support values were influenced by bin size. As expected, shorter bins resulted in weaker support for the topology shown in Figure 2.1. This is also highlighted in our STELLS analysis (Table 2.5). For this analysis we computed relative likelihood improvement as:

$$(2)\ r = \frac{(-ln_1 + ln_2)}{ln_2}$$

where lnL1 is the log likelihood of the 2nd most likely tree and lnL2 is the log likelihood of the species tree in Figure 2.1. We found that the relative likelihood improvement reduced from 0.57 in 10kb bins to 0.30 for 5kb and to 0.11 for 2kb Thus, both species tree methods (STELLS and eMRC) resulted in a weaker support compared to the support obtained with the supermatrix approach. The reduction was the strongest at node 3 (Figure 2.1) which dropped from 84% with 10kb bins to 60% in our eMRC analysis of 2kb bins. All other nodes support remained above 85% even for 2kb bins (Figure 2.1).

To further investigate why support decayed faster at node 3, we estimated the likelihood of three different topologies (Figure 2.2a; see Materials and Methods). We computed the difference in lnL (thereafter referred as ΔlnL) between the species tree (Figure 2.1) and the two alternative topologies. We repeated the same approach for node 6 (Figure 2.1&2.2b). Figure 2.3 shows the distribution of ΔlnL values for ILS1 and 2. For both ILS1 and 2, larger bins gave an overall better support for the species tree (ΔlnL<0). We found no upper outliers in ILS2, suggesting that there is no genuine incomplete lineage sorting at node 6, in comparison to node 3 (Figure 2.3). Interestingly, some upper outliers (strongly supported ILS) in bins of every size category were found using *S. scrofa* (Ssc, reference sequence) as an

ingroup (Figure 2.2a), but were not detected when using *S. celebensis* (NGS genome; Figure 2.3). This issue could arise from copy number variable (CNV) regions and/or non-annotated repetitive elements.

Table 2.2 Results of species tree analysis using STELLS. This table shows the log likelihood (lnL) of the species tree in Figure 2.1 and for the second most likely tree found by STELLS and the relative likelihood improvement (see Results) for different bin size

	lnL of tree in Figure 2.1	lnL of second most likely tree	Relative likelihood improvement (r)
10 kb	-27520.5	-43352.7	0.57
5 kb	-74061	-96789.4	0.30
2 kb	-263464	-293334	0.11

To test this hypothesis, we removed any locus that had a sequence depth higher than twice the average genome-wide depth in any species considered in this study. This approach removed the upper outliers only found when using *S. scrofa* as an ingroup (Figure 2.3; Ssc_O). However, removing regions of high depth did not remove all loci with ΔlnL values greater than 0 (Figure 2.3). This result suggests the presence of well-supported ILS rather than analytical artefact. In the ILS1 dataset, we found 975 10kb loci (9.75Mb), 2,290 5kb loci (11.45Mb) and 5,175 2kb loci (10.35Mb) that had a ΔlnL value over 2 (improvement of 2 in lnL score for an alternative topology). We found a significant increase in ILS count (considering bins with $\Delta lnL > 2$) in the 5kb dataset (5kb vs 10kb, chi-square, $X2 = 16.8$, $df = 1$, $p<0.001$) but not in the 2kb data set (2kb vs 10kb, chi-square, $X2 = 2.7$, $df = 1$, $p=0.095$). However, while some instances of ILS can be detected, there were more bins with ΔlnL values close to 0 ($-2<\Delta lnL<2$) than above 2. In the ILS1 dataset we found 4,391 10kb loci (43.91Mb), 17,335 5kb loci (85.67Mb) and 71,035 2kb loci (142Mb) with ΔlnL value close to 0. These results suggest that, while the overall lower support at node 3 reflects the presence of some ILS, it is primarily influenced by a lack of phylogenetic signal in our bins. In addition, our results show that while shorter bins increase the overall detection of ILS (especially 5kb bins), these also result in overall loss of phylogenetic signal.

We next tested which effect sequencing/genotyping errors could have on our phylogenetic results. Simulations of sequencing/genotyping errors revealed that a substantial number of sequencing or genotyping errors (5%) can affect power (Supplementary Material). This is especially true for short inter-nodes (Supplementary Figure 2). However, we show that errors are unlikely to result in

false positive topological inference (Supplementary Material). These results suggest that some of our finding described above (loss of power in 2kb at node 3) may have been affected by false genotype calls and/or sequencing errors.

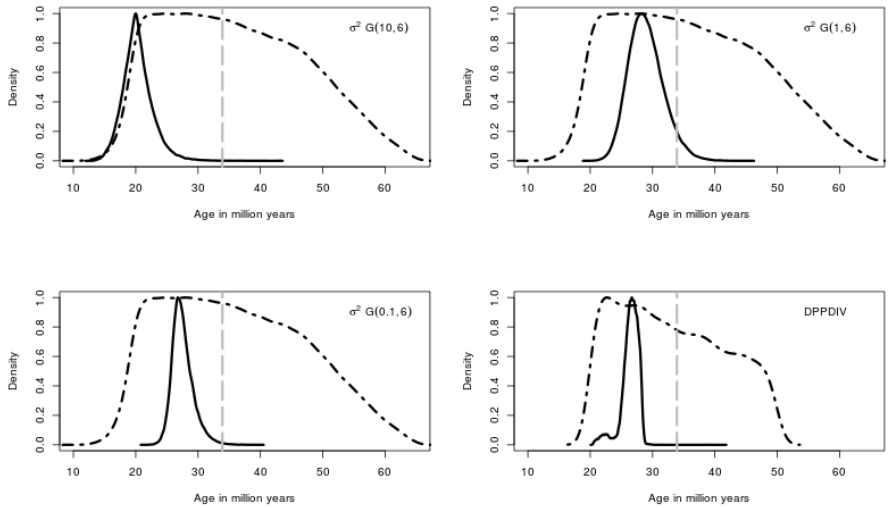


Figure 2.4: Posterior and prior age densities of the divergence *Suidae*/*Tayassuidae*. Dashed lines represent the marginal prior densities. Solid black line represents marginal posterior densities. All densities were scaled so that their maximum equals one. Vertical grey lines represent the Oligocene/Eocene boundary. The first 3 plots are derived from MCMCTREE analysis. Prior on rate-drift parameter ($\sigma 2$) are indicated on the right hand corner. The last plot is for the DPP analysis using DPPDIV.

Molecular clock analysis

Different priors for the rate-drift parameter in MCMCTREE altered the posterior time estimates (Figure 2.4; Supplementary Table 2.6). As shown previously (Inoue *et al.* 2010), we found that increasing the mean for this parameter reduced the time of divergence and vice-versa. The most affected age estimate was for the MRCA of *Tayassuidae* and Suidae. This age ranged from 20Ma (95% HPD = 16-26) to 27Ma (95% HPD= 25-35) (Figure 2.4; Supplementary Table 2.6). To further test the effect of substitution rate variation among branches, we used a DPP based approach as implemented in DPPDIV (Heath et al. 2012). Using the DPP approach to model substitution rate variation we found a divergence time of 26Ma (95% HPD = 24-28). The posterior mean number of categories of substitution estimated by DPPDIV was 2 (95% HPD=2-3). The MCMC never sampled values higher than 3, while the prior contained values up to 10.

We computed the mean logarithm of the variance in substitution rate (used as the rate-drift parameter [*σ2*] in MCMCTREE) among branches using the mean posterior value of rate estimated by DPPDIV and found that the logarithm of the variance of rates was approximately 0.028 in the DPPDIV analysis. The mean value used in the MCMCTREE runs that resulted in a younger age (*σ2* ~ G[10,6]; Fig 4.) for the divergence between *Tayassuidae* and *Suidae* was 1.6. The mean value used in the MCMCTREE run that resulted in the older divergence time (27Ma) was 0.017. Thus, the latter is much closer to the value estimated by DPPDIV. This finding likely explains the discrepancies found among the different MCMCTREE runs and the DPPDIV analysis (Figure 2.4). Nevertheless, these analyses strongly support an MRCA for *Tayassudiae* and *Suidae* more recently during the Oligocene (33.9-23 Ma) and not in the Eocene (56-33.9 Ma) (Figure 2.4). Other nodes were less affected by this prior (Supplementary Table 2.6).

Given the discrepancies among MCMCTREE runs, we discuss the results of the DPPDIV analysis (Figure 2.5). However, the discrepancies between the two analyses were limited for most nodes (Supplementary Table 2.6). The divergence between *B. babyrussa* and the other *Suidae* (sub-Saharan *Suidae* and *Sus*; node 2) most likely took place during the Late Miocene ~7.2 Ma (95% HPD= 6.7-7.9Ma). The divergence of sub-Saharan African *Suidae* and Eurasian *Sus* took place shortly after the divergence of *Babyrussa* ~5.7Ma (95% HPD= 5.8-6.1Ma). We found that the divergence between *Phacochoerus* and *Potamochoerus* (Sub-Saharan *Suidae*; node 4), took place during the early Pliocene, ~3.5Ma (95% HPD= 4-3.2 Ma). The divergence among species of the genus *Sus* was estimated to be ~2.3Ma (95% HPD= 2.5-2.1; node 5). This age is substantially younger than what was found in previous studies (~3.9 Ma; Frantz et al. 2013), however fits the time given by Gongora et al. (2011; ~2.49Ma). These discrepancies are likely the result of secondary contact between *Sus* species during the Pleistocene and the difficulty of specifying mutation rate prior distribution when fossil uncertainty is large as it is for *Sus* (Frantz et al. 2013). Lastly our analysis suggests that the divergence within *Potamochoerus* (river hog and bush pig), took place during the early / middle Pleistocene approximately 600Ka (95% HPD 450-760Ka; node 6).

2.4 Discussion

Reconstructing genomes from non-model species

Next-generation sequencing technology offers a cost effective method to identify millions of informative markers for phylogenetic analysis. Previous phylogenomic studies have used transcriptomes (*i.e.* Dunn *et al.* 2008; Hejnol *et al.* 2009; Smith *et al.* 2011), SNP arrays (*i.e.* Derek *et al.* 2010; White *et al.* 2009; Yoder *el. al.* 2013)

and RAD-Seq (*i.e.* Eaton and Ree 2013). However, few studies have used near-complete genome sequences for systematics. Consortia such as the Genome 10K project (https://genome10k.soe.ucsc.edu/), provide a large resource of assembled genomes that can be utilized to map low coverage reads to a foreign reference genome. Here we demonstrated that it is possible to reconstruct the sequence of a species using two distantly related reference genomes.

Reconstructing the genome of species from NGS short-reads alignment requires calling genotypes (or a *de-novo* assembly). We found that the choice of priors in GATK can have a strong effect on genotype calls (Supplementary Table 2.3). More precisely, our results show that the distance to the reference can be affected by the choice of prior values on heterozygous calls implemented in GATK. Little information may be available to specify this prior. This could be problematic for downstream analysis such as phylogenetics. Moreover, GATK computes the prior probability on homozygous non-reference call as half the probability of heterozygous call (set by the user). This assumption can be violated when aligning short-reads to a distant reference. However, even taking into account these possible issues, the GATK unified genotyper performed as well as our simple custom filtering scheme. We found that the congruence (Material and Methods) of genotype calls based on short read alignment to *S. scrofa* and *B. taurus* reference genomes was always above 97% for every species using both methods (Supplementary Table 2.3&2.2.4). In addition, we found that the congruence was higher in CDS than non-CDS (Supplementary Table 2.3&2.2.4). This suggests that some of these mismatches (<3%) are likely due to misalignment between the two assemblies. Lastly, our simulations showed that nucleotide miss-incorporation only affects power and did not result in wrong topological inference. This suggests that errors generated by prior miss-specification or sequencing (if random) can have an effect on difficult phylogenies with short inter-node distances.

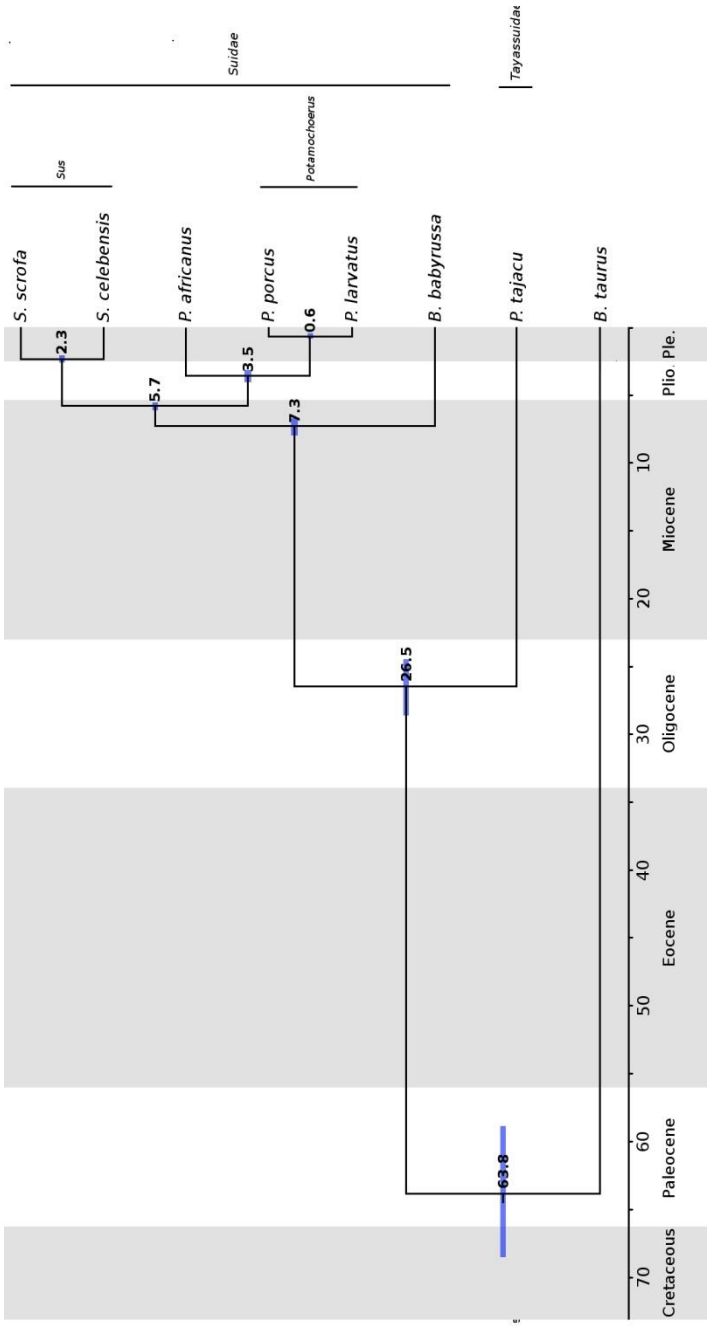


Figure 2.5: Species tree with mean posterior age. Species tree from Figure 2.1 containing time of divergence estimated using DPPDIV. Blue bars around nodes represent 95% HDP. Ages are in millions of years.

Genome-wide phylogenetic signal

The results of our phylogenetic analyses provide a well-resolved phylogeny with high support (Figure 2.1). However, our species tree analyses (eMRC and STELLS) show that the support for node 3 decreased with bin size (Figure 2.1). This phenomenon may result from ILS and/or lack of phylogenetic signal and large bins may concatenate multiple evolutionary histories. This could lead to spurious high support for a single topology (Kubakto and Degnan 2007) and explains the high support observed in our eMRC and STELLS analysis for 10kb bins at node 3 (Figure 2.1; Table 2.2). Moreover, smaller bins likely possess finer resolution and therefore may be enriched with genuinely incongruent trees resulting from, for example, ILS. Alternatively, this reduction in support could be the result of a lack of phylogenetic signal in smaller bins. To test these hypotheses, we simplified our tree taking only 3 ingroups and 1 outgroup taxa and computed the lnL of 3 different topologies for every bin size category (Fig 2&3). Our analysis showed that 5kb loci are significantly enriched with ILS compare to 10kb loci. This result suggests that the support obtained from the 10kb eMRC and STELLS analyses is an overestimation of the true support at this node due to 'unwilling' concatenation. We also investigated this phenomenon at node 6 (ILS2). We found no support for genuine incongruence at this node (Figure 2.3). Given that the inter-node preceding node 6 is much shorter than at node 3 (Figure 2.1&5), we believe that this result empirically demonstrates an excess of ILS for shorter inter-nodes.

We also identified a lack of phylogenetic signal as a prominent cause for the decrease of support at node 3. Indeed we show that the number of unresolved trees went up dramatically as we decreased bin size (Fig 3.). This shows that such reduction of support in 2-5kb bins was mainly the result of lack of phylogenetic signal. On the other hand, we show that this bin size still had enough information to resolve a younger node (ILS2; Figure 2.2&3). Taken together our results show that both ILS and lack of signal play a role in our support analysis. In addition, our results suggest that quantifying the 'real' conflict in phylogenomic data can be challenging and that reducing the size of the genomic fragments analysed, results in a higher ILS detection rate (as expected due to recombination) but also in a loss of phylogenetic signal. However, increasing the size may result in an overestimation of the support for a given node. Because recombination reduces the average size of fragments that carry a single evolutionary history at the deeper nodes, the overestimation problem is exacerbated, as inter-node intervals get shorter and deeper. Shorter inter-nodes, however, require fragments long enough to contain a sufficient number of informative sites to resolve the node.

Molecular clock analysis

The specification of priors in Bayesian molecular clock analyses can be difficult and misspecification of priors used to represent fossil age can lead to misleading results (Ho and Phillips 2008; Inoue et al. 2010). While the degree of uncertainty around the age of a fossil can be ascertained, the variation of rate among branches is more problematic. Fluctuations of demographic parameters or change in generation time but also natural selection can result in substantial variation in the time dependent substitution rate (Ho et al. 2005; Ho and Larson 2006) and information about these processes is rarely available. Thus, relaxed clocks have been developed to accommodate these fluctuations (Drummond *et al.* 2006; Yang and Rannala 2006). However, these clocks often require value specification of priors that can have a dramatic effect on the posterior distribution of the divergence time (Inoue *et al.* 2010). In this study we found that the age divergence between *Tayassuidae* and *Suidae* varied from 20Ma (95% HPD = 16-26) to 27Ma (95% HPD= 25-35) using different prior for the rate-drift parameter in MCMCTREE (Figure 2.4). However, we show that hyperpriors allow for a more flexible approach. Our results show that both MCMCTREE and DPPDIV give very similar estimates if similar variation of rate is assumed. However, the DPP approach combined with a hyperprior (Heath *et al.* 2012) has the advantage that it greatly reduces the burden of prior specification from the user. Moreover, DPP analyses provide information on the amount of substitution rate fluctuation within the tree. These findings can help to discriminate between reasonable and unreasonable priors using a more conventional approach like those implemented in MCMCTREE, increasing the confidence in molecular clock analyses.

Evolutionary history of Suiodae

This study produced a well-supported genome-wide molecular phylogeny for the Suiodae that allows us to draw conclusions related to their evolutionary history. Studies of basal Suiformes fossils suggested that both *Suidae* and *Tayassuidae* already diverged during the Eocene (*e.g.* Ducrocq 1994; Ducrocq et al. 1998; Liu 2001; Harris and Lui 2007; Prothero 2009). Indeed, multiple fossils such as *Perchoerus* (North America), *Doliochoerus* and *Palaeochoerus* (Eurasia) that share morphological features with Suoidae were found in mid-Eocene strata. These have been classified as crown *Suidae*, *Tayassuidae* or as stem Suoidae by different morphological analyses (Ducrocq 1994; Van der Made 1997; Ducrocq et al. 1998; Liu 2001; Geisler and Uhen 2003, 2005; Theodor and Foss 2005; Harris and Lui 2007; O'Leary and Gatesy 2008; Prothero 2009; Spaulding et al. 2009; Orliac 2013). The inclusion of Eocene taxa within *Tayassuidae* and/or *Suidae* suggests an early

split (>36 Ma) between the two families. In addition, the inclusion of Eurasian fossils within *Tayassuidae* implies the existence of two geographically distinct groups of peccary fossils: 1) the 'New World' peccaries appearing in the unequivocally in the Early Miocene of North America (Harris and Lui 2007) 2) the 'Old World' peccaries, which appear in the late Eocene of Eurasia (*e.g.* Van der Made 1997). Lastly, the inclusion of North American Eocene fossils such as *Perchoerus* within *Tayassuidae*, suggest single colonization event of the New World by Suiodae and an Eoecene MRCA for tayassuids and suids (Harris and Lui 2007).

Recent and comprehensive phylogenetic analyses of *Suiodae* fossils, however, argued that morphological convergence makes it difficult to resolve the basal phylogeny of *Suiodae (Orliac et al. 2010a)*. This analysis left the taxonomy of many Eocene Suiforme fossils, such as 'Old World' peccaries and *Perchoerus* (North American Eocene >36Ma), unresolved (Orliac *et al.* 2010a). However, other studies argued that mid-Eocene tayassuids (New and Old World) are basal Suiodae (Geisler and Uhen 2003, 2005; Theodor and Foss 2005; O'Leary and Gatesy 2008; Spaulding et al. 2009; Orliac 2013), implying paraphyly among New World Suiodae. In order to disentangle these hypotheses, we used a molecular clock approach and estimated the time of divergence between these two families.

Our marginal posterior distributions, from diverse analyses, showed little support for an Eocene split (Figure 2.4). Instead, we found that *Tayassuidae* and *Suidae* most likely diverged during the middle-late Oligocene or the very late Eocene (Figure 2.4). This suggests that extinct mid Eocene taxa such as 'Old World' peccaries (also known as Palaeochoeridae) are not crown *Tayassuidae* or *Suidae* but instead belong to basal Suiformes groups as argued by multiple authors (*e.g.* Harris and Lui 2007; Orliac 2013). The lack of peccary fossils that postdate the MCRCA of *Suiodae* in Eurasia supports the view that *Tayassuidae* is restricted to the American continent (Wright 1998). In addition, our divergence time estimate implies that mid Eocene fossils from North America such as *Perchoerus* are basal Suiodae. Our result suggests that tayassuids and suids would not yet have diverged during the time when *Perchoerus* fossils were deposited (at least 36Ma; Prothero 2009). This implies that *Perchoerus* are not crown tayassuids and supports the paraphyly of New World Suiodae (Geisler and Uhen 2003, 2005; Theodor and Foss 2005; O'Leary and Gatesy 2008; Spaulding et al. 2009; Orliac 2013). The paraphyly among New World Suiodae is puzzling as it implies at least two waves of colonization of America by Suiodae at a time when no land bridges existed between Eurasia and America. Lastly, clarifying the taxonomy of these early Suiformes fossils may in turn help our understanding of early radiation among

Cetartiodactlya families (*i.e. Hippopotamidae;* Orliac et al. 2010b) and help to resolve conflicting results between molecular and morphological analyses.

Our analysis reveals that the diversification of the family *Suidae* into *Babyrussa* (node 2), *Sus* and extant sub-Saharan African *Suidae* (node 3) took place during the late Miocene (Figure 2.4). Our molecular clock analysis suggests that these two splits (node 2 and 3) took place within a short time interval of approximately 1.5Ma. The short duration over which these lineages diverged likely affected lineage sorting and the absolute number of informative substitutions and the inflated rate of ILS found at this node (Figure 2.1; Figure 2.2).

Lastly, the monophyly of African *Suidae* has been contentious amongst taxonomists and the relationship among these species has been mainly assessed using morphological data (see Gongora *et al.* 2011 and reference therein). Our results revealed strong support for the monophyly of extant African *Suidae* that supports recent findings (Gongora *et al.* 2011). In addition, our analysis also demonstrated that African bush pig and river hogs (*P. larvatus* and *P. porcus*) are very closely related. We found that these taxa likely diverged during the Pleistocene approximately 600Ka (Figure 2.4), an interesting result given the distinct morphological differences between these two species.

Conclusions

Our analyses demonstrate that it is possible to use multiple reference genomes to reconstruct the sequence of species that do not yet have a reference genome using medium coverage short-reads sequences (10-25x in this study). This information can be used to reconstruct the phylogeny of a set of taxa with high confidence. Moreover, we show that identifying non-recombining fragments in a genome, which harbour a single evolutionary history, yet possess enough signal to recover the tree, is crucial to properly characterize genome-wide phylogenetic support. Thus, we argue that this task will be more challenging as the node investigated gets deeper and shorter. In addition, we show that concatenation can result in large overestimation of support values at these nodes, even using species tree methods. Together, these limitations could explain the conflicting results obtained by studies that attempt to recover short and deeper inter-nodes such as the root of Eutherain mammals (Meredith et al. 2011; Song et al. 2011; Teeling & Hedges, 2013; Romiguier et al. 2013; Morgan et al. 2013).

A complete characterization of phylogenetic support across the genome also provides valuable information about past population processes. Current and future methods at the forefront of population genetics and phylogenetics theories (i.e. STELLS, STEM, MPEST) will be able to utilize such information to reliably infer

ancestral population processes that took place tens or even hundreds of millions of years ago. In addition, we demonstrated that genome-scale data sets result in a sufficient reduction of confidence intervals around the time of divergence to allow for fine scale hypothesis testing (also see dos Reis *et al.* 2012), thereby feeding back chronological divergence information to palaeontologists. Our results support the paraphyly of tayassuids in the New World and suggests that America was colonized at least twice by Suoidae. We conclude that the on-going democratization of genome-based taxonomy will considerably improve our understanding not only of the tree of life but also of the underlying processes that have created it.

Acknowledgment:
The authors would like to thank Joshua Schraiber and Konrad Lohse for their numerous comments that greatly improve this work. This project was financially supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC Grant agreement no 249894 (SelSweep project)no. ERC-2009-AdG: 249894.

References

Altschul S.F., Gish W., Miller W., Maers E.W., Lipman D.J. 1990. Basic local alignment search tool. Journal of molecular biology. 215:403-10.

Bray N., Dubchak I., Pachter L. 2003. AVID: A global alignment program. Genome research. 13:97-102.

Brunet M., White T.D. 2001. Deux nouvelles espèces de Suini (Mammalia , *Suidae*) du continent Africain (Éthiopie ; Tchad). Comptes rendus de l'Academie des Sciences. 332:51-57.

Decker J.E., Pires J.C., Conant G.C., McKay S.D., Heaton M.P., Chen K., Cooper A., Vilkki J., Seabury C.M., Caetano A.R., Johnson G.S., Brenneman R. a, Hanotte O., Eggert L.S., Wiener P., Kim J.-J., Kim K.S., Sonstegard T.S., Van Tassell C.P., Neibergs H.L., McEwan J.C., Brauning R., Coutinho L.L., Babar M.E., Wilson G. a, McClure M.C., Rolf M.M., Kim J., Schnabel R.D., Taylor J.F. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proceedings of the National AcadeMaof Sciences of the United States of America. 106:18644-9.

Degnan J.H., Rosenberg N. a 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in ecology & evolution. 24:332-40.

Dewey C.N. 2007. Aligning multiple whole genomes with Mercator and MAVID. Methods in molecular biology 395:221-36.

dos Reis M., Inoue J., Hasegawa M., Asher R.J., Donoghue P.C.J., Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proceedings. Biological sciences / The Royal Society. 279:3491-500.

Ducrocq, S. 1994. An Eocene peccary from Thailand and the biogeographical origins of the Artiodactyl family *Tayassuidae*. Palaeontology, 37, 765–779.

Ducrocq, S., Chaimanee, Y., Suteethorn, V. & Jaeger, J.-J. 1998. The earliest known pig from the Upper Eocene of Thailand. Palaeontology, 41, 147–156.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745-9.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS biology. 4:e88.

Eaton D.A.R., Ree R.H. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Systematic biology. In press.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 5:163-166.

Gatesy J., DeSalle R., Wahlberg N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Systematic biology. 56:355-63.

Gatesy J., Springer M.S. 2013. Concatenation versus coalescence versus “concatalescence”. Proceedings of the National AcadeMaof Sciences of the United States of America. 110:E1179.

Geisler J.H., Uhen M.D. 2003. Morphological support for a close relationship between hippos and whales. Journal of Vertebrate Paleontology. 23:991-996.

Geisler J.H., Uhen M.D. 2005. Phylogenetic Relationships of Extinct Cetartiodactyls: Results of Simultaneous Analyses of Molecular, Morphological, and Stratigraphic Data. Journal of Mammalian Evolution. 12:145-160.

Gongora J., Cuddahee R.E., Nascimento F.F.D., Palgrave C.J., Lowden S., Ho S.Y.W., Simond D., Damayanti C.S., White D.J., Tay W.T., Randi E., Klingel H., Rodrigues-Zarate C.J., Allen K., Moran C., Larson G. 2011. Rethinking the evolution of extant sub-Saharan African suids (*Suidae*, Artiodactyla). Zoologica Scripta. 40:327-335.

Harris, J. & Liu, L.-P. 2007. Superfamily Suoidea. In D. R. Prothero & S. Foss (Eds) The Evolution of Artiodactyla (pp. 130–150). Baltimore, MD: John Hopkins University Press.

Heath T.A. 2012. A hierarchical Bayesian model for calibrating estimates of species divergence times. Systematic biology. 61:793-809.

Heath T.A., Holder M.T., Huelsenbeck J.P. 2012. A dirichlet process prior for estimating lineage-specific substitution rates. Molecular biology and evolution. 29:939-55.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñà J., Bailly X., Jondelius U., Wiens M., Müller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proceedings. Biological sciences / The Royal Society. 276:4261-70.

Ho S.Y.W., Phillips M.J., Cooper A., Drummond A.J. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Molecular biology and evolution. 22:1561-8.

Ho S.Y.W., Larson G. 2006. Molecular clocks: when timesare a-changin'. Trends in Genetics. 22:79-83.

Ho S.Y.W, Phillips M.J. 2009. Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. Systematic Biology. 58:367-380.

Inoue J., Donoghue P.C.J., Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Systematic biology. 59:74-89.

Kent W.J. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Research. 12:656-664.

Knowles L.L. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Systematic biology. 2009;58:463-467.

Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Systematic biology. 58:478-88.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic biology. 56:17-24.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods. 9:357-9.

Lee J.Y., Joseph L., Edwards S.V. 2012. A species tree for the Australo-Papuan Fairy-wrens and allies (Aves: Maluridae). Systematic biology. 61:253-71.

Li H. and Durbin R. 2009 Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542-3.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC evolutionary biology. 10:302.

Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Systematic biology. 58:468-77.

Liu, L.-P. 2001. Eocene suoids (Artiodactyla, Mammalia) from Bose and Yongle basins, China, and the classification and evolution of the Paleogene suoids. Vertebrata PalAsiatica, 39, 115–128

Lunter G., Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome research. 21:936-9.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Systematic biology. 55:21-30.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 20:1297-1303.

Meijaard E, d'Huart JP, Oliver WLR: Family *Suidae* (Pigs). In: Handbook of the Mammals of the World Vol 2. Edited by Wilson DE, Mittermeier RA. Lynx Edicions, Barcelona, Spain. 2011:248-291.

Meredith R.W., Janečka J.E., Gatesy J., Ryder O. a, Fisher C. a, Teeling E.C., Goodbla A., Eizirik E., Simão T.L.L., Stadler T., Rabosky D.L., Honeycutt R.L., Flynn J.J., Ingram C.M., Steiner C., Williams T.L., Robinson T.J., Burk-Herrick A., Westerman M., Ayoub N. a, Springer M.S., Murphy W.J. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science. 334:521-4.

Morgan C.C., Foster P.G., Webb A.E., Pisani D., McInerney J.O., O'Connell M.J. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Molecular biology and evolution. 30:2145-56.

O'Leary M. A., Gatesy J. 2008. Impact of increased character sampling on the phylogeny of Cetartiodactyla (Mammalia): combined analysis including fossils. Cladistics. 24:397-442.

Orliac M.J., Pierre-Olivier A., Ducrocq S. 2010a. Phylogenetic relationships of the Suidae (Mammalia, Cetartiodactyla): new insights on the relationships within Suoidea. Zoologica Scripta. 39:315-330.

Orliac M., Boisserie J.-R., Maclatchy L., Lihoreau F. 2010b. Early Miocene hippopotamids (Cetartiodactyla) constrain the phylogenetic and spatiotemporal settings of hippopotamid origin. Proceedings of the National AcadeMaof Sciences of the United States of America. 107:11871-6.

Orliac M.J. 2013. The petrosal bone of extinct Suoidea (Mammalia, Artiodactyla). Journal of Systematic Palaeontology. 11:925-945.

Ponstigl H. 2010. SMALTv 0.75 [http://www.sanger.ac.uk/resources/software/smalt/]

Prothero D.R. 2009. The early evolution of the North American peccaries (Artiodactlyla: Tayssuidae). Museum of Northern Arizona Bulletin 65 509-542.

Prüfer K., Stenzel U., Hofreiter M., Pääbo S., Kelso J., Green R.E. 2010. Computational challenges in the analysis of ancient DNA. Genome biology. 11:R47.

Rambaut A. 2009. Tracer v1.4 [http://beast.bio.ed.ac.uk/Tracer].

Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 13:235-238.

Randi E., Lucchini V., Diong C.H. 1996. Evolutionary genetics of the suiformes as reconstructed using mtDNA sequencing. Journal of Mammalian Evolution. 3:163-194.

Rannala B., Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Systematic biology. 56:453-66.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798-804.

Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular biology and evolution. 30:2134-44.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327-31.

Smith S.A., Wilson N.G., Goetz F.E., Feehery C., Andrade S.C.S., Rouse G.W., Giribet G., Dunn C.W. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*. 480:364-7.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences of the United States of America. 109:14942-7.

Spaulding M., O'Leary M. a, Gatesy J. 2009. Relationships of Cetacea (Artiodactyla) among mammals: increased taxon sampling alters interpretations of key fossils and character evolution. PloS one. 4:e7062.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688-90.

Sukumaran, J. and Mark T. Holder. 2010. DendroPy: A Python library for phylogenetic computing. Bioinformatics 26: 1569-1571.

Teeling E.C., Hedges S.B. 2013. Making the impossible possible: rooting the tree of placental mammals. Molecular biology and evolution. 30:1999-2000.

Theodor J.M., Foss S.E. 2005. Deciduous Dentitions of Eocene Cebochoerid Artiodactyls and Cetartiodactyl Relationships. Journal of Mammalian Evolution. 12:161-181.

Van der Made, J. 1997. Systematics and stratigraphy of the genera Taucanamo and Schizochoerus and a classification of the Palaeochoeridae (Suoidea, Mammalia). Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, 100, 127– 139.

White M. a, Ané C., Dewey C.N., Larget B.R., Payseur B. a 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS genetics. 5:e1000729.

Wright, D. B. 1998. *Tayassuidae*. In C. M. Janis, K. M. Scott & L. L. Jacobs (Eds) Evolution of Tertiary Mammals of North America (pp. 389–400). Cambridge, UK: Cambridge University Press.

Wu Y. 2011.Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66:763-775.

Yang Z., Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Molecular biology and evolution. 23:212-26.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 24:1586-91.

Yoder J.B., Briskine R., Mudge J., Farmer A., Paape T., Steele K., Weiblen G.D., Bharti A.K., Zhou P., May G.D., Young N.D., Tiffin P. 2013. Phylogenetic signal variation in the genomes of Medicago (Fabaceae). Systematic biology. 62:424-38.

Supplementary Material

Supplementary Methods

Genotype calling

Following the alignment step we used two different genotype calling approaches. We first used a simple scheme, selecting any allele at a site (variant or reference) covered by at least 3 reads with base quality (BQ) and mapping quality (MAPQ) higher than 20 (phred score; type 1 error p=0.01). In order to compare this simplistic approach to more model rich genotype calling methods we also inferred genotypes for reads mapped to both assemblies, using the popular GATK Unified Genotype Caller (McKenna *et. al.* 2010). GATK computes the probability of

observing every possible genotype (homozygous or heterozygous) given the reads covering the position and uses Bayes' rule to compute the posterior probability of each genotype. Thus, GATK requires the user to specify a prior on probability of observing a heterozygous position. The default value, scaled for human resequencing datasets, is set to 0.001, however this is too low for our study because the expected number of homozygous alternative calls of mismatches is much higher between our species and the reference genome of *S. scrofa* or *B. taurus*, than among humans. We therefore explored a range of priors (Supplementary Table 2.1). For the results presented in the main text, we assume a prior of p=0.01. To contrast the accuracy of these two approaches, we compared the congruence of genotypes of a single individual in 10kb bins as called on both assemblies in one-to-one *B.taurus/S.scrofa* orthologous regions identified from our WGA (see Methods).

To assess the possibility of reference bias in our query species, we computed the difference *D* between the raw nucleotide distance of *S.scrofa* vs. *B.taurus* (high-quality assembled *S.scrofa* genome) and query vs *B.taurus* (query species genotype, obtained from alignment against Ssc10.2) such as,

$$(1) \quad D = \left(\frac{M_{S,B}}{n_{S,B}}\right) - \left(\frac{M_{i,B}}{n_{i,B}}\right)$$

Where Mi,B represents the number of mismatch between species i (S = *S. scrofa*), and *B. taurus*, and ni,B represents the total number of sites considered. A positive value of D implies that the distance of our query species to the outgroup (*B. taurus*) is smaller than the distance between the two high quality reference genomes and suggests a bias toward *S. scrofa*, while a negative value would imply a bias toward *B. taurus*. We repeated the whole analysis only on coding sequence (CDS).

Simulation of sequencing errors

In order to explore the effect of sequencing/genotyping errors on phylogenetic inference, we simulated sequences from different phylogenetic tree shapes (Supplementary Figure 2.1b). We used 4 phylogenetic trees with the same total tree length, but that had different branch lengths at the same inter-nodes (from 0.01 to 0.001 substitutions/site). We then simulated one thousand 10kb alignments for 4 taxa using the software package seq-gen with a GTR+Γ4 model of substitution for each tree (4 inter-node length * 4 error rates = 16 data sets). For each alignment, we then randomly mutated 0.1, 1 and 5% of the nucleotides, using a custom perl script. We then compared the lnL (computed via RAxML) of the

topology that was used to generate the data with the lnL of 2 other alternative topologies, as done for the ILS1 and 2 data-set (see Methods).

Supplementary Table 2.1: Statistics for 2 million short-read pairs alignment against the *B. taurus* reference genome (UMD3.1). RM = read mapped. Site covered = total number of sites that had 1+ read aligned. MAPQ = mapping quality.

Species	Aligner	%RM	%RM (MAPQ>=13)	Site covered	CPU time (s)
Bbabyrussa	Bowtie2	0.50	0.21	23722682	53
Pafricanus	Bowtie2	0.53	0.21	23147709	33
Plarvatus	Bowtie2	0.50	0.20	23015173	56
Pporcus	Bowtie2	0.51	0.21	23391447	48
Ptajacu	Bowtie2	0.51	0.21	24176560	40
Scelebensis	Bowtie2	0.50	0.20	22568832	47
Bbabyrussa	BWA	0.02	0.01	2467890	74
Pafricanus	BWA	0.02	0.01	2345012	71
Plarvatus	BWA	0.02	0.01	2568302	77
Pporcus	BWA	0.02	0.02	2609595	82
Ptajacu	BWA	0.02	0.01	2487183	75
Scelebensis	BWA	0.02	0.01	2257181	77
Bbabyrussa	SMALT	0.56	0.24	32653195	58
Pafricanus	SMALT	0.55	0.23	29812010	61
Plarvatus	SMALT	0.55	0.23	31333495	69
Pporcus	SMALT	0.55	0.23	31743843	82
Ptajacu	SMALT	0.57	0.26	35256301	83
Scelebensis	SMALT	0.56	0.25	32981411	66
Bbabyrussa	Stampy	0.66	0.32	55882906	414
Pafricanus	Stampy	0.68	0.32	53761210	473
Plarvatus	Stampy	0.65	0.31	53419813	456
Pporcus	Stampy	0.66	0.32	54667244	445
Ptajacu	Stampy	0.65	0.32	55380902	496
Scelebensis	Stampy	0.67	0.32	52284930	418

Supplementary Results

Short-reads aligner benchmarking

As expected, the number of reads aligned and sites covered were much higher when aligning to *S. scrofa* than *B. taurus* (Table 2.1; Supplementary Table 2.1&2.2). In addition the distance to *S. scrofa* also influenced alignment statistics (Supplementary Table 2.2).

Supplementary Table 2.2: Statistics for 2 million short-read pairs alignment against the *S. scrofa* reference genome (Ssc10.2). RM = read mapped. Site covered = total number of sites that had 1+ read aligned. MAPQ = mapping quality.

Species	Aligner	%RM	%RM (MAPQ>=13)	Site covered	CPU time (s)
Bbabyrussa	Bowtie2	0.96	0.77	140273366	50
Pafricanus	Bowtie2	0.97	0.78	133902936	58
Plarvatus	Bowtie2	0.95	0.75	135749766	51
Pporcus	Bowtie2	0.95	0.77	138389455	44
Ptajacu	Bowtie2	0.87	0.55	91572001	50
Scelebensis	Bowtie2	0.97	0.76	132741646	82
Bbabyrussa	BWA	0.84	0.72	133990249	134
Pafricanus	BWA	0.87	0.74	134482912	154
Plarvatus	BWA	0.85	0.71	131858719	151
Pporcus	BWA	0.86	0.73	134297402	153
Ptajacu	BWA	0.09	0.06	5116554	141
Scelebensis	BWA	0.90	0.76	133631332	113
Bbabyrussa	SMALT	0.83	0.77	138620646	55
Pafricanus	SMALT	0.82	0.76	130383048	76
Plarvatus	SMALT	0.82	0.75	133689767	73
Pporcus	SMALT	0.82	0.75	135921227	71
Ptajacu	SMALT	0.75	0.59	101562481	70
Scelebensis	SMALT	0.82	0.77	133329655	93
Bbabyrussa	Stampy	0.96	0.78	143603263	581
Pafricanus	Stampy	0.97	0.77	140042305	562
Plarvatus	Stampy	0.96	0.77	140016106	549
Pporcus	Stampy	0.96	0.76	137766406	586
Ptajacu	Stampy	0.90	0.60	111649692	542
Scelebensis	Stampy	0.97	0.78	136212758	515

Indeed we can see that *P. tajacu*, the most distantly related species from *S. scrofa* displays lower mapping statistics (Supplementary Table 2.2). Overall, all aligners performed similarly when aligning against *S. scrofa,* expect BWA, that performed poorly for *P. tajacu*. Furthermore, BWA also performed poorly when aligning every species against *B. taurus*, suggesting that BWA is not a good choice when aligning very divergent data. The running time was also similar for all aligners except for Stampy.

Supplementary Table 2.3. Summary statistics for different priors settings in GATK. GATK-30 represents statistics for filter set at phred 30, all other use phred=20. GATK, default settings. GATK-01, for p(het)-0.01. GATK-05, for p(het)=0.05. GATK-1, for p(het)=0.1.

