Genome evolution of Suidae : a looking glass of speciation and hybridization

Langqing Liu

刘琅青
Langqing Liu

Genome evolution of *Suidae* :

a looking glass of speciation and hybridization

Propositions

1. Admixture is a driver in rapid range expansion and massive species replacement.
(this thesis)

2. Highly contiguous genome sequences are essential for a comprehensive understanding of evolution.
(this thesis)

3. The persistent model-species-centric bias in genomic studies means that advances in genomics research miss the forest for the trees.

4. Non-hypothesis driven research is the primary force for science.

5. Without comparison, culture uniqueness is solely an unconfirmed delusion.

6. The freedom of speech does not guarantee the authenticity of statements.

Propositions belonging to the thesis entitled:
“Genome evolution of Suidae: a looking glass of speciation and hybridization”

Langqing Liu
Wageningen, 17 February 2021

Genome evolution of *Suidae*: a looking glass of speciation and hybridization

Langqing Liu

Thesis committee

Thesis supervisor
Prof. Dr M.A.M. Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Thesis co-supervisors
Dr O. Madsen
Assistant Professor, Animal Breeding and Genomics
Wageningen University & Research

Other members
Prof. Dr A.L. Archibald, The University of Edinburgh, UK
Prof. Dr D. de Ridder, Wageningen University & Research
Prof. Dr M.E. Schranz, Wageningen University & Research
Dr K. Vrieling, Leiden University

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

Genome evolution of *Suidae*: a looking glass of speciation and hybridization

Langqing Liu

Thesis

submitted in fulfilment of the requirements for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday February 17 2021
at 4.00 p.m. in the Aula.

Langqing Liu
Genome evolution of *Suidae*: a looking glass of speciation and hybridization

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021)
With references, with summaries in English and Chinese

DOI: https://doi.org/10.18174/538553
ISBN: 978-94-6395-673-4

Abstract

Liu, L. (2021). Genome evolution of *Suidae*: a looking glass of speciation and hybridization. PhD thesis, Wageningen University, the Netherlands

Understanding the origin of species and biodiversity is one of the fundamental objectives in evolutionary biology. Developments in sequencing technology have revolutionized our understanding of evolutionary biology. In this thesis, with genomics information, I systematically describe the evolutionary history of the *Suidae* species and examine complex speciation scenarios with hybridization. In Chapter2, utilizing whole-genome sequence data, I show that pygmy hog should be classified into a distinct genus separated from other *Suinae* species, which initially emerged during the Miocene/Pliocene boundary. In this chapter, admixture analyses reveal at least two independent events of inter-species gene flow during wild boar range expansion across Eurasia. In Chapter3, I provide an in-depth analysis of the formation of the current pygmy hog population and demonstrated consequences of the historical demography. Our demographic analysis reveals that pygmy hog has remained at small population sizes with low genetic diversity since ~1 Mya. The long-term, extremely small population size may have constrained the purifying effect and led to the accumulation of genetic load. In Chapter 4, I present a chromosome level genome assembly of Visayan warty pig. The alignment of the Visayan warty pig assembly and Duroc pig assembly reveal a high degree of collinearity, but also chromosome fission and fusion. The similarity of the chromosome interaction maps may explain the absence of post-zygotic reproductive isolation among *Sus* species. Moreover, we investigated the evolution of olfactory and gustatory genes and report the genetic basis of species-specific sensation. In Chapter5, I describe a reference-guided assembly approach to generate genome sequences for three other *Sus* species and the outgroup species pygmy hog. With the near complete phylogenomic framework of *Sus* species, we were able to perform admixture analyses directly from genome sequences. We provide candidate genes that might have contributed to adaptive radiation and domestication of pigs. Together, thesis findings provide a comprehensive view of the evolutionary history of *Suidae*, describing species origin and demographic history afterwards. In this work I compared genome scale data from various pig species to refine the understanding of their evolutionary processes and provide new insights on the underlying genetic mechanism.

Contents

9 1. General introduction

31 2. Genomic analysis on pygmy hog (*Porcula salvania*) reveals multiple interbreeding during wild boar expansion

67 3. Genetic consequences of long-term small effective population size in the critically endangered pygmy hog

93 4. Genome assembly of Visayan warty pig (*Sus cebifrons*) provides insight into genome evolution of Sus during speciation

139 5. *Sus* reference-guided genome assemblies provide a high-resolution view of diversification, reticulation and adaptation during pig evolution

163 6. General discussion

183 Summary

187 概要

189 Acknowledgements

193 Curriculum Vitae

1

General introduction

1.1 Introduction

Introduction to genome evolution

Planet Earth is populated by an incredible variety of lifeforms, ranging from simple viruses and single celled bacteria to complex organisms like vertebrates and higher plants. The evolutionary trajectory of each organism is encoded within its genome. Understanding the genetic basis of evolutionary novelty is central to recognizing fundamental processes that drive biological diversity. Basic mechanisms of genome evolution can be broken down into two components: DNA mutation and selection. The DNA code is the core of the genome, and serves as a template for other components in the genome (Crick, 1970). Genomes evolve over time through accumulation of mutations in DNA, ranging from those at single base level like single nucleotide polymorphism (SNP) to large-scale rearrangements like inversions, duplications, and translocations (Figure 1.1).

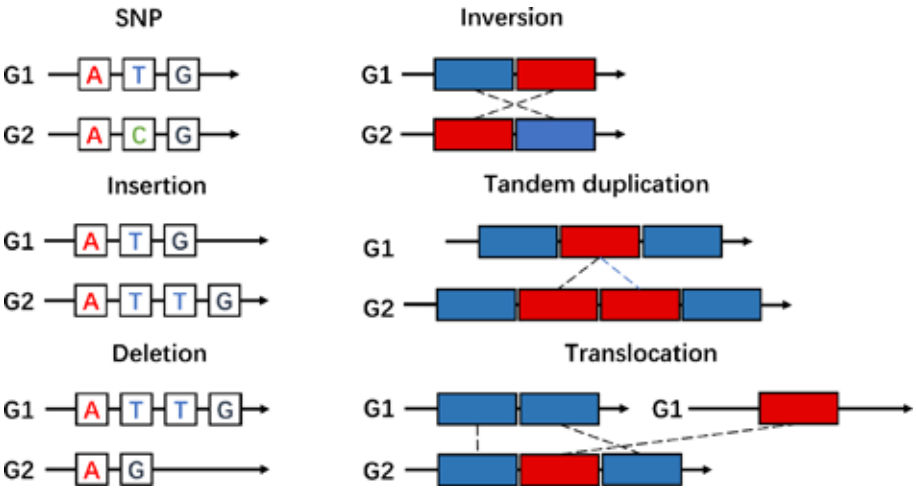


Figure 1.1 The most common types of genetic variants.

Mutations can emerge when the DNA sequence undergoes replication, is exposed to mutagenetic compounds, or even when an earlier error is being repaired (Kogoma, 1997; Wilson, 2014). Besides, recombination is another important molecular mechanism that promotes genome evolution. By exchanging sister chromatids during meiosis, recombination can introduce mutations from distinct populations or even different species. Not all mutations contribute to evolution. Only heritable mutations, which occur in germ line, can be passed to the next generation and lead to differentiation between parents and offspring.

Based on its effect on the organism, a mutation can be selected, counter-selected or unselected (a mutation is neutral and therefore there is no selection pressure). A mutation is selected if, for instance, it enhances the survival and reproductive ability of the individual in a particular environment. This could be through a substitution resulting in a better choice of amino acid, elevating the efficiency of protein product, or an increased production of a protein due to e.g., a duplication event. However, advantageous mutations are very rare. More commonly, mutations are counter-selected or neutral (Kimura, 2020). A counter-selected mutation may result in a lower viability, lower reproduction rate or death of the organism, by altering the function of the protein, or in the extreme case, damaging the gene product. Counter-selected mutations are likely to be rapidly removed from the population, because their phenotypic disadvantage prevents them from being passed on to the next generation. The vast majority of the remaining mutations in the genome are neutral, neither beneficial nor harmful enough to influence fitness. Although neutral mutations are usually not subjected to natural selection, they may reach fixation by random genetic drift. Moreover, neutral mutations can serve as a genetic repository, providing evolutionary potential to adapt to future environmental changes. Collectively, spontaneous mutations create differences in levels of variation across the genomic landscape. In the meantime, they can lead to both increases or decreases of genomic complexity. Directional selection or random drift on mutations has led to changes in the frequency of phenotypes, which gives rise to the diversity we see today.

The genome is made of alternately distributed coding regions and non-coding regions. Because protein coding genes can be directly linked to molecular phenotypes, until now, coding regions have been studied the most. Works from different ENCODE projects suggest that only ~2% of the mammalian genome codes for protein product. Nevertheless, ~80% of a genome, while not encoding a protein, also underlies evolutionary forces affecting the genome (Dunham et al., 2012). Non-coding regions are likely to have, often still unidentified, functional activities. Several lines of evidence indicate that non-coding regions are involved in pervasive transcription and epigenetic regulation throughout the genome (Carninci et al., 2005; Birney et al., 2007; Beagrie et al., 2017). Functional non-coding region can be a regulatory element which affect the expression of protein coding genes and RNA genes. It can also be a sequence involved in chromatin and chromosome structure (Pennacchio et al., 2006; Barrett et al., 2012). Phenotypic changes between both individuals and species have been shown to associate with alterations of non-coding sequences (Mattick, 2001). Moreover, the ratio of non-coding region to total genome size in complex multicellular organisms is usually higher than unicellular

organisms (Benjamini and Hochberg, 1995). This indicates that the complexity of higher organisms, which correlates with an increase in the proportion of non-coding regions of their genomes, arises from an increase in the number and complexity of regulatory pathways (Levine and Tjian, 2003).