Species	Genotyper	Coverage	n_Het	n_Hom	n_N	n_SNP	mis_prop	prop_cov
P. tajacu	Custom	3885120	14895	420874	6114866	435769	0.1122	0.3885
	GATK-30	4782983	22709	457583	5217009	480292	0.1004	0.4783
	GATK-01	4787859	24121	461102	5212133	485223	0.1013	0.4788
	GATK	4785261	22699	457589	5214731	480288	0.1004	0.4785
	GATK-05	3629989	24745	487297	6370003	512042	0.1411	0.3630
	GATK-1	3380010	24872	513158	6619976	538030	0.1592	0.3380
B. babyrussa	Custom	6239493	43479	185984	3760493	229463	0.0368	0.6240
	GATK-30	6503964	45213	184997	3496028	230210	0.0354	0.6504
	GATK-01	6506457	46901	185807	3493535	232708	0.0358	0.6506
	GATK	6508930	45198	185004	3491062	230202	0.0354	0.6509
	GATK-05	5834955	47894	189902	4165037	237796	0.0408	0.5835
	GATK-1	5613008	48167	193628	4386978	241795	0.0431	0.5613
P. africanus	Custom	5652306	33031	141609	4347680	174640	0.0309	0.5652
	GATK-30	6383535	31703	147829	3616457	179532	0.0281	0.6384
	GATK-01	6386221	33346	148861	3613771	182207	0.0285	0.6386
	GATK	6389379	31700	147834	3610613	179534	0.0281	0.6389
	GATK-05	5170381	34198	153474	4829611	187672	0.0363	0.5170
	GATK-1	4832671	34445	159059	5167315	193504	0.0400	0.4833
P. larvatus	Custom	6343866	30902	164183	3656120	195085	0.0308	0.6344
	GATK-30	6647822	31874	162557	3352170	194431	0.0292	0.6648
	GATK-01	6650479	33673	163406	3349513	197079	0.0296	0.6650
	GATK	6653042	31853	162564	3346950	194417	0.0292	0.6653
	GATK-05	5929819	34583	165780	4070173	200363	0.0338	0.5930
	GATK-1	5660460	34827	168424	4339526	203251	0.0359	0.5660
P. porcus	Custom	6331344	40795	158228	3668642	199023	0.0314	0.6331
	GATK-30	6639382	40655	155901	3360610	196556	0.0296	0.6639
	GATK-01	6642658	43134	156713	3357334	199847	0.0301	0.6643
	GATK	6644618	40663	155899	3355374	196562	0.0296	0.6645
	GATK-05	5857876	44192	159322	4142116	203514	0.0347	0.5858
	GATK-1	5549174	44456	162268	4450812	206724	0.0373	0.5549
S. celebensis	Custom	6359076	29707	83352	3640910	113059	0.0178	0.6359
	GATK-30	6814926	33374	82777	3185066	116151	0.0170	0.6815
	GATK-01	6816628	34462	83428	3183364	117890	0.0173	0.6817
	GATK	6818992	33352	82777	3181000	116129	0.0170	0.6819
	GATK-05	6140971	35239	84414	3859021	119653	0.0195	0.6141
	GATK-1	5955250	35425	85880	4044736	121305	0.0204	0.5955

Supplementary Table 2.4: Results of congruence analyses for different genotype callers on the full data set. First column represents Species name-Genotyper. Third column is the proportion of sites for which the genotype inferred was identical (congruent) between the two assemblies (standard deviation estimated from 10kb block size). Fourth and fifth column are the proportion of congruent polymorphic sites called against Ssc10.2 and UMD3.1, respectively. The fifth column represents the raw nucleotide distance between the two assemblies minus the nucleotide distance of the genotype of query species and *B. taurus* (see eq. 1). The last column corresponds to the same statistics but based solely on congruent sites.

Species-GenotypeCaller	Sites called In Mb	Congruence +- 1SD	Congruence SNP Ssc10.2 +- 1SD	Congruence SNP UMD3.1 +- 1SD	D all sites +-1SD	D congruent sites +-1SD
Bbabyrussa-custom	162.20	0.985+-0.014	0.913+-0.055	0.900+-0.058	-0.002+-0.004	0.006+-0.007
Bbabyrussa-GATK	240.30	0.982+-0.016	0.888+-0.065	0.864+-0.072	-0.002+-0.004	0.011+-0.009
Pafricanus-custom	164.25	0.985+-0.015	0.905+-0.063	0.899+-0.060	-0.002+-0.004	0.006+-0.007
Pafricanus-GATK	250.79	0.981+-0.017	0.875+-0.076	0.856+-0.077	-0.002+-0.003	0.011+-0.009
Plarvatus-custom	174.27	0.986+-0.014	0.911+-0.059	0.902+-0.059	-0.001+-0.003	0.005+-0.007
Plarvatus-GATK	266.39	0.982+-0.016	0.879+-0.071	0.861+-0.075	-0.002+-0.003	0.011+-0.009
Pporcus-custom	166.55	0.985+-0.014	0.886+-0.062	0.899+-0.059	-0.002+-0.003	0.005+-0.007
Pporcus-GATK	262.06	0.982+-0.016	0.847+-0.073	0.857+-0.075	-0.002+-0.003	0.011+-0.009
Ptajacu-custom	199.59	0.979+-0.014	0.926+-0.038	0.865+-0.058	-0.003+-0.009	0.010+-0.009
Ptajacu-GATK	264.90	0.975+-0.017	0.899+-0.046	0.821+-0.070	-0.005+-0.008	0.018+-0.011
Scelebensis-custom	269.78	0.980+-0.018	0.864+-0.091	0.875+-0.071	-0.002+-0.003	0.007+-0.008
Scelebensis-GATK	335.43	0.976+-0.020	0.831+-0.106	0.832+-0.086	-0.002+-0.003	0.012+-0.011

Despite the fact that Stampy provides better alignment statistics against both *S. scrofa* and *B. taurus*, its running time is almost 10 fold higher than other aligner (Supplementary Table 2.1&2.2), which make it too computationally expensive for large mammalian genomes. SMALT and Bowtie2 performed very similarly, in term of speed and sensitivity, when aligning against *S. scrofa* (Supplementary Table 2.1&2.2). However, SMALT provided a slight improvement over Bowtie2 when aligning against *B. taurus*. Therefore, we choose SMALT to align the complete data set.

Genotype calling

We compared the effect of different priors for the frequency of heterozygous calls in GATK with our custom method (see Material and Methods) based on our alignment to *Sus scorfa* reference genome. We found that this prior can have a strong impact on the divergence (Supplementary Table 2.3). Our results show that the mismatch proportion to the reference sequence increased with the prior. This phenomenon was more pronounced for most distantly related species.

Supplementary Table 2.5: Results of congruence analysis for different genotype caller only at coding sites. First column represents Species name-Genotyper. Third column is the proportion of sites for which the genotype inferred was identical (congruent) between the two assemblies (standard deviation estimated from 10kb block size). Fourth and fifth column are the proportion of congruent polymorphic sites called against Ssc10.2 and UMD3.1, respectively. The fifth column represents the raw nucleotide distance between the two assemblies minus the nucleotide distance of the genotype of query species and *B. taurus* (see eq. 1). The last column corresponds to the same statistics but based soley on congruent sites.

Species-Genotype Caller	Site called In Mb	Congruence	Congruence SNP Ssc10.2 +- 1SD	Congruence SNP UMD3.1 +- 1SE	D all sites +-1SE	D congruent Sites +-1SE
Bbabyrussa-custom	10.18	0.998+-0.014	0.984+-0.081	0.980+-0.085	-0.000+-0.012	0.000+-0.013
Bbabyrussa-GATK	11.79	0.997+-0.015	0.981+-0.089	0.968+-0.124	-0.000+-0.011	0.001+-0.015
Pafricanus-custom	7.87	0.998+-0.013	0.978+-0.094	0.976+-0.093	-0.001+-0.012	-0.000+-0.014
Pafricanus-GATK	10.38	0.997+-0.015	0.982+-0.087	0.965+-0.138	-0.000+-0.010	0.001+-0.013
Plarvatus-custom	11.31	0.998+-0.014	0.982+-0.086	0.979+-0.085	-0.000+-0.010	0.000+-0.013
Plarvatus-GATK	12.83	0.997+-0.015	0.978+-0.097	0.967+-0.127	-0.000+-0.010	0.001+-0.014
Pporcus-custom	11.02	0.997+-0.015	0.972+-0.106	0.975+-0.092	-0.001+-0.011	-0.000+-0.013
Pporcus-GATK	12.72	0.997+-0.015	0.970+-0.109	0.963+-0.134	-0.001+-0.010	0.001+-0.014
Ptajacu-custom	11.71	0.997+-0.016	0.986+-0.067	0.978+-0.089	-0.001+-0.022	-0.000+-0.023
Ptajacu-GATK	12.92	0.997+-0.017	0.978+-0.089	0.964+-0.125	-0.001+-0.022	0.002+-0.024
Scelebensis-custom	10.89	0.997+-0.017	0.977+-0.101	0.973+-0.102	-0.000+-0.009	-0.000+-0.012
Scelebensis-GATK	12.33	0.996+-0.017	0.978+-0.100	0.959+-0.141	-0.000+-0.008	0.000+-0.014

For example, the mismatch proportion for *P. tajacu* (most distantly related species to *S. scrofa*) varied from 0.1 using default settings (p[het]=0.001) to almost 0.16 using a prior of 0.1. Prior settings had less effect for less divergent species, for

example, the mismatch proportion for *S. celebensis* (most closely related species to *S. scrofa*) varied from 0.017 with default settings to 0.02 with p(het)=0.1. These different priors also yielded different coverage. We found that increasing p(het) slightly reduced the overall sequence coverage in every species (Supplementary Table 2.3). Thus, lower sequence coverage and a higher number of variant calls for higher values of p(het) probably result in a higher mismatch proportion.

Supplementary Table 2.6: Molecular clock analysis prior sensitivity. Each row contains the mean posterior age (10My) of a clade and the 95% HDP in bracket. The first three columns were obtained by merging two MCMCTREE runs using different priors for the rate-drift parameter. The shape and scale parameter of the gamma distribution used as prior are indicated for each column. Estimates in the last column were by merging two DPPDIV runs.

	MCMCTREE G(10,6)	MCMCTREE G(1,6)	MCMCTREE G(0.1,6)	DPPDIV
Root	5.10(4.72,5.)	6.09(5.17,6.55)	6.34(5.79,6.61)	6.38(5.89,6.85)
Node 1	2.03(1.60,2.58)	2.89(2.40,3.51)	2.75(2.47,3.16)	2.65(2.44,2.8575)
Node 2	0.60(0.48,0.77)	0.67(0.54,0.87)	0.58(0.51,0.72)	0.72(0.67,0.7942)
Node 3	0.49(0.39,0.63)	0.53(0.42,0.68)	0.45(0.40,0.56)	0.57(0.58,0.6102)
Node 4	0.31(0.24,0.41)	0.32(0.26,0.42)	0.27(0.24,0.34)	0.35(0.32,0.4019)
Node 5	0.24(0.19,0.31)	0.24(0.19,0.31)	0.21(0.18,0.26)	0.23(0.20,0.2564)
Node 6	0.05(0.04,0.08)	0.05(0.03,0.07)	0.04(0.03,0.05)	0.05(0.04,0.0759)

The overall congruence between the genotypes as called based on alignment to the *B. taurus* and *S. scrofa* reference assemblies was quite high (>97%; Supplementary Table 2.4). Moreover, the distance to *S. scrofa* barely affected the results. For example, *P. tajacu*, the most divergent species in our study, displayed on averaged slightly less congruence than other species, but these differences were marginal (Supplementary Table 2.4). Overall congruence was also slightly affected by the choice of the genotyper. Our simple scheme (see above) gave marginally higher congruence. However, these differences are likely the result of differences in overall sites called by the genotype caller (Supplementary Table 2.4). Congruence

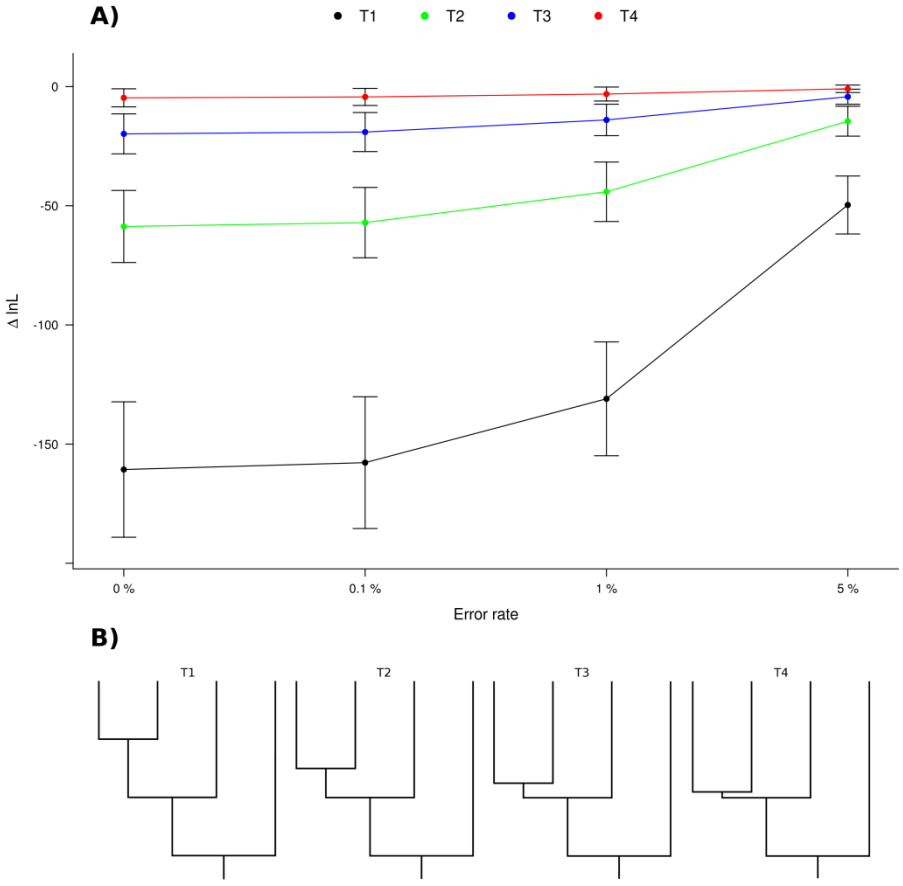
scores of genotypes at SNP sites, on both assemblies, were slightly lower than the overall congruence. Moreover, congruence at SNP sites was also on average lower when called on the UMD3.1 assembly than on Ssc10.2. Lastly, congruence at SNPs, called on Ssc10.2 was not affected by distance to *S. scrofa,* as illustrated by the higher congruence displayed by *P. tajacu* than by *S. celebensis* (the most closely related taxa to *S. scrofa*). Our metric of reference bias, *D*, (equation 1.), was very close to 0 when considering all sites called on alignments to the *S. scrofa* reference genome, suggesting little reference bias. Values for *P. tajacu* were marginally higher than for other taxa (0.003-0.005 vs 0.002-0.001). However, the SD values were higher in *P. tajacu* for each genotyping method (0.03-0.04 vs 0.009-0.008). In addition, the choice of genotyper did not seem to have much influence on this statistic (Supplementary Table 2.4). Indeed both mean and SD were the same for all individuals except for *P. tajacu* (Supplementary Table 2.4).

We found that the congruence (on SNPs and overall), in the CDS region was overall higher than when considering the total autosome dataset (Supplementary Table 2.4&5). *D* (see Material and Methods: eq 1.), using all sites called on the Ssc10.2 assembly, was marginally lower on the CDS data than on the overall data-set (Supplementary Table 2.4&5). Interestingly, *D*, computed on congruent sites was very close to zero when considering only the CDS but was overall positive when considering all sites (non-coding and coding; Supplementary Table 2.4&5). This result is expected as it suggests that our WGA between *S. scrofa* and *B. taurus* was more reliable in CDS than non-CDS regions.

Simulation of sequencing/genotyping error

Although, our results show that genotyping errors are not biasing the raw nucleotide distance to the outgroup, it is still possible that genotyping errors may play a role in the overall decrease in support observed in our phylogenetic analysis. To test this hypothesis we simulated 4000 DNA alignments, using seq-gen (Rambaut and Grass 1997), based on 4 trees with different inter-node lengths and applied different levels of random errors (0.1, 1 and 5%; Figure 2.4; see methods). We then compared the lnL score, obtained from RAxML under the GTR+Γ4 model, of the topology used to generate the alignments and two other alternative topologies. As expected, shorter inter-nodes resulted in overall lower power to distinguish the real topology from the two alternatives, as we previously demonstrated with empirical data (Figure 2.2). In addition, limited amounts of errors (0.1-1%) only marginally affected the power to differentiate these topologies (Supplementary Figure 2.1). The mean ΔlnL for inter-node length of 0.05 (T1) decreased from -160.7 (no error) to -130 (1% error) and decreased from -4.7 to -3.1

for inter-node length of 0.001 (T4). However, large amounts of errors (5%) substantially affected the power to distinguish between topologies for every category of inter-node length. The mean ΔlnL for inter-node length T1 (Supplementary Figure 2.1) decreased from -160.7 to -49.7, and from -4.7 to -0.9 for inter-node length T4 at 5% error rate. As the overall support decreased, the



Supplementary Figure 2.1: Results of error simulation. a) Overall support for the topologies in a. with different level of errors (x axis) and different inter-node lengths. Dots represent mean and black bar represents 1SD (obtained from 1000 replicates). b) Four different trees that were used to simulate sequences.

number of loci with ΔlnL > 0 only marginally increased. For example, in the dataset with the shortest inter-node length (T4) the number of 'supported' (ΔlnL>2) ILS

increased from 1/1000 (no-errors) to 14/1000 (5% error). This finding demonstrates that, while errors will result in overall decrease of support, it only marginally increases the false positive rate of ILS detection. Lastly, we computed correlations between error rate and ΔlnL for T1 and T4. We found that the correlation coefficient was almost twofold higher for D1 (Kendall's rank correlation, tau=0.60, $p<0.001$) than for D4 (Kendall's rank correlation, tau=0.35, $p<0.001$). This result suggests that errors result in a faster decrease in power for longer inter-nodes. This is expected, because longer inter-nodes result in a larger fraction of polymorphic sites, thus providing more targets for sequencing / genotyping errors.

3

Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*

Laurent AF Frantz [1], Joshua G Schraiber [2], Ole Madsen [1], Hendrik-Jan Megens [1], Mirte Bosse [1] , Yogesh Paudel [1],Gono Semiadi [3], Erik Meijaard [4,5], Ning Li [6], Richard PMA Crooijmans [1], Alan L Archibald [7], Montgomery Slatkin [2], Lawrence B Schook 8, Greger Larson [9] and Martien AM Groenen [1].

[1] Animal Breeding and Genomics Centre, Wageningen University, Droevendaalsesteeg 1, Wageningen, 6708 PB, The Netherlands. [2] Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA. [3] Puslit Biologi LIPI, Jl. Raya Jakarta-Bogor Km. 46, Cibinong 16911, Jawa Barat, Indonesia. [4:] People and Nature Consulting International, Vila Lumbung House no. 6, Jl. Kerobokan Raya 1000x, Badung 80361, Bali, Indonesia [5]: School of Archaeology & Anthropology, Australian National University, Canberra ACT 0200, Australia. [6:] State Key Laboratory for Agrobiotechnology, China Agricultural University, Beijing 100193, PR China. [7]: The Roslin Institute and R006Fyal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK. [8]: Department of Animal Sciences, University of Illinois, Urbana-Champaign, Illinois 61801, USA. [9]: Durham Evolution and Ancient DNA, Department of Archaeology, Durham University, Durham DH1 3LE, UK.

Genome Biology (2013) 14:R07

Abstract

Elucidating the process of speciation requires an in-depth understanding of the evolutionary history of the species in question. Studies that rely upon a limited number of genetic loci do not always reveal actual evolutionary history, and often confuse inferences related to phylogeny and speciation. Whole-genome data, however, can overcome this issue by providing a nearly unbiased window into the patterns and processes of speciation. In order to reveal the complexity of the speciation process, we sequenced and analysed the genomes of 11 wild pigs, representing morphologically or geographically well-defined species and subspecies of the genus Sus from insular and mainland Southeast Asia, and one African common warthog. Our data highlight the importance of past cyclical climatic fluctuations in facilitating the dispersal and isolation of populations, thus leading to the diversification of suids in one of the most species-rich regions of the world. Moreover, admixture analyses revealed extensive, intra- and inter-specific gene-flow that explains previous conflicting results obtained from a limited number of loci. We show that these multiple episodes of gene-flow resulted from both natural and human-mediated dispersal. Our results demonstrate the importance of past climatic fluctuations and human mediated translocations in driving and complicating the process of speciation in island Southeast Asia. This case study demonstrates that genomics is a powerful tool to decipher the evolutionary history of a genus, and reveals the complexity of the process of speciation.

Keywords: speciation, genomics, gene-flow, phylogenetics

3.1 Introduction

The diversity of life on Earth owes its existence to the process of speciation. The emergence of genetic techniques has allowed the relationships amongst hundreds of species to be investigated, and DNA studies have been invaluable in resolving long-standing taxonomic and phylogenetic questions [*i.e.* 1, 2]. The use of limited numbers of genomic markers, however, can result in misleading impressions of the phylogenetic relationships between organisms [3]. In addition, traditional bifurcating trees are constructed on the presumption that little or no gene-flow occurs following a split between two species, though gene-flow has been shown to occur during the splits between species [4, 5]. The recent advent of high-throughput sequencing allows inferences to be drawn from near-complete genomes, in turn offering an unprecedented understanding of organismal evolutionary history. The commensurate increase in resolving power has allowed numerous questions to be addressed including those related to genomic structure, deep phylogenetic relationships, the genetic variation responsible for specific phenotypes, and hybridisation patterns between ancient hominids [6, 7]. Few studies, however, have taken advantage of complete genomes to investigate the process of speciation.

Wallace [8] first recognized that Island Southeast Asia (ISEA) is an ideal natural laboratory to study speciation. Over the past 50 My (million years) tectonic activity has considerably altered the geography of this region. In addition, large-scale climatic fluctuations beginning in the early Pliocene [9] affected the region's biogeography [10]. Successive glacial and interglacial periods lowered and raised sea levels thus, alternately separating and connecting large landmasses. During cold periods, the Malay Peninsula, Borneo, Sumatra and Java formed the contiguous landmass known as Sundaland (Figure 3.1A), while in warmer periods these islands were isolated from each other. These alternating climatic conditions required frequent adaptation and induced intermittent allopatric and parapatric speciation processes. The fluctuations also created an ideal environment for diversification that has resulted in a complex and species-rich assemblage [10]. The development of models which explain the process of speciation in ISEA has been further complicated by anthropogenic factors that have influenced the dispersal and distribution of numerous species in the region [11].

The five biodiversity hotspots found in ISEA and Mainland Southeast Asia (MSEA) [12] are host to at least seven morphologically defined species of pig in the genus *Sus* [13]. Aside from *Sus scrofa (Eurasian wild boar and domestic pigs)*, which is distributed across most of Eurasia and parts of northern Africa, all other species of the genus *Sus* are restricted to MSEA and ISEA (Figure 3.1A). Because these species

are still capable of interbreeding and producing fertile offspring [14], the genus *Sus* presents an excellent model to study on-going speciation. Moreover, previous studies have found discrepancies between and among the phylogenies inferred from morphological and mtDNA marker [13, 15, 16]. Thus, the phylogeny of these species remains controversial. These discrepancies could be explained by either gene-flow between sympatric populations of different species, or by a rapid radiation that would have left little power to resolve the phylogeny.

The lack of post-zygotic reproductive barrier in pigs is not an isolated case. Indeed, many vertebrate taxa, recognized as different species, can still interbreed and produce fertile offspring. For example, it has been claimed that approximately 6% of European mammalian species can interbreed with at least one other species [17]. Additionally, while most of these species are young, there are examples of interbreeding species of birds that diverged over 55Mya [18]. Given the ease with which numerous closely related (and some distantly related) species can interbreed, it is important to develop and test methods that are not only robust to inter-specific gene-flow, but can also identify it. Speciation with gene-flow is expected to result in a richer phylogenetic history including periods of divergence (bifurcations) and periods of secondary contact (reticulations), and thus should leave genomic signatures.

In order to investigate the speciation history of these suids, and to assess the usefulness of whole-genome sequences to infer complex evolutionary histories, we sequenced and analysed the complete genomes of 11 individual pigs representing five *Sus* species and an African common warthog (*Phacochoerus africanus*; Additional file 1, Table S1). Our analysis of these 11 genomes demonstrates the power afforded by genomics to resolve a complex and controversial evolutionary history involving multiple reticulation events.

3. Speciation with gene-flow in *Sus*

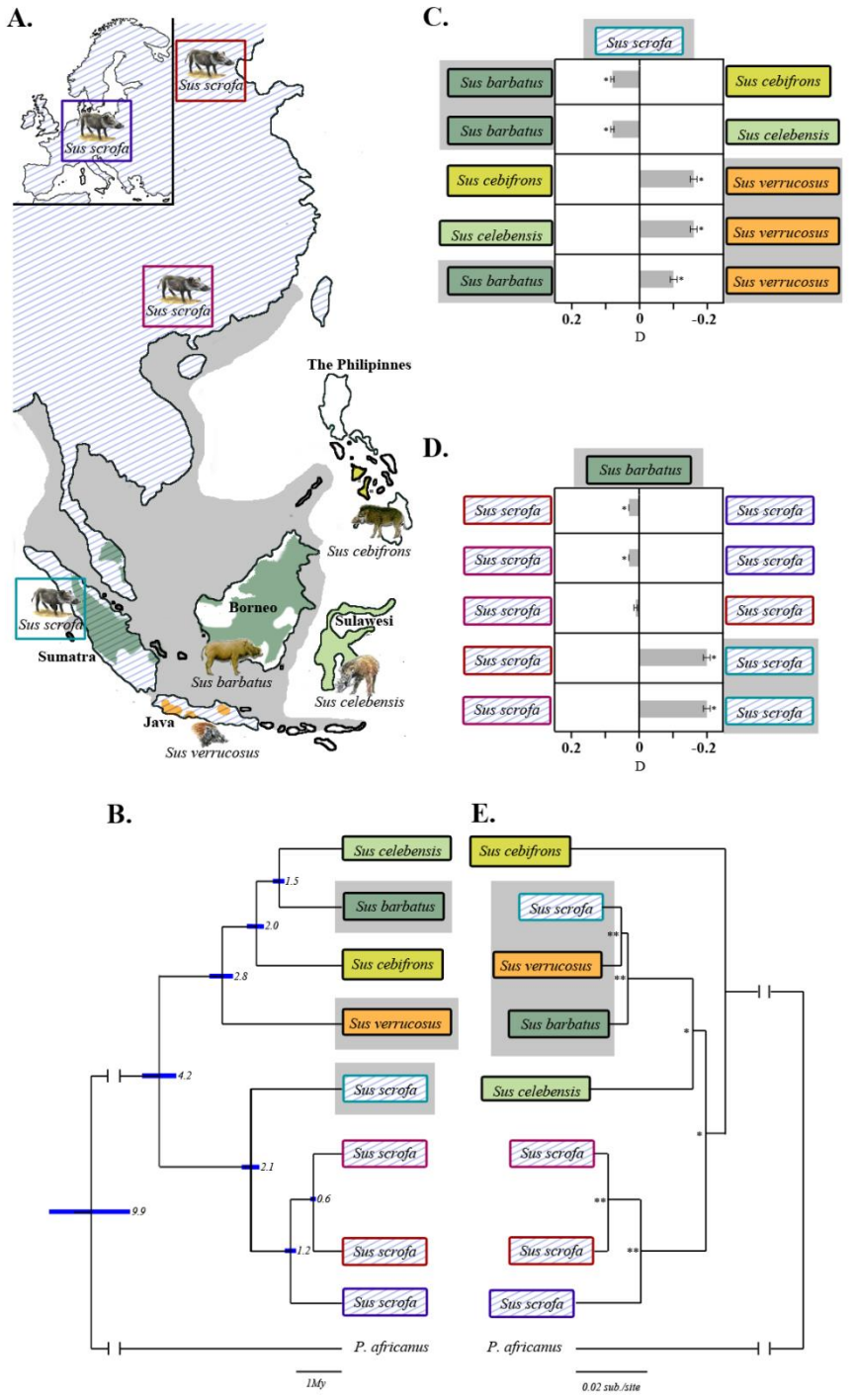


Figure 3.1: Geographic distribution, phylogenetic relationships and admixture between *Sus* lineages. A. A map of Island and Mainland Southeast Asia depicting the modern distributions of five *Sus* species. The grey shaded area represents the maximum geographical extent of Sundaland during periods of low sea level. **B.** Phylogenetic relationships among *Sus species* inferred from nuclear DNA. Node labels show age in My and 95% confidence interval. Grey shading highlights taxa living on Sundaland **C&D.** A diagram depicting the excess derived allele sharing when comparing sister taxa and outgroups. Each row contains the fraction of excess allele sharing by a taxon (left/right) with the top label/outgroup (*S. scrofa* or *S. barbatus*) relative to its sister taxon (left/right). The grey bar points in the direction of the taxon that shares more derived alleles with the outgroup than its sister taxon, and its magnitude indicates the amount of excess (D). Black bars represent 1 SE and stars indicate D values significantly different from 0 ($p<0.01$; see Material and Methods). **E.** A mtDNA Bayesian phylogenetic based tree with node labels that represent posterior probabilities (* > 0.85; ** = 1).

3.2 Results

SNP discovery and general divergence pattern across the genomes.

We aligned between 153 and 566 million reads per sample to the *Sus scrofa* reference genome (Sscrofa10.2) [19], resulting in an average read depth of 7.5 to 24x (Additional file 1, Table S2 ; Materials and methods). The number of SNPs discovered in each genome sequence (Additional file 1, Table S2) was higher in the *Sus* species than between *Sus scrofa* individuals, most of which were fixed differences between the *Sus scrofa* reference genome and the other species analysed. In order to understand how substitution rate within the genus varies across the genome, we computed the average sequence divergence from the Warthog to each *Sus* species in 1 Mb windows (see Material and methods). Our results demonstrated that the average sequence divergence to the outgroup (Warthog) was positively correlated with recombination rate (as estimated in *Sus scrofa*, [20]; tau=0.40, $p<0.001$) suggesting a relationship between recombination and divergence rate, as observed in other mammals [21, 22].

Phylogenomic analysis

Using near complete genome sequences, we applied several phylogenomic methods based on Maximum Likelihood (ML) implemented in RAxML 7.2 [23]. We used both supertree and supermatrix techniques (see Materials and Methods for details). Briefly, the supertree methodology involves computing a single tree per genomic locus in combination with an *ad-hoc* reconstruction of a consensus

phylogeny from the single trees whereby the stochastic behaviour of lineage sorting can be taken into account. In the supermatrix framework, a single tree is inferred from multiple loci assembled in multiple partitions.

We first identified regions in the genome, spanning a minimum of 5 kbp, that possessed less than 10% missing data (due to filtering) in all our samples (see Material and methods for details; Additional file 1, Table S3). We then built phylogenetic trees for every genomic bin identified and obtained a species tree using the supertree method STAR [24]. We also used a concatenation method by building multiple supermatrices. One hundred supermatrices, each spanning 1Mbp, were assembled by randomly joining genomic bins. We then computed a phylogenetic tree using RAxML, with 100 fast bootstrap replicates, for each supermatrix.

We found that the species tree topology depicted in Figure 3.1B was the most common across all of the genomic bins analyzed (Additional file 2, Figure S1), but several alternative topologies appeared in substantial numbers (Additional file 3, Table S5). This result is to be expected and can be caused by incomplete lineage sorting (in which deep coalescences occur in ancestral populations) and gene-flow (in which some genealogies cross species boundaries). The presence of such incongruence is created when recombination creates local gene trees; hence, we looked for a correlation between recombination rate and the frequency of alternative topologies. We found a positive correlation between mean pairwise Robinson-Foulds distance and recombination rate in 1Mbp windows ($tau=0.53$, $p<.001$; Materials and Methods). We also found a positive correlation with mean divergence to the outgroup ($tau=0.40$, $p<.001$). Together, these results suggest the importance of recombination in shaping the genomic landscape of speciation in suids.

To compare our results to earlier studies using mitochondrial DNA (matrilineal lineage), we carried out a Bayesian phylogenetic analysis using near-complete mitochondrial genomes (Materials and Methods). The resulting topology is consistent with previous studies [15, 16, 25] and shows a clear discordance with the phylogenetic tree obtained from autosomal chromosomes (Figure 3.1B & E). This discordance is expected given the wide range of topologies found in the autosomes, especially because mitochondrial DNA represents only one locus with no recombination.

The phylogenetic discordance found within the genome and between nuclear and mtDNA, could be the result of either incomplete lineage sorting (ILS) or post-divergence gene-flow.

Divergence time and admixture analysis.

In order to differentiate between ILS and gene-flow, we conducted an independent admixture analysis (using D-statistics) that directly addressed this issue [26] (see Materials and methods; Additional file 4). Overall, we found strong evidence of admixture among species living on Sundaland. Indeed, results of D-statistics (Material and Methods; Additional file 4; Additional file 5, Table S8) show that species living on Sundaland share a significant excess of derived alleles compared to what would be expected for a simple bifurcating scenario, as displayed in Figure 3.1B&C. In addition, we found further admixture signatures that involve species living outside of Sundaland. For a detailed discussion of these results, please refer to Additional file 4.

To put the admixture and divergence events in a temporal context, we first estimated molecular divergence times using a relaxed molecular clock as implemented in MCMCtree [27]. In order to account for the uncertainty in fossil dates, we used three separate fossil calibrations to place prior distributions on node age (see Additional file 6 for further discussion and references on the fossil calibrations used in this study). We then selected genomic loci supporting the main topology to obtain the date of original divergence between taxa (Figure 3.1B) thereby limiting the bias that arises from admixture between species (Additional file 4; Additional file 5, Table S8).

The correlation between the timing of the nodes on the phylogenetic tree and climate models [28] suggested that when global sea levels dropped during cold intervals, the resulting land bridges between islands allowed pigs to disperse across what were once sea barriers (Figure 3.1A; Figure 3.2). Warm periods raised sea levels, closed migration routes and isolated populations on individual islands leading to allopatric speciation. In addition, our admixture analysis revealed the existence of extensive inter-specific gene-flow that likely took place during cold intervals since these periods would have induced parapatric conditions via the connection of previously isolated islands.

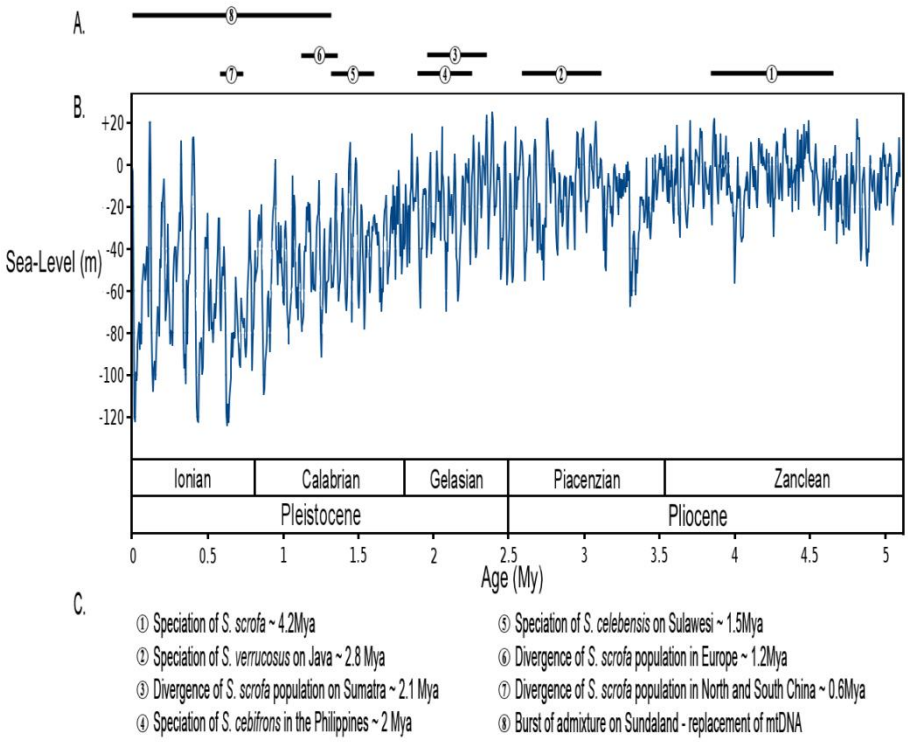


Figure 3.2: A eustatic curve adapted from ref. 18. A. Each black bar shows 95% CI of each divergence events as inferred from molecular clock analysis (see Figure 3.1B). **B.** Eustatic curve for the last 5 My. **C.** Legend of events represented as black bars in Figure 3.2A.

Demographic analysis

We used heterozygous SNP calls for demographic inference in a single individual genome sequence as implemented in PSMC (see Methods; Figure 3.3; Additional file 7, Figure S3). We found that the Pleistocene period led to a bottleneck in both ISEA (Figure 3.3) and MSEA populations (Additional file 7, Figure S3). These population size declines are consistent with the reduction of temperature observed during this period that would have reduced the overall forest cover in MSEA and ISEA [29, 30] (Figure 3.2). In addition, our results suggest that the populations from ISEA (Figure 3.3) have undergone a more severe bottleneck than populations of MSEA (Additional file 7, Figure S3).

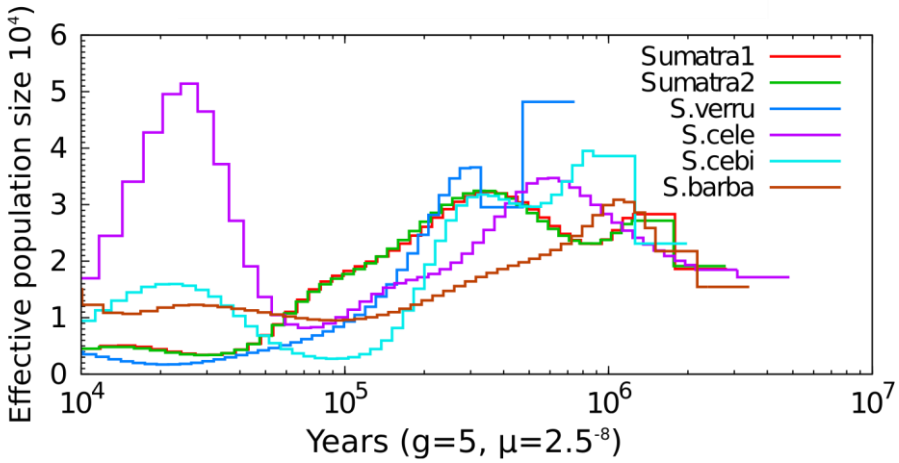


Figure 3.3: Population sizes of Sus in ISEA inferred from autosomes. ScSumatra=*S. scrofa* population from Sumatra. S.verru = *S. verrucosus*; S.cele = *S. celebensis*; S.cebi = *S. cebifrons*; S. barba= *S.barbatus.*

3.3 Discussion

Our results reveal that, unlike alternative strategies including SNP genotypes (from SNP microarrays), ascertained in a single species or population, that possess inherent biases in between species or population studies [31], whole-genome sequencing (leading to the detection of millions of polymorphisms) allow for phylogenetic relationships and admixture patterns, within the genus *Sus*, to be confidently resolved. Indeed, when attempting to recapitulate the analysis using the porcine 60K SNP chip [32] (Additional file 8, Figure S4), substantial differences in branch length estimates were found. These discrepancies are due to ascertainment bias demonstrating that a simple SNP array genotyping method, even for multiple individuals, would not have allowed the resolution afforded by a single complete genome. In addition, we show that there is a high degree of phylogenetic discordance across the genome. Such discordance could potentially lead to incorrect conclusions about the relationships between these species if only a subset of these loci were sampled [16]. While phylogenetic incongruence can frustrate taxonomic inference, it has the potential to test for the presence of inter-specific gene-flow. Our data demonstrate that the wealth of information extracted from these genomes allow for a thorough analysis (Additional file 4; Additional file 5, Table S8) that permits for the temporal reconstruction of the evolutionary history of *Sus* discussed below.

Evolutionary history of Sus

Our divergence time estimates suggest that the initial divergence of the Eurasian wild boar from a clade consisting of other *Sus* species took place during the Zanclean stage at the beginning of the Pliocene (Figure 3.1B; 5.3-3.5 Mya). Though the precise geographic location of this split (either in Sundaland or mainland Southeast Asia) remains unclear, the timing coincides with the divergence between other Sundaic and mainland Asian taxa [10]. The subsequent millions of years (from 3.5-2.5 Mya, the Piacenzian stage; Figure 3.1B; Figure 3.2) were marked by more intense cold periods that likely facilitated the emergence of a contiguous Sundaland landmass for prolonged periods (Figure 3.1A; Figure 3.2). Concomitant drops in sea levels are likely to have allowed the dispersal of the ancestor of *S. verrucosus* to Java (consistent with the fossil record, see Additional file 6). The deep split between *S. verrucosus* and other ISEA *Sus* demonstrates that this endangered species *S. verrucosus* represents a distinct lineage. Such a finding has implications for on-going ex- and in-situ conservation programs as it shows that this species represents an evident evolutionarily significant unit that deserves specific conservation strategies.

Our results provide evidence that following the divergence of the *S. verrucosus* lineage, the ancestor of *S. cebifrons* colonised the Philippines *during the* first stage of the Pleistocene approximately 2.4-1.6 Mya (Gelasian stage; Figure 3.1B; Figure 3.2). This date correlates with tectonic activity that led to the isolation of the Philippines from Sundaland even during periods of low sea levels [33]. This same period witnessed the divergence between *S. scrofa* populations on Sumatra and mainland East Asia (Figure 3.1B; Figure 3.2). However, it is unclear whether this divergence was the result of migration of *S. scrofa* from ISEA to the mainland or *vice-versa*. Moreover, this deep divergence between mainland and ISEA wild boars (*S. scrofa*) supports previous morphological studies that advocated the distinctiveness of these ISEA *S. scrofa* sub-species compared to other MSEA populations [13] (*i.e.* banded pig - *S. scrofa vittatus*).

Our results show that *S. celebensis* colonized Sulawesi, from the west (Borneo), during the latter stage of the Pleistocene (Calabrian; Figure 3.1B; Figure 3.2), approximately 1.6-0.8 Mya. It appears that this colonization occurred despite evidences that the Makassar Strait separating Sundaland and Sulawesi continued to exist even during periods of lowered sea levels, thus restricting dispersal during the Plio-Pleistocene [34]. Nonetheless, more frequent incidences of lower sea levels during this period [28] (Figure 3.2) would have reduced the distance between Sundaland and Sulawesi, thereby increasing the likelihood of a successful crossing of the strait. Our phylogenomic analysis implies that populations on Borneo acted

as the initial and main source for this dispersal even though the admixture analysis suggest that *S. verrucosus* on Java and *S. cebifrons* in the Philippines later also contributed to the *S. celebensis* gene pool (Additional file 4; Additional file 5, Table S8). These results may explain the existence of two well-supported but paraphyletic *S. celebensis* mtDNA clades present on Sulawesi [15, 25].

While the overseas dispersal of indigenous suids from Java and the Philippines into Sulawesi may have been the result of human-aided translocation, the initial divergence of *S. celebensis* from the Bornean population is too old to have been induced by modern humans. Thus, if overseas dispersal took place between Borneo and Sulawesi, it may also have been possible for pigs to disperse naturally from Java and the Philippines, within the last few million years (for example, by rafting or swimming). Further studies that can date these colonisation events from Java and the Philippines into Sulawesi, using multiple genomes from *S. celebensis*, could enable assessments of whether these migrations were in fact the result of human translocation.

The mainland divergence of *S. scrofa* into regionally discrete populations also started during the mid-Pleistocene (Figure 3.1B). Populations of *S. scrofa* from Asia migrated west approximately 1.2 Mya reaching Europe around 0.8Mya as suggested by the first appearance of *S. scrofa* in the fossil record (see Additional file 6 for details). The first divergence between Eastern and Western *S. scrofa*, as timed by our molecular clock analysis (Figure 3.1B), was likely the result of cooler climate during the Calabrian period that isolated populations in small refugia across Eurasia (Figure 3.2). Our data indicate that the split between Northern and Southern Chinese *S. scrofa* populations took place during the Ionian stage approximately 0.6 Mya (Figure 3.1B). This timing correlates with the most significant reduction in global temperature in the Plio-Pleistocene, characterised by long glacial intervals and short interglacial periods, that started approximately 0.8 Mya [35] (Ionian stage; Figure 3.1B; Figure 3.2). In this period forests contracted into small refugia, thereby isolating populations across MSEA [10].

Admixture and mtDNA replacement

Though we have presented the evolutionary history of *Sus* as speciation events resulting from simple bifurcations, D-statistics [26] and simulations challenge this view and suggest numerous instances of diversification and reticulation (**Additional file 4; Additional file 5, Table S8**). Our analysis shows that concomitant sea level fluctuations allowed for extensive intra- and inter-specific gene-flow during these periods, both within Sundaland and between Sundaland and MSEA (Figure 3.1 C & D; **Additional file 4; Additional file 5, Table S8**). Admixture fractions between

Sumatran and Chinese *S. scrofa* subpopulations were higher (9.5%-11%; Additional file 4) than those between Sumatran *S. scrofa* and other *Sus* species on Sundaland (1.3%-4.2%; Additional file 4). This finding suggest that, during the Pleistocene, more gene-flow took place between Chinese and Sumatran *S. scrofa* populations, than between Sumatran *S. scrofa* populations and other *Sus* species living on Sundaland. The geographic distance between Sumatran and Chinese *S. scrofa* populations is much larger than between Sumatran *S. scrofa* and the other *Sus* species that live on Sundaland (e.g. *S. verrucosus* and *S. barbatus*). Thus, this pattern supports a model of ongoing speciation with gene-flow in which interspecies relatedness is more closely correlated with a history of admixture than with current geographic proximity.