In recent years, a genomic revolution has taken place, which has been strongly stimulated by the developments in sequencing technology. This has provided unprecedented opportunities to unravel the complexities of the story of evolution. The increased feasibility of whole genome sequencing has allowed inference of the evolutionary history and identification of genetic loci associated with particular traits in any species at a very high resolution. The aim of my thesis is to utilize emerging sequencing technology and whole-genome scale data to provide new insights into the evolutionary process and mechanisms in *Suidae* evolution.

Glossary

DNA A linear chain of linked nucleotides, consisting of adenosine (A), thymine (T), guanine (G) and cytosine (C).

Genome Complete set of genetic instructions for an organism, including both the genes and the non-coding DNA. A genome can either consist of DNA or RNA (RNA virus).

Chromosome Thread-like DNA molecule and the histone proteins that support its structure.

Gene The basic physical and functional unit of heredity. The structure of a gene consists the sequence of DNA encoding a protein or RNA that performs a function, as well as regulatory sequence that is requited for its expression.

Recombination An event that occurs during meiosis, involving a crossing-over of homologous chromosomes. It can result in gene conversion, and it breaks and shuffles parental alleles to produce new combinations of alleles in the offspring.

Interbreeding Sexual reproduction between two distinct populations or species, thereby producing hybrid offspring.

Effective population size (Ne) Idealized population size that would be expected to experience the same rate of genetic diversity loss (due to genetic drift) as the population under study

Runs of homozygosity Two contiguous stretches of homologous DNA segments, whose parental haplotype are identical by descent.

1.2 Speciation genomics

1.2.1 Species diversification

A primary objective in evolutionary biology is to understand the origin of biodiversity via the process of species formation. Speciation is a process in which species originates or multiplicates by subdivision (Cook, 1906). Based on the classical definition, speciation could be initiated from the emergence of a barrier to dispersal, such as plate tectonics, human intervention, or mismatching mating seasons, which divides individuals into subpopulations and the interbreeding across the subpopulations become restricted (allopatric/peripatric, Figure1.2a). The speciation process finishes when postzygotic isolation (i.e., all hybrid offspring are sterile) is established between subpopulations. This spatial or temporal separation was previously thought to be a precondition for speciation and needed to persist during the whole speciation process.

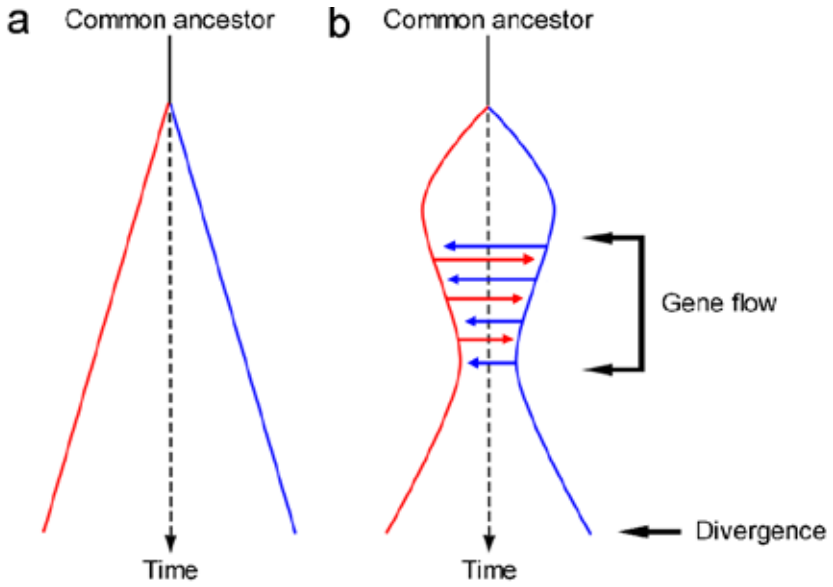


Figure 1.2 Illustration of two modes of speciation. a) Speciation without gene flow. Allopatric/peripatric speciation. Starting from a single ancestor population, which moves forward in time (from top to bottom of figure). An isolating event causes the population to split into two separate lineages (red and blue). **b) Speciation with gene flow.** Sympatric/parapatric speciation. After the first isolating event, the two diverging descendant populations have a secondary contact. This facilitates gene flow between populations and reduces their accumulated divergence by exchanging unique alleles from red to blue and vice versa. A second isolating event separates them once more and divergence can progress again.

However, the process of speciation can span a long evolutionary time scale. The beginning and end of this process can be considered as opposite ends of a continuous trajectory. The early stages of speciation can be marked by the time when interbreeding between two subpopulations starts to become restricted. Postzygotic isolation may not be established at this phase. Recent studies have shown that speciation can take place spontaneously in a shared or overlapping niche (sympatric/parapatric). The genetic compositions of two subpopulations can shift toward each other due to selection, genetic drift, and most sharply, gene flow. When gene flow accompanies speciation, a series of back and forth exchanges take place between divergent selection and interbreeding, in which the former builds up differentiation and the latter breaks down the difference and homogenizes subpopulations (Figure1.2b). These processes have influenced divergence and reproductive isolation in a dynamic context of geography and ecology. Over time, due to animals' preference of mating with members from their own subpopulations, new genetic variations will accumulate independently in each of the two subpopulations. Finally, when the two subpopulations have diverged sufficiently, speciation results in the formation of postzygotic isolation. The genomes of the two subpopulations have become reproductively incompatible, such that their hybrids are unable to develop into a fertile adult.

1.2.2 Tree of life and speciation genomics

In his enlightening book *On the Origin of Species*, Charles Darwin wrote in the first paragraph that he hopes to "throw some light on… that mystery of mysteries" (Darwin, 1859). Reconstructing the timing and pattern of phylogenetic relationships in the tree of life illuminates the mysterious processes of biodiversity origination. In a phylogenetic tree, organismal or molecular lineages coalesce back to a universal ancestor whose descendants are found in current or now extinct lineages. Using tree-like schemes for illustrating the evolutionary processes of speciation can be traced back as early as 1801 (Augier, 1801). After that, Darwin introduced the prototype of the modern phylogenetic tree, in which evolutionary lineages split and diverge from each other. The genomic revolution was brought about by stunning technical advancements since the beginning of 21st century. It has rapidly shifted the field of speciation studies towards the investigation of genome-wide patterns of diversity and differentiation in many species pairs, which are referred to as "speciation genomics". This added another dimension of complexity but also opportunity to understand how species develop. Given the massive observations of gene flow events in whole genome sequencing projects, there is growing appreciation that the tree of life may oversimplify the evolutionary history

(Patterson et al., 2012; Bridgett et al., 2016; Fan et al., 2016; Sankararaman et al., 2016). The reticular relationship necessitates modifying the model of the "tree of life" into, more precisely, the "web of life". Gene flow will cause different molecular lineages not to coalesce to the organismal common ancestors, but coalesce to different organismal lineages and to different times. Thus, it is imperative to not only reinterpret the concept of species, but to also test the complex speciation model under a reticulation framework.

Besides resolving the taxonomic question, speciation genomics also aims at revealing how genome evolution relates to speciation processes and the forces that govern it. For example, the variable effects of selection, random drift, as well as genome architecture and interactions between loci can all potentially reduce global and localized gene flow and stimulate species differentiation (Nosil and Feder, 2013; Seehausen et al., 2014). Regions of high inter-specific divergence or "islands of speciation" may result in hybrid incompatibilities, which can be resistant to gene flow if populations come into contact (Butlin, 2010; Ellegren, 2010; Harrison and Larson, 2014). In addition, differentiation of genome architecture, in the form of structural rearrangements and varied recombination rates, further facilitates divergence across the genome (Bush et al., 1977; Burri, 2017). As we study the evolution of the genome and understand the variations involved in it, we can start to understand the trajectory and duration of speciation, uncovering the potential patterns of demographic history and biogeographic context.

1.3 Comparative genomics

1.3.1 Principles of comparative genomics

Comparative genomics is a field of biological research in which the genomic features of different organisms are compared. Genomic features may include the genes, gene order, regulatory sequences, and other genomic structural landmarks. In this research branch of genomics, complete or large parts of genomes resulting from genome sequencing projects are compared to study basic biological similarities and differences in consideration of evolutionary relationships between organisms. The major principles of comparative genomics are straightforward. Common features of two organisms' genomes are often considered to be derived from their common ancestor, whereas different features reflect the distinct histories of each genome since they last shared a common ancestor. Based on incorporation of Darwinian theory and molecular coalescent theory, this rationale can be applied to any two or more species in a comparative analysis, because all cellular organisms on Earth can be connected to a single phylogenetic tree (i.e., all share the same common ancestry)

(Theobald, 2010). Comparative genomics becomes a fundamental tool of genome analysis that aims to (1) characterize the similarity of lineages, (2) understand the evolutionary forces, such as mutation and selection, that govern the changes of these genomic features, and (3) find out how genomic evolution leads to the divergence of phenotypes.

1.3.2 Translation between genotype and phenotype

After the first complete genomes of two organisms (the bacteria *Haemophilus influenzae* RD and *Mycoplasma genitalium* G37) were published in 1995 (Fleischmann et al., 1995; Fraser et al., 1995), biologists are able to compare whole genomes. Thus, the field of comparative genomics started to become booming. With the development of laboratory and analytical methods, large amounts of sequencing data can be produced efficiently. Up to now, genomes of nearly 100,000 species have been sequenced, of which over 20,000 are complete (Augustus 2020, https://gold.jgi.doe.gov). Comparing genomes of different evolutionary distances can address different questions. Highly similar sequences found in different species, in other words, evolutionarily conserved sequences, are expected to have critical functional roles (Siepel et al., 2005). General insights about classification and amount of shared genes can be obtained by genomic comparisons at very large phylogenetic distances, e.g., greater than 1 billion years since speciation. Over such large distances, the order of genes and the sequences regulating their expression are mostly not conserved. The evolutionary history of shared genes can be inferred by differences in the rates of synonymous and non-synonymous changes at the level of base-pairs within coding regions (Yang, 2007). At moderate phylogenetic distances (i.e., ~100 million years divergence, as between human and pig), both functional and nonfunctional DNA can be found within the conserved region. In these cases, functional sequences will show a signature of stronger purifying selection, with fewer changes than nonfunctional DNA (Jukes and Kimura, 1984). Furthermore, in genome comparisons within a species or between closely related species (i.e., wild boar and domesticated pig, or red jungle fowl and chicken), the different genomic features may be of interest for revealing the recent species-specific evolutionary trajectory, such as that caused by domestication.

Not only does comparative genomics aim to discriminate conserved from divergent and functional from nonfunctional DNA, this approach also contributes to identifying the general functional classes of certain DNA segments. Protein coding regions of genes are conserved across species in a highly specific pattern (3-bp code with a third degenerate base), and sequence conservation has proven to be an important feature for gene model prediction (Flicek et al., 2013). Due to the high variety in sequence

level and short length, identification of non-coding regulatory elements with traditional experimental approaches has generally been slow and laborious. Recent progress in comparative genomics has greatly improved the situation. Comparison of orthologous genomic sequences from different species identifies conserved non-coding elements (CNSs) as reliable candidates of regulatory element (Pennacchio et al., 2006). For example, CNSs identified by comparison of distant species such as human-zebra fish are highly enriched in functional enhancers (Clément et al., 2020). Subsequently, phylogenetic footprinting analyses searching for changes in sequence conservation upstream or downstream from a gene predicts promoter, e.g., transcription factor-binding sites. Furthermore, researchers have mapped conserved non-coding regions with histone modifications, chromatin structures, and transcription factor associations, which inferred the actual function of the regulatory elements (Feingold et al., 2004).

One fundamental question in biology is to understand what makes individuals, populations, and species different from each other. With the identified functional elements, approaches like genome-wide association study (GWAS), which use sequence variations in the whole genome together with the phenotype and pedigree information to perform association analysis, have identified numerous genes or regulatory elements that are important for the traits of interest (Stranger et al., 2011; Flint and Eskin, 2012; Buniello et al., 2019; Mills and Rahal, 2019). For a long time, genome research focused on inbred strains of model organisms with detailed phenotypic records. As genomic sequence data accumulates, a large collection of genomic resources comes from non-model organisms, whose phenotypes are often poorly recorded. Comparative genomics can transfer knowledge from one organism to that of another. It finds associations from known genes and variants in model organisms to unknown genes in non-model organisms and thereby infers functions for the unknown, thus offering opportunity to predict phenotypes using the genotypes of non-model organisms. This link between genomics and physiology contributes to, for example, advancing the discovery of gene function in natural contexts, and discovering mechanisms whereby organisms integrate cues from, respond to, and are ultimately shaped by the environment.

1.4 Population dynamic and conservation

1.4.1 Inferring demography history from genomic data

Lynch (2007) famously argued that “Nothing in evolutionary biology makes sense except in the light of population genetics”. While in the previous paragraphs I mainly described the evolutionary process on individual levels, there is also a long-standing

and ongoing interest in how genomes evolve on a population scale. The idea that molecular data contains information on the evolutionary history of populations traces back to the beginning of the twentieth century (Hirschfeld and Hirschfeld, 1919). In recent years, population genetic analyses, taking better advantage of genome-scale data, have received considerable attention. Quantifying changes in effective population size (Ne), admixture, inbreeding and outbreeding depression, and the genomic basis of fitness provides vital information for understanding organismal biology and informing resource conservation and management. Two metrics have been widely used in the field of population genetics, namely linkage disequilibrium (LD) and site frequency spectrum (SFS). With recombination rate and number of generations, the extent of LD in a population can be used to estimate the contemporary effective population size (Hill, 1981). A more detailed reconstruction of demographic history can be performed by using site frequency spectrum (SFS) (Nielsen, 2000; Gutenkunst et al., 2009). Different demographic scenarios change the shape (topology or branch length) of the underlying genealogies, which consequently change the SFS, and based on that, demographic history can be inferred. However, the use of LD and SFS requires large-scale of genomic data. This can result in many problems in practice, especially for studies on critically endangered species or multiple populations. Although a sufficient number of samples is not always available, a single genome can also contain information about the demography of the population. A recent popular demographic inference method is based on sequentially Markovian coalescence, e.g., PSMC, which uses a pair of diploid sequences from one individual to infer piecewise-constant population size histories (Li and Durbin, 2011). The key feature of PSMC is estimation of the time to the most recent common ancestor of two alleles at a given locus. Because the rate of coalescent events is inversely proportional to Ne, demographic history can be estimated. Even more exciting is the use of more informative statistical estimators such as runs of homozygosity (ROH), which measures inbreeding and effective population size changes from a single genome (Kirin et al., 2010; Kardos et al., 2015). Different demographic histories result in different ROH size (Kardos et al., 2017). Long ROH suggests a recent event of inbreeding whereas short segments indicate ancient inbreeding. From a molecular perspective, the bioinformatic tools, e.g., SIFT (Flanagan et al., 2010) and PROVEAN (Choi et al., 2012), can make predictions of the deleteriousness of the variations in protein coding genes. The recently developed CADD ranking systems can even evaluate the altering non-coding regulatory elements (Rentzsch et al., 2019; Groß et al., 2020a, 2020b). This will also increase our ability to understand the genomic architecture of inbreeding depression and to predict and compare populations for genetic load.

1.4.2 Population management

The list of endangered species has more than doubled in the past two decades, according to the International Union for Conservation of Nature (IUCN, https://www.iucn.org/). The genetic refugium of many endangered species is contained in a small number of individuals, typically with unknown kinship and uncertain demographic backgrounds. Small populations tend to lose genetic diversity rapidly, due to random drift and inbreeding, and finally fall into a so-called "extinction vortex", where the genetic variation is too small to sustain the population (Frankham et al., 2017; Laikre et al., 2020). Thus, managing population growth and maintenance of genetic variation are central features of captive management, such as that undertaken in zoos and free-ranging *ex-situ* populations of endangered species, for which recovery is actively being attempted. Genomic studies analyzing intra-specific and inter-specific differences can serve as effective and significant conservation tools. For example, in managed populations, genomic analysis can assist in estimating the relatedness of the wild founders that make up captive populations. Founders of captive populations are typically assumed to be unrelated to one another, which may be unrealistic for most species. Another immediate application of genomics for conservation management is to boost population fitness. Predicting the deleterious effects of the variants across the genome allows the estimation of the genetic load that are carried by a population. Combining this information with patterns of inbreeding across the genome enabled detection of candidate loci underlying inbreeding depression (Kardos et al., 2016). This finding underlines the fact that genomics analyses can aid conservation programs with optimal breeding strategies, where potential breeding candidates with high deleterious/mutation load can be excluded.

1.5 Evolution history of *Suidae* species

1.5.1 Brief introduction of the *Suidae* family

Pigs and hogs belong to the *Suidae* family, which comprises six extant genera, divided further by region: *Babyrousa* (deer hog) from Island Southeast Asia (ISEA), *Potamochoerus* (bush pig and river hog), *Phacochoerus* (warthog) and *Hylochoerus* (forest hog) from sub-Saharan Africa, *Porcula* (pygmy hog) from India, *Sus* (wild pig species and domestic pig) from Eurasia and ISEA (Figure 1.3). Previously, the taxonomic relationships among *Suidae* were assessed using morphological and molecular studies. *Suidae* emerged around 40 - 34 million years ago (Mya) (Gongora et al., 2011). In the middle Miocene (7.9 - 6.7 Mya), the first lineage to split from those early *Suidae* was *Babyrousa*, forming an ancient lineage endemic to the island

of Sulawesi (Frantz et al., 2018). Along with the global cooling during the late Miocene (Zachos et al., 2001), a new subfamily, the *Suinae*, emerged in the fossil record (Roth and Wagner, 1854; Geraads et al., 2008). The *Suinae* later diversified into multiple tribes (Roth and Wagner, 1854; Geraads et al., 2008). This was followed by a divergence of the African *Suidae* and the Eurasian *Sus* genus, at around the Miocene/Pliocene boundary (4.0 -3.2 Mya). Shortly thereafter, the *Sus* genus started to diverge during the early Pliocene (~3.9 Mya). Several *Sus* species on ISEA evolved during the early/mid Pliocene (Frantz et al., 2013). During the early Pleistocene, wild boars expanded from South Asia into almost every ecosystem across the old world (Fistani, 1996). Archaeological evidence suggests three major episodes of species replacement in Eurasia that took place during the evolutionary history of *Suidae*, including (i) replacement of almost all other subfamilies of *Suidae* (except *Babyrousa*) by Suinae around the Miocene/Pliocene boundary (van der Made et al., 2006; Orliac et al., 2010), (ii) replacement of all non-*Sus* species by *Sus* in Eurasia during the Pliocene, and (iii) replacement of most *Sus* species by *Sus scrofa* during the Pleistocene (Fistani, 1996; Guérin and Faure, 1997). These three events shaped the layout of modern *Suidae* species (Figure 1.3). However, many questions remain about the genetic consequence of *Sus scrofa*'s range expansion and species replacement. Further discussion will be provided in Chapters 2 and 5 of this thesis.