Despite these alternating periods of divergence and homogenisation, trees constructed using complete genomes recover the modern species designations. The same is not true of previously published mitochondrial phylogenetic trees of pigs from ISEA and MSEA that were able to distinguish geographically distinct populations of *S. scrofa in Eurasia, but were unable to recover the monophyly of* morphologically distinct species living on Sundaland [15, 16, 25, 36]. This paradox could result from either the limited phylogenetic information present in the short mitochondrial fragments used in previous studies, or from the complex pattern of admixture in Sundaland described above (Figure 3.1C & D).

Our phylogenetic tree based on near-complete mtDNA genomes (Figure 3.1E) is consistent with previous studies [15, 25], supporting a paraphyletic relationship among non *S. scrofa* species and a monophyletic clade of Sundaland taxa with short branch lengths. In addition our demographic analysis (Figure 3.3) shows that species living on Sundaland have undergone a long-term population decline, more extended than on MSEA (Additional file 7, Figure S3), during the Pleistocene. These results suggest that there was a replacement of mitochondrial haplotypes that took place across Sundaland during the latter part of the Pleistocene (1.5 Mya-Present; Additional file 4), after the divergence of *S. celebenisis* (Figure 3.1B & E; Additional file 4). The mtDNA replacement may have been facilitated by small population sizes (Figure 3.3). Taxa endemic to the Philippines and Sulawesi, isolated from Sundaland, were not involved in this admixture and harbour highly diverged mtDNA haplotypes of both complete mitochondrial sequences and fragments of the control region [15, 25] (Figure 3.1E). This phenomenon is unlikely to be an exception in pigs and has been recently observed in polar bears [3].

Human-mediated translocation

Though climate change has had the most dramatic and sustained influence on the speciation history of suids, humans have also affected this process. During the last 40Ky, humans have actively and passively translocated hundreds of species (as commensals, wild, or domestics) within ISEA, Wallacea and Australasia [11], and the signatures of the resulting admixture between suid lineages are evident in the genomic sequences. In addition, *S. scrofa* is an agriculturally important species that has been independently domesticated at least twice in mainland Eurasia (Near-east and China) [25]. The close relationship between humans and pigs make this species more prone to anthropogenic translocations. Indeed, our admixture analysis revealed the existence of inter-specific gene-flow that involved long distance dispersal across barriers that were unlikely to be the result of natural migration pathways.

Previous morphologic [37] and genetic [15] studies suggested that *S. celebensis* was kept captive and transported by humans from Sulawesi to Timor, Flora, Halmahera and Simeulue (North-West Sumatra). Admixture analyses support these claims by revealing gene-flow from *S. celebensis* into local *S. scrofa* populations on Sumatra and MSEA. Even during cold periods, Sulawesi and Sundaland were separated by a deep sea channel [34]. Thus, it seems unlikely that populations of *S. celebensis*, from Sulawesi, made it back to isolated islands around Sumatra and MSEA within the last 1.5 My since its divergence from *S. barbatus*. In their totality, these results provide evidences that human translocation of suids took place across the region and was not restricted to islands in close proximity to Sulawesi.

We also detected a strong signature of gene-flow from European *S. scrofa* populations into species in ISEA, consistent with a previous study that identified European mitochondrial haplotypes among populations in ISEA [15]. This gene-flow was most likely the result of human-induced dispersal of European pigs into ISEA within the past few hundred years. Some of these introduced pigs likely became feral and interbred with indigenous species.

While some of the admixture signals detected in this study are unequivocal, (*i.e.* admixture within Sundaland, supported by mtDNA and frequent merging of these islands during Plio-Pleistocene epoch), other signatures, including those involving long distance dispersal, are more difficult to interpret . For example, admixture involving un-sampled or extinct lineages can result in complex site patterns and could influence the results of the D-statistics [26]. For instance, the signal of gene-flow from European *S. scrofa* into species in ISEA could be the result of an admixture from an un-sampled sister lineage, and may not necessarily involve European pigs *per se*. Another limitation of the method can arise from ancestral

population subdivision as has been suggested to account for signatures of Neanderthal and Human admixture [38]. However, ancestral subdivision is unlikely to affect our analysis because of the evolutionary time frame investigated here (see Additional file 4).

Factors driving and reversing speciation in Sus

Our results suggest that Plio-Pleistocene climatic fluctuations had a significant impact on the diversification and homogenization of *Sus* in ISEA and MSEA. Speciation within *Sus* was mainly driven by dispersal across ISEA during the short glacial interval of the late Pliocene and early Pleistocene as suggested by evidence gleaned from other taxa [10, 39]. Rapid changes in climate and sea level resulted in population bottlenecks across ISEA (Figure 3.3). In addition, extensive intra- and inter-specific gene-flow led to instances of mtDNA replacement and a reversal (however temporary) of the speciation process.

Methodological Challenges

Our work demonstrates that the analysis of high-throughput sequencing data provides a powerful tool to investigate speciation history; but is unlikely to be devoid of sequencing errors, especially for low sequence coverage. However, the sequence coverage in our samples (7.5-25x) is expected to provide reliable genotype calls [40]. In addition, the major conclusions of this study are not expected to suffer from these biases as these analyses rely on non-singleton sites. Specifically, for a site to be phylogenetically informative the mutation must be shared by at least two taxa and the D-statistic analysis is explicitly designed to be robust to sequencing errors resulting in singletons [26]. Therefore, for a sequencing error to influence our phylogenetic or admixture analysis, it would have to be systematic and have occurred separately in different samples sequenced at different times in different sequencing centres. Thus, making the reasonable assumption that sequencing errors are independent between the samples, the probability of creating enough falsely informative sites to bias these analyses is exceedingly low.

Another limitation of our phylogenetic analysis could stem from recombination. Indeed, due to recombination, each of our genomic bins may represent a mosaic of different evolutionary histories. Nonetheless, theory and simulations suggest that our overall conclusions are relatively insensitive to the effects of recombination [41]. This insensitivity is because, moving along a sequence, different topologies are highly correlated and hence recombination is expected to have small effects over short recombination distances [42].

Lastly, it is important to take results of demographic history with caution. Indeed, while we believe that the general pattern described in Figure 3.3 is reliable, the magnitude of this bottleneck, in different species, is difficult to interpret. Differences in coverage among our samples likely result in variable power to call heterozygous sites, and could explain at least some of the differences in demographic history between different species.

3.4 Conclusion

The resolution afforded by complete genomes allowed us to infer not only ancient admixture episodes, but also those that took place as a result of more recent human-aided dispersal. Together, these findings provide insights related to the possible response to future climate and anthropogenic disturbances of mammalian taxa within ISEA.

Despite the challenges in building a single phylogeny from entire-genome sequences, we were able to obtain a well-resolved tree. In fact, the complexity of whole-genome data allows for a deeper appreciation of the complexities involved in the speciation process. Moreover, the substantial volume of data allows for robust time estimation. These findings reveal the power of multiple complete genomes from closely related species to comprehensively infer their speciation and evolutionary history and to resolve discrepancies between discordant trees constructed using smaller marker sets.

The complete genomes presented here provide compelling evidence that speciation in ISEA suids did not proceed according to a simple bifurcating model. Instead, our data indicate that the process involved numerous periods of both diversification and reticulation amongst several species and is on-going. Extensive inter-specific gene-flow has also been reported in fish [43, 44] and birds [45, 46]. The resolution afforded by complete genomes reveals that speciation is rarely as simple or linear as our traditional depictions, and that complex patterns of diversification and reticulation are likely the rule and not the exception.

The origin of new species often includes significant time periods during which closely related taxa in the initial stages of diversifying from one another can (and do) produce fertile offspring. The resolution provided by the use of whole genomes allows not only for an assessment of the current and past integrity of species, but also the elucidation of taxa specific speciation history. Genomics can thus reveal the molecular variability of life on earth, elucidate the process by which it emerged, and inform our attempts to preserve it.

3.5 Material and Methods

Sequencing, alignment and SNP calling

The samples used in this study were chosen from a larger pool of genotyped individuals (Illumina Porcine SNP60 chip) [32] in each species or population in order to ensure that each was representative of the genetic diversity of their respective species/populations (Additional file 8, Figure S4). DNA was extracted from blood or tissue using the DNeasy blood & tissue kits (Qiagen, Venlo, NL). Quality and quantity was measured with the Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA). Libraries of ~300 bp fragments were prepared using Illumina paired-end kits (Illumina, San Diego, CA) and sequenced with Illumina GAII or HiSeq (Additional file 1, Table S1).

Reads were trimmed for three consecutive base pairs (bp) with phred quality score equal or below 13, and discarded if they were shorter than 40 bp. We used Mosaik 1.1.0017 with unique alignment option to align reads to the Swine reference genome (Sscrofa10.2; GenBank GCA_000003025.4; Additional file 1, Table S2), together with the complete, mtDNA genome of *S. scrofa* (accession: AF486874) for all *Sus* species and the mtDNA genome of *Phacochoerus africanus (accession:* DQ409327) for *P. africanus*. *S. scrofa* and *P. africanus* mtDNA genome were aligned using ClustalW [47]. Mapping errors are unlikely to be problematic in this study, as the sequence mismatch to the reference genome was at max 3-4% (3-4 mismatch per 100bp read), a distance easily accommodated by short-read local aligners such as Mosaik. Mapped read depth ranged from 7.5-24x (Additional file1, Table S1), thus providing enough power to call genotype confidently [40]. The resulting BAM files were deposited on the EBI Sequence Read Archive under accession number ERP001813.

We used the pileup format (Samtools [48]) to call genotypes at sites covered by at least three reads with minimum base and mapping quality of 20. Additionally, we excluded any clusters of 3 or more SNPs within 10 bp or any SNP within 3 bp of an indel. We then identified genomic bins of 1 kbp that had an average depth under a maximum threshold (twice genome-wide average coverage) and 90% nucleotide sequence covered, to ensure maximum sequence coverage in every sample and exclude false positive SNPs resulting from copy number variation. These genomic bins were chained if adjacent.

Lastly, we computed the intersection of the genomic bins previously identified in each individual for further analysis using BedTools [49]. This resulted in an 11 way alignment with maximum sequence coverage and minimum false positive SNP calling in all our samples (~1.1Gbp; Additional file 1, Table S3).

We computed the distance to an outgroup (African Warthog) in 1Mbp windows for every *Sus* sample. Thereafter, we computed mean distances of all *Sus* to the outgroup. We obtained recombination rates from reference [20]. We used Kendall's rank test for correlation analysis as implemented in R.
Because the depth of coverage of mtDNA was highly variable across the different samples (Additional file 1, Table S4), we applied a different filtering strategy. For each position covered we computed the effective coverage of each allele as:

$$C(j) = \sum_{i=1}^{depth(j)} \left((1 - 10^{\frac{-mij}{10}}) * (1 - 10^{\frac{-qij}{10}}) \right)$$

where mij and qij refer to mapping quality and base quality score for read i at position j [50]. We filtered any sites were the major allele effective coverage did not represent at least 70% of the overall effective coverage at the position.

Phylogenetic analysis
First, we randomly selected genomic fragments (Additional file 1, Table S3) of at least 1 Kbp to make up 100 unique alignments of 1 Mbp (between 0.99 Mbp and 1.1 Mbp/each). We fitted a GTR+Γ4+I model of sequence evolution to each partition (genomic fragment) and ran 100 fast bootstrap replicates for each alignment and a thorough ML search using RAxML 7.1.2 [23]. We constructed a frequency consensus tree using all bootstrap replicates obtained from the 100 unique alignments using Phylip CONSENSE package [51]. These frequencies were then used as support for the species tree (Additional file 2, Figure S1).
To reconstruct the mtDNA tree we used a Bayesian tree reconstruction with 50,000,000 MCMC samples as implemented in MrBayes v3.2 [52]. We fitted a GTR+Γ4+I model suggested by AIC criterion as implemented in MrAIC [53]. We assessed the convergence of MCMC samples using TRACER [54]. The resulting phylogenetic tree is presented in Figure 3.1E.
To assess the robustness of these supermatrices we also applied more formal supertree methods by estimating a ML tree using RAxML with 100 fast bootstrap replicates for each genomic bin of at least 5 kbp (Additional file 1, Table S3). We used STAR [24] to reconstruct the species tree. Thereafter, we computed the relative frequency for each observed clade (Additional file 3, Table S5). Relative frequencies correspond to the proportion of each clade in the database of bootstrapped single locus trees.

In order to investigate how recombination affects phylogenetic concordance across the genome we computed the mean pairwise Robison-Foulds distance of trees, using Phylip [51], within 1Mbp windows. We obtained recombination rates from reference [20]. We used Kendall's rank test for correlation analysis as implemented in R.

Molecular clock analyses

We estimated divergence times using an approximate likelihood method as implemented in MCMCtree (PAML v.4), with an independent relaxed-clock and birth-death sampling [27]. To overcome difficulties arising from computational efficiency and admixture, we only used fragments (min. 5 kbp) that had a good bootstrap support (at least 70% bootstrap support for each node) for the main topology (Additional file 2, Figure S1). Although this is expected to bias estimates of divergence time toward the present, the amount of error is expected to be relatively small considering the deep time scale in this analysis. This resulted in 416 genomic bins and a 4.4 Mbp alignment. We fitted an HKY+Γ4 model to each partition (bin)and estimated a mean mutation rate by fitting a strict clock to each fragment setting a root age at 10.5 Mya, as suggested by previous studies [55]. This mean rate was used to adjust the prior on the mutation rate (rgene) modelled by a gamma distribution as G (1,125). The BDS and sigma2 values were set at 7 5 1 and G (1, 10) respectively. We ran two independent 40,000 (+10,000 burn in) MCMC samples for each combination of fossil calibration (Additional file 6) and assessed the convergence using TRACER [46] (ESS > 100).

Demographic analysis

We conducted a demographic analysis using a Hidden Markov Model (HMM) approach as implemented in PSMC [56] in our ISEA samples. We generated consensus sequences from bam files using the 'pileup' command in SAMtools. We used the following parameters: Tmax= 20; n = 64 ('4+50*1+4+6'). For plotting the results we used g=5 and a rate of 2.5x10-8 mutations per generation as in Humans.

Admixture analyses

To detect and quantify admixture among taxa we used D-statistics [6, 26] that take advantage of the large number of SNPs present in whole genomes. In short, the D-statistics provide a robust test for admixture by assessing the fit of a strictly bifurcating phylogenetic tree. For a triplet of taxa P1, P2 and P3, and an outgroup O, in which the underlying phylogeny is represented by the Newick string (((P1,P2),P3).O), one can compute the number of sites with mutations consistent

with incomplete lineage sorting: those where P1 and P3 (BABA) or P2 and P3 (ABBA) share the derived allele (B; assuming ancestral state, A, in the outgroup). Under a null hypothesis of no gene-flow (strict bifurcation), the ratio D=(ABBA-BABA)/(ABBA+BABA) is not expected to be significantly different from 0. This is because ABBA and BABA sites can only be created by coalescences in the common ancestor of P1, P2 and P3 and hence should happen with equal frequency. Alternatively, a significant excess of either ABBA or BABA site patterns is inconsistent with incomplete lineage sorting and provides evidence for a deviation from a phylogenetic tree, suggesting additional population structure or gene-flow.
To compute a standard error and assess the significance of the D-statistics we used a Weighted Block Jackknife approach. We divided the genome into N blocks and computed the variance of the statistics over the genome N times leaving each block aside and derived a standard error (SE) using the theory of the Jackknife (Supplementary Online Material 15 in ref. 6). We then computed the D-statistics for every possible combination of species (**Additional file 4; Additional file 5, Table S8**) using *P. africanus* as an outgroup. We corrected for multiple testing using a simple Bonferroni correction that involved multiplying our pvalues by the number of D calculation (**Additional file 4; Additional file 5, Table S8**). For additional details see Additional file 4.

Abbreviations:
ILS: Incomplete Lineage Sorting; ISEA: Island Southeast Asia; Kya: Thousands of years ago; Ky: Thousands years; ML: Maximum Likelihood; MSEA: Mainland Southeast Asia; Mya: Millions of years ago; My: Millions of years; Standard error: SE

Description of additional data files:
In this thesis I provide Additional file 4 because it contains very detailed additional results and discussion for the admixture analysis that will greatly help the reader. For sake of coherence and because I did not include all additional information in this thesis, I kept the original numbering of Supplementary Figures (Figure S1-X) as in the original paper. Please note that Additional File 4 contains its own referencing. The following paragraph provides a complete description of all additional files available with the online version of the paper:
The following additional data are available with the online version of this paper on the *Genome Biology* website (http://genomebiology.com/2013/14/9/R107). Additional file 1 contains Table S1-4, with information on sequence data and alignment results. Additional file 2 contains Figure S1, a species cladogram with support from various analyses. Additional file 3 is a table that contains results from

clade relative frequency analysis. Additional file 5 contains Table S8, which contains the full results from the D-statistics analysis. Additional file 6 is a text that contains information about fossil calibration. Additional file 7 contains Figure S3 describing the demographic history of population from MSEA. Additional file 8 contains Figure S4, a phylogenetic tree constructed using SNPs genotyped with the Illumina Porcine SNP60 array.

Acknowledgements:
We would like to thank Kelley Harris for her numerous comments that greatly improved this work. We thank Bert Dibbits and Lauretta Rund for sample acquisition and preparation, Dr. Oliver Raider for *Sus cebifrons DNA*, Dr. Alain Ducos for French wild boar DNA, and Dr. Sem Gemini for Italian wild boar DNA. We are also indebted to Alvaro G. Hernandez and Chris Wright at the University of Illinois Keck Center for Comparative and Functional Genomics for the sequencing. We also also thank Gus Rose and Konrad Lohse for their useful comments on earlier versions of this manuscript. Finally, we thank the Swine Genome Sequencing Consortium (SGSC) for the pre-release of the reference genome build10.2. This project was financially supported by European Research Council grant no. ERC-2009-AdG: 249894, a USDA grant 2007-04315, by NIH grants R01-GM40282 and T32-HG00047, and by BBSRC Institute Strategic Grants. Financial support was also provided by Illumina Inc.

References

1. Gatesy J, Hayashi C, Cronin MA, Arctander P: Evidence from milk casein genes that cetaceans are close relatives of hippopotamid artiodactyls. *Molecular biology and evolution* 1996, 13:954-63.
2. Stanhope MJ: Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. *Proceedings of the National Academy of Sciences* 1998, 95:9967-9972.
3. Hailer F, Kutschera VE, Hallstrom BM, Klassert D, Fain SR, Leonard J a., Arnason U, Janke a.: **Nuclear Genomic Sequences Reveal that Polar Bears Are an Old and Distinct Bear Lineage**. *Science* 2012, **336**:344-347.
4. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D: **Genetic evidence for complex speciation of humans and chimpanzees.** *Nature* 2006, **441**:1103-8.
5. Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton K, Presgraves DC: **Genome sequencing reveals complex speciation in the Drosophila simulans clade.** *Genome research* 2012 130922.111-.

6. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, et al.: **A draft sequence of the Neandertal genome.** *Science* 2010, **328**:710-22.
7. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Pääbo S: **Genetic history of an archaic hominin group from Denisova Cave in Siberia.** *Nature* 2010, **468**:1053-60.
8. Wallace AR: **On the Law Which Has Regulated the Introduction of New Species**. *Annals and Magazine of Nature History* 1855, **26**:184-196
9. Hall R, Asia SE, Holloway R, Sea P, Motion P: **Cenozoic plate tectonic reconstructions of SE Asia**. 1997:11-23.
10. Lohman DJ, Bruyn MD, Page T, Rintelen KV, Hall R, Ng PKL, Shih H-te, Carvalho GR, Rintelen TV: **Biogeography of the Indo-Australian Archipelago**. *Annual Review of Ecology and Systematics* 2011, **42**:205-228.
11. Heinsohn T: Animal translocation: long-term human influences on the vertebrate zoogeography of Australasia (natural dispersal versus ethnophoresy). *Australian Zoologist* 2003, 32:351-376.
12. Myers N, Mittermeier RA, Mittermeier CG, Fonseca GAB, Kent J: **Biodiversity hotspots for conservation priorities**. *Nature* 2000, **403**:853-858.
13. Meijaard E, d'Huart JP, Oliver WLR: **Family Suidae (Pigs)**. In: *Handbook of the Mammals of the World Vol 2.* Edited by Wilson DE, Mittermeier RA. Lynx Edicions, Barcelona, Spain. 2011:248-291.
14. Blouch RA, Groves CP: **Naturally occurring suid hybrids in Java**. *Zeitschrift für Säugetierkunde* 1990, 55:270-275.
15. Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, Rolett B, Spriggs M, Dolman G, Kim T-H, Thuy NTD, Randi E, Doherty M, Due RA, Bollt R, Djubiantono T, Griffin B, Intoh M, Keane E, Kirch P, Li K-T, Morwood M, Pedriña LM, Piper PJ, Rabett RJ, Shooter P, Van den Bergh G, West E, Wickler S, Yuan J, *et al.*: **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* 2007, **104**:4834-9.

16. Lucchini V, Meijaard E, Diong C: New phylogenetic perspectives among species of South-east Asian wild pig (Sus sp.) based on mtDNA sequences and morphometric data. *Journal of Zooglogy* 2005:25-35.
17. Mallet J: **Hybridization as an invasion of the genome.** *Trends in ecology & evolution* 2005, **20**:229-37.
18. Price TD, Bouvier MM: The evolution of F1 postzygotic incompatibilities in birds. *Evolution* 2002, 56:2083-9.
19. Groenen MAM, Archibald ALA, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, Li S, Larkin DM, Kim H, Frantz LAF, Caccamo M, Hyeonju A, Aken BL, Anselmo A, Anthon C, Auvil L, Badaoui B, Beattie CW, Bendixen C, Berman D, Blecha F, Blomberg J, Bolund L, Bosse M, Botti S, Bujie Z, *et al.*: **Analyses of pig genomes provide insight into porcine demography and evolution.** *Nature* 2012, **491**:393-398.
20. Tortereau F, Servin B, Frantz LAF, Megens H-J, Milan D, Rohrer G, Wiedmann R, Beever J, Archibald AL, Shook L, Groenen MAM: **A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content.** *BMC Genomics*, 2012 **13**:586.
21. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M: **A neutral explanation for the correlation of diversity with recombination rates in humans.** *American journal of human genetics* 2003, **72**:1527-35.
22. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen C-F, Thomas MA, Haussler D, Jacob HJ: **Comparative recombination rates in the rat, mouse, and human genomes.** *Genome research* 2004, **14**:528-38.
23. Stamatakis A: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006, 22:2688-90.
24. Liu L, Yu L, Pearl DK, Edwards SV: **Estimating species phylogenies using coalescence times among sequences.** *Systematic biology* 2009, **58**:468-77.
25. Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E, Rowley-Conwy P, Andersson L, Cooper A: **Worldwide phylogeography of wild boar reveals multiple centers of pig domestication.** *Science* 2005, **307**:1618-21.
26. Durand EY, Patterson N, Reich D, Slatkin M: **Testing for ancient admixture between closely related populations.** *Molecular biology and evolution* 2011, **28**:2239-52.
27. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Molecular biology and evolution* 2007, **24**:1586-91.

28. Miller KG, Kominz MA, Browning JV, Wright JD, Mountain GS, Katz ME, Sugarman PJ, Cramer BS, Christie-Blick N, Pekar SF: **The Phanerozoic record of global sea-level change.** *Science* 2005, **310**:1293-8.
29. Bird MI, Taylor D, Hunt C: Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland? *Quaternary Science Reviews* 2005, 24:2228-2242.
30. Wurster CM, Bird MI, Bull ID, Creed F, Bryant C, Dungait JAJ, Paz V: **Forest contraction in north equatorial Southeast Asia during the Last Glacial Period.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**:15508-11.
31. Albrechtsen A, Nielsen FC, Nielsen R: **Ascertainment biases in SNP chips affect measures of population divergence.** *Molecular biology and evolution* 2010, **27**:2534-47.
32. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu Z-L, Kerstens HH, Law AS, Megens H-J, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TPL, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MAM: **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PloS one* 2009, **4**:e6524.
33. Barrier E, Huchon P, Aurelio M: Geology Philippine fault : A key for Philippine kinematics. 1991.
34. Hall R: Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer-based reconstructions, model and animations. *Journal of Asian Earth Sciences* 2002, 20:353-431.
35. Zachos J, Pagani M, Sloan L, Thomas E, Billups K: **Trends, rhythms, and aberrations in global climate 65 Ma to present.** *Science* 2001, **292**:686-93.
36. Randi E, Lucchini V, Diong CH: **Evolutionary genetics of the suiformes as reconstructed using mtDNA sequencing**. *Journal of Mammalian Evolution* 1996, **3**:163-194.
37. Groves CP: Of mice and men and pigs in the Indo-Australian Archipelago. *Canberra Anthropology* 1984, 7:1-19.
38. Eriksson A, Manica A: Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences of the United States of America* 2012, 109:13956-60.

39. Nater A, Nietlisbach P, Arora N, van Schaik CP, van Noordwijk MA, Willems EP, Singleton I, Wich SA, Goossens B, Warren KS, Verschoor EJ, Perwitasari-Farajallah D, Pamungkas J, Krützen M: **Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant Orangutans (genus: Pongo).** *Molecular biology and evolution* 2011, **28**:2275-88.

40. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, Jorgensen T, Hansen T, Pedersen O, Wang J, Nielsen R: **Estimation of allele frequency and association mapping using next-generation sequencing data.** *BMC bioinformatics* 2011, **12**:231.

41. Lanier HC, Knowles LL: **Is recombination a problem for species-tree analyses?** *Systematic biology* 2012, **61**:691-701.

42. Wakeley, J. *Coalescent Theory: An Introduction.* Roberts & Company Publishers, Greenwood Village, Colorado. 2008.

43. Taylor EB, Boughman JW, Groenenboom M, Sniatynski M, Schluter D, Gow JL: Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (Gasterosteus aculeatus) species pair. *Molecular ecology* 2006, 15:343-55.

44. Vonlanthen P, Bittner D, Hudson a G, Young K a, Müller R, Lundsgaard-Hansen B, Roy D, Di Piazza S, Largiader CR, Seehausen O: **Eutrophication causes speciation reversal in whitefish adaptive radiations.** *Nature* 2012, **482**:357-62.

45. Grant BR, Grant PR: **Fission and fusion of Darwin's finches populations.** *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2008, **363**:2821-9.

46. Kraus RHS, Kerstens HHD, van Hooft P, Megens H-J, Elmberg J, Tsvey A, Sartakov D, Soloviev SA, Crooijmans RPMA, Groenen MAM, Ydenberg RC, Prins HHT: **Widespread horizontal genomic exchange does not erode species barriers among sympatric ducks.** *BMC evolutionary biology* 2012, **12**:45.

47. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 1994, 22:4673-80.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-9.

49. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, 26:841-2.

50. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A: Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics* 2011, 43.

51. Felsenstein J: PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989, 5:163-166.

52. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space.** *Systematic biology* 2012:sys029-.

53. Nylander J: **MrAIC** [http://www.abc.se/~nylander/mraic/pmraic.html]

54. Rambaut A: **Tracer v1.4** [http://beast.bio.ed.ac.uk/Tracer].

55. Gongora J, Cuddahee RE, Nascimento FFD, Palgrave CJ, Lowden S, Ho SYW, Simond D, Damayanti CS, White DJ, Tay WT, Randi E, Klingel H, Rodrigues-Zarate CJ, Allen K, Moran C, Larson G: **Rethinking the evolution of extant sub-Saharan African suids (*Suidae*, *Artiodactyla*)**. *Zoologica Scripta* 2011, **40**:327-335.

56. Li H, Durbin R: Inference of human population history from individual whole-genome sequences. *Nature* 2011, 475:493-6.

Additional File 4

Detecting admixture in single genome sequences

Several studies have estimated admixture fractions between individuals using a range of methods[1,2]. Programs such as SABER[3] can compute the admixture fraction from one population to another, using High-Density SNP genotypes. However, this method requires multiple samples from the sample population, to estimate Linkage Disequilibrium (LD), identify admixture blocks and estimate time of admixture. This approach is not applicable on single whole-genome sequences. Instead we decided to use D-statistics, originally implemented in the Neanderthal genome paper[2] and formalized by Durand *et al*. (2011)[4]. The D-statistics take advantage of the large number of SNPs present in whole genomes to infer admixture. Assume that we have sequenced one chromosome from 4 different populations P_1, P_2, P_3 and O, where P_1 and P_2 are sister taxa and O is an outgroup. One can compute number of derived alleles that match between P_1 and P_3 (ABBA count) and between P_2 and P_3 (BABA count). Under a null hypothesis of no gene flow between P_3 and either P_2 or P_1 we expect a similar count of ABBA and BABA patterns to arise from incomplete lineage sorting. Under an alternative scenario of admixture, ABBA counts may be significantly higher than BABA counts (or *vice versa*), which is indicative of gene flow between P_2 and P_3. For a full description of the method please refer to Durand *et al.* (2011). To compute a standard error on the D-statistics we used a Weighted

Block Jackknife approach. Briefly, we divided the genome into N blocks and computed the variance of the statistics over the genome N times leaving each block aside and derived a standard error (SE) using the theory of the Jackknife (For full approach see Supplementary Online Material 15 in Green *et al.* 2010). We then computed the D-statistics for every possible combination of species (**Additional file 5**) using *P. Africanus* as an outgroup. We corrected for multiple testing using a simple Bonferroni correction. Simply, we multiplied our p-values by the number of D calculation (360; **Additional file 5**). We tested the influence of different block sizes on the estimation on the SE (**Table S6**). Overall, the SE estimates were slightly higher at 5 and 10Mb blocks size than 2Mb. However, we did not observe significance levels higher than 0.01 (after correction) using 2Mb blocks size. Therefore, we used the 2Mb as block size for further analyses. We also assessed the effect of transition and transversion mutations on D estimates. Overall D-statistic computation using transitions or transversions resulted in the same outcome (**Table S7**). We also recomputed the D-statistics at higher coverage to test for the effect of false negative SNP calling. These results were similar to those presented here and show that our method is not sensitive to differences in coverage (data not shown). Lastly, the D-statistics may be sensible to different read length obtained from Illumina sequencing platforms[5]. The authors, found positive correlations between significance level of D-statistics and read-length, however these correlations were not significant. Thus, they noted these differences may lead to borderline significant false positive results. This phenomenon is unlikely to affect our results as read-length is uniform across our samples (raw length = 100bp) and our corrected p-values are always lower than 0.001. Furthermore, the authors also noted that different sequencing platforms (sanger, Illumina and 454) may also influence D-statistics. However, they do not mention if different Illumina technologies (GAII or HiSeq) may also influence. Our data comprise samples sequenced on both systems (**Table S1**). These differences in technologies did not influence our calculation between our Sumatran *S. scrofa* that were sequenced with either GAII or HiSeq (**Table S1&3**). Therefore, we think that this is unlikely to influence our results.

Admixture fraction

While the D statistics estimate the fraction of incomplete lineage sorting that is due to admixture, they are not linearly related to the proportion of admixture (Durand et al. 2011). To compute the admixture proportion, we require data from a taxon that is sister to the population that contributed the admixture (it is also possible to get an upper bound on the admixture proportion using two samples from the same

population). Consider a scenario where we have samples from the pairs of sister taxa, P_1 and P_2 and P_3 and P_4. If there is a significant D statistic indicating admixture from P_3 into P_2, we can compute the number of sites where P_2 and P_4 share the derived allele, $S(P_1,P_2,P_4)$. We can also compute the number of sites where P_3 and P_4 share the derived allele, $S(P_1,P_3,P_4)$. The portion of the genome of the sample from P_2 that comes from P_3 will then behave as if it were a member of P_3; therefore $S(P1,P_2,P_4)/S(P_1,P_3,P_4)$ = f the admixture proportion. Thus, while Durand *et al.* (2011) showed that $S(P_1,P_2,P_4)/S(P_1,P_3,P_4)$ = f represents an upper bound of the true admixture fraction in the case of simple admixture and constant population size, it is unclear how a more complex history effecst the estimation of f.

Confounded D-statistics

The D-statistics provide evidence of admixture between two populations. However, some of these admixtures might be confounded. For example, suppose we have 4 populations ((P_1, P_2), (P_3, P_4)), we detect admixture of P_3 with P_2 (P_3/P_2) and P_4 with P_2 (P_4/P_2) (P_3 and P_4 are sister taxa). **Figure S2** shows 5 possible models that can explain such a result. In the first model (**Figure S2 A**) the admixture is completely confounded because it comes from the common ancestor of P_3 and P_4. In the second and third model (**Figure S2 B, C**), the admixtures are partly confounded and involve $P_{3,4}$ and P_3 or P_4 only. In the fourth model (**Figure S2 D**) the admixtures independently involve P_3 and P_4. Finally, in the fifth model (**Figure S2 E**) there are 3 admixture events that are partly confounded and involve both P_3, P_4 but also $P_{3,4}$. One way to distinguish between these models is to compare the value of D statistics. Durand et al. (2011) showed that the value of D increases with the difference between time of admixture and time of divergence of the 3 taxa involved in the D calculation. In the case described here $D_1(P_1,P_2,P_3)$ tends to 1 as $t_{p1,2,3,4} - t_{GF1}$ becomes large (**Figure S2**). Moreover, D also increases with the admixture fraction f. Because, in this example, $t_{p1,2,3,4}$ is constant across D calculations, a significant increase of D_1 compared to D_2 is the result of, t_{GF1} being smaller than t_{GF2} and t_{GF3} such as $t_{p1,2,3,4} - t_{GF1} > t_{p1,2,3,4} - t_{GF2}$ and $t_{p1,2,3,4} - t_{GF1} > t_{p1,2,3,4} - t_{GF3}$ and/or due to a higher admixture fraction such as, $f_1 > f_2$. Therefore, there must be a more recent admixture event and/or higher admixture fraction between P_3/P_2 than P_4/P_2 which in turn indicates that the admixture P_3/P_2 is at least partly independent from $P_{3,4}/P_2$ and P_4/P_2. This rationale permits the rejection of models 1 and 3 (**Figure S2 A, D**). We assessed if a pair of D values were significantly different using a Z-test. Briefly, we found the difference between D-statistics of interest and used the sum of the Jackknife variance estimates as an estimate of the variance of this quantity. We then assessed if the difference was significantly

different from 0. However, this test does not allow us to distinguish between models 2, 4 and 5 (**Figure S2 B,D,E**). Because we can show that there must be an independent admixture event P_3/P_2 it does not mean that we can rule out the

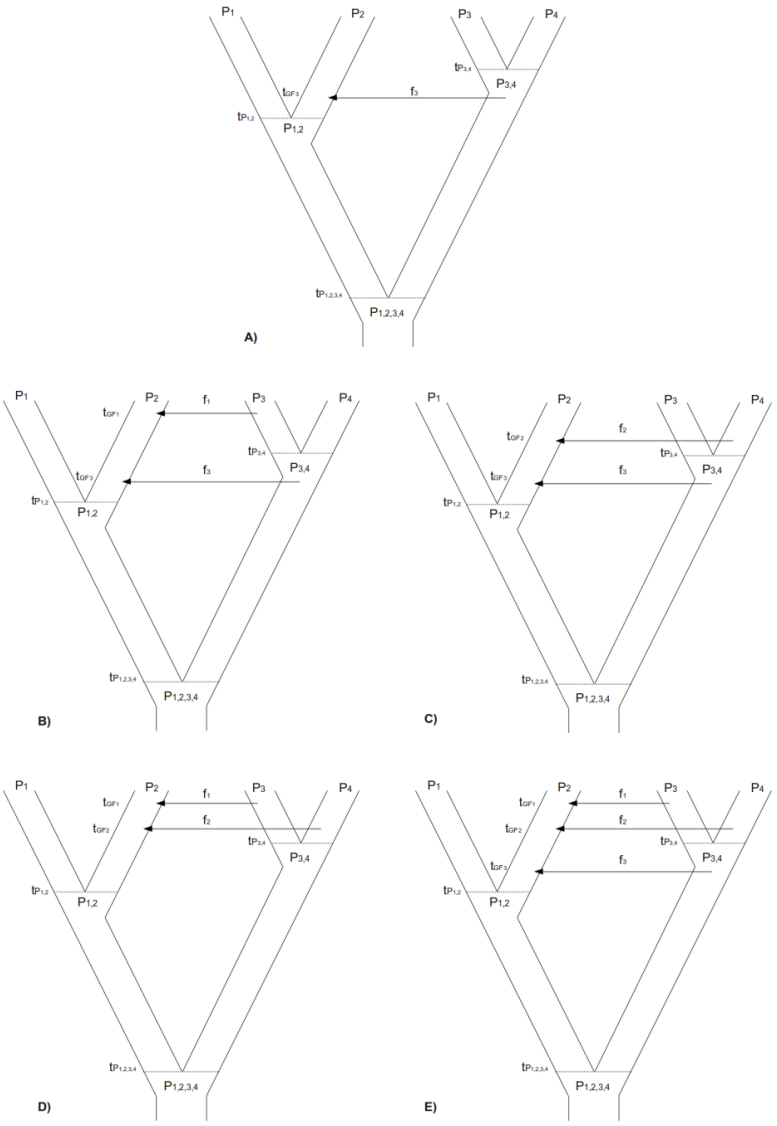


Figure S2: Examples of complex models of admixture, resulting in confounded D-statistics, involving 4 taxa P_1, P_2, P_3 and P_4.

possibility of independent admixture event P_4/P_2 or a confounded event $P_{3,4}/P_2$. To distinguish between these models, one can examine the overlap in sites that support the two D statistics. In the case of confounded admixture, there should be substantial overlap between sites that support a non-zero D-statistic. However, it is unclear how many overlapping ABBA or BABA can be expected under these models, as different population processes may influence their counts. Further studies should concentrate on deriving the number of expected overlapping ABBA and BABA under these three different models.

Table S6: Examples of the influence of different block size on SE of D-statistics. ScSuma1/2 = *S. scrofa* Sumatra; ScEuroIt = *S. scrofa* Italy; ScEurope = *S. scrofa* Europe; Sbarba = *S. barbatus*; Scebi= *S. celebensis*; Sverru= *S. verrucosus*; ScNChina = *S. scrofa* North China; ScSChina = *S. scrofa* South China.

P1, P2, P3	2Mbp, D +- SE	5Mbp, D +- SE	10Mbp, D +- SE
ScNChina, ScSuma1, Sbarba	0.2016 +- 0.0086	0.2016 +- 0.0118	0.2016 +- 0.0136
Sbarba, Scele, Sverru	0.1133 +- 0.0030	0.1133 + 0.0033	0.1133 + 0.0033
Sbarba, Scele, ScEurope	-0.2018 +- 0.0036	-0.2018 +- 0.0043	-0.2018 +- 0.0050
ScNChina, ScSuma1, Scele	0.2340 +- 0.0090	0.2340 +- 0.0122	0.2340 +- 0.0140

Interpretation of D-statistics

Sunda-shelf admixture

We detected admixture from Sumatran *S. scrofa* into other species living in the Sunda-shelf. The D-statistics reveal an excess of shared derived lineages between *S. barbatus* and Sumatran *S. scrofa* compared to both *S. cebifrons* and *S. celebensis* (D= 0.0813+-0.0042 ; D= 0.0795+-0.0042). This signal seems to be stronger in *S. verrucosus* (D= 0.1681+-0.0108; D = 0.1696+-0.0079). Moreover, this admixture was also detectable using *S. barbatus* and *S. verrucosus*, where Sumatran samples share more derived alleles with the latter (D= 0.1003+-0.0095). This pattern is consistent with the clustering of *S. verrucosus* and Sumtaran *scrofa* in the phylogenetic tree derived from complete mtDNA sequences (**Figure S2**).

This admixture pattern in the Sunda-shelf appears to be bidirectional. We found that derived alleles observed in *S. barbatus* matched more often Sumatran *S. scrofa* than *S. scrofa* from South China, North China and Europe (D=0.2000+-0.0084;

D=0.2016+-0.0086; D=0.2119+-0.0089). The same pattern was observed using *S. verrucosus* as admixting species(D=0.2996+-0.0120; D=0.3028+-0.0123 ;D=0.3311+-0.0116). Finally we found that *S. verrucosus* shares more derived lineages with *S. barbatus* than with *S. cebifrons* (D=0.1648+-0.0042). These findings show that admixture within the Sunda-shelf did not only involve gene flow between *S. scrofa* and non-*scrofa* species, but rather involved all species living on the Sunda-shelf. Thus, these results reinforce the conclusion that inter-specific gene flow resulted in mtDNA replacement in Sundaland and resulted in discordant phylogenetic signal between mtDNA and autosomal chromosomes.

Table S7: Examples of the influence of Transvertion (Tv) and Transitions (Ti) on D-statistics. ScSuma1/2 = S. scrofa Sumatra; ScEuroIt = S. scrofa Italy; ScEurope = S. scrofa Europe; Sbarba = S. barbatus; Scebi= S. celebensis; Sverru= S. verrucosus; ScNChina = S. scrofa North China; ScSChina = S. scrofa South China.

P_1, P_2, P_3	Ti, D+-SE	n. ABBA / BABA	Tv, D+-SE	n. ABBA / BABA
ScNChina, ScSuma1, Sbarba	0.2306 +-0.0102	350,113/ 237,240	0.1922 +- 0.0087	118,616 / 74,154
Sbarba, Scele, Sverru	0.1287 +- 0.0041	336,825 / 270,954	0.1083 +- 0.0030	116,775 / 90,130
Sbarba, Scele, ScEurope	-0.2377+- 0.0054	147,174 / 216,285	-0.1912 +- 0.0037	41,459 / 67,316
ScNChina, ScSuma1, Scele	0.2741 +- 0.0105	387,887 / 243,560	0.2285 +- 0.0090	132,497 / 75,486

Because we had two individuals from the *S. scrofa* Sumatra population we were able to obtain an upper bound of the admixture fraction from this population into Sunda-shelf species.

- Admixture fraction of Sumatran into *S. verrucosus*:

$$f_{ScSuma1,Sverru} = \frac{S(Scebi,Sverru,ScSuma1)}{S(Scebi,ScSuma2,ScSuma1)} = 0.042$$

$$f_{ScSuma2,Sverru} = \frac{S(Scebi,Sverru,ScSuma2)}{S(Scebi,ScSuma1,ScSuma2)} = 0.040$$

- Admixture fraction of Sumatran into *S. barbatus*:

$$f_{ScSuma1,Sbarba} = \frac{S(Scebi,Sbarba,ScSuma1)}{S(Scebi,ScSuma2,ScSuma1)} = 0.016$$

$$f_{ScSuma1,Sbarba} = \frac{S(Scebi,Sbarba,ScSuma1)}{S(Scebi,ScSuma1,ScSuma2)} = 0.016$$

$$f_{ScSuma1,Sbarba} = \frac{S(Scele,Sbarba,ScSuma1)}{S(Scele,ScSuma2,ScSuma1)} = 0.013$$

$$f_{ScSuma2,Sbarba} = \frac{S(Scele,Sbarba,ScSuma2)}{S(Scele,ScSuma1,ScSuma2)} = 0.013$$

We did not attempt to compute the admixture fraction from Sumatran *S. scrofa* into *S. verrucosus* using *S. celebensis* as non-admixting, because we found clear evidence of admixture between *S. verrucosus* and *S. celebensis*, which would bias the calculation (5.4.3). *S. barbatus* shows a higher admixture fraction (from Sumatran *S. scrofa*) when using *S. cebifrons* than *S. celebensis*. This is expected as *S. celebensis* is more closely related to *S. barbatus* than *S. cebifrons*. Thus, this result suggests that some admixture between Sumatran S. scrofa and S. barbatus may have taken place before the divergence of *S. barbatus* and *S. celebensis* (admixture into their common ancestor). However, these admixture fractions were very close (0.16 versus 0.13). This result suggests that most of the admixture from *S. scrofa* Sumatra into *S. barbatus* took place after the divergence of *S. barbatus* and *S. celebensis* .

We also found evidence of admixture from *S. cebifrons* into Sumatran *S. scrofa*. This observation can be the result of two possibilities. On one hand, independent admixtures from each both *S. barbatus* and *S. cebifrons* into *S. scrofa* Sumatra could explain this observation. On the other hand, gene-flow from their common ancestor could also explain this result. The latter hypothesis seems more plausible as *S. cebifrons* had no means of dispersal into the Sundaland after its divergence from *S. barbatus*. In addition, we found that, approximately 70% of the sites supporting an admixture from *S. cebifrons* into Sumatran *S. scrofa* (derived state in *S. cebifrons* and *S. scrofa* Sumatra and ancestral state in other *S. scrofa* populations) overlapped with sites supporting admixture from *S. barbatus* into

Sumatran *S. scrofa*, suggesting admixture from the common ancestor of *S. barbatus* and *S. cebifrons*. Moreover, D (into Sumatran *S. scrofa*) was significantly higher using *S. barbatus* than *S. cebifrons* as admixting taxa ($p < 0.01$). Thus, we interpret this result as a signal for admixtures into *S. scrofa* Sumatra, from the common ancestor of S. barbatus and S. cebifrons and an additional admixture from *S. barbatus* alone. However, although we cannot rule out the possibility of an additional independent event of admixture from *S. cebifrons* into the Sumatran population of *S. scrofa*, this scenario seems unlikely, as *S. cebifrons* had no means of dispersal to Sumatra.