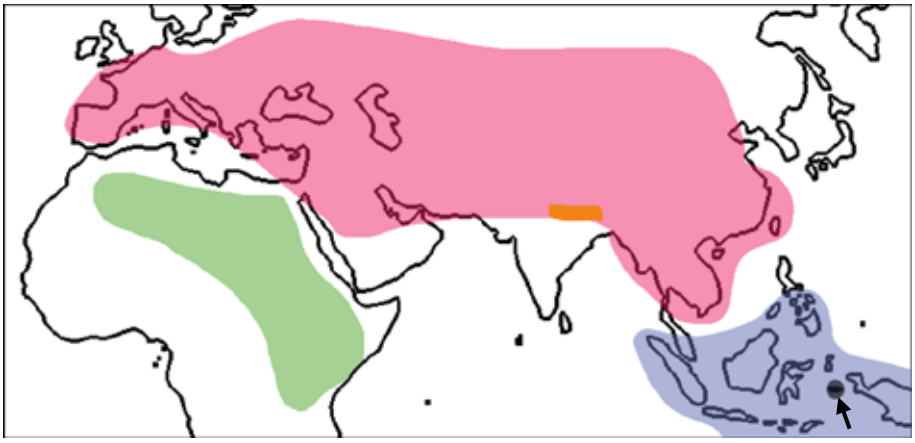


Figure 1.3 Geographic distribution of extant *Suidae* species. The area shaded black (black arrow) represents the distribution of *Babyrousa*. The green shaded area represents Sub-saharan suids. The orange shaded area represents *Procula*. The blue shaded area represents ISEA *Sus*. The red shaded area represents wild and domestic *Sus scrofa*.

1.5.2 Fast divergence of ISEA *Sus*

Sus is the largest genus of the *Suidae* family, comprising more than 7 species, most of which are restricted to ISEA. ISEA covers three (Sundaland, Wallacea and Philippines) of the 25 so-called global biodiversity hotspots, which are regions with an extremely high species diversity (Myers et al., 2000). Large-scale climatic fluctuations in the late Cenozoic induced fluctuations of sea levels in this region. The biogeography of this region was affected heavily when sea levels fluctuated with each of the Pleistocene glacial cycles. For instance, during the Last Glacial Maximum (LGM), the sea level dropped to approx. 118 m below the present level, enabling the Malay Peninsula, Java, Sumartra and Borneo to be connected by exposed seabeds, and a land area formed in Sundaland (Woodruff, 2010). In warmer periods, rising sea levels converted mountains into geographically isolated islands (Hall, 1998; Woodruff, 2010). These alternating climatic conditions required frequent adaptation and induced intermittent allopatric and parapatric speciation processes among the ISEA fauna and flora, resulting in a complex and species-rich assemblage (Whittaker and Fernandez-Palacios, 2007). The relatively recent divergence of species in the *Sus* genus appear to have been closely linked to sea level fluctuations (Frantz et al., 2013, 2014). The initial divergence of the Eurasian wild boar from a clade consisting of other *Sus* species took place during the beginning of the Pliocene (5.3-3.5 Mya, Zanclean stage). During the following millions of years (3.5-2.5 Mya, Piacenzian stage), consistent with fossil records, concomitant drops in sea levels likely allowed the dispersal of the ancestor of *Sus verrucosus* to Java (Aimi, 1989; Meijaard, 2004). Following the divergence of the *Sus verrucosus* lineage, the ancestor of *Sus cebifrons* colonized the Philippines, when tectonic activity led to the isolation of the Philippines from Sundaland during periods of low sea levels (2.4-1.6 Mya, Gelasian stage) (Barrier et al., 1991). During the latter stage of the Pleistocene (1.6-0.8 Mya, Calabrian stage), *Sus celebensis* colonized Sulawesi, coming from the west (Borneo). Although the Makassar Strait separating Sundaland and Sulawesi continued to exist during the Plio-Pleistocene, more frequent incidences of lower sea levels during this period would have reduced the distance between Sundaland and Sulawesi (Hall, 2002; Miller et al., 2005), thereby increasing the likelihood of a successful colonization. Sea level fluctuations also gave rise to open niches, leading to interspecific hybridization. Admixture analysis revealed that, because of the lack of a post-zygotic reproductive barrier in *Sus* species (Blouch and Groves, 1990), extensive inter-specific gene flow existed that likely took place during cold intervals (Frantz et al., 2013, 2014).

1.5.3 Domestication of Sus scrofa

Although sharing a common ancestor around 10 million years ago, today's *Suidae* species are all, like their extinct ancestor Entelodonts, gregarious, omnivorous, with short slender legs, bulky bodies and long snouts. What makes pigs (*Sus scrofa*) stand out from other *Suidae* is not only their global distribution (Meijaard et al., 2011), but also their irreplaceable role in human society. Pigs have been popular symbols dating back in many prehistoric cultures, often being a favorite animal for sacrifice as they implied richness (Sillar et al., 1961; Mizelle, 2012; Cucchi et al., 2016). Based on the excavations of early Neolithic sites found in the Near East and East Asia, it is clear that pigs had already been considered as a crucial source of food (Zeder, 1998; Zhang and Luo, 2008; Cucchi et al., 2011). It is widely accepted that pigs were domesticated independently in China and Anatolia ~10,000 years ago (Larson et al., 2005; Groenen et al., 2012). In the following thousands of years, extensive hybridization took place between wild and domesticated pigs, as well as between different populations of domesticated pigs (Ottoni et al., 2013; Bosse et al., 2014; Frantz et al., 2019). After the Industrial Revolution in the nineteenth century, pig production became industrialized. Breeding schemes, based on different production demands and cultural preferences, were made. Pigs were selected for various traits including litter size, carcass weight, back fat, and specific coat colors (Whittemore et al., 1998; McGlone and Pond, 2003). However, while pigs have been studied extensively as an important livestock and a model organism, a blanket of uncertainty descended on how they were domesticated initially and what was selected for, especially in terms of the underlying genetic mechanisms. Understanding the origin and distribution of variations in domesticated pigs is important for decoding the domestication process and conservation of genetic resources (Groeneveld et al., 2010).

1.5.4 Missing pieces in *Suidae* evolutionary history

Since the release of the *Sus scrofa* reference genome in 2012 (Groenen et al., 2012), the Animal Breeding and Genomics group at Wageningen University has generated genome sequences for nine Suidae species. Those genomic resources represent a wide range of collection of this family, but the pygmy hog (*Porcula salvania*) was still missing. The pygmy hog is the smallest and the most endangered suid found in the wild. The species was originally named *Porcula salvania* by Hodgson (Hodgson, 1847). This was amended as *Sus salvanius* based on morphological comparison (Garson, 1883; Groves, 1981; Corbet and Hill, 1991). However, a phylogenetic study using pygmy hog mitochondrial DNA (Funk et al., 2007) confirmed its original classification and the name *Porcula salvania* was regained. Currently, pygmy hogs are restricted to a small corridor of high grassland at the southern foothills of the

Himalaya, Assam, India. However, fossil remains suggest a far wider distribution of the pygmy hog in the middle Pleistocene, even covering current southwest China. In this thesis, I provide a comprehensive genomic analysis of the evolution of the pygmy hog in relation to the range expansion of *Sus scrofa*. Moreover, I investigated the formation of the current pygmy hog population.

1.6 Aims and outline

The main goal of my research is to gain a better understanding of the evolutionary process at a genomic level. On the basis of previous studies, this thesis provides a refined evolutionary history of *Suidae*. In **chapter 2 and 3**, I utilize whole genome re-sequencing data of the pygmy hog. With those, I reconstruct the evolutionary history from the origin of the species to the current population status. Specifically, in **chapter 2** I present a genome-scale phylogenetic and divergence tree for the *Suidae* family. Also, I evaluate complex models of wild boar range expansion during Pleistocene. **Chapter 3** describes the demographic trajectory of pygmy hog populations and evaluates the genetic consequences of long-term small population size. In **chapter 4** I report a high quality reference genome of the endangered Visayan warty pig. I characterize the evolution of genome architecture underlying the rapid speciation and adaptation of *Sus*. To take this one step further, in **chapter 5** I present reference-guided genome assemblies of an additional three *Sus* species and the outgroup pygmy hog. Thereafter, I provide a comparative system, which takes advantage of the genome collection to screen for genomic signatures of hybridization, domestication and adaptive radiation. Lastly, in **chapter 6** I provide an additional discussion on these topics as well as a synthesis of the work described in this thesis.

References

Aimi, M. (1989). A mandible of Sus stremmi Koenigswald, 1933, from Cisaat, Centra Java, Indonesia. *Publ. Geol. Res. Dev. Cent.*, 4–10.

Augier, A. (1801). *Essai d'une nouvelle classification des végétaux conforme à l'ordre de la nature*. Bruyset.

Barrett, L. W., Fletcher, S., and Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* 69, 3613–3634. doi:10.1007/s00018-012-0990-9.

Barrier, E., Huchon, P., and Aurelio, M. (1991). Philippine Fault: a key for Philippine kinematics. *Geology* 19, 32–35. doi:10.1130/0091-7613(1991)019<0032:PFAKFP>2.3.CO;2.

Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C. A., Chotalia, M., Xie, S. Q., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543, 519–524. doi:10.1038/nature21411.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi:10.1038/nature05874.

Blouch, R. A., and Groves, C. P. (1990). Naturally occurring suid hybrid in Java. *Z. Säugetierkd.* 55, 270–275. Available at: http://www.biodiversitylibrary.org/ [Accessed April 23, 2020].

Bosse, M., Megens, H. J., Frantz, L. A. F., Madsen, O., Larson, G., Paudel, Y., et al. (2014). Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nat. Commun.* 5, 1–8. doi:10.1038/ncomms5392.

Bridgett, M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., et al. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Sci. Adv.* 2, 1–13. doi:e1501714.

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi:10.1093/nar/gky1120.

Burri, R. (2017). Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. *Mol. Ecol.* 26, 3853–3856. doi:10.1111/mec.14167.

Bush, G. L., Case, S. M., Wilson, A. C., and Patton, J. L. (1977). Rapid speciation and chromosomal evolution in mammals. *Proc. Natl. Acad. Sci. U. S. A.* 74, 3942–3946. doi:10.1073/pnas.74.9.3942.

Butlin, R. K. (2010). Population genomics and speciation. *Genetica* 138, 409–418. doi:10.1007/s10709-008-9321-3.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). Molecular biology: The transcriptional landscape of the mammalian genome. *Science (80-.).* 309, 1559–1563. doi:10.1126/science.1112014.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* 7, e46688. doi:10.1371/journal.pone.0046688.

Clément, Y., Torbey, P., Gilardi-Hebenstreit, P., and Crollius, H. R. (2020). Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation. *Nucleic Acids Res.* 48, 2357–2371. doi:10.1093/nar/gkz1199.

COOK, O. F. (1906). FACTORS OF SPECIES-FORMATION. *Science (80-.).* 23.

Corbet, G. B., and Hill, J. E. (1991). *A world list of mammalian species. Third edition*. Natural History Museum Publications.

Crick, F. (1970). Central dogma of molecular biology. *Nature* 227, 561–563. doi:10.1038/227561a0.

Cucchi, T., Dai, L., Balasse, M., Zhao, C., Gao, J., Hu, Y., et al. (2016). Social Complexification and Pig (Sus scrofa) Husbandry in Ancient China: A Combined Geometric Morphometric and Isotopic Approach. *PLoS One* 11, e0158523. doi:10.1371/journal.pone.0158523.

Cucchi, T., Hulme-Beaman, A., Yuan, J., and Dobney, K. (2011). Early Neolithic pig domestication at Jiahu, Henan Province, China: Clues from molar shape analyses using geometric morphometric approaches. *J. Archaeol. Sci.* 38, 11–22. doi:10.1016/j.jas.2010.07.024.

Darwin, C. (1859). On the Origin of Species, 1859. doi:10.4324/9780203509104.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi:10.1038/nature11247.

Ellegren, H. (2010). Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol. Evol.* 25, 283–291. doi:10.1016/j.tree.2009.12.004.

Fan, Z., Silva, P., Gronau, I., Wang, S., Armero, A. S., Schweizer, R. M., et al. (2016). Worldwide patterns of genomic variation and admixture in gray wolves. *Genome Res.* 26, 163–173. doi:10.1101/gr.197517.115.

Fistani, A. B. (1996). Sus scrofa priscus (Goldfuss, de Serres) (Mammalia, Artiodactyla, Suidae) from the Middle Pleistocene layers of Gajtan 1 site, southeast of Shkoder (North Albania). *Ann. Palıontologie* 82, 177–229.

Flanagan, S. E., Patch, A. M., and Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomarkers* 14, 533–537. doi:10.1089/gtmb.2010.0036.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).* 269, 496–512. doi:10.1126/science.7542800.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., et al. (2013). Ensembl 2013. *Nucleic Acids Res.* 41, D48–D55. doi:10.1093/nar/gks1236.

Flint, J., and Eskin, E. (2012). Genome-wide association studies in mice. *Nat. Rev. Genet.* 13, 807–817. doi:10.1038/nrg3335.

Frankham, R., Ballou, J. D., Ralls, K., Eldridge, M. D. B., Dudash, M. R., Fenster, C. B., et al. (2017). "Inbreeding reduces reproductive fitness," in *Genetic Management of Fragmented Animal and Plant Populations* (Oxford University Press).

Frantz, L. A. F., Haile, J., Lin, A. T., Scheu, A., Geörg, C., Benecke, N., et al. (2019). Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proc. Natl. Acad. Sci. U. S. A.* 116, 17231–17238. doi:10.1073/pnas.1901169116.

Frantz, L. A. F., Madsen, O., Megens, H. J., Groenen, M. A. M., and Lohse, K. (2014). Testing models of speciation from genome sequences: Divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Mol. Ecol.* 23, 5566–5574. doi:10.1111/mec.12958.

Frantz, L. A. F., Rudzinski, A., Nugraha, A. M. S., Evin, A., Burton, J., Hulme-Beaman, A., et al. (2018). Synchronous diversification of sulawesi's iconic artiodactyls driven by recent geological events. *Proc. R. Soc. B Biol. Sci.* 285, 20172566. doi:10.1098/rspb.2017.2566.

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

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., et al. (1995). The minimal gene complement of Mycoplasma genitalium. *Science (80-.).* 270, 397–403. doi:10.1126/science.270.5235.397.

Funk, S. M., Verma, S. K., Larson, G., Prasad, K., Singh, L., Narayan, G., et al. (2007). The pygmy hog is a unique genus: 19th century taxonomists got it right first time round. *Mol. Phylogenet. Evol.* 45, 427–436. doi:10.1016/j.ympev.2007.08.007.

Garson, J. G. (1883). Notes on the anatomy of Sus salvanius (Porcula salvania, Hodgson). Part 1. External characters and visceral anatomy. in *Proceedings of the zoological society of London*, 413–418.

Geraads, D., Spassov, N., and Garevski, R. (2008). New specimens of <I>Propotamochoerus</I> (Suidae, Mammalia) from the late Miocene of the Balkans. *Neues Jahrb. für Geol. und Paläontologie - Abhandlungen* 248, 103–113. doi:10.1127/0077-7749/2008/0248-0103.

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

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi:10.1038/nature11622.

Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., et al. (2010). Genetic diversity in farm animals - A review. *Anim. Genet.* 41, 6–31. doi:10.1111/j.1365-2052.2010.02038.x.

Groß, C., Bortoluzzi, C., de Ridder, D., Megens, H.-J., Groenen, M. A. M., Reinders, M., et al. (2020a). Prioritizing sequence variants in conserved non-coding elements in the chicken genome using chCADD. *PLOS Genet.* 16, e1009027. doi:10.1371/journal.pgen.1009027.

Groß, C., Derks, M., Megens, H.-J., Bosse, M., Groenen, M. A. M., Reinders, M., et al. (2020b). pCADD: SNV prioritisation in Sus scrofa. *Genet. Sel. Evol.* 52, 4. doi:10.1186/s12711-020-0528-9.

Groves, C. (1981). Ancestors for the pigs: taxonomy and phylogeny of the genus Sus. *Dept. Prehist. Res. Sch. Pacific Stud. Aust. Nat. Univ., Tech. Bull.* 3, 1–96.

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

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

Hall, R. (2002). Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: Computer-based reconstructions, model and animations. *J. Asian Earth Sci.* 20, 353–431. doi:10.1016/S1367-9120(01)00069-4.

Harrison, R. G., and Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. in *Journal of Heredity* (Oxford University Press), 795–809. doi:10.1093/jhered/esu033.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* 38, 209–216. doi:10.1017/S0016672300020553.

Hirschfeld, L., and Hirschfeld, H. (1919). Serological Differences Between the Blood of Different Races. the Result of Researches on the Macedonian Front. *Lancet* 194, 675–679. doi:10.1016/S0140-6736(01)48686-7.

Hodgson, B. H. (1847). On a new form of the hog kind or Suidae. *J. Asiat. Soc. Bengal 16,* 16, 423–428.

J Whittaker, R., and Fernandez-Palacios, J. M. (2007). *Island biogeography: ecology, evolution, and conservation*. Oxford University Press.

Jukes, T. H., and Kimura, M. (1984). Evolutionary constraints and the neutral theory. *J. Mol. Evol.* 21, 90–92. doi:10.1007/BF02100633.

Kardos, M., Luikart, G., and Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity (Edinb).* 115, 63–72. doi:10.1038/hdy.2015.17.

Kardos, M., Qvarnström, A., and Ellegren, H. (2017). Inferring individual inbreeding and demographic history from segments of identity by descent in Ficedula flycatcher genome sequences. *Genetics* 205, 1319–1334. doi:10.1534/genetics.116.198861.

Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., and Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evol. Appl.* 9, 1205–1218. doi:10.1111/eva.12414.

Kimura, M. (2020). "The neutral theory and molecular evolution," in *My Thoughts on Biological Evolution* (Springer), 119–138.

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., Mckeigue, P. M., and Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PLoS One* 5, e13996. doi:10.1371/journal.pone.0013996.

Kogoma, T. (1997). Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol. Mol. Biol. Rev.* 61, 212–238. doi:10.1128/.61.2.212-238.1997.

Laikre, L., Hoban, S., Bruford, M. W., Segelbacher, G., Allendorf, F. W., Gajardo, G., et al. (2020). Post-2020 goals overlook genetic diversity. *Science (80-.).* 367, 1083.2-1085. doi:10.1126/science.abb2748.

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science (80-.).* 307, 1618–1621. doi:10.1126/science.1106927.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* 424, 147–151. doi:10.1038/nature01763.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi:10.1038/nature10231.

Lynch, M., and Walsh, B. (2007). *The origins of genome architecture*. Sinauer Associates Sunderland, MA.

Mattick, J. S. (2001). Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep.* 2, 986–991. doi:10.1093/embo-reports/kve230.

McGlone, J., and Pond, W. G. (2003). *Pig Production: Biological Principles and Applications*. Cengage Learning Available at: http://books.google.com/books?id=tc8UHbUSnAAC&pgis=1.