In addition to the signals of admixture described above, we find an excess of incomplete lineage sorting between the *S. scrofa* populations (Sumatra and MSEA populations) and the other *Sus* species. For example, we identified admixture between *S. barbatus* and Sumatran *S. scrofa* based on a D statistic of 0.2 (D(ScSChina, ScSuma1, Sbarba). This indicates that 20% of the incomplete lineage sorting between *S. barbatus and S. scrofa* is due to admixture. However, because of the deep divergence between *S. scrofa* and the non-*scrofa* species, very little incomplete lineage sorting is expected. Thus, the D statistic should be close to 1. Using simulations, we have seen that the observed D statistic is not possible without additional admixture between the ancestor *S. scrofa* and the ancestor of the non-*scrofa* species (J. Schraiber, unpublished observation). Thus, our results suggest continuous inter-specific gene-flow among population of the Sunda-Shelf throughout the Plio-Pleistocene epoch.

Natural dispersal in and out ISEA

The D-statistics revealed an excess of derived lineage shared between the *S. scrofa* Sumatran population and both South and North Chinese populations when comparing to the European population (D=0.1803+-0.0039; D=0.1938+-0.0036). Moreover, the D-statistics also support more admixture from Sumtaran *S. scrofa* into South Chinese population than into North Chinese population (D=0.0340+-0.0031). We interpret this pattern as isolation by distance, as Sumatra is closer to South China. These results show that admixture out of ISEA happened repeatedly before and after the divergence of North and South Chinese populations.

The admixture from ISEA into MSEA is not only restricted to within *S. scrofa*. We found signals of admixture from *S. barbatus* into both North and South Chinese *scrofa* compared to European population (D=0.0319+-0.0029; D=0.0339+-0.0029). This can be also found from *S. verrucosus*, (D=0.0654+-0.0035; D=0.0681+-0.0032), *S. cebifrons* (D=0.0354+-0.0035; D=0.0414+-0.0033) and *S. celebensis* (D=0.1029+-0.0030; D=0.1100+-0.0030).

We computed the admixture of Sumatran population into MSEA:

- Admixture fraction into North Chinese:

$$f_{ScSuma1,ScNChina} = \frac{S(ScEurope,ScNchina,ScSuma1)}{S(ScEurope,ScSuma2,ScSuma1)} = 0.096$$

$$f_{ScSuma2,ScNChina} = \frac{S(ScEurope,ScNchina,ScSuma2)}{S(ScEurope,ScSuma1,ScSuma2)} = 0.095$$

- Admixture fraction into South Chinese:

$$f_{ScSuma1,ScSChina} = \frac{S(ScEurope,ScSchina,ScSuma1)}{S(ScEurope,ScSuma2,ScSuma1)} = 0.110$$

$$f_{ScSuma2,ScSChina} = \frac{S(ScEurope,ScSchina,ScSuma2)}{S(ScEurope,ScSuma1,ScSuma2)} = 0.109$$

$$f_{ScSuma1,ScSChina} = \frac{S(ScNChina,ScSchina,ScSuma1)}{S(ScNChina,ScSuma2,ScSuma1)} = 0.016$$

$$f_{ScSuma2,ScSChina} = \frac{S(ScNChina,ScSchina,ScSuma2)}{S(ScNChina,ScSuma1,ScSuma2)} = 0.015$$

These results suggest a higher admixture fraction among *S. scrofa* populations than among Sunda-Shelf populations (see section 5.4.1). Moreover, the admixture fractions reveal that most of the admixture out of ISEA happened before the divergence between North and South Chinese populations (as these fractions are very close [0.95 vs. 0.11]).

Our results also suggest admixtures event from all MSEA *S. scrofa* into all ISEA species except *S. celebensis.* Counter intuitively, this pattern is stronger from European, rather than from South and North Chinese populations (**Additional file 5**). Two models could explain this result. Under the first model, only one admixture event took place from European pigs due to human translocation and the signal is

present in the Chinese population because of their relatedness. Alternatively, admixture happened before and after the divergence of *S. scrofa* populations on the mainland due to natural and human mediated migrations. Because we can show that there are migrations events from ISEA into the mainland (see section 5.2.3) we hypothesize that at least part of the admixture found from *S. scrofa* into the Sunda-shelf is due to a natural process that took place before and probably after the divergence of *S. scrofa* on the mainland. Moreover, we can show that there is more admixture from European than Chinese *S. scrofa* into ISEA species suggesting a distinct migration from Europe into ISEA which would be difficult to reconcile with natural migration (see section 5.2.4). It is also possible that part of the admixture from Chinese *scrofa* into ISEA species is due to Human-mediated dispersal of pigs (see section 5.2.4).

We computed the admixture proportion from mainland *S. scrofa* into ISEA species due to natural dispersal under this model:

$$f_{Mainland,Sbarba} = \frac{S(Scele,Sbarba,Mainland)}{S(Scele,ScEurope,China)} = 0.041$$

$$f_{Mainland,Sverru} = \frac{S(Scele,Sverru,Mainland)}{S(Scele,ScEurope,China)} = 0.040$$

$$f_{Mainland,Scebi} = \frac{S(Scele,Scebi,Mainland)}{S(Scele,ScEurope,China)} = 0.040$$

where Mainland represents the shared SNP between North, South China and Europe *S. scrofa* that supports these admixtures and China represents only the shared SNP between North and South China *S. scrofa.* This result supports the view that most of the admixture between continental Eurasia and the Sunda-shelf is due to natural migrations as most of the admixture from MSEA into ISEA seems to be confounded in the different MSEA S. scrofa.

Together these results show that natural migration from ISEA to MSEA and *vice-versa* took place throughout the mid / late Pleistocene. Thus, because we have no taxa that diverged prior to this period on the mainland, we cannot infer natural migration out ISEA during the late Pliocene and early Pleistocene. However, this is likely to be the case.

Natural dispersal into Sulawesi

Besides the signal for admixture between *S. verrucosus* and *S. barbatus*, we found evidence for admixture between *S. verrucosus* and *S. celebensis*. The D-statistics

supporting an admixture of *S. verrucosus* with *S. barbatus* is significantly lower than the value supporting an admixture with *S. celebensis* (using *S. cebifrons* as non introgressed; D=0.1650+-0.0042; D=0.2595+-0.0040; $p < 0.01$). Only two scenarios can explain such a result: a higher admixture fraction or more recent admixture from *S. verrucosus* into *S. celebensis* than into *S. barbatus* (see section 3). It is impossible, with our data, to distinguish between these hypotheses. However, there was a strong signal of admixture between *S. verrucosus* and *S. celebensis* using *S. barbatus* as a putatively non-introgressed species (D=0.1134+-0.0030). This result strongly supports the idea that *S. verrucosus* admixed with *S. celebensis* after its divergence from *S. barbatus*. Admixture into Sulawesi was also found from *S. cebifrons* (D(Scele,Sbarba,Scebi) = 0.0682 +- 0.0029). This finding shows that both *S. cebifrons* and *S. verrucosus* contributed to *S. celebensis*' gene pool.

Human-mediated admixture

The admixture found from MSEA into *S. cebifrons* in the Philippines can partly be explained by natural dispersal on the 'Sunda-shelf' before its divergence with *S. barbatus* and *S. celebensis*. However, we found that only 45% and 39% of SNP are shared between admixture of MSEA in *S. barbatus* and *S. cebifrons*, respectively. Therefore, we believe that these were, at least, partly independent. Moreover, North and South Chinese derived lineages are found more often in *S. cebifrons* than *S. celebensis*. Because *S. celebensis* is more closely related to *S. barbatus* than *S. cebifrons* it is unlikely that these observations were the result of an admixture from MSEA into the common ancestor of *S. cebifrons* and *S. barbatus*. In addition, we know that The Philippines have been completely separated from the Sundaland and MSEA during the latter part of the Pleistocene. Together, these results hint at a human mediated dispersal from MSEA into the Philippines. Such a Human-mediated dispersal of pigs may also have happened throughout ISEA. Again, because we can show that there are natural dispersals out of ISEA we assume that at least part of the admixture MSEA to ISEA is due to natural processes. However, although it is not possible, with our data, to reach a conclusion on the possibility of human-mediated dispersal, of *S. scrofa* of Asian origin in the rest of ISEA (particularly the Sunda-shelf), this hypothesis seems likely if it happened in the Philippines. Further studies, with multiple individuals in which admixture blocks can be identified may provide an answer to this question. We took these results with caution because we could not infer a significant excess of derived lineage shared between *S. cebifrons* with either North or South Chinese populations. This could be due to a power limitation as these two populations are very closely related, or simply because the admixture fraction was so small. However, we would expect a

significant difference between North and South Chinese pigs if this admixture was human-mediated as these populations would have diverged long before any humans reached the region. Therefore we could not conclude if this admixture was human-mediated or natural.

The admixture from *S. celebensis*, into Sumatran and Chinese *scrofa* seems to be independent from admixture by *S. barbatus* ($p < 0.01$; see section 5.4.1). This is difficult to reconcile with natural dispersal (see section 5.2.4). Previous studies have already found evidence for human mediated dispersal of *S. celebensis* to Flores and Timor[6,7]. Our analysis suggests that this translocation was probably more generalised to the whole Southeast Asian region rather than restricted to only Timor and Flores.

The D-statistics revealed a distinct admixture from European pigs into ISEA species (5.2.3). Moreover, European mtDNA haplotypes were found in domestic pigs in the Philippines[7]. Together these results support an admixture event from European pigs into ISEA, which is consistent with the idea that Europeans brought pigs to this area during the past few hundred years.

Furthermore, we found that the number of overlapping sites supporting an admixture between Chinese *S. scrofa* and *S. barbatus* and European *S. scrofa* and *S. barbatus* was lower in the latter comparison (217,801 vs 284,787). Simulations show that this is consistent with an admixture from the common ancestor of MSEA S. scrofra and an additional burst of admixture from European *S. scrofa* into *S. barbatus*.

Timing admixture

Because we cannot estimate the extent of LD in our different populations, we could not time directly admixture events. Further studies, using multiple individuals from different populations will provide the means to identify admixture blocks and assess the age of the admixture signals identified in this study.

References for Additional File 4

1. vonHoldt, B.M. *et al.* A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome research* 21, 1294-305 (2011).

2. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* 328, 710-22 (2010).

3. Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *American journal of human genetics* 79, 1-12 (2006).

4. Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular biology and evolution* 28, 2239-52 (2011).
5. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature*, in press.
6. Groves, C.P. Of mice and men and pigs in the Indo-Australian Archipelago. *Canberra Anthropology* 7, 1-19 (1984).
7. Larson, G. *et al.* 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 (2007).

4

Testing models of speciation from genome sequences: divergence and asymmetric admixture in Island Southeast Asian *Sus* species during the Plio-Pleistocene climatic fluctuations

Laurent A.F. Frantz[1], Ole Madsen[1], Hendrik-Jan Megens[1], Martien A.M. Groenen[1] and Konrad Lohse[2]

[1]: Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands.

[2]: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom

Molecular Ecology (2014) 23:5566-5574

Abstract

In many temperate regions, ice ages promoted range contractions into refugia resulting in divergence (and potentially speciation), while warmer periods led to range expansions and hybridization. However, the impact these climatic oscillations had in many parts of the tropics remains elusive. Here, we investigate this issue using genome sequences of three pig (*Sus*) species, two of which are found on islands of the Sunda-shelf shallow seas in Island Southeast Asia (ISEA). A previous study revealed signatures of inter-specific admixture between these *Sus* species (Frantz *et al.* 2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*. *Genome biology*, **14**, R107; **Chapter 3**). However, the timing, directionality and extent of this admixture remain unknown. Here we use a likelihood based model comparison to more finely resolve this admixture history and test whether it was mediated by humans or occurred naturally. Our analyses suggest that inter-specific admixture between Sunda-shelf species was most likely asymmetric and occurred long before the arrival of humans in the region. More precisely, we show that these species diverged during the late Pliocene but around 23% of their genomes have been affected by admixture during the later Pleistocene climatic transition. In addition, we show that our method provides a significant improvement over D-statistics which are uninformative about the direction of admixture.

Keywords: maximum likelihood, speciation, Island Southeast Asia, admixture.

4.1 Introduction

Over the last four million years, the Earth has undergone frequent climatic oscillations including many ice ages (Zachos *et al.* 2001; Miller *et al.* 2005). Genetic studies have revealed that these large scale climatic fluctuations played a critical role in the evolutionary history of contemporary species (Hewitt 2000, 2004). Recent studies making use of the increased power afforded by genome-scale data have allowed biologists to test increasingly finer hypotheses regarding the existence and the timing of post-divergence gene-flow (*i.e.* Rohland *et al.* 2010; Lawniczak *et al.* 2010; Cahill *et al.* 2013; Hearn *et al.* 2014).

The impact that quaternary climatic fluctuations had on speciation is highly dependent on the taxa and the geographic range (Stewart *et al.* 2010). In many temperate regions, range contractions into refugia during glacial periods likely promoted divergence (and speciation); while range expansions out of refugia during warm periods resulted in hybridization. However, we know a lot less about the Pleistocene history of less well-studied biodiversity hotspots in the tropics (Hewitt 2004; Hewitt 2011).

In this study we investigate the history of divergence and admixture of three species of pigs (genus *Sus*) from Island Southeast Asia (ISEA). The ISEA archipelago comprises thousands of islands on multiple tectonic plates (Hall 1998). While the islands of Borneo, Sumatra and Java and the Malay Peninsula form a large continental shelf known as the Sunda-shelf (Figure 4.1), other Island clusters such as the Philippines are on different plates. Islands on the same continental shelf are often separated by shallow seas and, given the large scale climatic fluctuations during the Pliocene and Pleistocene and the resulting sea-level changes, were connected by land bridges on many occasions (Hall 1998; Voris 2000). In particular, the sharp climatic transition in the mid Pleistocene (around 700KY) resulted in more frequent glacial cycles and hence exposure of the Sunda Shelf (Elderfield et al. 2012). However, what effect this had on forest cover and the history of those species that depend on it, remains controversial (Gathorne-Hardy *et al.* 2002; Bird et al. 2005; Cannon et al. 2009; Wurster et al. 2010; Silk et al. 2011).

The aim of this study is to characterize the speciation history of pig species in the genus *Sus* in ISEA. We focus on three species: *Sus verrucosus* (Java warty pig; Java, Indonesia), *Sus cebifrons* (Vishayan warty pig; The Philippines) and *Sus scrofa* (the Eurasian wild boar; mainland Eurasia, Sumatra and North Africa). Most species in the genus *Sus*, such as *S. verrucosus* and *S. cebifrons*, are endemic to a single or few islands of ISEA (Meijaard *et al.* 2011). In contrast, *S. scrofa* is a widely distributed species with a natural range extending to most of Eurasia, North Africa and part of

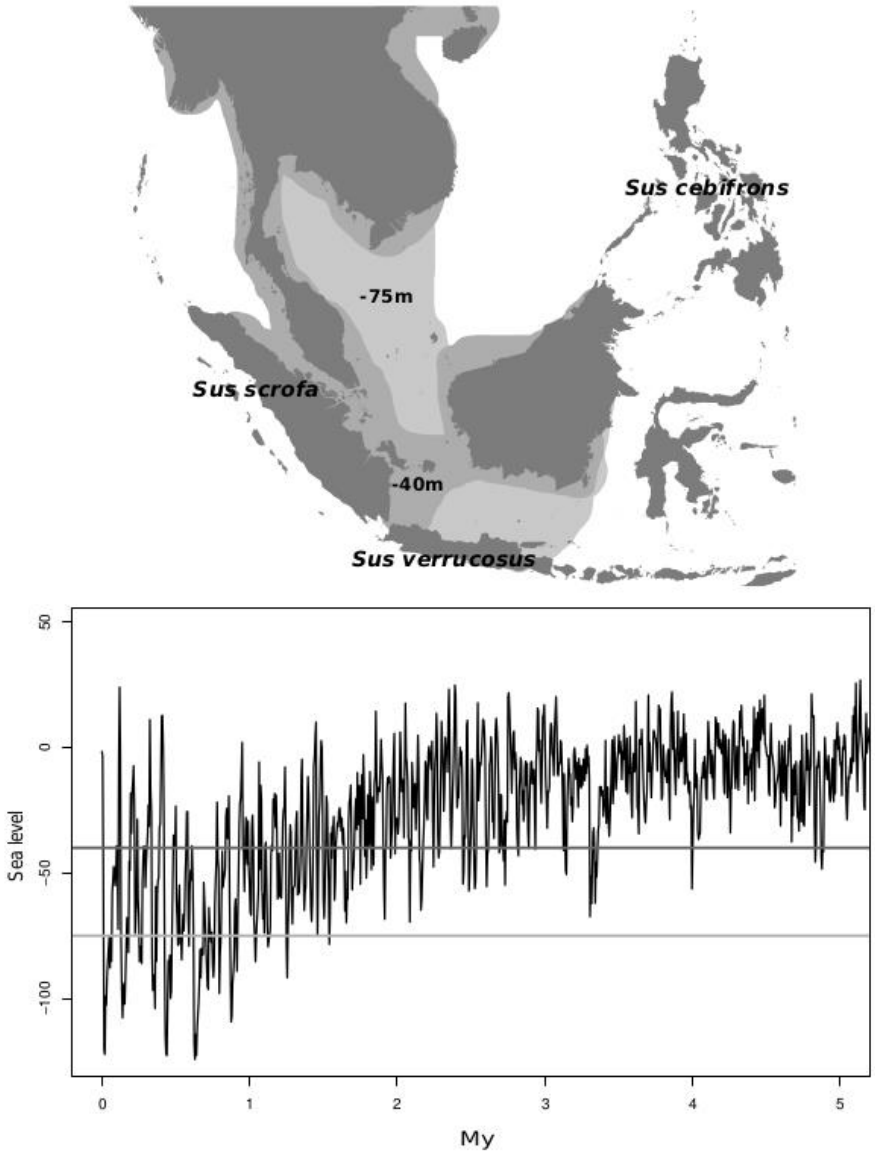


Figure 4.1: Map of Island South East Asia (upper part) with sea-level fluctuations (lower part) over the last 4My (adapted from Miller *et al.* 2005). Dark grey and light grey areas on the map represent the extent of the exposed Sunda-shelf when sea-level at 40m and 75m below current sea-level, respectively.

ISEA (Sumatra; Meijaard *et al.* 2011). In addition, this species has been introduced by humans into multiple regions of the world such as North America, Australia and

Java (Meijaard *et al.* 2011). A previous study showed that *S. verrucosus* from Java (Sunda-shelf) is more closely related to *S. cebifrons* (Visayan Warty pig) in the Philippines than to *S. scrofa* (Eurasian wild boar) on Sumatra (Sunda-shelf; Figure 4.1; Frantz *et al.* 2013). Moreover, this study showed that subsequent inter-specific admixture likely took place on the Sunda-shelf, between *S. scrofa* and *S. verrucosus* after *S. cebifrons* diverged in the Philippines (Frantz *et al.* 2013). However, the timing, magnitude direction of admixture remains unknown. Firstly, it is unclear whether this inter-specific admixture occurred naturally at all or, alternatively, whether it was the result of human-mediated translocation of pig species in ISEA (Groves 1984; Heinsohn 2003; Larson *et al.* 2005; Larson *et al.* 2007; Frantz *et al.* 2013) during the last 70 Ky (Mijares *et al.* 2010). This is crucial for conservation efforts such as *ex-situ* breeding programs, particularly for the endangered Java Warty pig *S. verrucosus* (Semiadi *et al.* 2008) and the critically endangered Visayan warty pig *S. cebifrons* (Oliver 2008). Secondly, if admixture was natural, we would like to understand its temporal context. For example, both divergence and hybridization could be the result of the mid-Pleistocene sharp climatic transition. Alternatively, these species may have diverged much earlier during the late Pliocene or early Pleistocene, when connections between islands on the Sunda-shelf were less frequent (Frantz *et al.* 2013; Figure 4.1) and admixed again during the more frequent and intense glacial period of the latter Pleistocene.

In this study we analysed three genomes of *Sus* from ISEA in a likelihood framework to i) determine if inter-specific admixture between *S. verrucosus* and *S. scrofa* is linked with recent human activities ii) quantify the timing, extent and directionality of this admixture.

4.2 Material and Methods

Data set

We used a genomic dataset from three species of South East Asian pigs that was previously analysed using phylogenetic methods and D statistics (Frantz et al., 2013). The dataset comprises a single unphased diploid genome sampled from a Eurasian wild boar *Sus scrofa* (Sumatran population; Figure 4.1) and the two island endemics *S. verrucosus* (Java; Figure 4.1) and *S. cebifrons* (Philippines; Figure 4.1). Triplet alignments were rooted using *Phacochoerus africanus* (common African warthog) as an outgroup. These genomes were sequenced at 10-20x depth of coverage and aligned to the *S. scrofa* reference genome (Ssc10.2; Bosse *et al.* 2012; Groenen *et al.* 2012; Frantz *et al.* 2013). The likelihood method we use fits explicit models of species divergence and admixture from multilocus data (see likelihood method) and requires short blocks of phased sequence (within which

recombination can be ignored) with equal length (Lohse *et al.* 2011; Hearn *et al.* 2014; Lohse and Frantz 2014). We divided the reference genome of *S. scrofa* into 500 and 1000bp blocks. To ensure enough coverage to call all heterozygous sites in each block and to remove possible CNV (Paudel *et al.* 2013) we filtered out, for each species, any block that had an average read depth of coverage lower than 7x or higher than twice the genome-wide average (Frantz el al. 2013) using the pileup format in SAMtools v0.1.12 (Li *et al.* 2009). Clusters of two or more single nucleotide polymorphisms (SNPs) in a 10bp window were filtered out as well as SNPs within 3bp of an indel. We removed blocks for which less than 90% of the sites were covered and excluded any site that had an effective coverage (Gronau *et al.* 2011) below 4. Lastly, we only selected blocks that passed the above filtering criteria in all 4 samples. We then randomly phased these diploid blocks as a previous study showed that Maximum likelihood estimates (MLE; Lohse and Frantz 2014) are robust to phasing error provided blocks are short. Although the data was phased at random, the low heterozygosity – only 0.12 % of sites were heterozygous in the *S. scrofa* individual from Sumatra, the most out-bred sample (Bosse et al. 2012; Frantz *et al.* 2013) – meant that the majority (67%) of 500bp sequences alignments contained at most one heterozygous site per individual and so were immune to phasing error. Violations of the 4-gamete criterion within a block can arise either due to recombination or back mutation, both of which are not compatible with the assumption of the model (Lohse and Frantz 2014). We therefore excluded blocks containing more than one type of shared derived mutation (6.6% and 15.6% in the 500bp and 1kb datasets, respectively). After applying these filtering steps to the entire pig autosome we are left with 232,373 and 190,692 of 500bp and 1kbp blocks, respectively.

Models

We compared the fit of five nested models to test different scenarios for the evolutionary history of these species (Figure 4.2). All our models assume the order of species divergence inferred by Frantz *et al.* 2013 as (*S. scrofa*, (*S. cebifrons*, *S. verrucosus*)) and have at least three parameters, the species divergence time T_1 (divergence of *S. verrucosus* and *S. cebifrons*), T_2 (the species divergence of *S. scrofa* and *S. verrucosus*/*S. cebifrons*) and a single N_e parameter (constant effective population size). Based on D-statistics analysis (Green et al. 2010; Durand et al. 2011) we assumed that inter-specific gene-flow takes place between *S. verrucosus* and *S. scrofa* after the divergence of *S. cebifrons* (Frantz et al. 2013; Data S1).

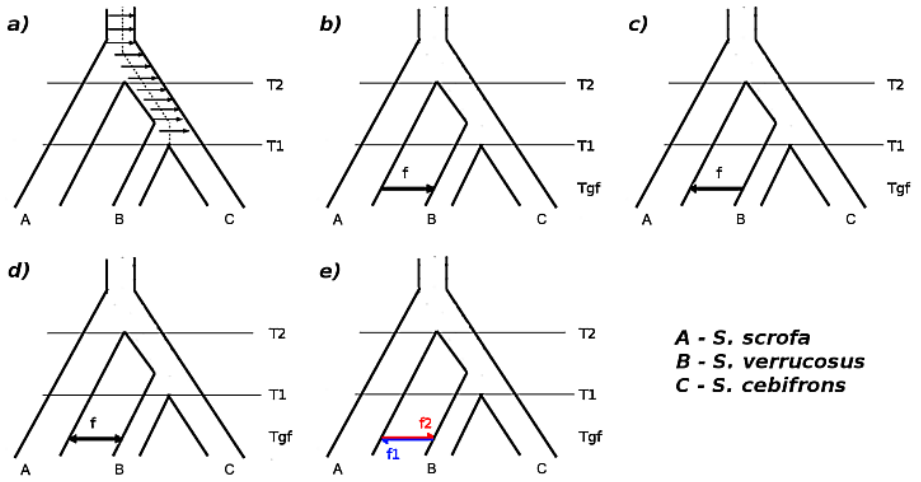


Figure 4.2: Schematic representation of the five models tested in this study. a) Strict divergence (DIV) b) Divergence with gene-flow from *S. scrofa* to *S. verrucosus* (IUA_SS) c) Divergence with gene-flow from *S. verrucosus* to *S. scrofa* (IUA_SV) d) Symmetrical admixture model with equal admixture fractions (ISA) e) Bi-directional admixture model with independent admixture fraction (IBA).

We first assessed the fit of the most complex/general history of instantaneous bi-directional admixture (IBA; Figure 4.2e), *i.e.* a scenario in which admixture between *S. scrofa* and *S. verrucosus* is assumed to happen in both directions. This model involves two admixture parameters (f_1 and f_2) and six parameters in total: T_1, T_2, f_1, f_2, T_{gf} (time of admixture) and N_e. We then assessed the fit of different model simplifications: i) a model of symmetrical admixture (ISA; Figure 4.2d, f_1=f_2=f) and ii) models of instantaneous unidirectional admixture (IUA)(IUA_SS and IUA_SV; Figure 4.2b & 2c) in which admixture goes only one way (either *S. scrofa* → *S. verrucosus* or *S. verrucosus* → *S. scrofa*). These are special cases of the IBA model in which we set either f_1=0 or f_2=0 (Figure 4.2) and so have five parameters. Lastly we evaluated the support of a simple divergence model (DIV; Figure 4.2a) with no inter-specific admixture, *i.e.* f_1=f_2=0.

The assumption of equal population size for all these species may be unrealistic. To test whether adding additional demographic parameters improved model fit, we also evaluated the support of the IUA models with different population size. We tested additional models in which we allowed either *S. scrofa* or both *S. verrucosus* and the ancestral population of *S. verrucosus* and *S. cebifrons* to have a different N_e than the common ancestor of all three species (note that given the sampling

scheme, we have no information about the N_e of *S. cebifrons).* Because a model with two N_e parameters and two admixture fractions is non-identifiable with minimal samples, we only assessed the influence of extra N_e parameters on IUA models (IUA_SS and IUA_SV; Table S1).

Likelihood analysis

Polymorphism information in a block of sequences from three species can be summarized as a vector of counts of mutations on different genealogical branches. For a polarized sample of three sequences there are six possible mutation types: three private and three shared mutations. We will hereafter refer to this vector as the mutational configuration of a block. Lohse *et al.* (2011, eq 1) have shown that the probability of an observed mutational configuration in a particular block can be expressed as a higher order derivative of the Generating Function of genealogical branch lengths. The generating function for triplet samples for the IUA models is described in Hearn *et al.* (2014) and Lohse and Frantz (2014). We give analogous results for the more general case of the IBA model in the supplementary information (Data S1). Assuming (initially) that blocks are unlinked (hence statistically independent observations) the logarithm of the likelihood (lnL) for a particular model is the sum of the lnL across blocks. We maximise the likelihood numerically using *Mathematica* v10. To correct for the effect of linkage when comparing models we re-scaled the difference in lnL between models as described in Lohse and Frantz (2014). We assumed that the effect of physical linkage between blocks separated by a distance of 100kb can be ignored (Tortereau *et al.* 2013). Further details of the general method for computing likelihoods and 95% CI of parameters are given in Lohse *et al.* (2011) and Lohse and Frantz (2014). For each model we computed ΔlnL, the difference in log likelihood to the best fitting model. We assessed statistical support between nested models in a likelihood ratio test and assumed that $2*\Delta lnL$ follows a χ^2 distribution with degree of freedom equal to the difference in the number of parameters of the two models (see Table 4.1).

To compare our approach with the D-statistics (Green et al. 2010; Durand et al. 2011) and to obtain a rough assessment of goodness of fit, we computed the expected counts of ABBA and BABA sites and E[D] from the generating function under the different admixture models (by fixing parameters to their MLE estimated from the data; see Data S1).

In order to scale relative time estimates into absolute values, we assumed an average divergence time between the African warthog and the ingroup of 10.5 MY and a generation time of 5 years (Gongora *et al.* 2011; Groenen *et al.* 2012; Frantz *et al.,* 2013).

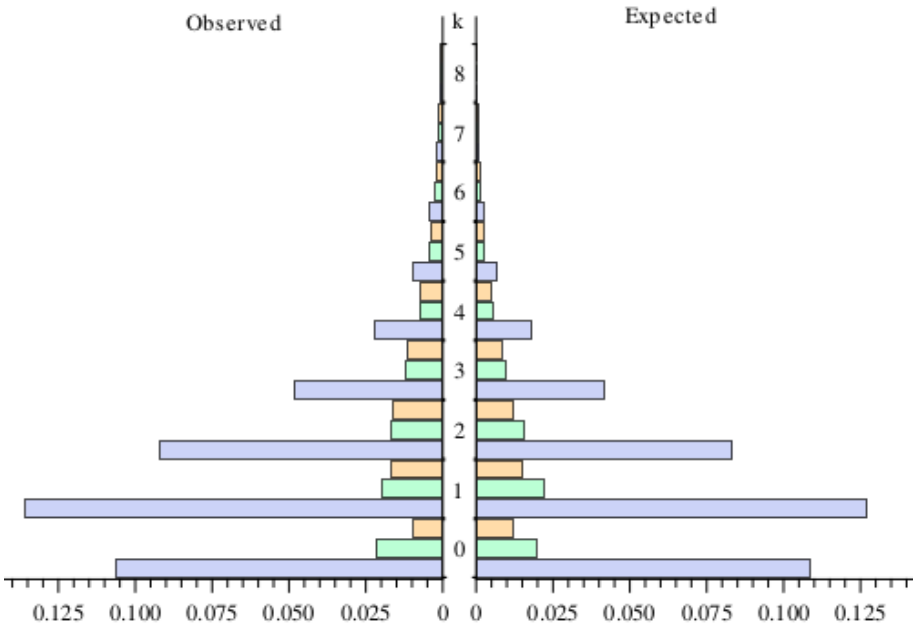


Figure 4.3: Expected (under IBA model with parameters fixed at their MLE) and observed mutational configuration. The X axis represents the proportion of blocks with k mutations (Y axis) for different topologies. Blue bars = (*S. scrofa*, (*S. verrucosus*, *S. cebifrons*)), green bars = (*S. verrucosus*, (*S. cebifrons*, *S. scrofa*)) and orange = (*S. cebifrons*, (*S. verrucosus*, *S. scrofa*)).

4.3 Results

Model comparison

By definition, the IBA model (Figure 4.2e) provided a better fit than any of the simpler nested models (Table 4.1). Setting admixture fractions to be equal (f_1=f_2=f; Figure 4.2d) significantly reduced the fit (ISA model; ΔlnL=-10.1 and -9.24 for 500bp and 1kb, respectively; Table 4.1). This difference in likelihood is highly significant assuming a $\chi 2$ distribution ($p<0.001$; Table 4.1). Likewise, a model in which f_1=0 (IUA_SS; Figure 4.2c), *i.e.* corresponding to a history with admixture only from *S. scrofa* into *S. verrucosus* also gave a significantly worse fit (ΔlnL=-22.8 and -26.8 for 500bp and 1kb, respectively; Table 4.1). In contrast, setting f_2=0 (IUA_SV model; Figure 4.2b) only marginally reduced the fit (ΔlnL=-1.77 [$p>0.05$] and -1.68 [$p>0.05$] for 500bp and 1kb, respectively; Table 4.1). Thus, a model of unidirectional admixture, with f_2=0, from *S. verrucosus* into *S. scrofa* cannot be excluded.

Lastly, a strict divergence model, *i.e*. f_1=f_2=0 (DIV model; Figure 4.2a) provided a significantly worse fit (*ΔlnL*=-49.3 and -80.1 for 500bp and 1kb, respectively; Table 4.1). These results demonstrate that this genomic data-set contains a strong signal of inter-specific admixture between *S. scrofa* and *S. verrucosus,* but surprisingly most of this admixture was from *S. verrucosus* into *S. scrofa,* so in the opposite direction than that assumed by previous studies (Frantz *et al.* 2013).
Including additional N_e parameters for different populations (see Methods) did not significantly improve the fit (Table S1). To get a sense of how well different admixture histories explain the data, we computed the expected D statistic (E[D]) under each admixture scenario and compared it to the observed value. This also allowed us to assess the sensitivity of D admixture in different directions. Constraining admixture to be from *S. scrofa* to *S. verrucosus* (the best fitting IUA_SV model) gives E[D]= 0.22, while limiting admixture to the opposite direction only (the IUA_SS model) gives E[D]=0.12. The observed D of 0.175 (for 0.5 kb data) is in between and matches E[D] under the estimated IBA model E[D]=0.16. Thus, the unidirectional models both fit the data worse than the bidirectional scenario (IBA). Comparing the number of mutations on external branches for each of the three possible topologies expected under the IBA model to the observed spectrum of mutation counts reveals a tight fit (Figure 4.3), suggesting that the IBA model explains most of the signal in the data.

Table 4.1: Model description and difference in likelihood support compared to best fitting model for 500bp and 1kb blocks. Significance was obtained using a likelihood ratio test (2**ΔlnL*) and a chi-square distribution (** p<0.001; * p<0.01).

Acronym (parameters)	Description	ΔlnL (500)	ΔlnL (1k)
DIV (3)	Strict divergence (no gene-flow) with or without ancestral substructure (Figure 4.2a)	-49.3**	-80.1**
IUA_SS (5)	Divergence with gene-flow from *S. scrofa* to *S. verrucosus* (Figure 4.2b)	-22.8**	-26.8**
IUA_SV (5)	Divergence with gene-flow from *S. verrucosus* to *S. scrofa* (Figure 4.2c)	-1.7 N.S.	-1.6 N.S.
ISA (5)	Symmetrical admixture model with equal admixture fractions. (Figure 4.2d)	-10.1**	-9.24**
IBA (6)	Bi-directional admixture model with independent admixture fraction (Figure 4.2e)	0	0

Parameter inference

Maximum likelihood estimates for each parameter (Figure 4.4) were obtained under the best fitting models (IBA and IUA_SV). Our first goal was to determine whether the admixture between *S. scrofa* and *S. verrucosus* could have been mediated by humans. The marginal curves for the time of admixture, under both IBA and IUA_SV models, show very little support for values of T_{gf} below 70Ky (ΔlnL<-6; Figure 4.4b), the time of the earliest human arrival in the region (Mijares *et al.* 2010), which strongly suggests that humans did not play a role in this admixture.

Our second goal was to put the initial divergence between these species in the context of the climate history during the Pleistocene. Our point estimate for T_2, *i.e.* the deeper speciation event in this study (*S. scrofa* and *S.verrucosus*/*S. cebifrons* split, T_2 in Figure 4.2) is approximately 4My (Figure 4.4b) for both block sizes, with lower 95% CI much greater than 2.5My (Plio-Pleistocene transition; Figure 4.4b). We estimated the split between the Javan warty pig (*S. verrucosus*) and the Visayan warty pig (*S. cebifrons*) (T_1) to be between 1.3-1.1My (Figure 4.4b) These divergence time estimates agree well with previous analyses based on the same molecular clock (Frantz *et al.* 2013).

4.4 Discussion

In this study we show that the extent and the directionality of inter-specific admixture between Sunda-shelf *Sus* species (Figure 4.1) is more complex than previously assumed (Frantz *et al.* 2013). Firstly, our likelihood approach allows us to rule out any major influence of humans in this admixture event. Secondly, our analysis suggests that ISEA *Sus* species diverged during warmer periods of the late Pliocene and hybridized during the more frequent glacial periods of the mid-Pleistocene. Finally, and perhaps surprisingly, our analyses suggest that admixture occurred mainly from *S. verrucosus* into *S. scrofa,* so in the opposite direction than that assumed previously.

Fine scale model testing using a likelihood approach

Models involving inter-specific admixture between *Sus* species on the Sunda-shelf (Figure 4.1) fitted our genomic data set significantly better than a simple divergence model without gene-flow (Table 4.1). We estimated a 23% admixture fraction from the Java Warty pig (*S. verrucosus*) into the *S. scrofa* population on Sumatra (Figure 4.4). This admixture can explain the large discrepancies found between nuclear and mtDNA phylogenetic analyses that found that *S. verrucosus* and Sumatran *S. scrofa* share very similar mtDNA haplotypes and form a monophyletic clade with short external branches (Larson *et al.* 2005; Larson *et al.*

2007; Frantz *et al.* 2013). Moreover, it seems that replacement of mtDNA in either *S. verrucosus* or *S. scrofa* from Sumatra was complete as no divergent mtDNA haplotypes have been found by any previous study (Larson *et al.* 2005; Larson *et al.* 2007).

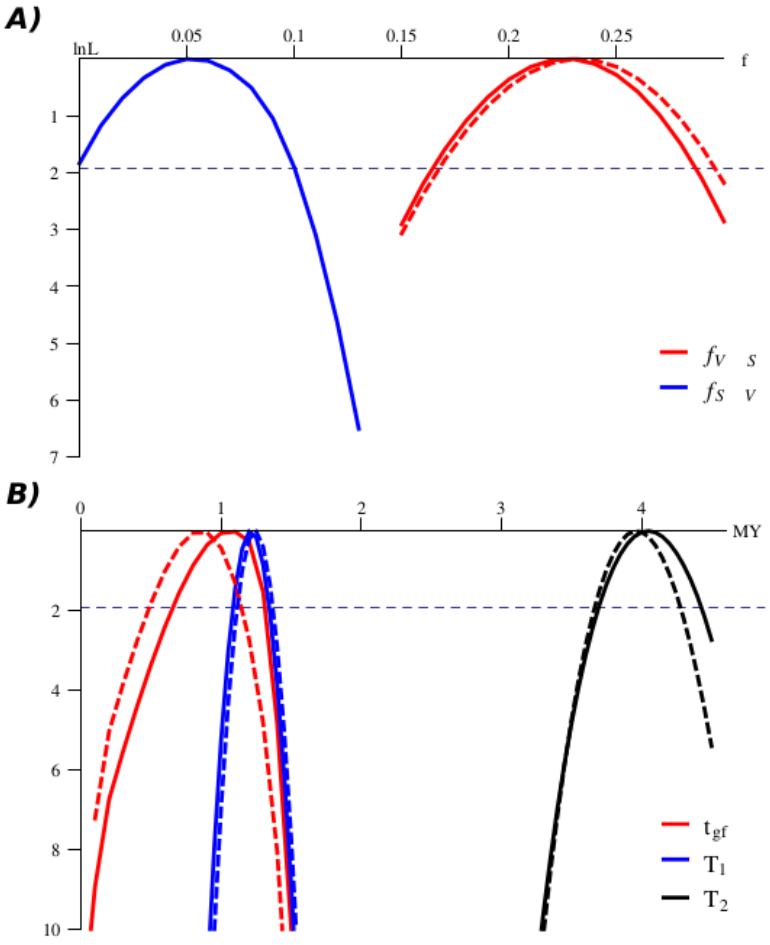


Figure 4.4: Marginal support (*ΔlnL*) for a) admixture fraction (f) and b) divergence and admixture times. Solid and dashed lines represent *ΔlnL* curves under the IUA_SV and IBA models respectively. The black dashed horizontal lines delimit the 95% confidence interval. a) admixture fraction from *S. verrucosus* into *S. scrofa* and from *S. scrofa* into *S. verrucosus* are shown in red and blue represent respectively b) T_2, T_1 and T_{gf} are shown in black, blue and red respectively.

This suggests that interspecific admixture can lead to complete mtDNA replacement even with an admixture fraction of ~ 23% and illustrates that phylogenetic and phylogeographic studies that rely solely on mtDNA can be highly misleading.

Our results also show that previous analyses of these genomes (Frantz *et al.* 2013), using D-statistics (Green *et al.* 2010; Durand *et al.* 2011), only incompletely resolved this interspecific admixture. This is because the D-statistics are uninformative about the direction of admixture (Green *et al.* 2010). Frantz *et al.* 2013 had assumed a history unidirectional admixture from *S. scrofa* into *S. verrucosus* to be able to use ABBA/BABA counts as a mean to compute an upper bound for the admixture fraction under this model (see Additional File 6 in Frantz et al. 2013). In contrast, the joint distribution of branch lengths, used in our likelihood approach, contains additional information about the direction of admixture and our analysis reveals that *S. verrucosus* is largely the source of admixture rather than the recipient. While we can show that the IBA model fitted slightly better than the IUA_SV model, the difference was not significant (Table 4.1). In other words, although our estimate for *f* (IBA model) from *S. scrofa* to *S. verrucosus* was very similar to the fraction estimated by Frantz *et al.* 2013 (4% versus 5% in this study) our 95% CI for this parameter also include 0 (Figure 4.4a). Therefore, while our results unequivocally support that *S. verrucosus* was the most important source of inter-specific admixture we cannot rule out that this species was also a recipient. Together our analyses show that it is important to interpret admixture fractions computed based on D-statistics with caution when the direction of the admixture is unknown.

Natural inter-specific admixture on the Sunda-shelf

Our analysis showed that most of the inter-specific admixture between the Sunda-shelf species took place before humans arrived in the region and so anthropogenic disturbances are unlikely to explain this phenomenon. However, while useful, our models are necessarily over-simplistic. For example, we assumed that admixture was a single, instantaneous event. However, the Sunda shelf was likely exposed during multiple glacial cycles in the mid-Pleistocene (Voris 2000), which in turn could have lead to many admixture events. Therefore, although we ruled out humans as the cause for most of this admixture, this does not exclude the possibility that a small amount of admixture occurred more recently as a result of anthropogenic disturbances. This is especially true given that previous studies have found that humans most likely translocated *Sus* species in the region (Groves 1984; Heinsohn 2003; Larson *et al.* 2005, 2007; Frantz *et al.* 2013). Thus, the genomic

signature left by human mediated translocation of species may be confounded with the signal of large scale naturally occurring admixture.

We also evaluated the support of a model of symmetrical admixture (ISA; Table 4.1). Our model comparison clearly demonstrates that this scenario provides a significantly poorer fit than the IBA model. This is unsurprising given the large difference in admixture fraction under the IBA model (5% versus 23%). This asymmetry could arise in at least two ways: Firstly, a low effective population size (N_e) of Sumatran *S. scrofa* – perhaps as a result of a founder event when this species colonized Sumatra from the mainland at the time of admixture could explain this observation. However, the origin of *S. scrofa* (on the mainland or on ISEA) remains controversial due to the difficulty of inferring demographic events that took place more than 2My ago (Frantz *et al.* 2013). Alternatively, this discrepancy in the admixture fraction could be the result of greater mate discrimination against hybrids by *S. verrucosus*. This interpretation is difficult to assess given the very sparse ecological and behavioural data available for the Javanese warty pig *S. verrucosus* (Blouch 1993). However, such information is available for *S. scrofa*, an invasive generalist that can easily colonize new environments (Barrios-Garcia and Ballari 2012). *S. scrofa* can be found natively all over Eurasia, in Sumatra and parts of North Africa. Moreover, feral *S. scrofa* have been able to colonize new ecosystems in Australia, Hawaii, Java, North America and many other parts of the globe in recent years. Thus, given the generalist behaviour of this species, its wide range and its ability to colonize new environments and the relatively narrow range of *S. verrucosus* (restricted to a few areas on Java) mate discrimination against hybrids by the latter would appear more probable. Such an asymmetry in sexual selection against hybrids has been suggested in mice (*i.e.* Latour *et al.* 2014). Further research on the ecology and the behaviour of the Java warty pig is needed to better interpret these results, especially given its endangered status. More efficient mate discrimination against hybrids by *S. verrucosus* would have important consequences for on-going conservation effort. Indeed, one of the major threats to *S. verrucosus*, listed by the IUCN Wild Pig Specialist Group (Semiadi *et al.* 2008), is the hybridization with the potentially recently introduced *S. scrofa* on Java. However, hybrids are difficult to identify in the wild and the extent of this threat remains unknown (Semiadi *et al.* 2008). Disentangling these two hypotheses (hybrid recruitment versus founder effect) would provide crucial information for the conservation of *S. verrucosus*.

Adding parameters to model population size difference between these species did not improve the fit of our models (Table S1). This does of course not imply that these species have the same effective population size, but rather demonstrates

that there is little information in the block-wise data to fit more realistic histories (Figure 3). In contrast, the information contained in linkage across longer stretches of the genome suggests that these species have experienced substantial changes in N_e. For example both *S. verrucosus* and *S. cebifrons* carry long runs of homozygosity and have a low current N_e that was attributed to very recent bottlenecks possibly due to anthropogenic disturbances (Bosse *et al.* 2012; Frantz *et al.* 2013). In addition, all three species showed demographic signals consistent with long term bottlenecks during the Pleistocene (Frantz *et al.* 2013). Lastly, as discussed above, in the IBA model f and N_e are almost entirely confounded (Lohse *et al.* 2011), thus it is not surprising that there is no additional power to estimate variation in the latter.

The impact of Plio-Pleistocene glaciations on the evolutionary history of Sus

Our analyses show that the divergence between Sunda-shelf *Sus* species (Figure 4. 4b) took place during the Pliocene around 4My ago (Figure 4. 4b). Moreover, we found that the divergence between *S. verrucosus* and *S. cebrifrons* took place over 1.2My ago, during the early Pleistocene before the sharp climatic transition ~700Ka (Figure 4. 4b). This suggests that the milder climatic fluctuations of the Pliocene (Figure 4.1) allowed for dispersal between 'islands' during short glacial period and subsequent isolation during long warm periods (due to high sea-level), while the longer and more frequent ice ages of the late Pleistocene which resulted in longer exposure of the Sunda-shelf (due to low sea-level) led to a partial merging of gene pools. Thus, if true, the effect of the Plio-Pleistocene climatic fluctuation may act in reverse in ISEA when compared to more temperate regions such as Europe, in which glacial maximas induce divergence (assemblage of refugia) and inter-glacial periods induce range expansions and hybridization (*i.e.* Hewitt 2000, 2004; Schmitt 2007; Hewitt 2011). However, our large confidence intervals around time parameters (divergence and admixture) as well as our model assumptions (single instantaneous admixture) do not provide the necessary resolution to correlate these events with individual glacial cycles during the Plio-Pliestocene era (Zachos et al. 2005; Elderfield et al. 2012). Further studies using larger data sets, increasingly sophisticated methods and combining historical inferences from multiple species will shed light on the mechanisms that generated and erased biodiversity in this mega biodiverse region of the world (Myers et al. 2001).

Acknowledgments

This paper is dedicated to the memory of our colleague William Oliver, who devoted his life to conservation biology. We would like to thank members of the IUCN Wild Pig Specialists Group, in particular, Erik Meijaard and Gono Semiadi for their fruitful insights into the conservation status and challenges of S. verrucosus. This project was financially supported by a European Research Council grant (ERC-2009-AdG: 249894) and a junior research fellowship from the Natural Environmental Research Council (NE/I020288/1) to KL.

Supplementary Information

Supplementary informations are available with the online version of the manuscript at http://onlinelibrary.wiley.com/doi/10.1111/mec.12958/full. This includes the derivation of the IBA model as well as Table S1.

Data accessibility

BAM files: EBI Sequence Read Archive ERP001813

Mathematica notebook and input files (mutational configurations for 500bp and 1kbp blocks): Dryad doi:10.5061/dryad.1q0h6

References

Barrios-Garcia MN, Ballari SA (2012) Impact of wild boar (*Sus* scrofa) in its introduced and native range: a review. *Biological Invasions*, **14**, 2283-2300.

Bird MI, Taylor D, Hunt C (2005) Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland? *Quaternary Science Reviews*, **24**, 2228-2242.

Blouch RA (1993). The Java Warty Pig (*Sus verrucosus*). In: Pigs, peccaries and hippos: status survey and conservation action plan (eds, Oliver WLR), pp 129-136. IUCN, Gland, Switzerland.

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.

Cahill JA, Green RE, Fulton TL *et al.* (2013) Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. (MW Nachman, Ed,). *PLoS genetics*, **9**, e1003345.

Cannon CH, Morley RJ, Bush ABG (2009) The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 11188-93.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, **28**, 2239-52.

Elderfield H, Ferretti P, Greaves M *et al.* (2012) Evolution of ocean temperature and ice volume through the mid-Pleistocene climate transition. *Science*, **337**, 704-9.

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

Gathorne-Hardy FJ, Davies RG, Eggleton P, Jones DT (2002) Quaternary rainforest refugia in south-east Asia: using termites (Isoptera) as indicators. *Biological Journal of the Linnean Society*, **75**, 453-466.

Gongora J, Cuddahee RE, Nascimento FFD *et al.* (2011) Rethinking the evolution of extant sub-Saharan African suids (Suidae, Artiodactyla). *Zoologica Scripta*, **40**, 327-335.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710-22.

Groenen MAM, Archibald AL, Uenishi H *et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, **491**, 393-8.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, **43**.

Groves CP (1984) Of mice and men and pigs in the Indo-Australian Archipelago. *Canberra Anthropology*, **7**, 1-19.

Hall R (1998). The plate tectonics of Cenozoic SE Asia and the distribution of land and sea. In: Hall, R. & Holloway, J. D. (eds.) Biogeography and Geological Evolution of SE Asia. Backhuys Publishers, Leiden, The Netherlands, 99–131

Hearn J, Stone GN, Bunnefeld L *et al.* (2014) Likelihood-based inference of population history from low-coverage de novo genome assemblies. *Molecular ecology*, **23**, 198-211.

Heinsohn T (2003) Animal translocation: long-term human influences on the vertebrate zoogeography of Australasia (natural dispersal ver*sus* ethnophoresy). *Australian Zoologist*, **32**, 351-376.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907-13.

Hewitt GM (2004) Genetic consequences of climatic oscillations in the Quaternary. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **359**, 183-95; discussion 195.

Hewitt GM (2011) Quaternary phylogeography: the roots of hybrid zones. *Genetica*, **139**, 617-38.

Larson G, Cucchi T, Fujita M *et al.* (2007) 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**, 1618-21.

Latour Y, Perriat-sanguinet M, Caminade P *et al.* (2014) Sexual selection against natural hybrids may contribute to reinforcement in a house mouse hybrid zone. *Proceedings. Biological sciences / The Royal Society*, **281**, 20132733.

Lawniczak MKN, Emrich SJ, Holloway AK *et al.* (2010) Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. *Science*, **330**, 512-4.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078-9.

Lohman DJ, Bruyn MD, Page T *et al.* (2011) Biogeography of the Indo-Australian Archipelago. *Annual Review of Ecology and Systematics*, **42**, 205-228.

Lohse K, Harrison RJ, Barton NH (2011) A general method for calculating likelihoods under the coalescent process. *Genetics*, **189**, 977-87.

Lohse K, Frantz LAF (2014) Neandertal Admixture in Eurasia Confirmed by Maximum Likelihood Analysis of Three Genomes. *Genetics*, genetics.114.162396-.

Meijaard E, d'Huart JP, Oliver WLR. (2011) Family Suidae (Pigs). In Handbook of the Mammals of the World. Volume 2. Edited by Wilson DE, Mittermeier RA. Barcelona, Spain; Lynx Edicions; 2011:248-291.

Mijares AS, Détroit F, Piper P *et al.* (2010) New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines. *Journal of human evolution*, **59**, 123-32.

Miller KG, Kominz MA, Browning JV *et al.* (2005) The Phanerozoic record of global sea-level change. *Science*, **310**, 1293-8.

Myers N, Mittermeier RA, Mittermeier CG, Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. Nature , **403**, 853-858.

Oliver, W. (2008). *Sus cebifrons*. In: IUCN 2014. IUCN Red List of Threatened Species. Version 2014.1. <www.iucnredlist.org>. Downloaded on 22 May 2014.

Paudel Y, Madsen O, Megens H-J *et al.* (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC genomics*, **14**, 449.

Rohland N, Reich D, Mallick S *et al.* (2010) Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS biology*, **8**, e1000564.

Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in zoology*, **4**, 11.

Slik JWF, Aiba S-I, Bastian M *et al.* (2011) Soils on exposed Sunda shelf shaped biogeographic patterns in the equatorial forests of Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 12343-7.

Semiadi G, Meijaard E, & Oliver W (2008). *Sus* verrucos*us*. In: IUCN 2013. IUCN Red List of Threatened Species. Version 2014.1. <www.iucnredlist.org>. Downloaded on 22 May 2014.

Stewart JR, Lister AM, Barnes I, Dalén L (2010) Refugia revisited: individualistic responses of species in space and time. *Proceedings. Biological sciences / The Royal Society*, **277**, 661-71.

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.

Voris HK (2000) Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *Journal of Biogeography*, **27**, 1153-1167.

Wurster CM, Bird MI, Bull ID *et al.* (2010) Forest contraction in north equatorial Southeast Asia during the Last Glacial Period. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15508-11.

Zachos J, Pagani M, Sloan L, Thomas E, Billups K (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, **292**, 686-93.

5

Speciation and domestication history of *Sus scrofa*

Laurent AF Frantz [1], Hendrik-Jan Megens [1], Mirte Bosse [1] , Joshua G Schraiber [2], Yogesh Paudel [1], Richard PMA Crooijmans [1], and Martien AM Groenen [1]

[1] Animal Breeding and Genomics Centre, Wageningen University, Droevendaalsesteeg 1, Wageningen, 6708 PB, The Netherlands. [2] Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA.

Published as part of Nature (2012) 491:393–398

Abstract

Molecular genetic evidence indicates that *Sus scrofa* emerged in Southeast Asia during the climatic fluctuations of the early Pliocene 5.3-3.5 MYA. Then, beginning ~10,000 years ago, pigs were domesticated in multiple locations across Eurasia. However, many aspects of the evolutionary history of the species remain unknown. In this paper we use genome data from over 55 samples of wild and domestic pigs to investigate multiple aspects of the evolutionary history of this widely spread species. The demographic history of this widespread species is remains unknown. Thirdly, we know little about the time of divergence of the Asia and European subtypes of wild boars that have been domesticated independently. Lastly, the evolutionary history of domestic pigs and wild boar is poorly known. For example, we do not know how common interbreeding was between wild and domestic pigs or between Asia and European domestic pigs or how domestication affected demography. Phylogenomic analyses of complete genome sequences from these wild boars and six domestic pigs revealed distinct Asian and European lineages that split during the mid-Pleistocene 1.6-0.8 MYA (Frantz et al. 2013; Calabrian stage). Our demographic analysis on the whole genome sequences of European and Asian wild boars, revealed an increase in the European population after pigs arrived from China. During the Last Glacial Maximum (LGM; ~20KYA), however, Asian and European populations both suffered through bottlenecks. These bottlenecks were more pronounced in Europe than Asia suggesting a greater impact of glaciation on higher latitude regions. In addition, our admixture analysis revealed European influence in Asian breeds, and a ~35% Asian fraction in European breeds. These results are consistent with the known exchange of genetic material between European and Asia pig breeds. We also observed that European breeds form a paraphyletic clade, which cannot be solely explained by varying degrees of Asian admixture. Within each continent, our analysis revealed different degrees of relatedness between breeds and their respective wild relatives.

Keywords: demography, domestication, population genetics

5.1 Introduction

The domestic pig (*Sus scrofa*) is a eutherian mammal and a member of the Cetartiodactyla order, a clade distinct from rodent and primates, that last shared a common ancestor with man between 79 and 97 million years ago (MYA). Molecular genetic evidence indicates that *Sus scrofa* emerged in Southeast Asia during the climatic fluctuations of the early Pliocene 5.3-3.5 MYA. Then, beginning ~10,000 years ago, pigs were domesticated in multiple locations across Eurasia. However, many aspects of the evolutionary history of the species remain unknown. The geographic origin of the species is puzzling. Indeed the wide range that the species inhabit makes it difficult to pinpoint the location of an ancestral population from which the species colonized Eurasia and North Africa. In addition, little is known about their demographic history. Moreover, the time of divergence of the two main subtypes of wild boars that were independently domesticated (Larson 2005). Such knowledge will provide clues upon the genetic differentiation between Asia and European domestic pigs. Lastly this work investigates the relationship between wild and domestic pigs in Europe and Asia. This approach presented in this study, aim at characterizing the extent of gene-flow between wild and domestic *S. scrofa.* In addition such analysis is expected to provide clues upon the history and the importance of breed trading within and between Europe and Asia. In this study we use the whole-genome sequence of ten domestic and wild boars and an African Warthog as an outgroup (Table 1). Our analysis provides answers to many interesting aspects on the Evolutionary history of *S. scrofa*. In particular, we provide clues upon the speciation and domestication history of this important livestock species.

5.2 Material and Methods

Prior to whole-genome sequencing of the animals used for the present study, 60K single nucleotide polymorphism (SNP) genotype data of candidate animals were compared against a large dataset of 60K SNP genotype data for *Sus scrofa* and related species from all continents with the exception of Antarctica. We have genotyped over 3000 individuals with the Illumina PorcineSNP60 chip (Ramos et al., 2009) and sequenced the D-loop of the mtDNA for all individuals. A representative selection of these individuals is shown in Figure 5.1. The dendrogram demonstrates that the wild boars sampled for the present study are representative for the geographic extremes of continental Eurasia. In addition, it demonstrates that the domesticated animals used for the population genetic analysis are highly representative for pigs of Europe and China. The analysis was based on 50,492 SNPs from the Illumina PorcineSNP60 chip that were mapped to the autosomal

chromosomes in Sscrofa10.2. The design of the 60K SNP assay was mainly based on SNPs discovered in European pigs, and the very deep genetic divide between European and East-Asian *S. scrofa* that is evident from the current study (see also (Megens et al. 2008)is expected to result in a high ascertainment bias. Note that with increasing topological distance to European pigs the branch lengths decrease, which is a clear manifestation of ascertainment bias.

The analysis represents most of the major Eurasian areas as defined by mitochondrial analysis by Larson et al., 2005. For context, *Sus scrofa* from Sumatra was included that represents the 'basal clade' as defined by Larson et al., 2005. Genotyping using the PorcineSNP60 chip on other species in the genus *Sus* and African Warthog would yield relatively high genotyping success (>90%, compared to typically >97% for *Sus scrofa*). Despite assay success, the SNP loci as defined for European pigs would usually not be polymorphic in these species, with only a few percents of SNPs actually found to be polymorphic. Further discussion on the other species in the genus *Sus* are presented elsewhere (Frantz et al., 2013; Chapter 3). Note that the two wild boar populations from the Netherlands actually form two distinct populations. Both are more related to the French wild boar than to the Wild boar from the Italian Peninsula.

The European pigs sequenced for this study were all derived from commercial breeds. To demonstrate that these animals are good representatives of the entire breed, animals from at least two distinct populations were selected. Invariably, animals from the same breed cluster together, which shows that despite many generations of selective breeding the breed concept – at least at the genetic level – remains intact as expected (c.f. Megens et al., 2008). Note that the Duroc breed, that also includes TJ Tabasco (Duroc 2-14), tends to cluster basally to all other European or European-derived Sus scrofa. This has been observed previously (e.g. Megens et al., 2008). The documented history of the breed is rather sketchy, and despite its clear genetic and historic relationships to European pigs, its precise origin is mostly unknown.

Sequencing, alignment and SNP calling

Genome re-sequencing was targeted at a depth of around 8-10x. All sequencing was performed on Illumina HiSeq2000 sequencers. Library construction and re-sequencing of the individual samples was performed with 1-3 µg of genomic DNA according to the Illumina library prepping protocols (Illunima Inc.).

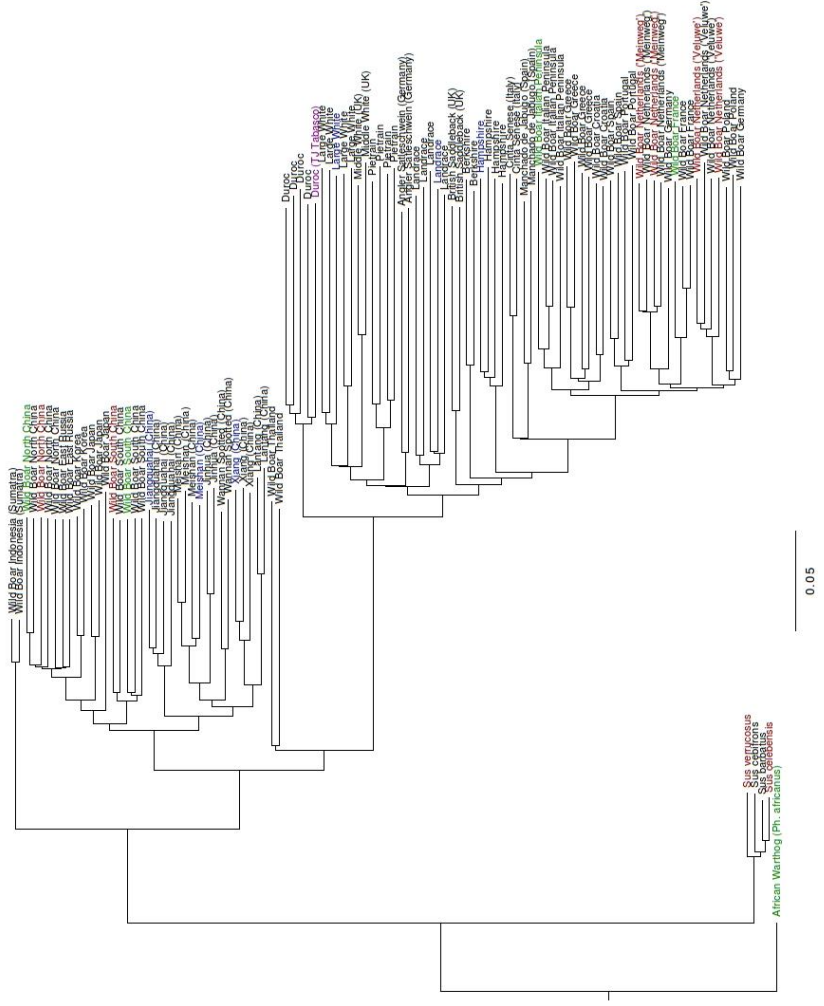


Figure 5.1 Dendrogram showing the variation that we sampled by re-sequencing, placed in a larger context of animals from the same population. For further context, other animals from different populations were added. The animals used in this study have a blue label if used for the population genetic analysis (Nature (2012) 491:393–398, Supplementary material section 9), a red label if used for the selective sweep analysis (Nature (2012) 491:393–398, Supplementary material section 10), or green if used in both. Genotype data from the animal used for the genome assembly (the Duroc sow named 'TJ Tabasco'), was also included with a purple label. A selection of animals from the same population was included, and in addition representatives of other populations of European and East-Asian wild and domestic *Sus scrofa* were included, with black labels. Pigs, wild boar, and outgroups were genotyped using the Illumina PorcineSNP60 chip (Ramos et al., 2009) per

the manufacturer's instructions. Pairwise IBS scores were calculated using PLINK v1.07 (Purcell, 2009). Hierarchichal clustering was done using the 'Neighbor' program, which is part of the Phylip phylogenetic analysis package (Felsenstein 2009).

The library insert sizes ranged for 300-500 bp and sequencing was performed with the 100 paired-end sequencing kit. Reads were quality trimmed prior to sequence alignment. The trimming strategy involved a 3 bp sliding window, running from 5` to 3`, with sequence data upstream being discarded if the 3-bp window average quality dropped below 13 (i.e. average error probability equal to 0.05). Only sequences 45 bp or more in length were retained. In addition, sequences with mates <45 bp after trimming were also discarded. During trimming, quality scores were recoded to follow the Sanger fastq format to standardize upstream processing.

Sequence alignment was done against the *Sus scrofa* genome, build 10.2, using Mosaik 1.1.0017. We initially used both BWA and Mosaik in our SNP detection pipeline. After evaluation of the false discovery rate in regions of the genome where individuals are homozygous for a single haplotype, it was decided to use Mosaik for our population study. Aligning was done using a hash-size of 15, with maximum of 10 matches retained, and using a 7% maximum mismatch score, for all populations and the outgroup species. Post aligning, alignment files were sorted using the 'Mosaiksort' function, which entails removing ambiguously mapped reads that are either orphaned or fall outside a computed insert-size distribution. Alignment archives were converted to BAM format (Li et al., 2009) using the Mosaiktext function. Manipulations of BAM files, such as merging of alignments archives pertaining the same individual, were done using samtools v. 1.12a (Li et al., 2009).

Variant calling was performed per individual using the 'pileup' function in samtools, and variations were initially filtered to have minimum quality of 50 for indels, and 20 for SNPs. In addition, all variants that had a higher than 3x the average read density estimated from the number of raw sequence reads obtained were also discarded, to remove false positive variant calling originating from off-site mapping as much as possible.

To obtain genotype calls for all the polymorphic sites identified across all individuals, every individual was interrogated for the genotype call for each of the sites found to be polymorphic, including the species-specific differences. Sequence depth, SNP and consensus quality were retrieved for these sites using the samtools pileup function. These *de facto* genotype calls were subsequently filtered based on sequenced depth (minimum sequence depth of 4, and maximum of 2x the average

Table 5.1 Number of filtered SNPs, in autosomal chromosomes and the X chromosome, per individual after filtering for non-uniquely mapping reads.

Sample	Read depth	Fixed SNPs against reference	Heterozygous SNPs
Landrace (LR)	10.4x	2,420,631	2,420,631
Large White (LW)	10.8x	2,616,584	2,254,121
Hampshire (HA)	12.3x	2,875,911	2,004,188
European - NL (WBNL)	11.8x	3,163,655	1,376,164
European – IT (WBIT)	15.1x	3,238,530	1,294,633
Meishan (MS)	9.3x	5,560,909	2,836,716
Xiang (XI)	9.2x	5,481,531	2,696,464
Jiangquhai (JQ)	11.2x	5,124,983	2,750,918
North Chinese (WBNC)	10.7x	4,999,191	3,034,822
South Chinese (WBSC)	10.5x	4,967,382	4,090,363
Phacochoerus Africanus (Pafri)	13.5x	23,000,541	2,159,994

genome-wide depth), where in this case the average sequence depth was established based on the actual sequence depth for each individual separately. Further filtering was done on SNP and consensus quality (in case the individual was homozygous, either SNP or consensus quality > 2, in case the individual was heterozygous, both consensus and SNP quality > 20). All indels were removed.

After the filtering, genotype calls were established for a total of 66,668,635 single nucleotide positions in the genome.

For phylogenetic analysis, we identified genomic bins in each sample separately that had an average depth below 2x the genome-wide average depth. We then excluded clusters of 3 SNP in 10 bp and within 3 bp of an indel, in each bin, as these variations are likely to be caused by misalignments. Finally, we calculated the intersect using BedTools (Quinlan & Hall, 2010), of the genomic bins previously identified for each individual for further analysis (Table 5.2). This resulted in an 11 way alignment with maximum sequence coverage and low false positive variation calling in all our samples.

Phylogenomic analysis

We estimated ML locus trees (bins) using RAxML 7.1.2 (Stamatakis, 2006) with 100 fast bootstrap replicates for each genomic fragment of at least 5 kbp (Table 5.2) to ensure that enough phylogenetic signal was retained in each bin to obtain a reliable tree.

We then built 100 species trees, with one bootstrap replicate from each genomic bin, using STAR (Liu, Yu, & Edwards, 2010). Then, we reconstructed a final

frequency consensus species tree, from our 100 STAR replicates, using consense from the Phylip package (Felsenstein, 1989).
We computed a concordance factor for each observed clade (Table 5.3). Concordance factor correspond to proportion of each possible clade in the database of bootstrapped single loci trees. Overall the concordance factor supports the main topology (Table 5.3; Figure 5.2).

Table 5.2. Summary of fragmented 11-way alignment.

	Total Size	Average Size	%genome
All	1,232,373,599	2,948	~ 45%
Less than 5kb	626,231,249	1,807	~ 23%
Over 5kb less than 10kb	378,416,929	6,864	~ 14%
Over 10kb	227,725,421	13,864	~ 8%

Finally, we randomly selected genomic bins (Table 5.2) of minimum 1 kbp to make up 100 non-overlapping alignments of 1Mbp (between 0.99 Mbp and 1.1 Mbp). In each alignment, we fitted a separate GTR+G+I model to each partition (bin) as implemented in RAxML 7.1.2. Thereafter, we ran 100 fast bootstrap replicates for each alignment and a thorough ML search using RAxML 7.1.2. We then constructed 100 frequency consensus trees using one bootstrap replicate from each jackknife replicate and a final frequency consensus tree using all 100 previous consensus using the consense method as implemented in Phylip. This last frequency value was then used as support for species tree (Figure 5.2).

Demographic analysis

We conducted a demographic analysis using a Hidden Markov Model (HMM) approach as implemented in PSMC (Li & Durbin, 2011). PSMC requires diploid consensus sequences. The consensus was generated from the 'pileup' command of SAMtools software package. We applied the same filtering approach as in S1. Then we used the tool 'fq2psmcfa' from the PSMC package to create the input file for the HMM.
We used Tmax= 20, n = 64 ('4+50*1+4+6'). Plotting the results requires input of generation time and mutation rate. Because there are no convincing data on a different mutation rate in pigs compared to Human we used the default value of 2.5x10-8 mutation per generation that is the mutation rate in Human. For generation time, we used our best guess and assumed a generation time of 5 years. Results are presented in Figure 5.3.

Table 5.3: Concordance factors, that represents the number of time each clade is observed in our database of bootstrapped locus trees. For breed abbreviations, see Table 5.1.

Clade	All	X
WBNL,WBIT	0.251	0.180
HA,LR,LW,WBNL,WBIT	0.251	0.409
JQ,MS	0.200	0.284
WBNC,WBSC	0.153	0.081
JQ,MS,WBNC,WBSC,XI	0.143	0.183
LR,LW	0.140	0.123
HA,WBNL,WBIT	0.123	0.123
HA,LR,WBNL,WBIT	0.119	0.126
HA,WBIT	0.118	0.118
HA,WBNL	0.118	0.147
HA,LR	0.113	0.116
WBNC,XI	0.112	0.191
JQ,WBSC	0.109	0.079
MS,XI	0.108	0.125
JQ,XI	0.108	0.111
LR,WBIT	0.104	0.089
HA,LW,WBNL,WBIT	0.103	0.165
LR,WBNL	0.102	0.104
HA,LW	0.102	0.151
LR,LW,WBNL,WBIT	0.102	0.072
MS,WBSC	0.099	0.060
WBSC,XI	0.098	0.065
LW,WBIT	0.094	0.098
LW,WBNL	0.093	0.095
LR,WBNL,WBIT	0.091	0.067
MS,WBNC	0.091	0.082
JQ,WBNC	0.090	0.122
LW,WBNL,WBIT	0.090	0.066
JQ,MS,XI	0.080	0.151
JQ,MS,WBSC	0.066	0.097
HA,LR,LW	0.064	0.082
JQ,MS,WBNC,XI	0.061	0.293
HA,JQ,LR,LW,MS,WBNL,WBIT,WBSC,XI	0.059	0.082

Admixture analysis – D-statistics:

To detect admixture among our samples we computed D-statistics (Durand, Patterson, Reich, & Slatkin, 2011; Green et al., 2010). Briefly, with sequence data from one chromosome in 4 different populations P1, P2, P3 and O, where P1 and P2 are sister taxa and O is an outgroup, it is possible to infer the state of each allele (derived or ancestral) using the outgroup. Then one can compute the number of derived alleles common between P1 and P3 (ABBA count) and between P2 and P3

(BABA count). Under the null hypothesis of solely incomplete lineage sorting and no gene flow between P3 and either P2 or P1 we expect a similar count of ABBA

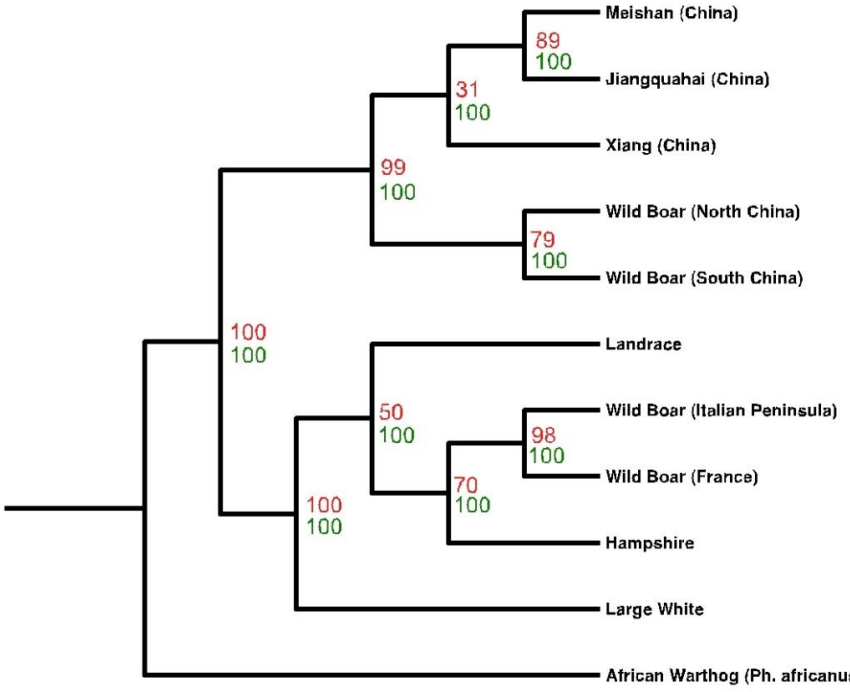


Figure 5.2. Cladogram representing phylogenetic relationship between sequenced pigs. Green values at nodes represent support from STAR analysis, Red values represent support from 100 1 Mbp supermatrices.

and BABA patterns. Under an alternative scenario of gene flow, the count of ABBA must be significantly higher than BABA counts (or vice versa). For a full description of the method please refer to Durand *et al.* (2011). A standard error (SE) of the D-statistics was computed using a Weighted Block Jackknife approach. We divided the genome into N blocks and computed the variance of the statistics over the genome N times leaving each block aside and derived a standard error (SE) using the theory of the Jackknife (For full approach see Green *et al.* Supplementary Online Material 15). We then computed the D-statistics for every possible combination of individuals, using *P. Africanus* as an outgroup. A Bonferroni correction was used to correct for multiple testing by simply multiplying our p-values by the number of D calculations. Because SE may vary greatly depending on block size, we recomputed SE for different block sizes (Table 5.4). Overall these SE estimates were very similar across block sizes. Therefore we used 2 Mb as the block

size for further analyses. Finally, we assessed the effect of transition and transversion mutations on D. Overall these resulted in the same outcome (Table 5.5).

The D statistics are not linearly related to the proportion of admixture (Durand et al. 2011). Computing admixture proportion requires data from a sister taxon to the population that contributed the admixture (an upper bound can be computed with two samples from the same population). In a scenario where we have two sister samples, P1,P2 and P3,P4. If there is an excess of derived lineage from P3 into P2, it is possible to compute the number of common derived alleles between P2 and P4, S(P1,P2,P4). In addition, we can also compute the amount of common derived lineage between P3 and P4, S(P1,P3,P4). The portion of the derived lineage common between P2 and P3 will then behave as if it were a member of P3; hence, S(P1,P2,P4)/S(P1,P3,P4) = f (admixture proportion).

Table 5.4: Examples of SE estimation from jackknife analysis using different bin sizes.

P_1, P_2, P_3	2Mbp, D +- SE	5Mbp, D +- SE	10Mbp, D +- SE
WBIT,LW, MS	0.1993 +-0.0118	0.1993 +-0.0140	0.1993 +-0.0167
WBNC MS, LR	0.0676 +-0.0074	0.0676 +-0.0090	0.0676 +-0.0103
WBNC, WBSC, WBIT	- 0.0996 +-0.0046	- 0.0996 +-0.0054	- 0.0996 +- 0.0062
WBNC, WBSC, WBNL	- 0.0979 +-0.0045	- 0.0979 +-0.0051	- 0.0979 +-0.0058

5.3 Results and Discussion

Divergence between Asian and European wild boar

We investigated the evolution within Sus scrofa in Eurasia by sequencing 10 individual unrelated wild boars from different geographical areas. In total 17,210,760 single nucleotide polymorphisms (SNPs) were identified amongst these 10 wild boars. The number of SNPs segregating in the 4 Asian wild boars (11,472,192) was much higher than that observed in the 6 European wild boars (6,407,224) with only 2,212,288 shared SNPs. This higher nucleotide diversity was visible in the distribution of heterozygous sites of the Asian compared to the European wild boar genomes. Phylogenomic analyses of complete genome sequences from these wild boars and six domestic pigs revealed distinct Asian and European lineages (Figure 5.2) that split during the mid-Pleistocene 1.6-0.8 MYA (Frantz et. al. 2013; Calabrian stage). Colder climates during the Calabrian glacial intervals likely triggered isolation of populations across Eurasia. Admixture analyses within Eurasian Sus scrofa disclosed gene flow between the Northern Chinese and

European populations consistent with pig migration across Eurasia, between Europe and Northern China throughout the Pleistocene. Our demographic analysis on the whole genome sequences of European and Asian wild boars (Figure 5.3), revealed an increase in the European population after pigs arrived from China. During the Last Glacial Maximum (LGM; ~20KYA), however, Asian and European populations both suffered through bottlenecks. The drop in population size was more pronounced in Europe than Asia (Figure 5.3) suggesting a greater impact of the LGM in Northern European regions and likely resulting in the observed lower genetic diversity in modern European wild boar.

Table 5.5: Examples of the influence of Tv/Ti on D calculation – using 2Mb bins.

P_1, P_2, P_3	Ti, D+-SE	n. ABBA / BABA	Tv, D+-SE	n. ABBA / BABA
WBIT,LW, MS	0.2032+-0.0125	130136 86176	0.1979+-0.0119	349653 234094
WBNC,MS, LR	0.0684+-0.0078	154610 134797	0.0659+-0.0075	416916 365357
WBNC, WBSC, WBIT	-0.1015+-0.0053	119693 146759	-0.0990 +-0.0046	330635 403304
WBNC, WBSC, WBNL	-0.0992+-0.0052	119829 146237	-0.0975+-0.0045	330704 402186

North Eurasian biogeographic zone.

We found a clear signal for admixture between North Chinese and European populations of wild boars that we interpret as migrations across Eurasia during the later stage of the Pleistocene (Table 5.6). Moreover, this hypothesis is further supported by the high value of concordance factor on the X chromosomes (Table 5.3). The demographic analysis shows that the last glacial maximum (LGM)-induced bottleneck had similar magnitude in Europe and North China (Figure 5.3). Together, these evidences suggest the existence of another (besides Asian + European) biogeographic zone for pigs, extending across North Eurasia.

Table 5.6: Results from D-statistics analysis. First column displays trios involve in D computation. P3 is the population from which we query derived allele into P1 and P2. Second column represents derived sites considered. The third column displays D value±standard error; and significance level (** p <0.001; * p <0.05; NS non-significant). A positive D value mean admixture in P2, while negative values mean admixture in P1.

P1 P2 P3	ABBA BABA	D±SE
WBNC WBSC WBNL	548423 450533	-0.0980±0.0045 **
WBNC WBSC WBIT	550063 450328	-0.0997±0.0047 **
HA WBNL MS	412264 294590	-0.1665±0.0103 **
HA WBIT MS	417559 297334	-0.1682±0.0105 **
LR WBNL MS	470045 313587	-0.1997±0.0102 **
LR WBIT MS	473987 315140	-0.2013±0.0104 **
WBNL LW MS	315903 472494	0.1986±0.0115 **
WBIT LW MS	320270 479789	0.1994±0.0119 **
HA WBNL JQ	407239 302651	-0.1473±0.0116 **
HA WBIT JQ	410441 305295	-0.1469±0.0115 **
LR WBNL JQ	470015 316332	-0.1954±0.0110 **
LR WBIT JQ	473030 318655	-0.1950±0.0111 **
WBNL LW JQ	316752 478129	0.2030±0.0116 **
WBIT LW JQ	322237 484284	0.2009±0.0116 **
HA WBNL XI	411976 275888	-0.1978±0.0108 **
HA WBIT XI	414583 278569	-0.1962±0.0110 **
LR WBNL XI	471840 291850	-0.2357±0.0102 **
LR WBIT XI	473972 294110	-0.2342±0.0100 **
WBNL LW XI	296517 468500	0.2248±0.0098 **
WBIT LW XI	301643 473609	0.2218±0.0103 **
HA WBNL MS	412264 294590	-0.1665±0.0103 **
HA WBIT MS	417559 297334	-0.1682±0.0105 **
LR WBNL MS	470045 313587	-0.1997±0.0102 **
LR WBIT MS	473987 315140	-0.2013±0.0104 **
WBNL LW MS	315903 472494	0.1986±0.0115 **
WBIT LW MS	320270 479789	0.1994±0.0119 **
WBSC XI MS	625181 708687	0.0626±0.0052 **
WBNC XI MS	647153 671457	0.0184±0.0053 NS
WBSC XI JQ	625402 703092	0.0585±0.0054 **

Table 5.6: continued

WBNC XI JQ	656865 659913	0.0023±0.0053 NS
MS JQ WBSC	562037 547462	-0.0131±0.0063 NS
MS JQ WBNC	562326 562472	0.0001±0.0069 NS
MS JQ XI	558052 537347	-0.0189±0.0070 NS
JQ XI WBNC	656865 594843	-0.0495±0.0058 **
MS XI WBNC	647153 585888	-0.0497±0.0047 **
JQ XI WBSC	625402 624882	-0.0004±0.0053 NS
MS XI WBSC	625181 610281	-0.0121±0.0046 NS
MS WBSC XI	708755 610282	-0.0747±0.0051 **
WBSC JQ XI	625026 703159	0.0588±0.0059 **
WBNC JQ XI	594947 659974	0.0518±0.0060 **
WBNC MS XI	586000 671512	0.0680±0.0054 **
HA LR WBIT	528296 458122	-0.0711±0.0211 NS
HA LW WBIT	554700 449809	-0.1044±0.0209 **
HA LR WBNL	529654 455864	-0.0749±0.0210 NS
HA LW WBNL	553525 452552	-0.1004±0.0212 **
HA LR JQ	393507 442038	0.0581±0.0136 **
HA LW JQ	401830 458169	0.0655±0.0142 **
HA LR MS	398784 436610	0.0453±0.0132 NS
HA LW MS	409494 448504	0.0455±0.0144 NS

Breed trading.

There was a strong signal for admixture from Asian into European breeds. We found that European domestic breeds such as Landrace and Large White have a significant amount of Asian genetic material (Table 5.6). This admixture is likely to be due to importation of Chinese breeds into Europe (especially UK) at the onset of the 'agricultural' revolution in the late 18th and 19th century.

Within Asia.

The difficulty of building a phylogenetic tree for the breeds within Europe and China is puzzling. Many aspects such as incomplete lineage sorting (ILS), breed trading, multiple domestication origin, husbandry practices and biogeographic pattern could explain these results. In Asia, the clustering of breeds illustrated in Figure 5.2, appears to be much more complex. The Meishan and Jiangquhai pigs do not share significantly more derived alleles with Chinese wild boars or Xiang (Table

5.6), which suggests that these two breeds have a common wild ancestor and did not undergo admixture since their separation. This is not surprising as these breeds are from very similar geographic areas. Thus, bootstrap values and concordance factors can be explained solely by ILS. However, this is not the case for the Xiang breed. We found that North Chinese wild boar derived alleles match Meishan or Jiangquhai significantly more often than Xiang (Table 5.6). This is expected as Meishan and Jiangquhai are from Northern regions of China. Surprisingly, Xiang do not share significantly more derived allele with Southern Chinese wild boar than with Northern wild boar, MS and JQ, which is in agreement with the South Chinese origin of Xiang. In addition, the Xiang's derived alleles are found significantly more often in Jiangquhai and Meishan than in both Northern and Southern Chinese wild boar (Table 5.6). Lastly, we found Jiangquhai derived alleles 6% more often in Southern Chinese wild boar than Xiang. This pattern highlights the composite origin of the Xiang breed. Such a finding can be explained by complex breed trading and admixture with local wild boar populations within China and / or multiple origins of domesticated pigs in China. Thus, our analyses do not allow us to distinguish between these hypotheses. Further studies that capture more genetic diversity within Asia may be able to provide an answer to this question.

Within Europe

In Europe, the clustering of breeds and wild boar seems even more complex. Breeds do not form a monophyletic group as one would expect if they shared a common wild ancestor. For example, the derived lineages from Dutch wild boar match the Hampshire lineage 10% more often than the Large White lineage (Table 5.6), thus, supporting our phylogeny (Figure 5.2). As we showed above, these breeds have different degrees of Asian genetic material, imported during the agricultural revolution. Thus, this may solely explain the paraphyly of European breeds. Under this null hypothesis we expect that alleles coming from Asian admixture will influence the topology and D value. Let us suppose that the Large White has more Asian derived alleles than Hampshire, when querying the alleles of the Dutch wild boar in these breeds as D(HA,LW,WBNL), we expect that the excess of Asian alleles in Large White compared to Hampshire influences our calculation, thus making Hampshire closer to the Dutch wild boar. To test this hypothesis we re-computed the Dstat pulling out every derived allele common between (LW, MS+JQ) and (HA, MS+JQ), thus, minimizing the Asian influence in our European calculations. We found that Asian alleles had a very minor influence on our calculation. For D(HA, LW, WBNL) we considered 1,006,195 derived sites. When removing sites where either Large White or Hampshire matched Meishan and

Jiangquhai rather than the Dutch wild boar, this number fell to 1,006,172. Moreover, removing those sites did not influence our estimated of D (D(HA, LW, WBNL) = 0.1003). In addition, we found that both D(LW,HA,JQ) and D(LW,HA,MS) show an excess of match between Large White and Jiangquhai or Meishan, however this was not significant using Meishan (Table 5.6). Thus we hypothesise that Asian admixture is not solely responsible for the paraphyly of European breeds. Other factors such as husbandry practices and / or multiple domestication origin in Europe probably played an important role.

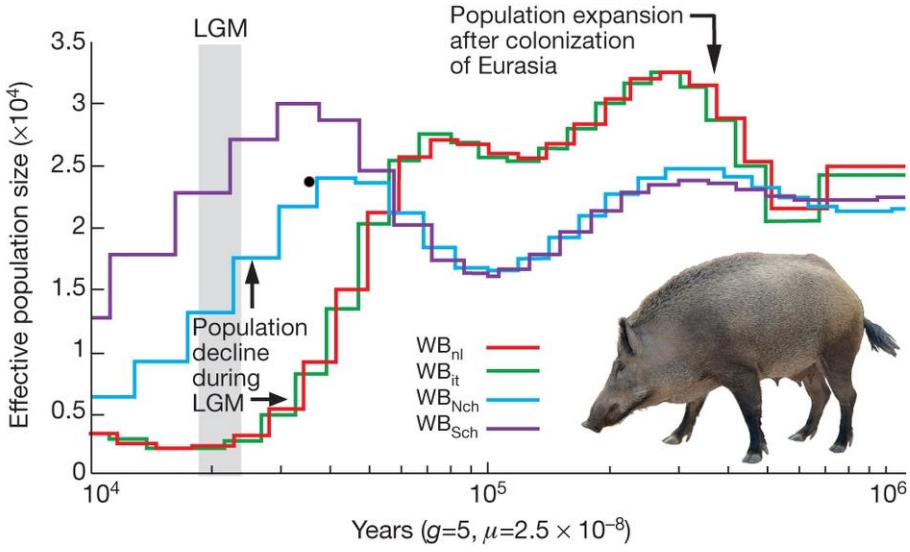


Figure 5.3: Demographic history of wild boars Demographic history was inferred using a Hidden Markov Model (HMM) approach as implemented in PSMC[45]. In the absence of known mutation rates for pig, we used the default mutation rate for human of 2.5×10^{-8}. For the generation time we used, an estimate of 5 years. The last glacial maximum (LGM) is highlighted in grey. WB_{nl}=Wild boar Netherlands; WB_{it}=Wild boar Italy; WB_{Nch}=Wild boar North China; WB_{Sch}=Wild South China. Adapted from Groenen *et al.* (2012).