Meijaard, E. (2004). Solving Mammalian Riddles : A reconstruction of the Tertiary and Quaternary distribution of mammals and their palaeoenvironments in island South-East Asia. *Sch. Archaeol. Anthropol.* PhD, 349. Available at: http://hdl.handle.net/1885/47989.

Meijaard, E., d'Huart, J. P., and Oliver, W. L. R. (2011). Family Suidae (Pigs). *Handb. Mamm. world* 2, 248–291.

Miller, K. G., Kominz, M. A., Browning, J. V., Wright, J. D., Mountain, G. S., Katz, M. E., et al. (2005). The phanerozoic record of global sea-level change. *Science (80-.).* 310, 1293–1298. doi:10.1126/science.1116412.

Mills, M. C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. *Commun.*

Biol. 2, 1–11. doi:10.1038/s42003-018-0261-x.
Mizelle, B. (2012). *Pig*. Reaktion Books.
Myers, N., Mittermeler, R. A., Mittermeler, C. G., Da Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403, 853–858. doi:10.1038/35002501.
Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms.
Nosil, P., and Feder, J. L. (2013). Genome evolution and speciation: Toward quantitative descriptions of pattern and process. *Evolution (N. Y).* 67, 2461–2467. doi:10.1111/evo.12191.
Orliac, M. J., Pierre-Olivier, A., and Ducrocq, S. (2010). Phylogenetic relationships of the Suidae (Mammalia, Cetartiodactyla): New insights on the relationships within Suoidea. *Zool. Scr.* 39, 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., et al. (2013). Pig domestication and human-mediated dispersal in western eurasia revealed through ancient DNA and geometric morphometrics. *Mol. Biol. Evol.* 30, 824–832. doi:10.1093/molbev/mss261.
Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi:10.1534/genetics.112.145037.
Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502. doi:10.1038/nature05295.
Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi:10.1093/nar/gky1016.
Roth, J., and Wagner, J. A. (1854). *Die fossilen Knochenüberreste von Pikermi in Griechenland: Gemeinschaftlich bestimmt u. beschrieben nach d. Materialien, welche durch die von dem Erstgenannten im Winter 1852/3 dortselbst vorgenommenen Ausgrabungen erlangt wurden*. Verlag d. Akad.
Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* 26, 1241–1247. doi:10.1016/j.cub.2016.03.037.
Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., et al. (2014). Genomics and the origin of species. *Nat. Rev. Genet.* 15, 176–192. doi:10.1038/nrg3644.
Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi:10.1101/gr.3715005.
Sillar, F. C., Meyler, R. M., and Holt, O. (1961). The Symbolic pig: an anthology of pigs in literature and art. *Symb. pig. An Anthol. pigs Lit. art.* Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Symbolic+pig+:+an+anthology+of+pigs+in+literature+and+art#0.
Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187, 367–383. doi:10.1534/genetics.110.120907.
Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. *Nature* 465, 219–222. doi:10.1038/nature09014.
van der Made, J., Morales, J., and Montoya, P. (2006). Late Miocene turnover in the Spanish mammal record in relation to palaeoclimate and the Messinian Salinity Crisis. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 238, 228–246. doi:10.1016/j.palaeo.2006.03.030.
Whittemore, C., and others (1998). *The science and practice of pig production.* Blackwell Science Ltd.
Wilson, S. H. (2014). The dark side of DNA repair. *Elife* 2014, 2001. doi:10.7554/eLife.02001.
Woodruff, D. S. (2010). Biogeography and conservation in Southeast Asia: How 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. *Biodivers. Conserv.* 19, 919–941. doi:10.1007/s10531-010-9783-3.
Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088.
Zachos, J., Pagani, H., Sloan, L., Thomas, E., and Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science (80-.).* 292, 686–693. doi:10.1126/science.1059412.

Zeder, M. A. (1998). Pigs and emergent complexity in the ancient Near East. *MASCA Res. Pap. Sci. Archaeol.* 15, 109–122.
Zhang, J., and Luo, Y. (2008). Restudy of the Pigs' Bones from the Jiahu Site in Wuyang County, Henan. *Archaeology* 1, 90–96.

2

Genomic analysis on pygmy hog (*Porcula salvania*) reveals multiple interbreeding during wild boar expansion

Langqing Liu[1*], Mirte Bosse[1], Hendrik-Jan Megens[1], Laurent AF Frantz[2,3], Young Lim Lee[1], Evan K Irving-Pease[3], Goutam Narayan[4,5], Martien AM Groenen[1], Ole Madsen[1*]

[1]Animal Breeding and Genomics, Wageningen University & Research, 6708PB Wageningen, the Netherlands; [2]School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, United Kingdom; [3]Palaeogenomics and Bioarcheology Research Network, Research Laboratory for Archeology and History of Art, University of Oxford, Oxford OX1 3QY, United Kingdom; [4]Durrell Wildlife Conservation Trust, Les Augrès Manor, Jersey JE3 5BP, Channel Islands, United Kingdom; [5]Pygmy Hog Conservation Programme, EcoSystems-India, Indira Nagar, Basistha, Guwahati, Assam 781029, India

Nat Commun 10, 1992 (2019)

Abstract

Wild boar (*Sus scrofa*) drastically colonized mainland Eurasia and North Africa, most likely from East Asia during the Plio-Pleistocene (2-1Mya). In recent studies, based on genome-wide information, it was hypothesized that wild boar did not replace the species it encountered, but instead exchanged genetic materials with them through admixture. The highly endangered pygmy hog (*Porcula salvania*) is the only suid species, in mainland Eurasia, known to have outlived this expansion and therefore provides a unique opportunity to test this hybridization hypothesis. Analyses of pygmy hog genomes indicate that despite large phylogenetic divergence (~2My), wild boar and pygmy hog did indeed interbreed as the former expanded across Eurasia. In addition, we also assess the taxonomic placement of the donor of another introgression, pertaining to a now-extinct species with a deep phylogenetic placement in the Suidae tree. Altogether, our analyses indicate that, the rapid spread of wild boar was facilitated by inter-specific/inter-generic admixtures.

Key words: Genetic hybridization, Genomics, Phylogenetics, Phylogenomics, Population genetics

2.1 Introduction

The expansion of species into novel habitats can have tremendous impacts on the native fauna. If the native fauna contains species that are closely related to the invasive population, hybridization may also threaten the integrity and survival of native species (Rhymer and Simberloff, 1996). Observation of admixture in expanding populations has led to speculation whether admixture has an important role in driving the success of those populations (Kolbe et al., 2008; Lawson Handley et al., 2011; Nuijten et al., 2016). Although the old-world pigs, the *Suidae*, are distributed throughout Africa and Eurasia, only two species are recognized across mainland Eurasia: wild boar (*Sus scrofa*) and pygmy hog (*Porcula salvania*) (Groves, 1981; Hardjasasmita, 1987; Oliver, 2001; Groenen et al., 2012; Munz, 2017).

This, however, was not always the case and extensive fossils records suggest that Eurasia hosted a highly diverse set of *Suidae* species that originated during the Miocene (Andrews, 1988; Pickford, 1988; Pickford Senut, B., Hadoto, D., 1993; Made, 1999) (Fig 2.1a). In middle Miocene, the first isolated lineage to split from those early *Suidae* was *Babyrousa* which forms an ancient lineage endemic to the island of Sulawesi (Fig 2.1b) (Frantz et al., 2016, 2018). Along with the global cooling during the late Miocene (Zachos et al., 2001), a new subfamily, the *Suinae*, emerged in the fossil record (Roth and Wagner, 1854; Geraads et al., 2008a) and replaced almost all other subfamilies of *Suidae* present at that time (van der Made et al., 2006; Orliac et al., 2010) (Fig 2.1c). The *Suinae* later diversified into multiple tribes (Roth and Wagner, 1854; Geraads et al., 2008b). This was followed by a divergence of the African *Suidae* and the Eurasian *Sus* genus, at around the Miocene/Pliocene boundary (Fig 2.1d). Shortly thereafter, the divergence within the *Sus* genus started during the early Pliocene (Fig 2.1f). Several *Sus* species on the islands of southeast Asia (ISEA) evolved during the early/mid Pliocene (Frantz et al., 2013). Relatively high levels of species diversity were likely maintained across Eurasia, until the early Pleistocene, when wild boar expanded out of East Asia into almost every type of ecosystem across the old-world. This expansion was highly efficient mirroring the great human expansion during the late Pleistocene (Schiffels and Durbin, 2014), and previous study have suggested that it is the reason for the disappearance of most suid species across Eurasia (Fistani, 1996; Guérin and Faure, 1997) (Fig 2.1g&h). With this, the layout for the modern *Suidae* species became settled.