Conclusions

A phylogenetic tree constructed using four European wild boar and domestic pigs and six East Asian wild boar and domestic pigs revealed a clear distinction between European and Asian breeds, thus substantiating the hypothesis that pigs were independently domesticated in western Eurasia and East Asia. An admixture analysis revealed European influence in Asian breeds, and a ~35% Asian fraction in European breeds. These results are consistent with the known exchange of genetic

material between European and Asia pig breeds. We also observed that European breeds form a paraphyletic clade, which cannot be solely explained by varying degrees of Asian admixture. Within each continent, our analysis revealed different degrees of relatedness between breeds and their respective wild relatives.

During domestication, pigs were often allowed to roam in a semi-managed state and recurrent admixture between wild and domesticated individuals was not uncommon, especially in Europe. Thus, the most likely explanation for the paraphyletic pattern seen in domestic individuals is a long history of genetic exchange between wild and domestic pigs.

References

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, *28*(8), 2239-52. doi:10.1093/molbev/msr048

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*(2), 163-166. doi:10.1111/j.1096-0031.1989.tb00562.x

Frantz, L. A., Schraiber, J. G., Madsen, O., Megens, H.-J., Bosse, M., Paudel, Y., Semiadi, G., et al. (2013). Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome biology*, *14*(9), R107. doi:10.1186/gb-2013-14-9-r107

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., et al. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710-22. doi:10.1126/science.1188021

Groenen AM, Alan L Archibald, Hirohide Uenishi, Christopher K Tuggle, Yasuhiro Takeuchi, Max F Rothschild, Claire Rogel-Gaillard, Chankyu Park, Denis Milan, Hendrik-Jan Megens, Shengting Li, Denis M Larkin, Heebal Kim, Laurent AF Frantz, Mario Caccamo (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491(7424), 393-398.

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., Lowden, S., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science*, *307*(5715), 1618-21. doi:10.1126/science.1106927

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. *Data Processing*, 1-2.

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, *10*(1), 302. doi:10.1186/1471-2148-10-302

Megens, H.-J., Crooijmans, R. P. M. A., San Cristobal, M., Hui, X., Li, N., & Groenen, M. A. M. (n.d.). Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genetics, selection, evolution : GSE*, *40*(1), 103-28. doi:10.1051/gse:2007039

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841-2. doi:10.1093/bioinformatics/btq033

Ramos, A. M., Crooijmans, R. P. M. A., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., Bendixen, C., et al. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. (L. Orban, Ed.)*PloS one*, *4*(8), e6524. Public Library of Science. doi:10.1371/journal.pone.0006524

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, *22*(21), 2688-90. doi:10.1093/bioinformatics/btl446

6

Analyses of Eurasian wild and domestic pig genomes reveal long-term gene-flow and selection during domestication

Laurent A.F. Frantz[1], Joshua G. Schraiber[2,3], Ole Madsen[1], Hendrik-Jan Megens[1], Alex Cagan[4], Mirte Bosse[1], Yogesh Paudel[1], Richard PMA Crooijmans[1], Greger Larson[5] and Martien AM Groenen[1]

1 Animal Breeding and Genomics Centre, Wageningen University, Droevendaalsesteeg 1, Wageningen, 6708 PB, The Netherlands. 2 Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA. 3 Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, USA. 4 Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. 5 Durham Evolution and Ancient DNA, Department of Archaeology, Durham University, Durham DH1 3LE, UK.

Submitted

Abstract

The process of domestication led to one of the most important transitions in human evolution. Traditionally, this process is assumed to be strongly human-directed, with few individuals initially selected to be domesticated and reproductive isolation between wild and domestic forms. However, zooarchaeological evidence depicts animal domestication as a geographically restricted, long-term process without reproductive isolation or strong intentional selection. Here, we ask whether pig domestication follows a traditional, linear model or a complex, reticulate model as predicted by zooarchaeologists. To do so, we fit models of domestication to whole genome data from over 100 wild and domestic pigs. We found that the assumptions of traditional models, such as reproductive isolation and strong domestication bottlenecks, are incompatible with the genetic data and provide support for the zooarchaeological theory of a complex domestication process in pigs. In particular, gene-flow from wild to domestic pigs was a ubiquitous feature of domestication and post-domestication processes in pigs. In addition, we show that despite gene-flow, the genomes of domestic pigs show strong signatures of selection at loci that affect behaviour and morphology. Specifically, our results are consistent with independent parallel sweeps in two independent domestication areas (China and Anatolia) at loci linked to morphological traits. We argue that recurrent selection for domestic traits likely counteracted the homogenising effect of gene-flow from wild boars and created "islands of domestication" in the genome. Overall, our results suggest that genomic approaches that allow for more complex models of domestication to be embraced should be employed, and that results from these studies will have significant ramifications for studies that attempt to infer the chronology and geographic origin of domesticated animals.

Keywords: domestication, approximate bayesian computation (ABC), population genetics, zooarcheology, artificial selection.

6.1 Introduction

The rise of agriculture, which occurred approximately 10,000 years ago, was one of the most important transitions in human history. During the Neolithic revolution, the domestication of plant and animal species led to a major subsistence shift, from hunter-gatherers to sedentary agriculturalists that ultimately resulted in the development of complex societies. The process of animal domestication led to striking morphological and behavioural changes in domesticated organisms compared with their wild progenitors (1). Traditionally, this process has often been viewed as human-directed, involving strong bottlenecks in the domestic population (*i.e.* founder events due to the selection of only a few individuals at the beginning of domestication) and reproductive isolation between wild and domestic forms (2–6). This straightforward model provides an attractive theoretical framework for geneticists, because key events such as the geographic origin and timeframe of domestication are well defined. Thus, the assumption of reproductive isolation eases the interpretation of genetic data from domestic and wild forms (7). For instance, under this model, geneticists have interpreted phylogenetic affinities of domestic animals with multiple, geographically divergent wild populations as evidence of independent domestication origins in multiple species (8–13).

However, this view conflicts with zooarchaeological evidence that suggests that domestication events are rare, and that domesticated forms diffused out from a limited number of core regions (7, 14, 15). Moreover, there is a growing body of empirical and theoretical archaeological work that challenges the simplicity of traditional models (3, 4, 16). In these new, more complex models, pre-historic domestication of animals is viewed as mainly unintentional (3, 4, 7) and neither reproductive isolation nor strong intentional selection are thought to be as crucial and widespread as previously thought. Instead, domestication is seen as a long-term, diffuse process (17), involving gene-flow (during as well as post-domestication) between wild and domestic forms (18) and with emphases on multiple, taxon specific, human-animal relationships (3, 4). The possibility of post-domestication gene-flow between domestic animals and their wild progenitors, as well as a lack of strong domestication bottlenecks, are key predictions from this novel framework that contrast with more traditional models of domestication (18). Moreover, extensive gene-flow between wild and domestic forms violates the assumptions of traditional models of domestication and has significant ramifications for studies that attempt to infer the spatial and chronological origin of domestication using genetic data.

Here, we focus on pig domestication using genome-wide datasets of modern domestic pigs and wild boars. Pigs were most likely domesticated independently

once in Anatolia (16) and once in the Mekong valley around 9,000 BP (19). Furthermore, ancient mtDNA analyses found that the first domestic pigs in Europe were transported by early farmers from the Levant into Europe around 5,500 BC, concordant with zooarchaeological evidence for a single domestication origin of Western Eurasian domestic pigs (20, 21). However, a few thousand years after their introduction, domestic pigs in Europe had completely lost the original mtDNA signatures and instead acquired mtDNA haplotypes typically found in local European wild boars (20, 21). These findings suggest that early domestic populations experienced post-domestication gene-flow from wild boar populations that were not involved in the Anatolian domestication process.

Further mtDNA analyses of ancient Anatolian material demonstrated that, by 500 BC, local mtDNA haplotypes were also replaced by haplotypes from European wild boars. This result suggests extensive mobile swineherding throughout Europe and Anatolia (21), consistent with both archaeological and historical evidence, as well as limited management and selection up until the industrial revolution in the 19th century (22, 23). Thus, under a complex model of domestication, mtDNA replacement in ancient European and Anatolian pigs is the result of post-domestication gene-flow, loose pig management and mobile swine herding. We therefore expect such phenomenon to have left a strong signal of gene-flow from wild boars in the genome of modern domestic pigs. However, while unsupported by any zooarchaeological evidence, the observed mtDNA turnovers could also be interpreted as a *de-novo* domestication of a population of European wild boars rather than the result of post-domestication gene-flow from wild boars. Moreover, because of its mode of inheritance and limited resolution, small mtDNA markers provide a very limited impression of gene-flow, making it impossible to test these hypotheses. Thus, the hypothesis of complex domestication in pigs has yet to be tested with the resolution and confidence afforded by unlinked, nuclear markers. In addition, unlike horses and donkeys, intentional interbreeding between pigs and wild boars confers no clear productive advantage and is thought as being mainly unintentional (18). Lastly, there is a clear morphological and behavioural dichotomy between wild and domestic pigs that is evident in modern animals as well as in the zooarchaeologic record (24–27). Thus, the possibility of unintentional gene-flow between wild and domestic pigs also raises questions regarding the mechanisms behind the maintenance of traits that differentiate domestic and wild forms.

Here, we fit models of domestication to a genome-wide dataset from over 100 wild and domestic pigs. Our main aim is to ask whether pig domestication follows a traditional, linear model or a complex, reticulate model. More precisely, we want to assess if the zooarchaeological evidences for a single, geographically restricted,

domestication of (Western) pigs in Anatolia (7, 15, 20) are compatible with the assumption of a traditional model of domestication involving reproductive isolation and strong bottlenecks.

6.2 Results and Discussion

We evaluated the support of multiple models for the domestication of pigs in Europe and Asia. Our analysis focused on 103 genomes from European wild boars (EUW) (8) and European commercial / historical domestic pigs (EUD) (Table S6.1). In addition, this data set comprises multiple populations of Asian wild boar (ASW) and Asian domestic pigs (ASD; Table S6.1). In order to better understand the early process of domestication, we sampled a range of wild boar populations, from Asia and all major European Pleistocene refugia, rare historical European and Asian breeds, as well as modern commercial breeds. To test key predictions of the complex domestication framework described above, we fit simple but informative models to these genomic data sets using Approximate Bayesian Computation (ABC) (see Materials and Methods).

Testing models of domestication from genome sequences

We first tested the hypothesis of gene-flow between wild and domestic pigs. More precisely, we asked whether reproductive isolation between wild and domestic pigs is compatible with zooarchaeologic evidence that pig were domesticated only twice, independently in Anatolia and China. To do so, we first evaluated the fit of the traditional model in which domestication is modelled as two parallel events in Asia and Europe. In this model, domestication takes place at time T1 in Europe and T2 in Asia and involves no gene-flow between wild and domestics (reproductive isolation) or between domestics from Asia and Europe (Figure 6.1a). We then compared this null model to 5 other models involving different patterns of continuous mixture: within wild, within domestic, between wild and domestic, etc. (Figure S6.1). By comparing these six models (Figure S6.1), we found that a model involving gene-flow between domestic and wild (within Asia and Europe) as well as between domestic and domestic (between Europe and Asia) provided a large improvement of fit (Bayes Factor [BF] > 14) when compared to any other model tested in this study (Figure 6.1a; Figure S6.1). Thus, our explicit modelling framework provides very strong evidence that reproductive isolation between wild and domestics was not maintained during and after domestication in Asia and Europe.

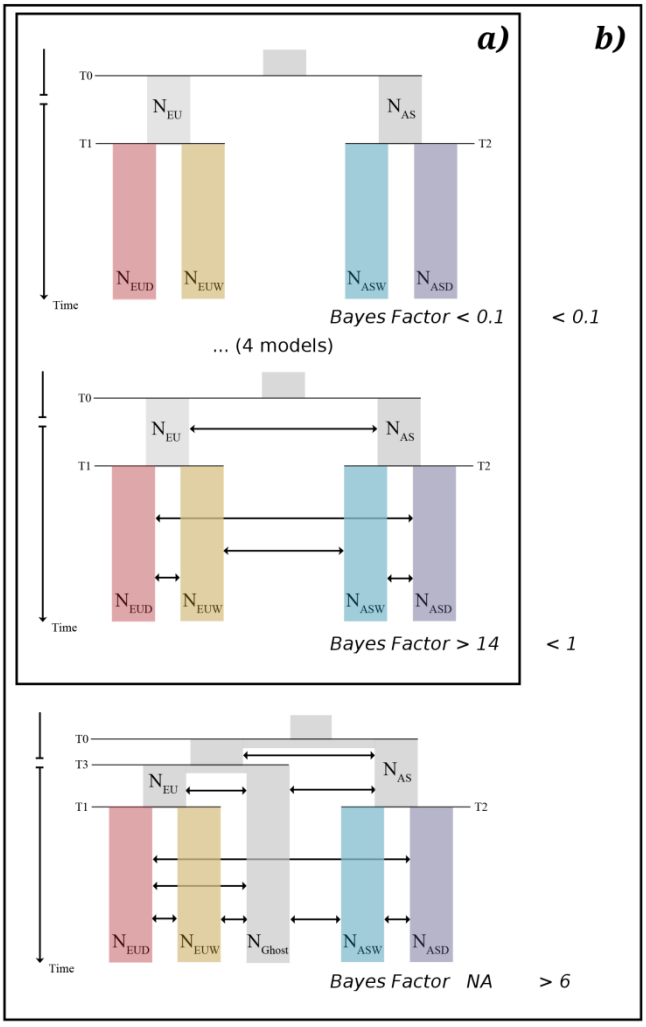


Figure 6.1: Schematic representation of models tested in this study. All migration parameters (depicted as a single double headed arrow) are in fact modelled as two independent continuous migration parameters. **a)** Model testing approach comparing six models. Two models, one without gene-flow (null model; top) and one model with gene-flow between wild (ASW and EUW) and domestic (ASD and EUD) as well as between domestic and wild (full; middle) are displayed. The four additional model tested in this study are displayed in Supplementary Figure S1. Bayes Factors in this square were computed without the Ghost model (6 models in total; see b)). **b)** Same as a) but with the Ghost model (bottom). Bayes Factors were computed with all 7 models together.

We further assessed this finding using a data-set of over 600 pigs (from the same populations as in the genome-wide data) that were genotyped on the Porcine SNP60 array (Supplementary Information). We investigated the historical relationship among these populations using *TreeMix* (28). Our analysis showed that EUD and ASD were both paraphyletic while EUW was monophyletic (Supplementary Information; Figure S6.2). The paraphyly of EUD and ASD is difficult to reconcile with the assumptions of a linear spatially restricted model of domestication. Instead, this finding provides further evidence of a complex domestication process that involved gene-flow between wild and domestic pigs. Moreover, we found that gene-flow between wild and domestic in Europe was strongly asymmetrical, with EUW sending more migrants than EUD (Supplementary Information; Figure S6.3). Lastly, we saw that Asian and European domestic pigs exchanged genetic material. This is consistent with previous studies and is most likely the result of European importations of Chinese pigs during the 19th century to improve European commercial breeds (23, 29, 30). However, the migration parameters between wild and domestics (in Europe and Asia) were much higher than between ASD and EUD (Figure S6.3). This demonstrates that this intercontinental admixture had no influence on our conclusion of gene-flow between wild and domestics (see Supplementary Information).

Together, these findings demonstrate that domestic pigs do not form a tight, homogeneous genetic group, as expected under a simple human-driven model of domestication. Instead, domestic pigs are a genetic mosaic of different wild boar populations. Thus, the assumption of reproductive isolation between wild and domestic pigs is incompatible with the zooarchaeological evidence of a single domestication of pigs in Asia and Anatolia. Rather, our results demonstrate that modern genetic data from domestic pigs can only be reconciled with zooarchaeological evidence if modelled with continuous gene-flow between wild and domestic pigs.

Demography of pig domestication

We also tested whether the genomic sequences revealed an absence of a strong bottleneck associated with domestication. To do so, we estimated the posterior distribution of demographic parameters using 10,000 retained simulations out of 2,000,000. Under the assumption of a simple linear model of domestication with no gene-flow and strong intentional selection by humans, we would expect a strong bottleneck in domestic populations. Overall, we found a population decline in EUW and EUD (Figure 6.2). This is consistent with previous results demonstrating that Pleistocene glaciation resulted in long-term population decline in European

wild boars (23, 31–33). However, this population decline was more pronounced in EUW than in EUD (Fig 2). In addition, we found that the effective population size of EUD (Ne-EUD ~=20,563) was more than twice as large as the effective population size of EUW (Ne-EUW =~8,497). This is most likely due to a series of strong bottlenecks in the European wild population, caused by over-hunting and loss of suitable habitat (23, 31-33). Together these results do not support the existence of a strong domestication bottleneck in European domestic pigs and instead support the contention that continuous gene-flow from multiple genetically and geographically distinct wild boar populations likely increased the effective population size of EUD.

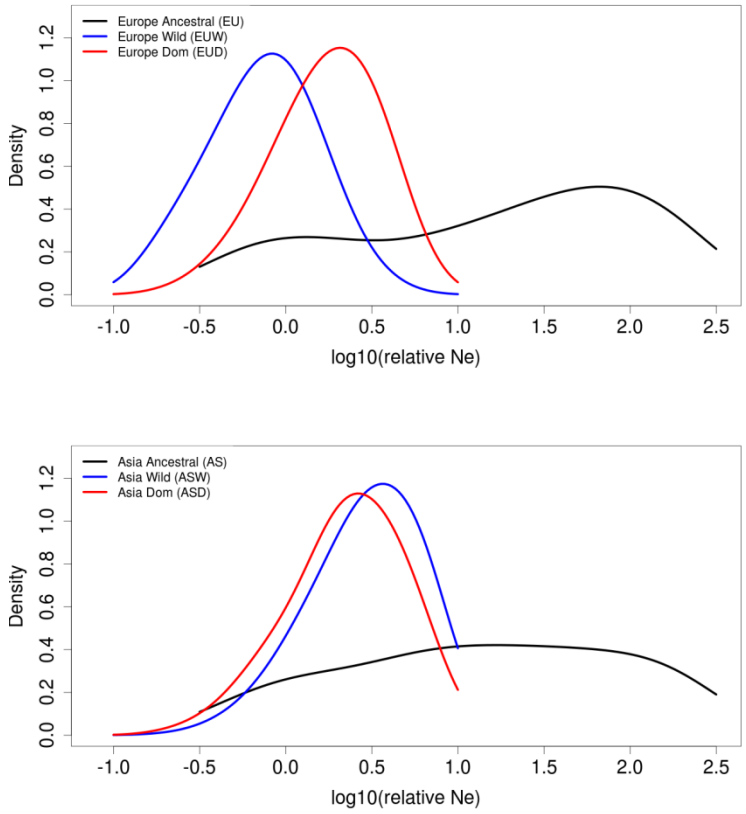


Figure 6.2: Posterior density distribution of demographic parameters. Population size are of the relative population size (the ratio of the current population size over the population size before T0 [Figure 6.1]).

Gene-flow from wild boar populations not involved in the domestication process

We showed that a model incorporating continuous gene-flow between wild and domestic pigs is significantly more compatible zooarchaeological evidence compared to a traditional hypothesis of reproductive isolation. Despite this fact, we only modelled gene-flow from a population of wild boar that we assumed to be derived from the source of domestication. Here, we test the hypothesis that another population of wild boars either extinct (due to hunting pressure and habitat loss (23)) or un-sampled in our analysis (*i.e.* our sampling does not cover Central Eurasia) also contributed to the gene-pool of domestic pigs. To do so, we used a model that is similar to our best fitting model (see above; Figure 6.1a) but with an additional ghost population that splits from EUW/EUD during the Pleistocene (Figure 6.1b) and act as a step between ASW and EUW (migration ASW → Ghost → EUW; Figure 6.1b). This model provided a substantial improvement of fit (BF>6). This result shows that mobile herding of domestic pigs across Europe most likely resulted in gene-flow from a least one wild boar population that was genetically divergent from the population involved in the domestication process in Anatolia.

Positive selection in domestic pigs

Our analysis shows that gene-flow between wild and domestic forms was a ubiquitous feature of domestication and post-domestication processes in pigs. Thus, extensive gene-flow from wild boars into domestic pigs during and after domestication raises questions regarding the mechanisms behind the maintenance of the clear morphological and behavioural differences observed between domestic and wild pigs. Intentional or unintentional selection by humans could have counteracted the effect of gene-flow and resulted in morphological and behavioural differentiation between wild and domestics. In order to assess the importance of selection in the genome of domestic pigs in face of gene-flow we conducted a scan for positive selection using SweeD (34, 35). SweeD computes the composite likelihood ratio (CLR) of a sweep model over a neutral model. Such a test can be very sensitive to demography and migration (36). To correct for effects of demography and migration we used the 10,000 closest simulations (out of 2,000,000) under our best fitting model (see above) to generate an expected cumulative distribution function (CDF) of neutral CLR and to compute the p-value for all empirical CLR value in the genome (Materials and Methods; Supplementary Information). We identified 249 and 136 10kbp regions with $p<0.01$ in the genome of European and Asian domestics, respectively.

First, we examined sweeps private to each population (Supplementary Information). These sweeps in domestic pigs (EUD and ASD) were significantly enriched with GO terms related to multiple developmental processes of bones, teeth and nervous system (Table S6.2&S3). These terms comprise multiple gene candidates related to height (*PLAG1*, *NCPAG*, *PENK*, *RPS20* and *LYN* in EUD; Figure 6.3a; *LEMD3* and *UPK1* in ASD) in pigs (37, 38) and cattle (39, 40), nervous system development and maintenance (*NRTN, SEMA3C, PLXNC1, AAK1, RAB35, FRS2*) (41–52) as well as genes directly influencing behaviour (*i.e.* aggressiveness and feeding; *APBA2*, *MC4R*, *RCAN1*, *BAIPA3*) (53–60). These results suggest that domestication and/or post-domestication selection for behavioural and morphological traits was important in Asian and European domestic pigs and most likely counteracted the effect of continuous gene-flow in certain parts of the genome.

However, the mechanism behind this maintenance remains unknown. One possibility is that there was recurrent selection for similar traits. This phenomenon may have resulted in parallel sweeps at the same loci. To investigate this possibility, we looked for signals of parallel adaptation between the two independent domestication events (ASD and EUD). To do so we identified genes with CLR above the significance threshold in both ASD and EUD but not in ASW and EUW. In order to rule out admixture between ASD and EUD as the cause for observing overlapping significant signal we conducted a phylogenetic analysis in each region separately (Supplementary Information). The genealogy of some of these regions shows a signal that is consistent with introgression between EUD and ASD (*e.g.* Figure S6.7). However, we found one region of particular interest seems to have swept independently in EUD and ASD (Figure 6.3). Phylogenetic analysis in this region (Figure 6.3b) reveals that ASD and ASW as well as EUD and EUW are monophyletic (Figure 6.3c), suggesting an independent sweep in ASD and EUD. Interestingly, while this sweep does not overlap with genes, the region is just a few kbp upstream of the highest CLR in EUD (among others; Figure 6.3a). This region has been shown to have a strong effect on body height and stature in pigs (37, 38) and cattle (39, 40). In particular, variation in this region explains up to 18% in body length difference between wild boars and commercial EUD (37). Given the importance of this region for morphology in commercial EUD (38) it is possible that human-mediated selection for similar traits in Asian and European domestic pigs resulted in parallel sweeps at the same loci. Parallel selection of this form may be the responsible for some of the morphological convergence in the two independent domestication events in Europe and Asia. Thus, while the phenotypic effect of this sweep is still unclear, this region provides a particularly interesting candidate to further study the possibility of convergence between ASD and EUD.

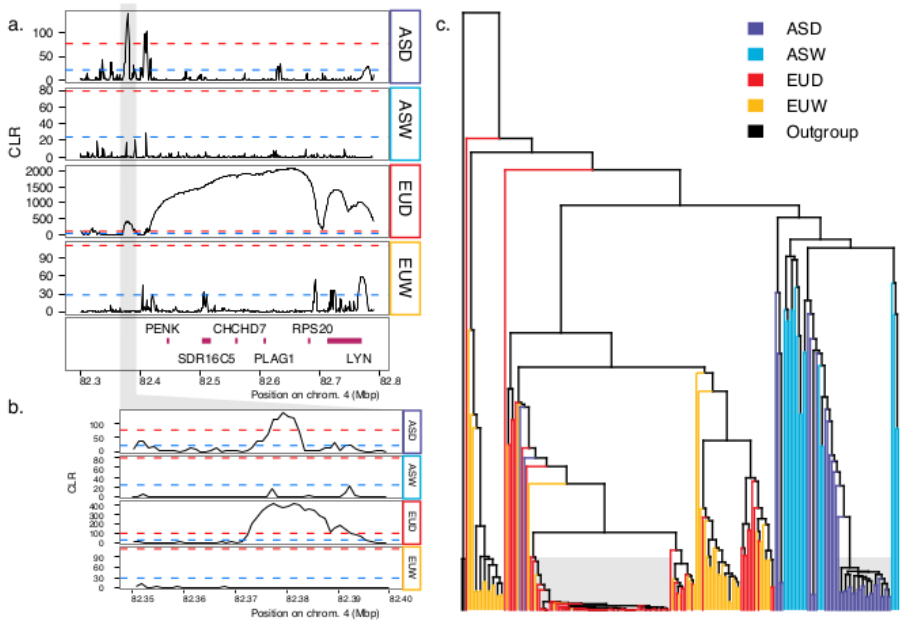


Figure 6.3: Example of a parallel sweep in ASD and EUD. a) Composite likelihood ratio (CLR) values in the PLAG1 region. Dashed blue and red lines represent p=0.05 and 0.01 thresholds respectively. Grey shaded area is the position of the parallel sweep (see b). **b)** CLR values in the parallel sweep region few kbp upstream of the PLAG1 region. **c)** Genealogy of phased haplotypes for the region in Figure 6.3b. Shaded area highlights the very short branch lengths that are the result of a sweep. The shaded area on the left side (Europe) contains 64 haplotypes from EUD (>72% of total EUD haplotypes) and 2 haplotypes from EUW (<4% of total EUW haplotypes). The shaded are on the right side (Asia) contains 24 haplotypes from ASD (>54% of total ASD haplotypes) and no ASW haplotype.

Conclusions

The generation of larger amounts of genomic data with ever-greater resolution is allowing us to embrace the complexity of domestication. The commensurate advancements in theoretical and empirical perspectives is allowing for more sophisticated models to be tested and for a greater understanding of animal domestication. In this study we demonstrated that the assumptions of traditional models, such as reproductive isolation and strong domestication bottlenecks, are incompatible with the zooarchaeological evidence of a geographically restricted domestication process in pigs. Instead our model testing approach revealed that continuous gene-flow from wild boars to domestic pigs is necessary to reconcile modern genetic data with zooarchaeological evidence. Moreover, we saw that in

Western Eurasia, gene-flow most likely involved at least a second, possibly extinct population of wild boars that was not the source of domestication, likely the result of mobile domestic swine herding, as predicted by zooarchaeologists (18, 21). Thus, our results support a model in which the replacement of Anatolian mtDNA haplotypes, in domestic pigs from the late European Neolithic, by local wild boars haplotypes (20) was most likely the result of post-domestication gene-flow rather than independent domestication of pigs in Europe as it was previously suggested (7, 15, 16).

Such extensive gene-flow from wild boars raises questions regarding the maintenance of morphological and behavioural traits in domestic pigs. Our study reveals extensive evidence of selection at candidate genes that influence anatomical and nervous system development, suggesting that selection may have counteracted the homogenizing effect of gene-flow and maintained the genetic basis for the morphological and behavioural dichotomy observed between wild and domestic pigs. In addition, our results show that regions close to genes governing morphological traits have been subject to selection in parallel in Asia and Europe. Such parallel selection may have resulted in "islands of domestication" (*sensu* "islands of speciation" (61)), that we define as regions in the genome containing variations that affects domestic traits and are thus less affected by gene-flow from wild boars. However, it is unclear whether these sweeps involved recurrent selection of different haplotypes from standing genetic variation in wild boars or are the result of selection from *de-novo* mutations. Thus, our results highlight a list of candidate genes that will provide further studies with the means to further test these hypotheses.

Lastly, it is important to underline the limitations of modern DNA and traditional domestication models to determine the origin and time of domestication of animals, as well as to identify the genes involved in during domestication. Indeed, extensive gene-flow clearly violates the assumption of traditional models and likely eroded most of the signal to infer time and geographic parameters (62, 63). Moreover, signal selection during early domestication may be confounded with signal from strong post-domestication selection as was shown in chicken (64). It is therefore important to apply caution when conducting comparative analyses of modern genetic material from wild and domestic animals. However, future sequencing of ancient DNA, together with more realistic modelling frameworks, such as the one presented here, will provide the necessary information not only to determine the origin and time of domestication of animals but also to identify genes involved during domestication and will ultimately significantly enhance our knowledge of this fascinating and important process.

6.3 Materials and Methods

Sample collection and DNA preparation

Blood samples were collected from a total of 622 individuals, 403 European domestics, 92 Asian domestics, 103 European wild boars and 23 Asian wild boars and a Javan Warty pig (*S. verrucosus*), used as an outgroup (33). For full description of the samples see Table S6.1. DNA was extracted from the blood samples with QIAamp DNA blood spin kits (Qiagen Sciences). Quality and quantity of DNA extraction was checked on a Qubit 2.0 fluorometer (Invitrogen). Single nucleotide polymorphism (SNP) genotyping was performed with the Illumina Porcine 60K iSelect Beadchip. For the genome re-sequencing, we used 1–3 ug of genomic DNA to construct libraries (insert size range 300–500 bp). Library preparation was conducted according to the Illumina library preparation protocol (Illumina Inc.). Sequencing was done on Illumina Hi-Seq with 100 and 150 paired-end sequencing kits.

Alignment and variant calling

All samples selected for genome sequencing were sequenced to approximately 10x coverage (Table S6.1). Reads were trimmed for a minimum phred quality > 20 over three consecutive base pairs and discarded if shorter than 45 base pairs. Alignment was performed with Mosaik Aligner (V. 1.1.0017) with the unique alignment option to the Porcine reference genome build 10.2. Variants were called using GATK Unified Genotyper version 2.8 (65). We used a prior of 0.01 for the probability of heterozygous calls (32).

Approximate Bayesian Computation (ABC)

104 genomes were used for the ABC analysis. Simulations were performed on 100 10kbp unlinked loci. Backward coalescent simulations with recombination were performed using *ms* (66) under 7 models (Figure S6.1). For model testing purposes, we ran 200,000 simulations per model. Summary statistics were computed on observed and simulated data using libsequence (67). We compared all models simultaneously (68) using a standard ABC-GLM approach as implemented in *ABCtoolbox* (69). For parameters inference we ran 2,000,000 simulations under the best fitting model. We extracted 10 Partial Least Square (PLS) components from the 93 summary statistics in the observed and simulated data (70). We retained a total of 10,000 simulations closest to the observed data and applied a standard ABC-GLM (71).

Exploratory analysis using SNP array

We used *TreeMix* (28) to build a maximum likelihood (ML) population tree from the 60K SNP dataset. We generated 10 replicates (with different seeds) and selected the run with the highest likelihood score. The PCA analysis was performed using *flashpca* (72).

Selection scan

We used SweeD to detect sweeps (35). To obtain critical threshold values (p-values), we used a posterior predictive simulation (PPS) approach. We simulated 2 replicates of 3Mbp each using the parameters of the 10,000 closest retained simulations from our ABC analysis (20,000 simulations). Simulations were run using macs (73). We derived a critical threshold for observed CLR in each population using the cumulative descriptive function (CDF) derived from the CLR distribution that was obtained from the PPS. All regions with $p<0.01$ were selected for further analysis.

Acknowledgments

We thank Ben Peter for his help and guidance during the model-fitting step of the analysis as well as for kindly sharing his code. We are also indebted to Daniel Wegmann for providing us with the latest version of *ABCtoolbox*. We also thank Konrad Lohse for his insights during the conception of the project. This project was financially supported by the European Research Council under the European Community's 256 Seventh Framework Programme (FP7/2007–2013)/ERC Grant agreement no 249894. JGS were supported by National Institutes of Health grant R01-40282 and National Science Foundation postdoctoral fellowship DBI-1402120.

Supplementary figures, tables and text

Supplementary text as well as figures and text are included (below the references) but Supplementary Table S6.1 because of formatting issues (too large). This table is available upon request from Laurent Frantz (laurent.frantz@gmail.com) or from the online version of the paper when published. Moreover, a summary of the populations analysed in this study can be obtained from Figure S2.

References

1. Darwin C (1868) The Variation of Animals and Plants Under Domestication (Murray, John, London).
2. Price EO (2002) Animal Domestication and Behavior (CABI Publishing, New York).

3. Zeder MA (2011) in Harlan II: Biodiversity in Agriculture: Domestication, Evolution and Sustainability, eds Damania A, Gepts P (Univ California Press, Davis), pp 227-229.
4. Vigne J-D (2011) The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. Comptes rendus biologies 334:171-81.
5. Driscoll CA, Macdonald DW, O'Brien SJ (2009) From wild animals to domestic pets, an evolutionary view of domestication. Proceedings of the National Academy of Sciences of the United States of America 106 Suppl :9971-8.
6. O'Connor TP (2007) Wild or domestic? Biometric variation in the catFelis silvestris Schreber. International Journal of Osteoarchaeology 17:581-595.
7. Larson G, Fuller DQ (2014) The Evolution of Animal Domestication. Annual Review of Ecology, Evolution and Systematics in press.
8. Larson G et al. (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science 307:1618-21.
9. Hanotte O et al. (2002) African pastoralism: genetic imprints of origins and migrations. Science 296:336-9.
10. Luikart G et al. (2001) Multiple maternal origins and weak phylogeographic structure in domestic goats. Proceedings of the National Academy of Sciences of the United States of America 98:5927-32.
11. Naderi S et al. (2008) The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. Proceedings of the National Academy of Sciences of the United States of America 105:17659-64.
12. Pedrosa S et al. (2005) Evidence of three maternal lineages in Near Eastern sheep supporting multiple domestication events. Proceedings Biological sciences / The Royal Society 272:2211-7.
13. Vilà C et al. (2001) Widespread origins of domestic horse lineages. Science (New York, NY) 291:474-7.
14. Macneish RS (1992) The origins of agriculture and settled life ed University of Oklahoma Press (Norman).
15. Zeder MA (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proceedings of the National Academy of Sciences of the United States of America 105:11597-604.
16. Ervynck A, Hongo H, Dobney K, Meadow R (2001) Born Free? New Evidence for the Status of Sus scrofa at Neolithic Çayönü Tepesi (Southeastern Anatolia, Turkey). Paléorient 27:47-73.
17. Dobney K, Larson G (2006) Genetics and animal domestication: new windows on an elusive process. Journal of Zoology 269:060222013030001-???

18. Marshall FB, Dobney K, Denham T, Capriles JM (2014) Evaluating the roles of directed breeding and gene flow in animal domestication. Proceedings of the National Academy of Sciences of the United States of America 111:6153-8.
19. Cucchi T, Hulme-Beaman A, Yuan J, Dobney K (2011) Early Neolithic pig domestication at Jiahu, Henan Province, China: clues from molar shape analyses using geometric morphometric approaches. Journal of Archaeological Science 38:11-22.
20. Larson G et al. (2007) 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.
21. Ottoni C et al. (2013) Pig Domestication and Human-Mediated Dispersal in Western Eurasia Revealed through Ancient DNA and Geometric Morphometrics. Molecular biology and evolution:mss261-.
22. Albarella U, Manconi F, Trentacoste A (2011) in Ethnozooarchaeology. The present and past of human-animal relationships, eds Albarella U, Trentacoste A (Oxbow Books, Oxford), pp 143-159.
23. White S (2011) From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. Environmental History 16:94-120.
24. Owen J et al. (2014) The zooarchaeological application of quantifying cranial shape differences in wild boar and domestic pigs (Sus scrofa) using 3D geometric morphometrics. Journal of Archaeological Science 43:159-167.
25. Evin A et al. (2013) The long and winding road: identifying pig domestication through molar size and shape. Journal of Archaeological Science 40:735-743.
26. Albarella U, Davis SJM, Detry CP, Rowley-Conwy P (2005) Pigs of the " Far West ": the biometry of Sus from archaeological sites in Portugal. Anthropozoologica 40:27-54.
27. Rowley-Conwy P, Albarella U, Dobney K (2012) Distinguishing Wild Boar from Domestic Pigs in Prehistory: A Review of Approaches and Recent Results. Journal of World Prehistory 25:1-44.
28. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics 8:e1002967.
29. Groenen MAM et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393-398.
30. Bosse M et al. (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature communications 5:4392.
31. Groenen MAM et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393-8.

32. Bosse M et al. (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS genetics 8:e1003100.
33. Frantz LA et al. (2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome biology 14:R107.
34. Nielsen R et al. (2005) Genomic scans for selective sweeps using SNP data. Genome research 15:1566-75.
35. Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Molecular biology and evolution 30:2224-34.
36. Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping It Local: Evidence for Positive Selection in Swedish Arabidopsis thaliana. Molecular biology and evolution:msu247-.
37. Andersson-Eklund L et al. (1998) Mapping quantitative trait loci for carcass and meat quality traits in a wild boar x Large White intercross. Journal of animal science 76:694-700.
38. Rubin C-J 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 109:19529-36.
39. Karim L et al. (2011) Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nature genetics 43:405-13.
40. Setoguchi K et al. (2009) Cross-breed comparisons identified a critical 591-kb region for bovine carcass weight QTL (CW-2) on chromosome 6 and the Ile-442-Met substitution in NCAPG as a positional candidate. BMC genetics 10:43.
41. Quartu M et al. (2005) Neurturin, persephin, and artemin in the human pre- and full-term newborn and adult hippocampus and fascia dentata. Brain research 1041:157-66.
42. Simanainen U et al. (2013) Evidence for increased tissue androgen sensitivity in neurturin knockout mice. The Journal of endocrinology 218:151-63.
43. Oschipok LW, Teh J, McPhail LT, Tetzlaff W (2008) Expression of Semaphorin3C in axotomized rodent facial and rubrospinal neurons. Neuroscience letters 434:113-8.
44. Hernández-Montiel HL, Tamariz E, Sandoval-Minero MT, Varela-Echavarría A (2008) Semaphorins 3A, 3C, and 3F in mesencephalic dopaminergic axon pathfinding. The Journal of comparative neurology 506:387-97.

45. Gonthier B et al. (2007) Functional interaction between matrix metalloproteinase-3 and semaphorin-3C during cortical axonal growth and guidance. Cerebral cortex (New York, NY: 1991) 17:1712-21.
46. Ruediger T et al. (2013) Integration of opposing semaphorin guidance cues in cortical axons. Cerebral cortex (New York, NY: 1991) 23:604-14.
47. Niquille M et al. (2009) Transient neuronal populations are required to guide callosal axons: a role for semaphorin 3C. PLoS biology 7:e1000230.
48. Pasterkamp RJ, Kolk SM, Hellemons AJCGM, Kolodkin AL (2007) Expression patterns of semaphorin7A and plexinC1 during rat neural development suggest roles in axon guidance and neuronal migration. BMC developmental biology 7:98.
49. Brown CB et al. (2001) PlexinA2 and semaphorin signaling during cardiac neural crest development. Development 128:3071-3080.
50. Ultanir SK et al. (2012) Chemical genetic identification of NDR1/2 kinase substrates AAK1 and Rabin8 Uncovers their roles in dendrite arborization and spine development. Neuron 73:1127-42.
51. Chevallier J et al. (2009) Rab35 regulates neurite outgrowth and cell shape. FEBS letters 583:1096-101.
52. Ong SH et al. (2000) FRS2 Proteins Recruit Intracellular Signaling Pathways by Binding to Diverse Targets on Fibroblast Growth Factor and Nerve Growth Factor Receptors. Molecular and Cellular Biology 20:979-989.
53. Sokol DK et al. (2006) High Levels of Alzheimer Beta-Amyloid Precursor Protein (APP) in Children With Severely Autistic Behavior and Aggression. J Child Neurol 21:444-449.
54. Grayton HM, Missler M, Collier DA, Fernandes C (2013) Altered social behaviours in neurexin 1α knockout mice resemble core symptoms in neurodevelopmental disorders. PloS one 8:e67114.
55. Bhoiwala DL et al. (2013) Overexpression of RCAN1 isoform 4 in mouse neurons leads to a moderate behavioral impairment. Neurological research 35:79-89.
56. Dierssen M et al. (2011) Behavioral characterization of a mouse model overexpressing DSCR1/ RCAN1. PloS one 6:e17010.
57. Kim KS, Larsen N, Short T, Plastow G, Rothschild MF (2000) A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. Mammalian Genome 11:131-135.
58. Xu P et al. (2013) Double deletion of melanocortin 4 receptors and SAPAP3 corrects compulsive behavior and obesity in mice. Proceedings of the National Academy of Sciences of the United States of America 110:10759-64.

59. Valette M et al. (2013) Eating behaviour in obese patients with melanocortin-4 receptor mutations: a literature review. International journal of obesity (2005) 37:1027-35.
60. Wojcik SM et al. (2013) Genetic markers of a Munc13 protein family member, BAIAP3, are gender specifically associated with anxiety and benzodiazepine abuse in mice and humans. Molecular medicine (Cambridge, Mass) 19:135-48.
61. Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in Anopheles gambiae. PLoS biology 3:e285.
62. Pickrell J, Reich D (2014) Towards a new history and geography of human genes informed by ancient DNA (Cold Spring Harbor Labs Journals)
63. Larson G, Burger J (2013) A population genetics view of animal domestication. Trends in genetics: TIG 29:197-205.
64. Girdland Flink L et al. (2014) Establishing the validity of domestication genes using DNA from ancient chickens. Proceedings of the National Academy of Sciences of the United States of America 111:6184-9.
65. McKenna A et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-1303.
66. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.
67. Thornton K (2003) libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325-2327.
68. Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS genetics 8:e1003011.
69. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC bioinformatics 11:116.
70. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182:1207-18.
71. Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. Genetics 184:243-52.
72. Abraham G, Inouye M (2014) Fast principal component analysis of large-scale genome-wide data. PloS one 9:e93766.
73. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome research 19:136-42.

Supplementary Information

ABC

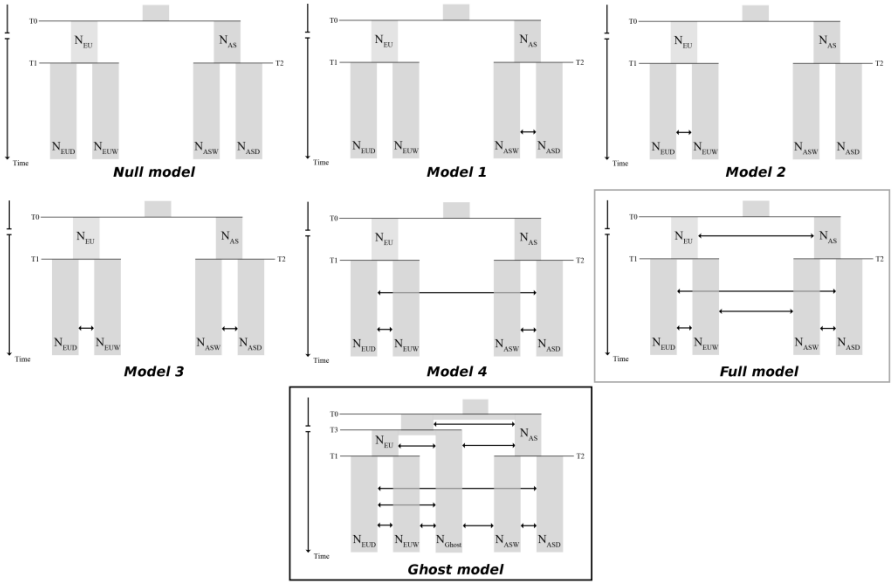
104 genomes were used for the ABC analysis. Simulations were performed on 100 10kbp unlinked loci. To match these simulations we filtered out 10kb loci with more than 10% missing data (from the variant calling step) in all 104 genomes. We also filtered out any loci containing CpG islands (1) and within 100kb of coding sequences. We then required that all loci were separated by at least 100kb to limit the effect of linkage. We polarized mutations using the genome of a Java Warty pig (*S. verrucosus*)(2). Lastly we randomly selected 100 loci that met these criteria. Backward coalescent simulations with recombination were performed using *ms* (3) under 7 models (Figure S6.1). Table S6.4 recapitulates the priors used for the model parameters. For model testing purposes, we ran 200,000 simulations per model. For each simulation we computed summary statistics, solely based on allele frequency to avoid phasing issues, using libsequence (4). For each population, we computed the number of segregating sites (S), number of private mutations (n1), nucleotide diversity (pi), ThetaW, ThetaH, Tajima's D, and Fay and Wu's H. In addition, we computed Fst as well as all other statistics for each pair of populations. For model testing we choose a set of informative summary statistics with a Partial Least Squares Discriminant Analysis as in (5) using the '*plsda*' function in R (6). We compared all models simultaneously using a standard ABC-GLM approach as implemented in *ABCtoolbox* (7).

For parameters inference we ran 2,000,000 simulations under the full migration model (Fig 1a; Fig S1). We did not use the ghost model for parameter inference because of the higher number of parameters in this model (6 extra: 1 Ne, 1 time, 4 migrations) that increases parameter space. Moreover, given we have no data about this ghost population, these parameters cannot be accurately estimated with the current approach (8). We extracted 10 Partial Least Square (PLS) components from the 93 summary statistics in the observed and simulated data (9). We retained a total of 10,000 simulations closest to the observed data and applied a standard ABC-GLM (10).

We checked for bias in the prior using 1,000 pseudo observed data (POD) sets with known parameters value (11). We then computed the coverage properties of the posterior distribution using our 10,000 closets simulations. Uniformity was assessed using a classical Kolmogorov-Smirnov test for each parameter independently (11) (Figure S6.4).

We evaluated the power of our approach to infer each parameter using the 1,000 POD by computing root mean square error of the mode (RMSEmode; Table S6.4) for known parameters (11). In order to check if the data is in agreement with the

assumed model we computed the distribution of the marginal densities of the 10,000 retained simulations for posterior estimation and computed the fraction of simulation with smaller marginal densities than the observed data set (11).



Supplementary Figure S6.1: All models investigated in this study. Schematic of all models tested in this study. The upper 6 models were first compared together. In this comparison, the Full model (circled with a grey square) was the best fitting model. When all 7 models were tested together, the Ghost model fitted best (circled with a black square). All priors and support values are reported in Supplementary Table S6.1.

Validation of ABC procedure

To validate our model testing procedure, we used 1,000 pseudo-observed datasets (POD). We found that our approach can recover the right model for 899 out of 1,000 POD. In addition, we found that under all models but model 4, the full model and the ghost model (Figure S6.1), all retained simulation had higher marginal likelihood than the observed data for all models. This suggests that these models provided a very poor fit to this genomic data-set. In contrast, we found that the fraction of simulation with lower marginal likelihood was 0.009 for model 4, 0.043 for the full model and 0.1 for the ghost model. This suggests that these models are capable of reproducing the observed summary statistics (10 PLS components; Figure S6.2) (5, 11). We also used 1,000 POD under the full model to check for biased prior during parameter estimation. To do so, we checked the uniformity of the posterior quantile distribution using a Kolomogorov-Smirnov test for each

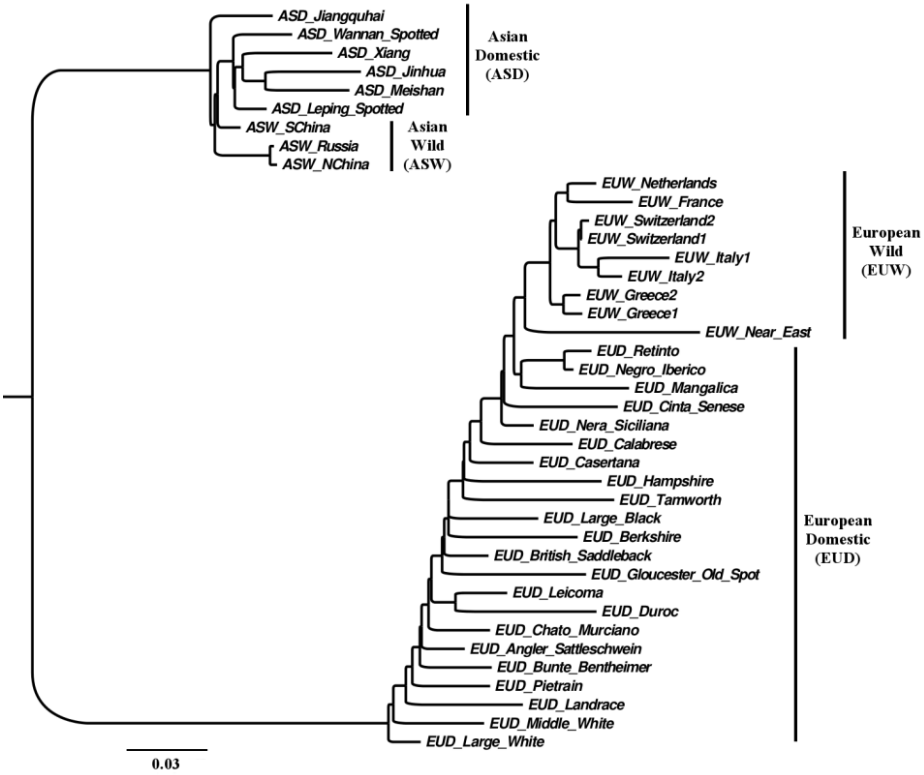
parameter as suggested by (9, 11). We found that most parameter had a uniform distribution (Table S6.4).

Table S2: List of enriched term in EUD

Gene ontology term	Gene count	P(FDR)
developmental process	26#3347	<0.0001
cellular component organization and biogenesis	24#3277	0.0008
anatomical structure development	17#2005	0.0013
multicellular organismal development	18#2299	0.0027
urogenital system development	3#40	0.0224
cellular developmental process	14#1810	0.0224
cell differentiation	14#1810	0.0224
multicellular organismal process	23#3822	0.0245
cell communication	30#5560	0.0245
vesicle-mediated transport	8#606	0.0252
signal transduction	28#5142	0.0292
multicellular organismal development#system development	13#1605	0.0394
positive regulation of cell adhesion	2#15	0.0394
anatomical structure morphogenesis	10#1047	0.0424
positive regulation of biological process	10#1062	0.0424
nervous system development	8#716	0.0424
blood circulation	4#160	0.0424
circulatory system process	4#160	0.0424
biological regulation	33#6731	0.0440
neuropeptide signaling pathway	4#168	0.0456
mesenchymal cell development	2#24	0.0575
cell development	10#1242	0.0585

Table S6.3: List of enriched term in ASD

Gene ontology term	Gene count	P(FDR)
cellular component organization and biogenesis	18#3277	0.0002
establishment of protein localization	8#922	0.0309
protein localization	8#961	0.0309
protein complex assembly	5#340	0.0309
nucleotide metabolic process	5#340	0.0309
macromolecule localization	8#1012	0.0309
nucleobase, nucleoside and nucleotide metabolic process	5#367	0.0313
cellular localization	8#1126	0.0392
protein transport	7#866	0.0392
odontogenesis	2#25	0.0392
positive regulation of transcription	4#279	0.0455
base-excision repair, AP site formation	1#1	0.0455
optic placode formation involved in camera-type eye	1#1	0.0455
optic placode formation	1#1	0.0455
calcium-independent cell-matrix adhesion	1#1	0.0455
DNA catabolic process	2#35	0.0455
positive regulation of nucleobase and nucleic acid metabolic process	4#289	0.0455
macromolecular complex assembly	6#756	0.0550
anatomical structure morphogenesis	7#1047	0.0591
cellular component assembly	6#813	0.0607
establishment of cellular localization	7#1098	0.0607
regulation of nitrogen compound metabolic process	1#2	0.0607
heme oxidation	1#2	0.0607
nitrogen utilization	1#2	0.0607
regulation of nitrogen utilization	1#2	0.0607
organ morphogenesis	4#362	0.0632



Supplementary Figure S6.2: Result of the *TreeMix* analysis for the 602 pigs genotyped on the porcine 60SNP array data set.

Table S4: Prior and posterior distribution for full migration model

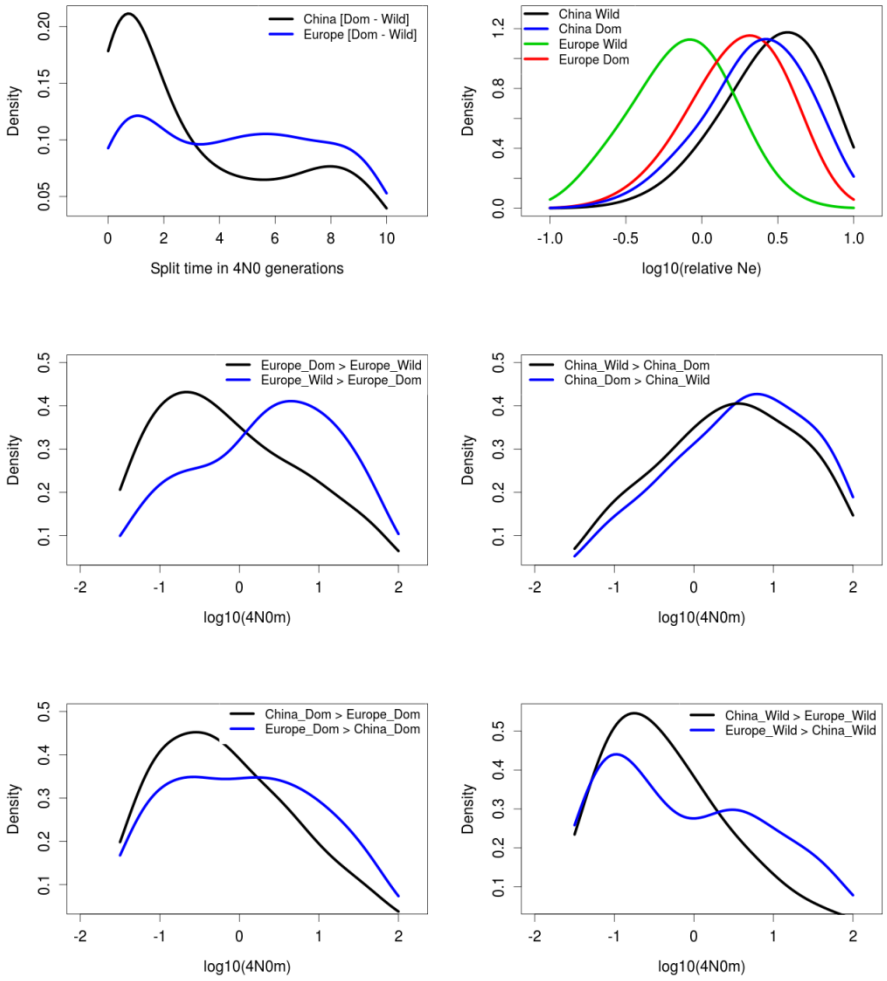
Parameter	Prior_lower	Prior_upper	P_value_KS	RMSE	mode	HDI50_lower	HDI50_upper	HDI90_lower	HDI90_upper	HDI95_lower
N_AS	-0.5	2.5	1	0.72	1.22727	0.750015	1.9697	-0.0729657	2.39394	-0.181818
N_EU	-0.5	2.5	1	0.7	1.83333	1.17008	2.27273	-0.0902323	2.42424	-0.237376
N_ASD	-1	1	<0.001	0.2	0.414142	0.191426	0.666667	-0.19746	0.929293	-0.309252
N_ASW	-1	1	1	0.22	0.555556	0.343435	0.795971	-0.0408476	0.989899	-0.199722
N_EUD	-1	1	<0.001	0.19	0.313131	0.0614713	0.525252	-0.316608	0.787878	-0.428899
N_EUW	-1	1	0.001	0.19	-0.070707	-0.337659	0.141414	-0.722642	0.424242	-0.80808
m_ASW_ASD	-1.5	2	1	0.67	0.550505	-0.0328282	1.28176	-0.916225	1.87626	-1.05808
m_EUD_ASD	-1.5	2	1	0.7	-0.580808	-0.878541	0.568182	-1.38927	1.45202	-1.44697
m_ASD_ASW	-1.5	2	0.748	0.67	0.79798	0.233987	1.48737	-0.718329	1.98232	-1.02233
m_EUW_ASW	-1.5	2	0.745	0.73	-0.969697	-1.4563	-0.103535	-1.44697	1.38789	-1.44697
m_ASD_EUD	-1.5	2	1	0.7	-0.545455	-1.09495	0.0732324	-1.44697	1.12186	-1.44697
m_EUW_EUD	-1.5	2	0.018	0.66	0.656566	0.0732324	1.38113	-1.10469	1.73485	-1.27587
m_ASW_EUW	-1.5	2	1	0.75	-0.757576	-1.12879	-0.14448	-1.44697	0.832851	-1.48232
m_EUD_EUW	-1.5	2	0.166	0.66	-0.651515	-1.16414	0.0862117	-1.44697	1.29853	-1.44697
T0	20	70	1	14.07	54.8485	36.6545	59.6465	23.4734	66.2121	22.2727
T1	0.0001	10	0.389	2.51	0.707165	0.151615	3.10774	0.0506056	8.46701	0.0506056
T2	0.0001	10	0.136	2.48	1.1112	0.353634	5.06885	0.151616	8.82958	0.151616

Exploratory analysis SNP array

To further support our claim of gene-flow between wild and domestic pigs, we analysed a 622 pigs from the same population as above that were genotyped using the Porcine SNP60 array (Table S6.1; (12). We first performed a Principal Component Analysis (PCA) as implemented in *flashpca* (13) to investigate the relationship among these populations. Unsurprisingly, we found that the first PC discriminates between Asian and European pigs (Fig S5). This is in line with previous studies that found that European and Asian wild boar populations likely diverged around 1My ago (2). In addition, we found that none of the PCs discriminate among Asian populations (Fig S5-6), while PC2-4 show clear differentiation among most European breeds (Figure S6.6). This result is due to the fact that the Porcine SNP60 chip was ascertained in European commercial pigs (12). To further investigate historical relationship among these populations we used *TreeMix* (14) to fit a bifurcating graph to this dataset. Surprisingly, we found that EUD and ASD are paraphyletic, while EUW are monophyletic (Figure S6.2). Such a finding is difficult to reconcile with a simple model of domestication that involves a single source population and/or little gene-flow between wild and domestics. However, paraphyly and complex ancestry in domestic pigs could be the result of multiple events of ascertainment bias as well as interbreeding between Asian and European domestics during the 19th century industrial revolution (15–17). Nevertheless, our samples include many non-commercial breeds which are unlikely to be heavily admixed with Asian domestics (15).

Migration rates

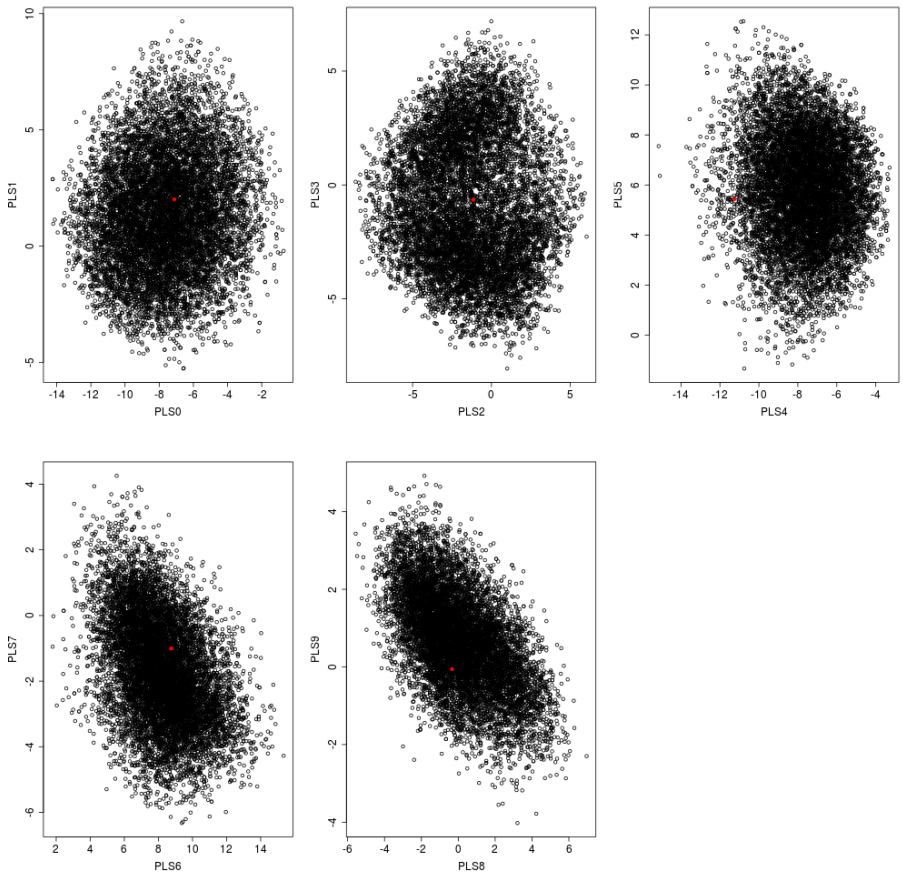
To further test the hypothesis that the gene-flow ASD <-> EUD did not influence our findings we simulated 2 million samples under the best fitting model and used ABC to estimate the posterior distribution of migration rates. We found that rate of gene flow EUW → EUD was quite high. We estimate mEUD,EUW (fraction of the EUD population made up of EUW migrants each generation) to be ~ 1.1x10-4 (Table S6.3), which corresponds to 2.33 migrants per generations. On the other hand we found that the rate of gene-flow EUD → EUW was quite low with mEUW,EUD ~= 5.577x10-6, which corresponds to 0.047 migrants per generation. This pattern was reversed in Asia, with mASW,ASD ~= 1.5x10-4 (ASD → ASW; ~5.64 migrants/generations) and mASD,ASW ~= 8.9x10-5 (ASW → ASD; ~2.3 migrants/generations).



Supplementary Figure S6.3: Posterior distribution of all parameters in the Full model (Figure S6.1). Population size are of the relative population size (the ratio of the current population size over the population size before T0 [Figure 1]).

Lastly, the rate of migration between the two domestic populations (ASD and EUD) was much lower, with mASD,EUD ~= 6.56x10-6 (EUD → ASD; ~ 0.17 migrants per generation) and mEUD,ASD ~= 7.12x10-6 (ASD → EUD; ~= 0.14 migrants per generation). This result shows that the gene-flow between ASD and EUD did not

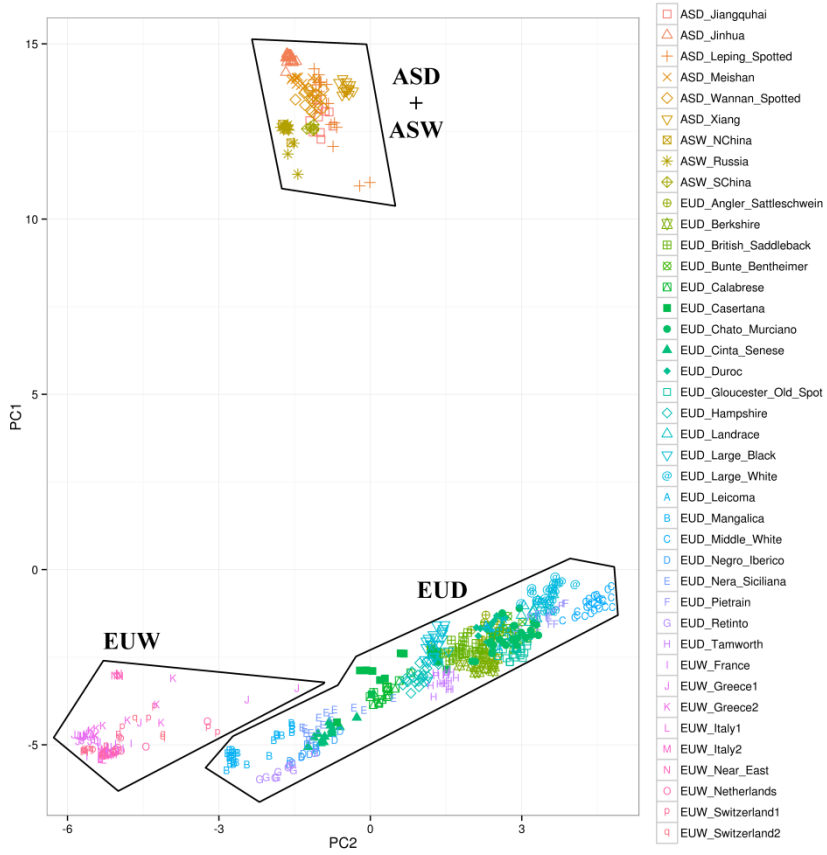
affect our finding that wild boars significantly contributed to the gene-pool of domestic pigs



Supplementary Figure S6.4: PLS distribution of 10,000 (out of 2,000,000) retained simulations (black) and observed data (red) under the Full model.

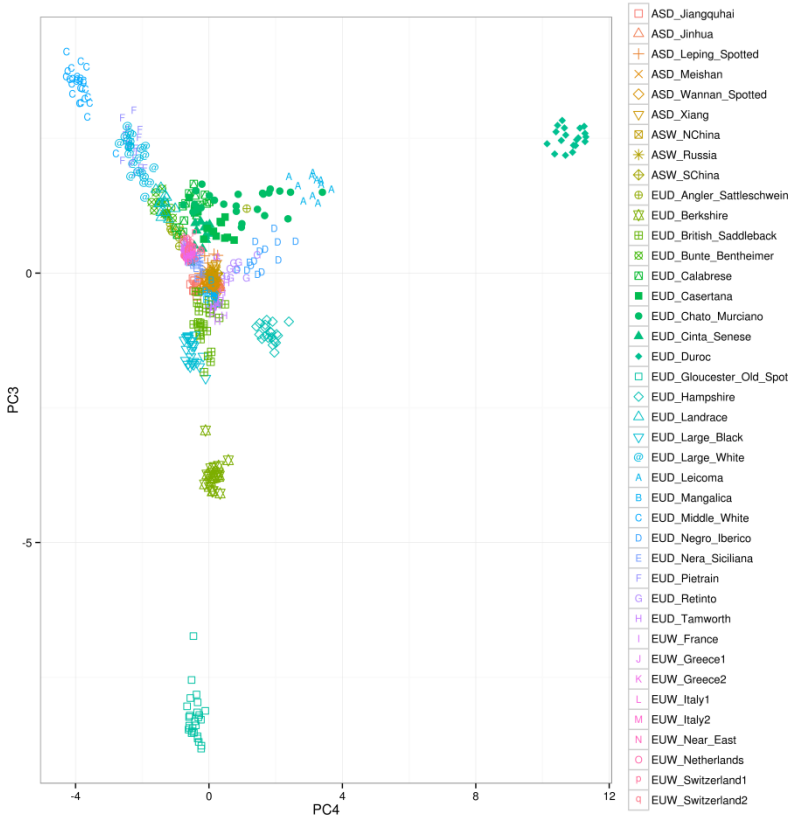
Demography of Asian pigs

The same possible population decline as highlighted in the main text was observed in ASW and ASD (Figure S6.3). This is also consistent with Pleistocene glaciation-induced population decline (33–35). Nevertheless, we found that contrary to European pigs, ASW had a higher effective population size (Ne-ASW=~35,933) than ASD (Ne-ASD=~25,947).



Supplementary Figure S6.5: Result of PCA analysis (PC1-2) based on 602 genotyped pigs.

However, for modelling purpose as well as due to the limited number of wild boars from Asia available in the study, we made the assumption that all Asian wild boars form a single population. This assumption likely influenced our demographic analysis in China; as it was shown previously that population from North and South China show much greater genetic differentiation than between any modern European wild boar population (2, 16, 18, 19). In addition, North and South Chinese wild boars did not form a monophyletic clade in our 60K SNP analysis (Figure S6.2). Such un-accounted long-term substructure likely influenced the results of the demographic analysis in China and makes it difficult to draw strong conclusions about demography of domestication in Chinese pigs.



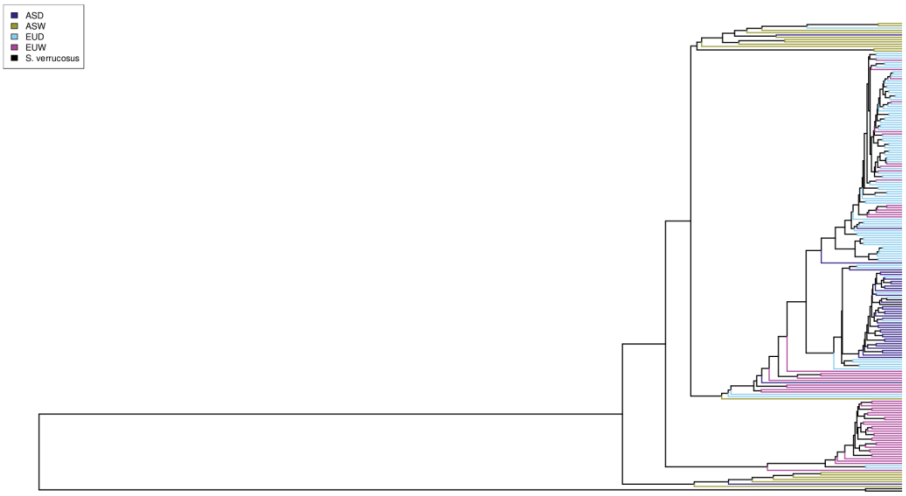
Supplementary Figure S6.6: Result of PCA analysis (PC3-4) based on 602 genotyped pigs.

However, such finding is not expected to have any influence on the conclusion from our gene-flow analysis. Indeed, substructure within ASD could result in artificial migration ASD -> ASW, if one of the ASW subpopulation was closer to ASD. However, substructure cannot explain the migration ASW → ASD if ASD was genetically isolated.

Selection scan

We used SweeD to detect sweeps (20). The program was run for each population separately using all available SNPs. The highest composite likelihood ratio (CLR) score for every 10kb interval was used for further analysis. To obtain critical threshold values (p-values), we used a posterior predictive simulation (PPS) approach. We simulated 2 replicates of 3Mbp each using the parameters of the 10,000 closest retained simulations from our ABC analysis (20,000 simulations).

Simulations were run using macs (21). We derived a critical threshold for the observed CLR in each population using the cumulative descriptive function (CDF) derived from the CLR distribution that was obtained from the PPS. All regions with p<0.01 were selected for further analysis. We computed the overlap of these regions between populations and defined set of regions unique to each population as well as overlapping only between ASD and EUD. These sweep coordinates were then overlapped with the Ensembl (v75) gene annotation. We tested for enrichment of gene ontology term in each population using a fisher-exact test with a Benjamini-Hochberg correction for multiple testing as implemented in the Gostat program (22). We only considered genes with human orthology (Goa-human).



Supplementary Figure S6.7: Example of a genealogy at a sweep region that could be explained by admixture ASD<->EUD.

To perform phylogenetic analyses, we first extracted 20bk basepairs around putative parallel sweeps. We then phased these regions in BEAGLE 4 (23) using default settings. We then built trees using UPGMA as implemented in the R package Phangorn (24) after computing Kimura-2-parameter model corrected distances using the R package ape (25).

References for Supplementary Information

1. Tortereau F 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.
2. Frantz LA et al. (2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome biology 14:R107.
3. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.
4. Thornton K (2003) libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325-2327.
5. Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS genetics 8:e1003011.
6. Lê Cao K-A, González I, Déjean S (2009) integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics 25:2855-6.
7. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC bioinformatics 11:116.
8. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences of the United States of America 108:15123-8.
9. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182:1207-18.
10. Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. Genetics 184:243-52.
11. Wegmann D, Excoffier L (2010) Bayesian inference of the demographic history of chimpanzees. Molecular biology and evolution 27:1425-35.
12. Ramos AM 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.
13. Abraham G, Inouye M (2014) Fast principal component analysis of large-scale genome-wide data. PloS one 9:e93766.
14. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics 8:e1002967.
15. White S (2011) From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. Environmental History 16:94-120.

16. Groenen MAM et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393-8.
17. Bosse M et al. (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature communications 5.
18. Larson G et al. (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science (New York, NY) 307:1618-21.
19. Bosse M et al. (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS genetics 8:e1003100.
20. Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Molecular biology and evolution 30:2224-34.
21. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome research 19:136-42.
22. Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics (Oxford, England) 20:1464-5.
23. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American journal of human genetics 81:1084-97.
24. Schliep KP (2011) phangorn: phylogenetic analysis in R. Bioinformatics (Oxford, England) 27:592-3.
25. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289-290.

7

General discussion

7.1 Introduction

Modern evolutionary genetics in combination with high-throughput sequencing have revolutionised our understanding of evolutionary biology. The unprecedented amount of data together with advanced modelling of complex evolutionary processes have allowed biologists to make increasingly finer predictions about the complex evolutionary history of life on earth as well as providing a better understanding of the mechanisms that forged it.

In this thesis I have described many aspects of the evolutionary history of *Suiformes*. Using modern evolutionary genetics techniques together with large genomic data-sets, I described many novel aspects of the complex evolutionary history of this fascinating group and provided many novel methodological insights. In the following paragraph I will discuss these findings in a broader context to provide, a comprehensive evolutionary history of *Suiformes*, from speciation to domestication.

7.2 Phylogenies and divergence time from genome sequences.

Molecular phylogenetics provides the back bone of most of the work described in this thesis. Phylogenetics, not only allows biologists to infer relationships among a group of organisms but it also provides a framework to draw and test hypotheses. In Chapter 2 I have shown how time calibrated phylogenetics and molecular clock analyses on genome scale sequence data allows one to test hypotheses on the time of divergence of a clade. In particular, in Chapter 2 I have shown how molecular data suggest that *Suiformes* dispersed overseas to North America at least twice in the last 30My. However, such inference could not be possible without robust methods to build phylogenetic trees from whole genome sequences. Thus, in this section I will discuss some of the findings from Chapter 2 that highlight some of the weaknesses of current phylogenomics practices.

The development of parametric phylogenetics (see General Introduction) has allowed for robust tree inference from DNA sequences. However, as we are scaling from single genes to genomes, new limitations are arsing. Large scale phylogenomic studies, involving hundreds of informative characters, as done in Chapter 2, are becoming common. Nevertheless, multiple nodes in the tree of life are yet to be resolved. One particularly striking example is the root of Eutherian mammals which has received considerable attention, but still remains controversial. Indeed multiple methods have been applied to genome-wide mammalian datasets giving widely different results (*i.e*. McCormack et al., 2012; Morgan et al., 2013; Romiguier et al., 2013; Song, et. al 2012). These large scale

phylogenetic studies can be divided into two main paradigms. In the first paradigm, known as concatenation, genes are treated as partitions and assembled into a 'supermatrix' that is used to compute a single phylogeny. However, this methodology can be misleading as it ignores the inherent genealogical incongruence in the genome (Kubatko & Degnan, 2007). Indeed common evolutionary processes such as incomplete lineage sorting (ILS), or horizontal gene transfer (HGT) can lead to incongruence (see General Introduction). The second paradigm, also known as supertree, proposes a very different strategy in which locus-trees or gene-trees are computed separately, for each locus, and reconciled into a species tree. There is a wide range of methods available for the reconciliation step in a supertrees analysis. These methods involve simple consensus based inference that summarises the support for a species tree given a set of gene-trees (Jermiin et al., 1997; Liu et al., 2009). More complex methods modelling biological processes such as ILS (Kubatko, 2009; Liu et al., 2010; Maddison & Knowles, 2006) to compute the probability of a species tree given a set of gene-trees under the coalescence model with recombination (Wakeley, 2008) are also widely used.

These supertree methods almost always rely on assumptions that intra loci recombination is absent and complete between loci. The lack of recombination within a locus is a key concept because all supertree methods assume that each genetic segment analysed evolved under a single evolutionary history (represented by a single gene-tree). However, in Chapter 2 I argue that this assumption is likely to be violated in many cases, especially for deep phylogenies, that involve divergence time on the order of millions of generations or in the case of species with large effective population size. For example, assuming that 1Mbp = 1 centiMorgan (cM) we can compute the recombination rate, per generation, between two adjacent base pairs as:

$$r = \frac{1-e^{-2m}}{2} = 1e^{-7}.$$

where m is the distance in Morgan. Using this recombination rate one can compute the expected number of recombination in a random mating population with an effective population size N_e as:

$$C = 4N_e r \quad \text{(Hudson, 1983).}$$

Thus, for deep nodes and/or very large effective population sizes, it is easy to recognise that we can expect at least one recombination event between every

adjacent base in the genome since the most recent common ancestor (MRCA) of the species of interest. Nevertheless, recombination is not expected to bias topology search by species tree methods, as long as the recombining fragments share the same evolutionary history (hence the same topology). However recombination events that assemble fragments with incongruent topologies may pose a problem as this phenomenon violates the assumptions of free recombination between loci and no intra-locus recombination shared by most species tree methods (Gatesy & Springer, 2013; Kubatko & Degnan, 2007).

The probability of observing incongruent gene trees depends on the population size and the length of internal branches (time between two divergence events). The probability of coalescence, *P*, for a random mating population in an interval of time t (in generations) is:

$$P = \left(\frac{1}{2N_e}\right)^t$$

where N_e is the effective population size. This simple equation shows that, short internal branches and large populations will have a drastic effect on rate of coalescence. Therefore, if N_e is large or t is short, we would expect a large number of incomplete lineage sorting (ILS). These ILS can then be assembled with so called 'sorted' lineages (species tree like) on the same haplotype via recombination. If recombination, over the evolutionary depth under study, has acted between every base pair in the alignment this problem may have some important impact on phylogenetic analyses.

Few studies have so far investigated the impact of recombination on phylogenetic reconstruction (Lanier & Knowles, 2012; Posada & Crandall, 2002; Schierup & Hein, 2000). While some studies have found that recombination can have an effect on single loci tree inference (Posada & Crandall, 2002; Schierup & Hein, 2000) others have found that the rate of recombination has very little effect on phylogenetic incongruence (Lanier & Knowles, 2012). Thus, the fact that recombination rate *per-se* does not affect tree reconstruction is not surprising because of the expected saturation of recombination events for deep phylogenies and/or for species with large effective population size. However, no studies have yet investigated the joint effect of internal branch length, recombination and *Ne* on supertree and supermatrix methods. The interplay of these parameters can have a drastic impact on deep phylogenetic reconstruction that uses conventional phylogenetics (supermatrix as well as supertree). This problem is starting to be roughly

characterised (*i.e.* Chapter 2; Gatesy and Springer, 2013) and it is a crucial issue that needs to be addressed in more details if we are to resolve all nodes of the tree of life such as the root of mammals (see General introduction).

7.3 The role of climatic fluctuation and the complex speciation history of Sus in Island South East Asia (ISEA)

Another problematic phenomenon for phylogenetics is the possibility of inter-specific admixture. In the previous paragraphs I mentioned horizontal gene transfer (HGT) as a source of incongruence. This definition is often used for genetic material exchanged via non-traditional reproduction such as plasmid conjugation in bacteria. These can lead to incongruence among genealogies if the HGT was between two non-monophyletic species. Exchange of genetic material between non-monophyletic species is also possible in mammals through traditional reproduction. Common wording for higher organisms include introgression, admixture, evolutionary loops or reticulation. In Chapter 3 and 4 we showed, using methods based on phylogenetics and population genetics, that reticulation is common during speciation in *Sus*. Together these studies provide the most comprehensive evolutionary history of the genus to date.

Previous to this work the evolutionary history of these species had only been assessed using mtDNA (Larson et al., 2005, 2010; Lucchini et al. 2005; Randi et al. 1996). These analyses have shown that it was possible to distinguish between Eurasian *S. scrofa* populations but not between these different species in ISEA. This is particularly peculiar given the wide morphological differences between these species. Two hypotheses could explain this pattern.

Hypothesis 1: The speciation of these species was rapid and recently induced by late Pleistocene climatic fluctuations and resulted in large incongruence and little internal substitution leaving the phylogeny impossible to resolve also at the nuclear level.

Hypothesis 2: The speciation of these species has taken place over the early climatic fluctuations of the late Pliocene/early Pleistocene; when islands were separated during long inter glacial periods. However, long glacial periods during the late Pleistocene resulted in land bridges between islands, parapatric conditions and led to inter-specific admixture. Such phenomenon would also have resulted in complex phylogenies and conflict between mtDNA and nuclear DNA.

In chapters 3 and 4 we conclusively show that hypothesis 2 is much more likely. This demonstrates the power afforded by a single genome sequence to resolve complex evolutionary histories compare to multiple short DNA fragments. Moreover our analyses reveal multiple aspects of the evolutionary history of these

species. We show that a bifurcating model is too simplistic to explain the evolution of the genus *Sus*. For example, we found multiple ancestry origins for the Sulawesi warty pig, which mainly finds its root in Borneo but also has ancestry from Java (Javanese Warty pigs) and the Philippines. The endogenous species of The Philippines, *S. cebifrons* was also likely the result of multiple colonisation waves, one from Borneo, and one from mainland China. These species have clear hybrid origins, and it would be interesting to know if hybridization provided the means for these species to be more adaptive, or alternatively, if reticulation reduce adaptivity.

Besides these reticulations events, our main finding is that the Sunda-shelf is a cradle for inter-specific admixture. Indeed we show that this continental shelf, combined with the recurrent glacial periods, had a strong effect on speciation. This is highlighted by the mtDNA that shows the monophyly of the Sunda-shelf taxon. It would be interesting to see how that has affected other species in this area, especially if the exposed Sunda-shelf also acted as a barrier for obligate forest taxon. Indeed, the existence of a savannah corridor during glacial periods has been advocated by numerous analyses (Cannon et al., 2009; Gathorne-Hardy et al., 2002; Nater et al., 2011; Slik et al., 2011). For example, such a corridor would have had different impact on taxon such as pigs and orang-utan. Thus, it will be interesting to know how this 'cradle' for inter-specific admixture impacted the biodiversity of these massively biodiverse regions of the world (Myers et al., 2000).

7.4 Fitting complex models to genome sequences

The genus *Sus* is not an isolated case. Indeed many other cases of complex speciation, involving reticulation have been reported (Brandvain et al. 2014; Cahill et al., 2013; Eaton & Ree, 2013; Ellegren et al., 2012; Green et al., 2010; Miller et al., 2012). These studies have an important impact on our conception of the process of speciation. In my general Introduction I discussed simple models of speciation, such as allopatry, parapatry and sympatry. Thus, these reticulate speciation events are incompatible with the assumption of allopatric speciation. However, how does one detect these reticulations events with confidence from molecular data? How can we disentangle these from ILS in genomic data?

Fitting complex models of speciation to genetic data is a difficult task. Different methods have been applied, such as simple summary statistics (Patterson et al., 2012) (i.e. D-statistics, F-statistics), and approximate Bayesian computation (Nadachowska-Brzyska et al., 2013; Wegmann & Excoffier, 2010). Summary statics based methods such as D-statistics or Fst only provide limited information about specific models. Indeed, these methods lack the ability to implicitly test for specific

historical scenarios, thus limiting their usefulness. In Chapter 4 I show that such statistics can be difficult to interpret and can lead to biased inferences. In particular these methods often do not have an implicit assumption about the direction of gene-flow. As we showed in Chapter 4, this can lead to an erroneous conclusion.

More parametric approaches have been proposed such as the Isolation with Migration (IM) model (Hey & Nielsen, 2007). These approaches have the particularity that they use an analytical solution for the likelihood of different scenarios, thus allowing for explicit hypothesis testing. However, one of the main problems with these approaches is the difficulty to find an analytical solution to the likelihood of complex models involving many populations/species. In addition, these methods often rely on MCMC simulations to infer the posterior distribution of parameters in the model, making these computationally demanding. "Short-cuts" such as ABC permits to compute the marginal likelihood of complex models for which an analytic solution to the likelihood is not available (Leuenberger & Wegmann, 2010). However, these are even more computationally intensive and require large-scale simulations (see Chapter 6). Recent approaches such as the one used in Chapter 4 or in Lohse & Frantz (2014) provide a fast and robust alternative to evaluate the likelihood of different models of evolution from three genome sequences obtained from different populations/species. This approach allows for detailed model testing. We demonstrated its usefulness to unravel the evolutionary history of archaic humans (Lohse and Frantz 2014) and of *Sus* (Chapter 4). However, this approach is still limited to few populations (three at max) and requires long enough evolutionary depth to obtain meaningful distribution of branch length, while short enough not to violate infinite site assumption (no back mutations; Lohse et al., 2011). Thus, this method was not applicable in the context of domestication (see Chapter 6). Lastly, other methods rely on simulations and composite likelihood to infer the demographic parameters and model support using site frequency spectrum (SFS) (Excoffier et al. 2013; Gutenkunst et al. 2009). These SFS methods are very promising and their accuracy and computational intensity are continuously improved. Thus, it is needless to say that this is a very exciting area of science and the constant improvement of these methods will allow for finer and more accurate model testing and parameter information from whole genome sequences, ultimately providing decisive information on the process of speciation.

7.5 Dissecting genomes to understand complex speciation mechanisms

It is important to note the limitations of the methods described above when it comes to answer basic questions about speciation. Indeed, while we now have a pretty clear idea that the process of speciation is often more complex than a simple bifurcating model, the mechanisms that allow species to diverge with gene flow is puzzling. There is now a growing body of work that investigates this issue, scrutinising the genomes of multiple species for regions that are permeable and impermeable to inter-specific admixture. This often involves methods such as admixture mapping. However, detangling regions that are admixed from regions that are sharing ancestral polymorphisms is difficult. Moreover, past population processes such as ancestral substructure, which is very difficult to detect, can result in false impression of admixture at a specific region. One very straightforward example is the Neanderthal / Human admixture (Eriksson & Manica, 2012). Thus, it is important to develop new methods that are robust to a wide range of biases. Novel methods using complex machine learning algorithms provide an attractive way to solve this problem (Sankararaman et al., 2014). However, much work is still needed to test these methods under realistic models.

Identifying divergent and admixed regions is crucial to understand the process of speciation. One clear example is the *Heliconus* butterfly, that shows interspecific admixture which can be mapped to genes link to colour in wing patterns (Martin et al., 2013). Thus, it has been hypothesised that non-monophyletic species can share similar wing patterns due to adaptive introgression following secondary contact (Martin et al., 2013). Most often the regions that we need to identify to better characterise the process of speciation are regions that show high inter-specific divergence. For example, we recently proposed that species specific olfactory receptors (OR) copy numbers (CN) are highly divergent among *Sus* species (Paudel et. al. in prep.). These OR may provide the means for these species to reduce inter-specific gene-flow. Indeed, if species specific CN at OR are primordial for a species (*i.e.* for mate recognition or foraging), we expect these to be highly divergent. Understanding which regions are involved in speciation will help to characterise this fundamental process and to figure out how reproductive isolation arises. However, it is still difficult to disentangle divergence arising from neutral processes (*i.e.* demographic induced genetic-drift) and divergence arising from natural selection. The development of methods to infer demographic parameters and to evaluate increasingly more complex models as well as methods that accurately measure divergence in genome sequences will play a decisive role in speciation

research and will provide the necessary means to understand how sympatric or parapatric species stay genetically isolated.

7.6 Domestication of *S. scrofa*, theories concept and novel genomic insights

The second major theme of this thesis is domestication. Since Darwin (Darwin, 1868), the process of domestication has fascinated Evolutionary Biologists. The incredible varieties of domestic breeds from different species (i.e. pigeons and dogs) provide an incredible example of evolutionary plasticity. Traditionally, this process is viewed as human-directed, involving strong bottlenecks in the domestic population (i.e. founder events) and reproductive isolation between wild and domestic forms (O'Connor, 2007; Price, 2002; Vigne, 2011; Zeder, 2011). Under this view, domestication is similar to the process of speciation in which reduction of gene-flow between populations (and selection) eventually leads to phenotypic differentiation and reproductive isolation (Driscoll et al. 2009). Such a view is also common in people's mind. Indeed most people clearly distinguish a dog from wolf or a pig from a wild boar and often consider the wild and domestic form as two, clearly defined, entities that imply orthology of domestic traits.