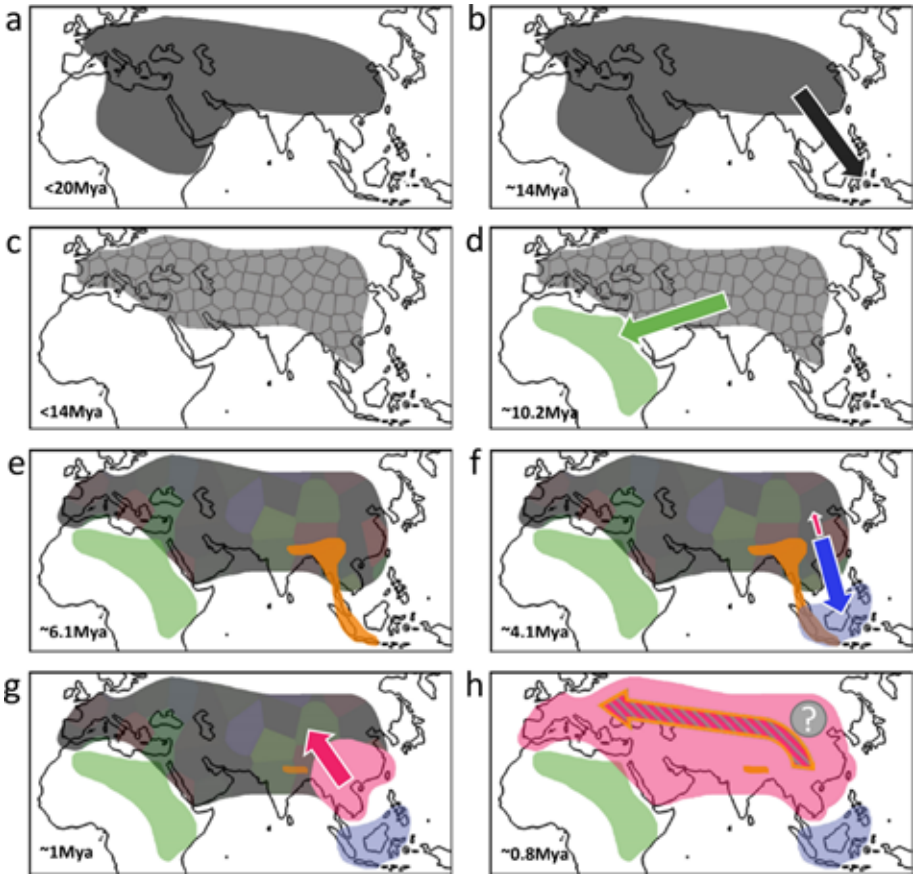


Fig 2.1: A series of schematic models depicting the geographic evolution of *Suidae* species over the past 20Mya. (**a**) Emergence of Suidae across Eurasian and Northern Africa. (**b**) The black arrow depicts the hypothesized trajectory of *Babyrussa* migrating to ISEA. (**c**) Emergence of *Suinae* (grey collage pattern) that replaced other *Suidae*. (**d**) The green arrow indicates the diversification of the ancestor of Sub-Saharan suids. (**e**) Eurasian *Suinae* split into multiple genera (multi-color collage pattern), including pygmy hog (orange shade). (**f**) The blue arrow depicts the migrations of *Sus* to ISEA. The red arrow indicates emergence of *Sus scrofa*. (**g-h**). The spread of *Sus scrofa* from southern Asia to Europe and replacement of all Suinae species except pygmy hog. During the replacement, Sus scrofa populations introgressed at least three times with one of these Suinae species (ghost linage), pygmy hog and ISEA *Sus*, respectively. The colors correspond to those used in Fig.2 and represent the cluster on the tree to which the samples belong.

As the only reminiscence of the once highly diverse Pliocene suid fauna, the pygmy hog represents not only an important species to conserve but also the key to our understanding of the expansion of wild boar. Indeed, previous studies have suggested that the rapid and highly efficient expansion of wild boar was facilitated by inter specific adaptive gene-flow (Frantz et al., 2013, 2014). Under this scenario, wild boar absorbed rather than replaced species, a process paralleled to some extent in humans (Green et al., 2010; Huerta-Sánchez et al., 2014; Slon et al., 2018). Although pygmy hog is now highly endangered and restricted in a small corridor of high grassland at the southern foothills of the Himalaya (Breeding et al.; von Koenigswald, 1963; PHCP, 2008), it was far more widespread in the past (Pickford, 2013) (Fig 2.1e). According to middle Pleistocene fossil remains from southwest China, the geographical range of pygmy hog and wild boar did overlap (HAN and D.-F., 1975, 1987), implying that the temporal and geographical proximity of pygmy hog and wild boar could have resulted in hybridization. Therefore, pygmy hog provides an excellent comparative framework to study the evolutionary processes that occurred during a fast and extensive radiation. We analyzed 6 genomes of pygmy hog in combination with genomes of related suid species and found strong support for an important role of inter-species hybridization during range expansion. The results suggest that wild boar hybridized with pygmy hog and a now-extinct suid species during the rapid spread across Eurasia and North Africa.

2.2 Results

Phylogenetic relationships and divergence of the *Suidae* species.

We sequenced the genomes of six pygmy hogs and one *Babyrousa babyrussa* and analyzed these with the genome sequences of 31 individuals, in total representing 10 of the extant *Suidae* species (See Material and methods and Supplementary data 1 for details). We first assessed the phylogenetic relationship of these species. The concatenation and consensus methods resulted in the same main topology (Fig 2.2a, Supplementary Fig 2.1-2.2). The phylogenetics analyses clearly show that the most basal split within the *Suinae* are sub-Saharan suids followed by a highly supported split of pygmy hogs (BS=100 in supermatrix and CF=1 in supertree) from all *Sus* species. To compare our phylogenetics results to an earlier study using fragments of mitochondrial DNA (Funk et al., 2007), we also carried out a Bayesian phylogenetic analysis using complete mitochondrial genomes (Supplementary Fig 2.3). The resulting topology is consistent with previous studies confirming pygmy hog as basal to *Sus*. Thus, both the genome wide autosomal phylogenetic analysis

and complete mitochondrial genome analysis support, with very high confidence, that pygmy hog is a monophyletic sister taxon of *Sus*.

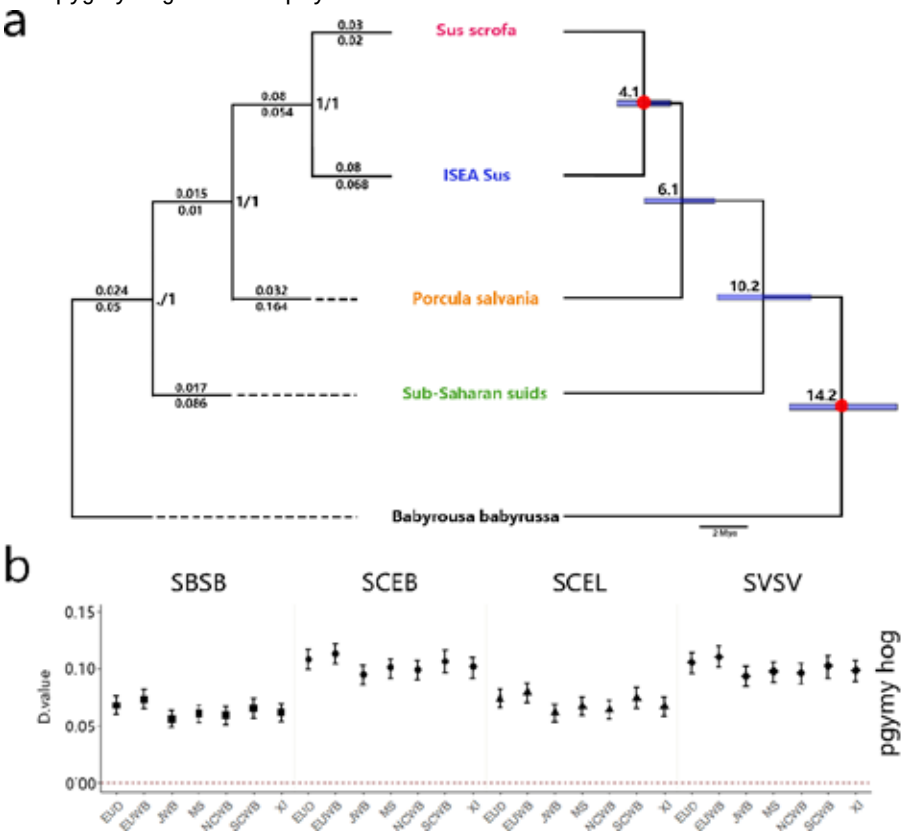


Fig 2.2 Phylogenetic relationships and divergence of the *Suidae* species used in the current study and admixture event between pygmy hog and *Sus scrofa*. (a) The tree on the left is the ML tree of the Suidae family based on consensus and concatenation methods. Branch length of the consensus analysis above branches, concatenation below branches. Node labels show bootstrap values of the concatenation analysis and concordance factors of the consensus analysis respectively. The tree on the right is the time tree of divergence. Node labels show age in million years. Blue bars indicate 95% confidence interval and red dots show the calibration points (See Supplementary material figures 1-3 for full trees). (b) A diagram depicting the excess derived allele sharing when comparing sister taxa and outgroups. Each column contains the fraction of excess allele sharing by a taxon (up/down) with the pygmy hog/outgroup compared to its sister taxon (up/down). We computed D statistics of the form D (X, Y, Pygmy hog, warthog). Error bars correspond to three standard errors. (Note: the Fig 2.2b in this thesis is the corrected version, see Additional file)

We selected autosomal genomic loci supporting the main topology to obtain the basal divergence between the studied taxa (Fig 2.2a, Supplementary Fig 2.4). Molecular clock analyses indicated that the divergence between Sub-Saharan suids and Eurasian *Suidae* (pygmy hog and *Sus*) took place shortly after the divergence of *Babyrousa babyrussa* ~10.2 Mya (95% HPD = 12.7-7.9). Pygmy hog separated from the common ancestor with *Sus* during the early Pliocene, ~6.1 Mya (95% HPD = 7.8-4.2) and the Eurasian wild boar split from other *Sus* species during the early Pliocene ~4.1 Mya (95% HPD = 5.5-2.7). (Fig 2.2a, Supplementary Fig 2.4, Supplementary Note)

Admixture between pygmy hog and wild boar.