The major discovery of Chapter 6 is that this preconception has no genetic basis in pigs. Indeed, our analysis in Chapter 6 demonstrates that pig domestication was more complex than depicted by traditional models. More precisely, we showed that domestication was a complex process that involved large amount of gene-flow from wild boars to domestic pigs and thus violates the assumption of reproductive isolation. This has multiple implications. Firstly, this demonstrates that domestication of pigs was a complex long term process, as opposed to a fast genetic differentiation induced by human maintained reproductive isolation between wild and domestic pigs. The fact that there was a limited maintenance of isolation between wild and domestic forms also suggests that domestication of pigs was not necessarily initiated voluntarily by humans (see commensal domestication in Chapter 1; Larson & Fuller, 2014; Zeder, 2011). Indeed, the finding that gene-flow was important during domestication stands in stark contrast with more recent voluntary (direct) domestication of species such as mink and rabbits. Involuntary domestication is a fascinating idea as it suggests that humans may not have been as active in this process as previously thought. Thus, leaving aside the anthropocentric view of domestication, one could argue that pigs domesticated humans as well. Indeed, association with humans provides a large boost of fitness in the target species. Using humans as a vector for increased fertility and stable

food supply is an efficient evolutionary strategy. Such strategy must be as beneficial for the fitness for both domesticated species and humans.

The second implication of our finding is that reproductive isolation between wild and domestic forms was not maintained, even after domestication. This strongly violates assumptions of traditional domestication models and has very important consequences for studies that infer time and origin of domestication in pigs, dogs or any pre-historic domestic animal (Gerbault et al., 2014; Larson & Burger, 2013). This is highlighted by the finding that at least one population of wild boars, not involved in the original domestication process, most likely contributed to the gene-pool of domestic pigs (Chapter 6). Thus, while there are unequivocal proofs that pigs were first domesticated in Anatolia (Dobney & Larson, 2006; Ervynck et al., 2001; Larson & Fuller, 2014), the transportation across Europe of early domestic pigs resulted in interbreeding with local wild boar populations (Larson et al., 2007; Marshall et al., 2014). Under a traditional model of domestication, observations of greater genetic similarity between domestic and wild boar than between domestic and domestic is interpreted as the result of *de-novo* (independent) domestication (given the assumption of reproductive isolation between wild and domestic forms). However, post-domestication gene-flow and *de-novo* domestication are very different phenomena. Indeed, gene-flow between domestic pigs and wild-boars, even early after domestication, cannot be considered as a domestication event simply because domestication preceded gene-flow. Thus, this raises the question: what is domestication? We can define domestication as "morphological and or behavioural changes induced by human-mediated involuntary or voluntary selection that results in direct control over breeding to improve traits that are beneficial for humans (the last step in the process)." Domestication in the case of pigs and dogs is a long term association that became final when humans started to acquire control over breeding. Therefore, I believe that the result of Larson et al. 2007 that showed that Near-eastern pigs were first brought by Anatolians into Europe and subsequently replaced, few thousand years afterwards, by genetically European pigs is the consequence of post-domestication gene-flow. Indeed, clear control over breeding was already acquired, to some degree, by early farmers moving through Europe (Larson & Fuller, 2014). The recruitment of local wild boars by early European farmers is therefore by no mean a domestication event but simply the consequence of loose pig management (White, 2011). It is important to highlight the nuance difference between domestication and post-domestication gene-flow to understand how modern human civilizations arose. Indeed, under this model, domestication has been achieved only in a few centres. This suggests that knowledge and technology are not necessarily generated in parallel around the

globe. For example, domestication of goats and sheep took place also in Anatolia few thousand years before pigs. So was the domestication of these species a necessary step before domestic pigs? Was the Levant a gold mine for "domesticable" species? I believe that underlying available resources in a geographic area (such as the number of species that can be domesticated) greatly influenced the development of modern societies and this phenomenon is responsible for part of technological differences in the modern world (Diamond 1997).

It is interesting to note that the definition of domestication, given above, also fits to more recent, cases (i.e. rabbits and minks), in which domestication is direct (Zeder, 2011). Direct domestication is fast because humans take control of breeding in a single generation and apply much stronger voluntary artificial selection. In my opinion these direct domestication episodes are similar to more ancient domestications (i.e. dogs, pigs, cattle, sheep etc.) but differ in the order of events in the process. In a direct pathway, control over breeding is achieved before behavioural and morphological alteration. In these direct domestications, traits are most likely fixed at a much faster rate due in part to the fact that human maintained reproductive isolation between domestic and wild forms. Moreover, direct and more ancient domestications also differ due to technological advances. Indeed, in a direct domestication pathway, it is clear that humans already know that domestication is possible and have a good understanding of how the process could be achieved in a relatively fast manner. In case of more ancient domestication, this process most likely took much longer due to the lack of clear examples of previous domesticated species. In other words, technological advances not only allowed for faster domestication but also allowed to reduce the criterion necessary for a species to be domesticated (*i.e.* see Diamond, 1997). Indeed, early domestication may have been solely possible in species that had the necessary plasticity to engage in a long term domestication process with humans. For example, highly sociable wolves must have had large variation in behaviour to maintain social hierarchy within packs. Such large variation may have provided the necessary phenotypic plasticity for domestication to take place without strong involvement by humans. On the other hand species such as rabbits and mink clearly display less behavioural plasticity (from a human perspective) and thus should be less adapted to long term association with humans.

Lastly, the realisation that gene-flow was ubiquitous during and after domestication of pigs raises questions regarding the maintenance of domestic traits in face of gene-flow (Marshall et al., 2014). Indeed, how are domestic traits maintained if domestication and post-domestication processes involved much gene

influx from wild forms? This question was addressed, to some extent, in Chapter 6. In this study we showed results that are consistent with the existence of parallel sweeps in two independent domestication processes (Asian and European). This finding suggests that selection for similar traits may result in sweeps at the same loci. Thus, given this finding, I propose the existence of genomic regions, governing domestic traits in pigs, which are impermeable to gene-flow from wild boars. Similarly to islands of speciation (maintained via natural selection), these would act as "island of domestication". However, the existence of such islands is purely speculative and requires further investigation. In addition, I propose two mechanisms for the maintenance of these islands. Firstly, these islands could be obtained from a single haplotype. Under this hypothesis, all domestic pigs should share an MRCA in the sweep region, that is younger that the MRCA of any pig and wild boar. On the other hand, these islands could be the results of recurrent selection from standing genetic variation in wild boar populations and leave soft sweep signatures. I think the latter is more likely given our results in Chapter 6 that showed great heterogeneity in genealogies at sweep regions found in the genome of domestic pigs (paraphyly of domestic pigs). The possibility of soft sweeps (from standing genetic variation) during domestication has important implications for conservation of wild boars. Indeed, if true, this hypothesis would imply that wild boars are an incredibly valuable genetic resource that could provide the necessary means to adapt domestic pigs in our changing world. Lastly, it would be interesting to test hypotheses such as hybrid vigour as a possible explanation for intentional breeding between pigs and wild boars. Indeed, intentional breeding local wild boars may have been necessary to provide flexible immune response in diverse area of the globe.

The rise of multidisciplinary domestication research combining expertise from Evolutionary Biologists and Zooarchaeologists as well as new methods such as ancient DNA and Geometric Morphometric (GMM) (*i.e.* Evin et al., 2013; Owen et al., 2014*)* together with testable hypotheses, such as the one proposed above, will most likely shed light on the mechanisms that allowed early farmers to maintain domestic traits while allowing large-scale gene-flow between wild and domestic pigs as well as refine our understanding of this fascinating process.

References

Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G., & Sweigart, A. L. (2014). Speciation and introgression between Mimulus nasutus and Mimulus guttatus. PLoS genetics, 10(6), e1004410. doi:10.1371/journal.pgen.1004410

Cahill, J. A., Green, R. E., Fulton, T. L., Stiller, M., Jay, F., Ovsyanikov, N., Salamzade, R., et al. (2013). Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS genetics, 9(3), e1003345. doi:10.1371/journal.pgen.1003345

Cannon, C. H., Morley, R. J., & Bush, A. B. G. (2009). The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. Proceedings of the National Academy of Sciences of the United States of America, 106(27), 11188-93. doi:10.1073/pnas.0809865106

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

Diamond, J. (1997). Gun, Germs, and Steel. W. W. Norton & Company.

Dobney, K., & Larson, G. (2006). Genetics and animal domestication: new windows on an elusive process. Journal of Zoology, 269, 261-271. doi:10.1111/j.1469-7998.2006.00042.x

Driscoll, C. A., Macdonald, D. W., & O'Brien, S. J. (2009). From wild animals to domestic pets, an evolutionary view of domestication. Proceedings of the National Academy of Sciences of the United States of America 9971-8. doi:10.1073/pnas.0901586106

Eaton, D. A. R., & Ree, R. H. (2013). Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Systematic biology, in press. doi:10.1093/sysbio/syt032 Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T.,

Künstner, A., et al. (2012). The genomic landscape of species divergence in Ficedula flycatchers. Nature, 491(7426), 756-60. doi:10.1038/nature11584

Eriksson, A., & Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proceedings of the National Academy of Sciences of the United States of America, 109(35), 13956-60. doi:10.1073/pnas.1200567109

Ervynck, A., Hongo, H., Dobney, K., & Meadow, R. (2001). Born Free ? New Evidence for the Status of Sus scrofa at Neolithic Çayönü Tepesi (Southeastern Anatolia, Turkey). Paléorient, 27(2), 47-73. doi:10.3406/paleo.2001.4731

Evin, A., Cucchi, T., Cardini, A., Strand Vidarsdottir, U., Larson, G., & Dobney, K. (2013). The long and winding road: identifying pig domestication through molar size and shape. Journal of Archaeological Science, 40(1), 735-743. doi:10.1016/j.jas.2012.08.005

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS genetics, 9(10), e1003905. Public Library of Science. doi:10.1371/journal.pgen.1003905

Gathorn-Hardy, F. J. Davies, R. G. Eggleton, P., & Jones, D. T.. (2002). Quaternary rainforest refugia in south-east Asia: using termites (Isoptera) as indicators. Biological Journal of the Linnean Society, 75(4), 453-466. doi:10.1046/j.1095-8312.2002.00031.x

Gatesy, J., & Springer, M. S. (2013). Concatenation versus coalescence versus "concatalescence". Proceedings of the National Academy of Sciences of the United States of America, 110(13), E1179. doi:10.1073/pnas.1221121110

Gerbault, P., Allaby, R. G., Boivin, N., Rudzinski, A., Grimaldi, I. M., Pires, J. C., Climer Vigueira, C., et al. (2014). Storytelling and story testing in domestication. Proceedings of the National Academy of Sciences of the United States of America, 111(17), 6159-64. doi:10.1073/pnas.1400425111

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., et al. (2010). A draft sequence of the Neandertal genome. Science, 328(5979), 710-22. doi:10.1126/science.1188021

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics, 5(10), e1000695. doi:10.1371/journal.pgen.1000695

Hey, J., & Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences of the United States of America, 104(8), 2785-90. doi:10.1073/pnas.0611164104

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. Theoretical population biology, 23(2), 183-201.

Jermiin, L. S., Olsen, G. J., Mengersens, K. L., & Easteap, S. (1997). Majority-Rule Consensus of Phylogenetic Trees Obtained by Maximum-Likelihood Analysis. Molecular biology and evolution, 14(12), 1296-1302.

Kubatko, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. Systematic biology, 58(5), 478-88. doi:10.1093/sysbio/syp055

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic biology, 56(1), 17-24. doi:10.1080/10635150601146041

Lanier, H. C., & Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? Systematic biology, 61(4), 691-701. doi:10.1093/sysbio/syr128

Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Schibler, J., Tresset, A., Vigne, J.-D., et al. (2007). 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(39), 15276-81. doi:10.1073/pnas.0703411104

Larson, G., & Burger, J. (2013). A population genetics view of animal domestication. Trends in genetics, 29(4), 197-205. doi:10.1016/j.tig.2013.01.003

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., Lowden, S., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science, 307(5715), 1618-21. doi:10.1126/science.1106927

Larson, G., & Fuller, D. Q. (2014). The Evolution of Animal Domestication. Annual Review of Ecology, Evolution and Systematics, in press.

Larson, G., Liu, R., Zhao, X., Yuan, J., Fuller, D., Barton, L., Dobney, K., et al. (2010). Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. Proceedings of the National Academy of Sciences of the United States of America, 107(17), 7686-91. doi:10.1073/pnas.0912264107

Leuenberger, C., & Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. Genetics, 184(1), 243-52. doi:10.1534/genetics.109.109058

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC evolutionary biology, 10(1), 302. doi:10.1186/1471-2148-10-302

Liu, L., Yu, L., Pearl, D. K., & Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. Systematic biology, 58(5), 468-77. doi:10.1093/sysbio/syp031

Lohse, K, Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. Genetics, 189(3), 977-87. doi:10.1534/genetics.111.129569

Lohse, Konrad, & Frantz, L. A. F. (2014). Neandertal Admixture in Eurasia Confirmed by Maximum Likelihood Analysis of Three Genomes. Genetics, genetics.114.162396-. doi:10.1534/genetics.114.162396

Lucchini, V., Meijaard, E., & Diong, C. (2005). New phylogenetic perspectives among species of South-east Asian wild pig (Sus sp.) based on mtDNA sequences and morphometric data. Journal of Zooglogy, (266), 25-35. doi:10.1017/S0952836905006588

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. Systematic biology, 55(1), 21-30. doi:10.1080/10635150500354928

Marshall, F. B., Dobney, K., Denham, T., & Capriles, J. M. (2014). Evaluating the roles of directed breeding and gene flow in animal domestication. Proceedings of the National Academy of Sciences of the United States of America, 111(17), 6153-8. doi:10.1073/pnas.1312984110

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., et al. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome research, 23(11), 1817-28. doi:10.1101/gr.159426.113

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elemenss are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome research, 22(4), 746-54. doi:10.1101/gr.125864.111

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., Kim, H. L., et al. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proceedings of the National Academy of Sciences of the United States of America, 109(36), E2382-90. doi:10.1073/pnas.1210506109

Morgan, C. C., Foster, P. G., Webb, A. E., Pisani, D., McInerney, J. O., & O'Connell, M. J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. Molecular biology and evolution, 30(9), 2145-56. doi:10.1093/molbev/mst117

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. Nature, 403, 853-858.

Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013). Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. PLoS genetics, 9(11), e1003942. doi:10.1371/journal.pgen.1003942

Nater, A., Nietlisbach, P., Arora, N., van Schaik, C. P., van Noordwijk, M. A., Willems, E. P., Singleton, I., et al. (2011). Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant Orangutans (genus: Pongo). Molecular biology and evolution, 28(8), 2275-88. doi:10.1093/molbev/msr042

Owen, J., Dobney, K., Evin, A., Cucchi, T., Larson, G., & Strand Vidarsdottir, U. (2014). The zooarchaeological application of quantifying cranial shape differences in wild boar and domestic pigs (Sus scrofa) using 3D geometric morphometrics. Journal of Archaeological Science, 43, 159-167. doi:10.1016/j.jas.2013.12.010

O'Connor, T. P. (2007). Wild or domestic? Biometric variation in the catFelis silvestris. International Journal of Osteoarchaeology, 17(6), 581-595. doi:10.1002/oa.913

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., et al. (2012). Ancient admixture in human history. Genetics, 192(3), 1065-93. doi:10.1534/genetics.112.145037

Posada, D., & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. Journal of molecular evolution, 54(3), 396-402. doi:10.1007/s00239-001-0034-9

Price, E. O. (2002). Animal Domestication and Behavior. New York: CABI Publishing. Randi, E., Lucchini, V., & Diong, C. H. (1996). Evolutionary genetics of the suiformes as reconstructed using mtDNA sequencing. Journal of Mammalian Evolution, 3(2), 163-194. doi:10.1007/BF01454360

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., & Douzery, E. J. P. (2013). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Molecular biology and evolution, 30(9), 2134-44. doi:10.1093/molbev/mst116

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., et al. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature, 507(7492), 354-7. doi:10.1038/nature12961

Schierup, M. H., & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. Genetics, 156(2), 879-91.

Slik, J. W. F., Aiba, S.-I., Bastian, M., Brearley, F. Q., Cannon, C. H., Eichhorn, K. A. O., Fredriksson, G., et al. (2011). Soils on exposed Sunda shelf shaped biogeographic patterns in the equatorial forests of Southeast Asia. Proceedings of the National Academy of Sciences of the United States of America, 108(30), 12343-7. doi:10.1073/pnas.1103353108

Song, S., Liu, L., Edwards, S. V., & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences of the United States of America, 109(37), 14942-7. doi:10.1073/pnas.1211733109

Vigne, J.-D. (2011). The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. Comptes rendus biologies, 334(3), 171-81. doi:10.1016/j.crvi.2010.12.009

Wakeley, J. . . R. (2008). Coalescent Theory: An Introduction. Roberts and Company Publishers.

Wegmann, D., & Excoffier, L. (2010). Bayesian inference of the demographic history of chimpanzees. Molecular biology and evolution, 27(6), 1425-35. doi:10.1093/molbev/msq028

White, S. (2011). From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History. Environmental History, 16(1), 94-120. doi:10.1093/envhis/emq143

Zeder, M. A. (2011). Pathways to animal domestication. In A. Damania & P. Gepts (Eds.), Harlan II: Biodiversity in Agriculture: Domestication, Evolution and Sustainability (pp. 227-229). Davis: Univ California Press.

Summary

Summary

The evolutionary history of *Suiformes* is poorly known. However, pigs are widely spread all across the world, from pests, to key species in various ecosystems but also a major source of animal protein for millions of humans across the world. For millenniums, humans have used these species in various ways, and it is needless to say that modern societies owe part their development to pigs and related species. Moreover, because of their peculiar evolutionary history (complex speciation and domestication) *Suiformes* are a perfect model to study complex evolutionary processes such as domestication and speciation. The main aim of this thesis is to provide a compressive view of the evolutionary history of *Suiformes*, from speciation to domestication and to refine our understanding of these fundamental evolutionary processes. In this work I use genome scale data to unravel many aspects of the peculiar evolution of these fascinating species and to provide new insights on basic evolutionary processes. Secondary aims of this work focus on developing and testing methods to better characterise the evolutionary history of species.

In **Chapter 1** I provide some basic information about *Suiformes* and genomics. In particular, I introduce many concepts that are recurrent in this thesis, such as theoretical definitions of speciation and domestication as well as a brief introduction of state of the art methods used in this thesis. This chapter should provide the reader with the necessary background to comprehend the motivation and the conclusions of this thesis.

Phylogenetics provides a basic evolutionary toolkit to study speciation. However, the trees of Suiformes as well as methods to construct genome wide phylogenies are controversial. Indeed, processing next-generation short-read genome sequences requires either a reference genome for alignment or a *de-novo* assembly, the latter of which is often prohibitively expensive for large genomes. In **Chapter 2** we present the results of our analyses of the genome sequences of six species representing all the genera of the *Suiodae* superfamily for which only distant reference genomes are available. To do so, we first evaluated the performance of multiple aligners to align reads to a distant reference genome. We then tested the effect of different variant calling methods. Our results show that while local aligners perform well over large hamming distances, different methods to call variants can have strong effects on nucleotide distance to the reference. However, we show that it is possible to overcome this issue using two reference genomes. Thereafter we simulated DNA sequences with sequencing errors under multiple phylogenetic tree shapes. We found that while errors have a strong effect on phylogenetic power, these are unlikely to positively bias phylogenetic analyses.

We then investigated phylogenetic support across the genome by comparing the likelihood of different trees at multiple genomic scales (2, 5 and 10kb). We show that concatenation approaches lead to overly optimistic support values, whereas supertree approaches can lead to overly pessimistic support. We show that the latter is the result of incomplete lineage sorting and lack of phylogenetic signal in small genome segments. Thus, while we empirically demonstrate the presence of ILS at shorter inter-nodes our analysis also reveals that it is difficult to divide the genome in blocks small enough to detect ILS yet long enough to keep enough signal. We expect this phenomenon to be more problematic as inter-nodes get shorter and older. Lastly we perform a thorough molecular clock analysis to time the divergence of the two families *Suidae* and *Tayassuidae*. Our results support the view that New World *Suiodae* is not a monophyletic clade and suggest two wave of colonization of America.

Phylogenies provide the backbone of speciation research. However, elucidating the process of speciation requires an in-depth understanding of the evolutionary history of the species in question. Studies that rely upon a limited number of genetic loci do not always reveal actual evolutionary history, and often confuse inferences related to phylogeny and speciation. Whole-genome data, however, can overcome this issue by providing a nearly unbiased window into the patterns and processes of speciation. We address these questions in **Chapter 3**. In order to reveal the complexity of the speciation process, we sequenced and analysed the genomes of 10 wild pigs, representing morphologically or geographically well-defined species and subspecies of the genus Sus from insular and mainland Southeast Asia, and one African common warthog. Our data highlight the importance of past cyclical climatic fluctuations in facilitating the dispersal and isolation of populations, thus leading to the diversification of suids in one of the most species-rich regions of the world. Moreover, admixture analyses revealed extensive, intra- and inter-specific gene-flow that explains previous conflicting results obtained from a limited number of loci. We show that these multiple episodes of gene-flow resulted from both natural and human-mediated dispersal. Our results demonstrate the importance of past climatic fluctuations and human mediated translocations in driving and complicating the process of speciation in island Southeast Asia. This case study demonstrates that genomics is a powerful tool to decipher the evolutionary history of a genus, and reveals the complexity of the process of speciation.

In **Chapter 3** we thus demonstrate the complexity of the process of speciation in *Sus* and showed that inter-specific admixture is common in the genus. However, the timing, directionality and extent of this admixture remain unknown. **In Chapter**

4 we use a likelihood based model comparison to more finely resolve the admixture and test whether it was mediated by humans or occurred naturally. Our model testing approach suggests that inter-specific admixture between Sunda-shelf species was most likely asymmetric and occurred before the arrival of humans in the region. In addition, we show that our method provides a significant improvement over previous methodology (D-statistics) to characterize the direction of admixture. Our analysis reveals that these species diverged during the late Pliocene but around 23% of their genomes have been affected by admixture during the later Pleistocene climatic transition.

In Chapters 3 and 4 we showed that *Sus scrofa* (wild and domestic pigs) emerged in Southeast Asia during the climatic fluctuations of the early Pliocene 5.3-3.5 MYA. Then, beginning ~10,000 years ago, pigs were domesticated in multiple locations across Eurasia. However, many aspects of the evolutionary history of the *Sus scrofa* remain unknown. In **Chapter 5** we use genome data from over 55 samples of wild and domestic pigs to investigate multiple aspects of the evolutionary history of this widely spread species. The demographic history of this widespread species is remains unknown. Thirdly, we know little about the time of divergence of the Asian and European subtypes of wild boars that have been domesticated independently. Lastly, the evolutionary history of domestic pigs and wild boar is poorly known. For example, we do not know how common was interbreeding between wild and domestic pigs or between Asia and European domestic pigs or how domestication affected demography. Phylogenomic analyses of complete genome sequences from these wild boars and six domestic pigs revealed distinct Asian and European lineages that split during the mid-Pleistocene 1.6-0.8 MYA (Frantz et al. 2013; Calabrian stage). Our demographic analysis on the whole genome sequences of European and Asian wild boars, revealed an increase in the European population after pigs arrived from China. During the Last Glacial Maximum (LGM; ~20KYA), however, Asian and European populations both suffered through bottlenecks. These bottlenecks were more pronounced in Europe than Asia, suggesting a greater impact of glaciations on higher latitude regions. In addition, our admixture analysis revealed European influence in Asian breeds, and a ~35% Asian fraction in European breeds. These results are consistent with the known exchange of genetic material between European and Asia pig breeds. We also observed that European breeds form a paraphyletic clade, which cannot be solely explained by varying degrees of Asian admixture. Within each continent, our analysis revealed different degrees of relatedness between breeds and their respective wild relatives.

Chapter 6 focuses on domestication of pigs and tests many hypothesises drawn in Chapter 5. The process of domestication led to the most important transitions

during the Neolithic era. Traditionally, this process is assumed to be strongly human-directed, with few individuals initially selected to be domesticated and reproductive isolation between wild and domestic forms. However, zooarchaeological evidence depicts animal domestication as a geographically restricted, long-term process without reproductive isolation or strong intentional selection. Here, we ask whether pig domestication follows a traditional, linear model or a complex, reticulate model as predicted by zooarchaeologists. To do so, we fit models of domestication to whole genome data from over 100 wild and domestic pigs. We found that the assumptions of traditional models, such as reproductive isolation and strong domestication bottlenecks, are incompatible with the genetic data and provide support for the zooarchaeological theory of a complex domestication process in pigs. In particular, gene-flow from wild to domestic pigs was a ubiquitous feature of domestication and post-domestication processes in pigs. In addition, we show that despite gene-flow, the genomes of domestic pigs show strong signatures of selection at loci that affect behaviour and morphology. Specifically, our results are consistent with independent parallel sweeps in two independent domestication areas (China and Anatolia) at loci linked to morphological traits. We argue that recurrent selection for domestic traits likely counteracted the homogenising effect of gene-flow from wild boars and created "islands of domestication" in the genome.

The General discussion in **Chapter 7** provides additional discussion on these topics as well as synthesis of the work described in this thesis. More precisely, this section aims at providing a compressive evolutionary history of *Suiformes* but also a reflection on complex evolutionary processes such as speciation, domestication as well as on the effect of these processes on the genome. Lastly, this section also aims at discussing many methodological aspects that were developed or tested in this thesis.

Samenvatting

Over de evolutionaire geschiedenis van varkensachtigen is weinig bekend. Toch zijn 2toont de admixture analyse intra- en interspecifieke gene-flow aan die vroegere conflicterende resultaten op basis van een klein aantal loci kan verklaren. We laten zien dat deze episodes van gene-flow het resultaat zijn van zowel natuurlijke als mens-gedreven dispersie. Onze resultaten demonstreren het belang van vroegere klimatologische fluctuaties alsmede door de mens gedreven translocaties zijn in de complexiteit van de soortvorming op de Zuidoost-Aziatische eilanden. Deze case study laat zien dat genomica een krachtig middel is om de evolutionaire geschiedenis van een soort te ontcijferen en toont de complexiteit van het proces van soortvorming.

3In hoofdstuk 3 demonstreren we de complexiteit van soortvorming in Sus en het veelvuldig optreden van inter-specifieke admixture binnen dit genus. Toch is de timing, directionaliteit en omvang van deze admixture onbekend. In **hoofdstuk 4** gebruiken we een waarschijnlijkheids-gebaseerd model voor een nauwkeuriger analyse van de admixture en om te testen of humane interferentie daarbij een rol speelde. Onze resultaten suggereren dat inter-specifieke admixture tussen Sunda-schaal soorten meest waarschijnlijk asymmetrisch was en plaats vond voor de komst van de mens in de regio. Bovendien laten we zien dan onze methode een significante verbetering geeft over voorgaande methoden (D-statistics) voor het bepalen van de richting van de admixture. Onze analyse onthult dat deze soorten gedurende het late plioceen divergeerden, maar dat ongeveer 23% van hun genomen beïnvloed zijn door admixture gedurende de latere pleistocene klimatologische overgang.

4In hoofdstukken 3 en 4 beschreven we dat de soort *Sus scrofa* (wilde en gedomesticeerde varkens) in Zuidoost Azie is ontstaan tijdens de klimatologische fluctuaties van het vroege plioceen 5.3-3.5 Mya. Vervolgens begint ~10.000 jaar geleden de domesticatie van deze soort in verschillende gebieden in Eurazië. Toch zijn vele aspecten van de evolutionaire geschiedenis van Sus scrofa onduidelijk gebleven**. In hoofdstuk 5** gebruiken we daarom de genomische data van meer dan 55 individuele wilde en gedomesticeerde varkens om meerdere aspecten van de evolutionaire geschiedenis van deze wijdverspreide soort te onderzoeken. De demografische geschiedenis van deze wijdverspreide soort is nog steeds onbekend. Ook weten we weinig over de tijd van divergentie van de Aziatische en Europese subtypes van wilde zwijnen die onafhankelijk gedomesticeerd zijn. Tenslotte is er weinig kennis over de evolutionaire geschiedenis van gedomesticeerde varkens en wilde zwijnen. Wij weten bijvoorbeeld niet hoe gebruikelijk paringen tussen wilde en gedomesticeerde varkens of tussen Aziatische en Europese gedomesticeerde varkens waren, of hoe domesticatie de demografie heeft beïnvloedt.

Fylogenomische analyses van complete genoomsequenties van deze wilde zwijnen en zes gedomesticeerde varkens hebben gescheiden Aziatische en Europese lijnen onthuld die gedurende het mid-pleistoceen, 1.6-0.8 MYA, zijn gesplitst (Frantz et al. 2013; Calabrische fase). Onze demografische analyse op de volledige genoomsequenties van Europese en Aziatische wilde zwijnen vertonen een toename van de omvang van de Europese populatie nadat varkens vanuit China arriveerden. Desalniettemin, gedurende het hoogtepunt van de laatste ijstijd (LGM; ~20KYA) leden zowel Aziatische als Europese populaties onder bottlenecks. Deze bottlenecks waren sterker in Europa dan in Azië wat een groter effect van ijstijden op populaties in hogere hoogtegraad suggereert. Bovendien onthult onze admixture analyse Europese invloed in Aziatische rassen, en een Aziatische fractie van ~35% in Europese rassen. Deze resultaten zijn consistent met de bekende uitwisseling van genetisch materiaal tussen Europese en Aziatische varkensrassen. We hebben ook gevonden dat Europese rassen een parafyletische clade vormen, die niet alleen verklaard kan worden door een verschillende mate van Aziatische admixture. Binnen elk continent laat onze analyse verschillende waarden van verwantschap tussen rassen en hun respectieve wilde soortgenoten zien.

5De focus in **Hoofdstuk 6** ligt op de domesticatie van varkens en het testen van diverse hypothesen die **in** hoofdstuk 5 zijn opgeworpen. Het proces van domesticatie heeft tot de meest belangrijke overgangen gezorgd gedurende het neolithische tijdperk. Traditioneel gezien is dit proces sterk beïnvloed door de mens, met slechts een klein aantal individuen die geselecteerd zijn voor domesticatie en reproductieve isolatie tussen wilde en gedomesticeerde vormen. Desalniettemin tekent zooarcheologisch bewijs dierdomesticatie als een geografisch besloten, langdurig proces zonder reproductieve isolatie of sterke bedoelde selectie. We vragen ons hier af of varkensdomesticatie een traditioneel, lineair model volgt, of een complex reticulair model zoals de zooarcheologen voorspellen. Hiervoor passen we modellen van domesticatie toe op volledige genoomdata van meer dan 100 wilde en gedomesticeerde varkens. De resultaten tonen aan dat de veronderstellingen van traditionele modellen, zoals reproductieve isolatie en sterke domesticatie bottlenecks niet compatibel zijn met de genetische data en we verstrekken bewijs voor de zooarcheologische theorie van een complex domesticatie proces in varkens. Voornamelijk, gene-flow van wilde naar gedomesticeerde varkens was een veelvoorkomend aspect van domesticatie en post-domesticatie processen in varkens. Bovendien laten we zien dat ondanks gene-flow, de genomen van gedomesticeerde varkens sterke selectie signatuur vertonen op loci die gedrag en morfologie beïnvloeden. Meer specifiek, onze resultaten zijn consistent met onafhankelijke parallelle ‘sweeps’ in twee

onafhankelijke domesticatie regio's (China en Anatolië) op loci gelinkt aan morfologische kenmerken. We argumenteren dat herhaaldelijke selectie voor gedomesticeerde kenmerken waarschijnlijk het homogeniserende effect van gene-flow van wilde zwijnen heeft tegengewerkt, en dat dit 'eilanden van domesticatie' in het genoom heeft gecreëerd.

6De algemene discussie in **hoofdstuk 7** gaat vervolgens dieper in op deze onderwerpen en geeft een synthese van het in dit proefschrift uitgevoerde werk. Preciezer nog beoogt deze sectie een beknopte evolutionaire geschiedenis van varkensachtigen te verstrekken maar ook een reflectie op complexe evolutionaire processen zoals soortvorming, domesticatie alsmede het effect van deze processen op het genoom. Tenslotte wordt verder geprobeerd om de vele methodologische aspecten te bespreken die in dit proefschrift ontwikkeld of getest zijn.

Acknowledgments

I am grateful for the help and support of many people that have contributed directly or indirectly to the completion of my PhD. I would first like to thank Martien for his scientific vision that allowed him to secure the funds to finance this work. Without your great scientific vision none of this work would have been possible. I would also like to thank you for the freedom you have given me over these four years. Your style of supervision, your enthusiasm and your expertise in many areas allowed me to develop my scientific profile to such an extent that would have been unimaginable to me four years ago. Thank you for allowing me to work on my interests, for your support during failures and congratulations for my successes. I am sure this is not the end of our scientific collaboration; we have many more scientific accomplishments to come.

Thank you, Ole for your extensive support during these four years. Your daily supervision has been invaluable to me. Thank you for your knowledge, listening skills and for being so patient. Thanks also to your Danish roots that allowed you to follow me during all these heavy drinking sessions at various conferences.

Thank you, Hendrik-Jan for your invaluable contributions on all my scientific achievements. Your knowledge on the subject has no equivalent and I am grateful for your daily supervision, your honesty and your constructive help during these four years. Thank you for you for being so dedicated, your help was really invaluable.

Thanks to Greger for all these very fruitful collaborations. Thank you for your help at every stage of this thesis, for all these amazing discussions on domestication and speciation and for offering me this new job in Oxford. Thank you for introducing me to all these new people and for your trust and faith in my work.

Yogesh, it is difficult to thank you enough in a few sentences. Thank your support and your friendship during these years. Our friendship and collaboration has been very productive for both of us. I am sad you decided to sell out your soul to a pharmaceutical company so easily; however there is always cow urine to redeem your Karma!

Mirte, it was great to work with such a smart and lively person. Thank you for your support and for the great work we have done together. Merci pour ta joie de vivre et ton éternel enthousiasme pour la biologie. Merci aussi pour avoir supporté mon manque de manières et ma mauvaise humeur pendant ces quatre ans.

Richard and Bert, it is needless to say that all your hard work in the lab have made all of this possible. Thank you for your amazing organization and your management of the lab.

Konrad, thank you for your these awesome collaborations during these four years and for all our future collaborations. I am sure we can figure out more about cave men sex and pig speciation.

I would also like to thanks Johsua for his invaluable help and support during this work. Most of these projects would not have been possible without your help. Your passion for biology, statistics, academia and death metal has been invaluable.

Gus, these few sentences are never going to be enough to say how grateful I am for your help and support during these four years. Life in the Netherlands would not have been the same without you. Thank you for your friendship, all the discussions and beers, but also for your help with my work. It always amazed me how much you could comment and understand my work having no background in evolutionary biology! Also, thank you for marrying an evolutionary biologist (and thank you Rea for saying yes!).

Thanks to many people at ABGC, my colleagues and friends Mahmoud, Mathieu, Andre, Gabriel, Juan and others for making Wageningen a fun place to work. Thanks to the staff and the secretaries for all your various day to day help.

Guiz, pas facile de te remercier en quelque mot. En tous cas, merci d'avoir passé toute ces années au Pays-Bas avec moi, pour ton soutien, pour m'avoir supporté au jour le jour et pour me supporter encore. Merci à ton Fréro Reno et les autres Montpelliérain / Nîmois (Alex, Anna, Eloi, Léa, Louis, Julian, Clem, Gui, Tib, Audrey etc.) pour vos encouragements pendant toutes ces années loin de France et pour vos nombreuses visites aux Pays-Bas.

Davide, thanks for all these years in Wageningen that made my life so much better. Thanks for all these awesome discussions, all these beer drinking sessions and this awesome friendship.

Papa et Sylvianne, merci pour votre infaillible soutien pendant toutes ces années loin de la maison, même durant la maladie. Maman et Michel, merci aussi pour vos encouragements pendant toutes ces années (bien avant le doctorat) et pour votre soutien à la fois financier et morale.

Curriculum Vitae

About the author

Laurent Frantz was born on December 3rd 1985 in Nîmes, France. In 2004 he obtained his baccalauréat in Science in the lycée Dhuoda in Nîmes. In 2005 Laurent studied Life and Earth Science in the Université de Montpellier II to pursue his interests in paleontology and evolutionary biology. In 2008 he joined the Zoology course of the University of Manchester to complete the 3rd year of his bachelor as an Erasmus student. He obtained his bachelor in 2009 with thesis entitled "Phylogeography of the widely-spread tree frog, *Racophorus bipunctatus*" under the supervision of Dr. Catherine Walton (University of Manchester). Between 2009 and 2010 Laurent undertook the Master course in Evolutionary Genetics and Genomics at the University of Manchester for which he was awarded a BBSRC training award. During the course of his master's degree he investigated the effect of plants, herbivores and parasites interactions using quantitative genetics and wrote a first MSc thesis entitled "Genotypic Diversity in Hosts-Plants Mediates the Indirect Ecological Effects between the Hemi-Parasitic Plant *Rhinanthus minor* and the Insect Herbivore *Sitobion avenae*" under the supervision of Dr. Richard Preziozi. He also learned bioinformatics and genomics under the supervision of Dr. Casey Bergman with whom he investigated the effect of natural selection on regulatory elements in the genome of *Drosophila melanogaster* and wrote a second MSc thesis entitled "Evolution of *cis*-regulatory modules in *Drosophila melanogaster*". In 2010 he obtained his MSc with distinction from The University of Manchester. In November 2010 he joined Prof. Martien Groenen's group at the Animal Breeding and Genomics centre in Wageningen where he investigated the process of speciation and domestication in the genome of suids (pig and related species) and wrote a PhD thesis entitled "Speciation and Domestication in *Suiformes*: a genomic perspective". In December 2014 he moved to the University of Oxford to undertake a postdoctoral research assistant position in Dr. Greger Larson's group with the aim to carry on with his work on domestication and speciation using ancient DNA.

Peer reviewed publications

- Groenen M.A.M., Archibald A.L., Uenishi H., Tuggle C.K., Takeuchi Y., Rothschild M.F.,Rogel-Gaillard C., Park C., Milan D., Megens H.J., Li S., Larkin D.M., Kim H., **Frantz L.A.F.** *et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491(7424):393-398.
- Tortereau F., Servin B., **Frantz L.A.F**, Megens H.J., Milan D., Rohrer G., Wiedmann R., Beever J., Archibald A.L., Schook L.B., Groenen M.A.M. (2012) Sex specific differences in recombination rate in the pig are correlated with GC content. *BMC Genomics* 13:586.
- Bosse M., Megens H.J., Madsen O., Paudel Y., **Frantz L.A.F**, Schook L.B., Crooijmans R.P.M.A and Groenen M.A.M. (2012) Regions of Homozygosity in the Porcine genome: interactions between demography and the recombination landscape. *PLoS Genetics* 8(11):e1003100.
- Ottoni C., Flink L.G., Evin A., Geörg C., Cupere B., Neer W., Bartosiewicz L., Linderholm A.,Barnett R., Peters J., Decorte R., Waelkens M., Vanderheyden N, Ricaut F., Çakırlar C., Çevik O, Hoelzel R., Mashkour M., Karimlu A.F.M., Seno S.S., Daujat J., Brock F., Pinhasi P, Hongo H., Perez-Enciso M., Rasmussen M., **Frantz L.A.F.**, *et al.* (2012) Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics. *Molecular Biology and Evolution* 30(4):824-832.
- Paudel Y., Megens H.J., Madsen O., **Frantz L.A.F**, Bosse M., Bastiaansen J.W.M., Crooijmans R.P.M.A and Groenen M.A.M. (2013) Evolutionary dynamic of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* 14:449.
- **Frantz L.A.F**, Schraiber J., Madsen O., Megens H-J., Bosse M., Yogesh P., Semiadi G., Meijaard E., Li N., Crooijmans R.P.M.A, Archibald A.L., Slatkin M., Schook L.B. Larson G. and Groenen M.A.M. (2013) Genome sequencing reveals fine scale diversification and reticulation history in Sus. *Genome Biology* 14(9):R107.
- Lohse K. and **Frantz L.A.F.** (2014) Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes. *Genetics* 196(4):1241-1251.
- Bosse M., Megens H-J., Madsen O, **Frantz L.A.F.**, Paudel Y., Crooijmans R.P.M.A and Groenen M.A.M. (2014) Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. *Molecular Ecology*, 23(16), 4089-4102.
- Zytynska S.E., **Frantz L.A.F.**, Hurst B., Johnson A., Preziosi R.F. and Rowntree J.K. (2014) Hostplant genotypic diversity and community genetic interactions mediate aphid spatial distribution. *Ecology and evolution* 4(2):121-131.

- Rowntree J.K., Zytynska S.E., **Frantz L.A.F.**, Hurst B., Johnson A., Preziosi RF. (2014) The genetics of indirect ecological effects - plant parasites and aphid herbivores. *Frontiers in genetics*, 5:72.
- Bosse M., Megens HJ., **Frantz L.A.F**, Madsen O., Larson G., Paudel Y., Duijvesteijn N., Harlizius B., Hagemeijer Y., Crooijmans R.P.M.A and Groenen M.A.M. (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature Communications*, 5:4392.
- **Frantz L.A.F,** Megens HJ., Madsen O., Groenen M.A.M. and 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*, 23 (22), 5566-5574.
- **Frantz L.A.F.**, Madsen O., Megens HJ., Bosse M., Schraiber J.G., Paudel Y., Crooijmans P.M.A and Groenen M.A.M. (2014) The evolution of Tibetan wild boars. *Nature Genetics*, in press.
- **Frantz L.A.F.**, Madsen O., Megens HJ., Paudel Y., Bosse M., Crooijmans R.P.M.A and Groenen M.A.M. (2014) Next-generation phylogenomics: A case study in Suiformes. *Molecular Biology and Evolution*, submitted.
- **Frantz L.A.F.**, Schraiber J., Madsen O., Megens HJ., Paudel Y., Bosse M., Cagan A., Crooijmans P.M.A, Larson G. and Groenen M.A.M. (2014) Analyses of Eurasian wild and domestic pig genomes reveals long term gene-flow and selection during domestication. *Nature Genetics*, under review. Available online from *BioRxiv* (http://dx.doi.org/10.1101/010959).
- Bosse M., Madsen O., Megens H-J, **Frantz L.A.F.**, Paudel Y., Crooijmans P.M.A and Groenen M.A.M. (2014) Hybrid origin of European commercial pigs examined by an in-depth haplotype analysis on chromosome 1. *Frontiers in genetics*, 5:442.

Training and supervision plan

The Basic Package (3 credits)	year
WIAS Introduction Course	2010
Ethics and Philosophy of Animal Science	2011

Scientific Exposure (21 credits)

International conferences

ESEB 2011	2011
SMBE 2012	2012
8th Biennial Conference of the Systematics Association	2011
SMBE 2013	2013
Evolution 2012	2012
PAG 2014	2014
SAGE 2013	2013

Seminars and workshops

MAGE 2011	2011
Institute of evolutionary Biology (IEB) seminar, Edinburgh	2012

Presentations

Otto Warburg International Summer School and Research Symposium 2011(Poster)	2011
ESEB 2011 (Poster)	2011
WIAS science day 2014 (oral)	2012
SMBE 2012 (Poster)	2012
Evolution 2012 (Poster)	2012
PAG 2014 (Invited Speaker)	2014
SAGE 2013 (Oral)	2013
IEB Seminar, Edinburgh (Invited Speaker)	2012
SMBE 2013 (Oral)	2013
Evolutionary genomics seminar, Manchester (Oral)	2011
SMBE 2014 (Oral)	2014

In-Depth Studies (9 credits)

Disciplinary and interdisciplinary courses

MPI Otto Warburg International Summer School and Research Symposium on Evolutionary Genomics (14th -22nd September) – Berlin Germany	2011
EMBO Next-generation sequencing course (6-13 August) – Reykjavik Iceland	2011
Workshop on Molecular Evolution (27 July – 6 August) -Woods Hole, USA	2012

Professional Skills Support Courses (3 credits)

Scientific Publishing	2014
Data management	2014
Writting grant proposal	2015

Research Training Skills (6 credits)

Preparing own PhD proposal	2014

Didactic Skills Training (12 credits)

Supervising practicals and excursions

Genomics WUR	2011
Genomics WUR	2012
Genomics WUR	2013

Supervising theses

Animal Breeding and Genomics Msc student	2011
Animal Breeding and Genomics Msc student	2012
Animal Breeding and Genomics Bsc student	2013

Total (54 credits)

This research was financed by the European Research council (ERC).

Artwork on cover by Laurent Frantz.

Printed by GVO drukkers en vorngevers B.V./Ponsen & Looijen, Ede, The Netherlands.