In order to test whether temporal and geographical proximity of closely related species could have resulted in hybridization, we looked for interspecific admixture signal within our genome-wide data. Several studies have reported that in diverged species sex-linked markers may show evolutionary histories incongruent to other sex-linked and/or autosomal markers (Chan and Levin, 2005; Ropiquet and Hassanin, 2006; Nakagome et al., 2008; Yannic et al., 2010; Ai et al., 2015). Thus, we analyzed autosomes and the X chromosome separately. To investigate whether any of the sequenced pygmy hogs showed evidence for autosomal introgression from ancestors of present-day *Sus* species, or vice-versa, the pygmy hog was compared to representatives of eleven *Sus* populations using D-statistics. We found a significant overrepresentation of derived alleles between the pygmy hog and mainland Eurasian wild boar at autosomal chromosomes (Fig 2.2b, Additional file, Supplementary data 2), indicative of admixture. This signal of admixture was also supported by TreeMix analysis. The best fitting model suggests an ancient admixture between ancestral wild boar and pygmy hog (Supplementary Fig 2.5).

To further examine this autosomal genome-wide pattern of admixture between pygmy hog and wild boar, we combined D-statistics and fd to infer regions of introgression. We also calculated DNA sequence divergence (d_{xy}) for each window to reduce false positive signals (Smith and Kronforst, 2013). This resulted in 636 putative introgression intervals between pygmy hog and wild boar from Europe, North China and South China, of which 427 (67.1%) are shared within wild boars (Supplementary Fig 2.6, Supplementary Note). This suggests the admixture occurred before the divergence within Eurasian wild boar, further sustaining the evidence of an ancestral gene flow between pygmy hog and the common ancestor of wild boar.

Wild boar harbors genetic introgression from a ghost linage.

We next addressed evidence of admixture from the X chromosome starting with reconstructing a maximum-likelihood tree for the X chromosome. This tree displays a different topology compared to our main phylogenetic tree (Supplementary Fig 2.8). Previous study has reported two distinct haplotypes in European pigs and South Asian pigs, and proposed that this might be derived from a now extinct Suid (ghost) lineage(Ai et al., 2015). To investigate the existence of genealogical discordance, we carried out a sliding-window D-statistic and a machine-learning based detection of local phylogenetic incongruent regions (Zamani et al., 2013). Both approaches supported that there is a ~40.6Mbp (46.2-86.8 Mbp) region on the X chromosome, where pygmy hog clusters with ISEA *Sus* and South China pigs, while the European pigs and North China wild boars appear to be basal to this cluster. (Fig 2.3, Supplementary data 3, Supplementary Fig 2.9-2.15, Supplementary Note). In addition, an ambiguous pattern was also observed in northern Chinese domestic pigs, where the region from 46.2-57.1 Mbp support a clustering of northern Chinese domestic pigs with European pigs/North Chinese wild boars while from 57.1-86.8 Mbp support northern Chinese domestic pigs clustering with southern Chinese pigs (Fig 2.3, Supplementary Fig 2.13). The signature of the introgression regions was also supported by maximum-likelihood trees (Supplementary Fig 2.16). Taken together, our results show that within this genomic region, sequences of European/North Chinese pigs have an ancient origin. With the inclusion of pygmy hog genome, we could locate this ghost lineage to be older than the split of pygmy hogs but younger than the split to the Sub-Saharan suids. So far, there is no molecular or fossil evidence for this ancient linage, which was probably extinct long time ago. We further looked for evidence of this introgression in the autosomes. Comparison between different wild boar populations further identified autosomal regions supporting the X-chromosome introgression signal (Supplementary Fig 2.10-2.15). The amount, length and magnitude of ghost introgression in autosomes are similar among the wild boar populations (Supplementary Fig 2.17&2.18), which suggests that this hybridization likely occurred early within the evolutionary history of wild boar.

It has been reported that the region around the centromere on the X chromosome in pigs has an extremely low recombination rate (Ma et al., 2010; Fernández et al., 2014). Also, as a consequence of the global reduction in effective population size of wild boar in the past ~1 My to the end of the Last Glacial Maximum (Groenen et al., 2012), wild boar went through processes like incomplete lineage sorting and positive natural selection. The joint effect likely resulted in distinct distribution of ancient introgressed haplotype between European/northern Chinese and southern

Chinese wild boar populations (Ai et al., 2015). Finally, the genealogical discordance on the X chromosome became fixed in pigs from different regions. Recombination rates are highly variable on the autosomes and this hybridization probably happened at least 1 Mya. The long period of recombination, in addition with post-divergence gene flow between wild boar population (Groenen et al., 2012), highly truncated autosomal haplotype blocks. This would lead to many very short introgression segments scattered over the autosomes, which is what we observed in our analyses (Supplementary Fig2. 18).

For the male individuals, we also reconstructed the phylogeny based on the non-recombining part of the Y-chromosome (4.8~43.5 Mb), which resulted in a topology consistent with our main phylogenetic tree (Supplementary Fig 2.19).

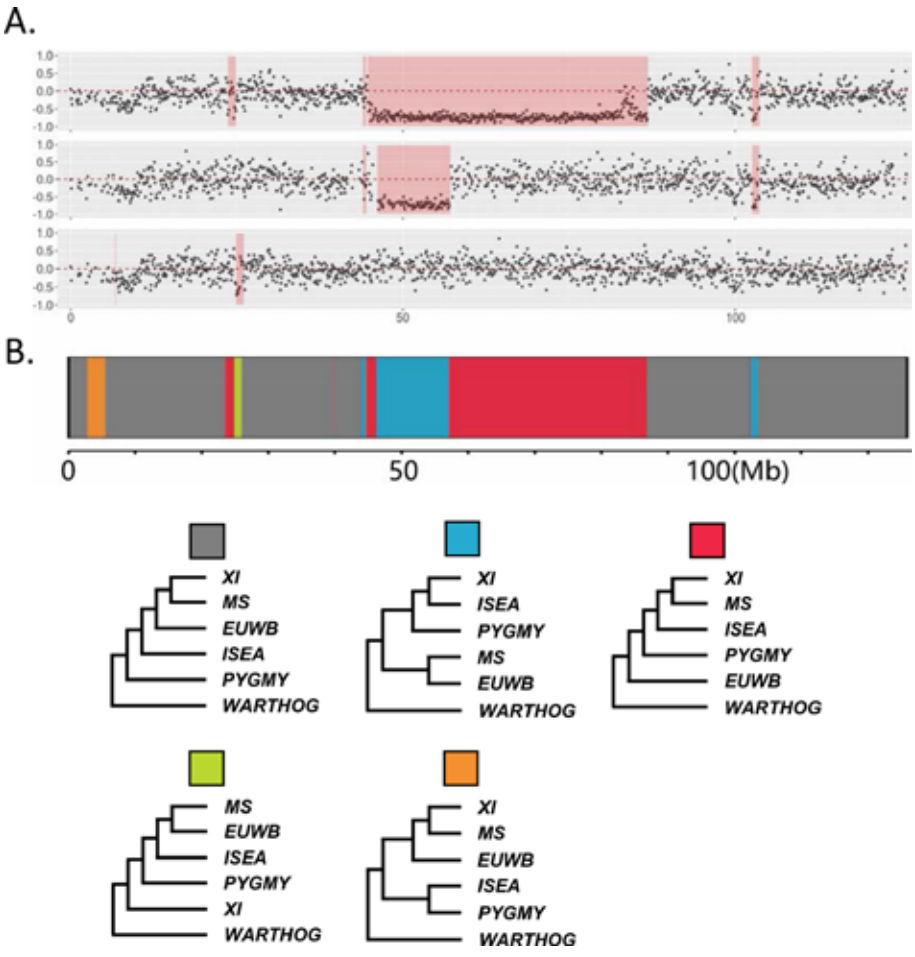


Fig 2.3 Genealogical discordance in Chromosome X. (**a**) The D-statistic for testing introgression for 100-kb windows on chromosome X for tree topology (((ISEA, X), pygmy hog), warthog) (X = EUWB, MS and XI). Excesses of BABA in 46 – 86 Mb indicated higher genetic similarity between ISEA and pygmy hog. See supplementary Figure 9-14 for D-statistic testing introgression for 100-kb windows on autosome chromosomes. (**b**) SAGUARO plot illustrating the distribution of the 5 frequent rooted local topologies over the X chromosome. The red unrooted topologies support introgression from a missing linage into European/North China wild boar. Note that the blue unrooted topology includes the discrepant haplotype within Chinese population (see text for further explanation). ISEA=Islands of Southeast Asia Sus; MS=Meishan; XI=Xiang; EUWB=European wild boar; See Supplementary data 3 for further details and supplementary Figure 8 for results for autosomal chromosomes.

Demographical modeling of the Suidae evolutionary history

We then used an automated qpGraph approach (Patterson et al., 2012; Leathlobhair et al., 2018) to evaluate the fit of various admixture graphs to our data (see Material and methods). We then estimated the marginal likelihood of 37 models that left no *f4*-outliers using a MCMC approach (Leppälä et al., 2017) (Supplementary Fig 2.21). According to these models, both pygmy hog and ISEA *Sus* contributed to wildboar. Interestingly, these models suggest that the ISEA *Sus* only contributed to south Chinese wildboar. This discrepancy is likely the result of the near simultaneous divergence of all three wildboar lineages, as well as the fact that a population closer to the South Chinese wildboar potentially also contributed to ISEA *Sus* (Frantz et al., 2013). The best admixture graph, however, is slightly different from those obtained from *TreeMix* and phylogenetic analyses, as it finds a signal of ISEA *Sus* admixture into pygmy hog population. This suggest that pygmy hog actually interbred with an unsampled admixing/separating population which was intermediate between ISEA *Sus* and *Sus scrofa*. However, our parsimonious model may be too simple to reflect the complexity of the reticulate admixture in these populations and to disentangle the ancestral variants between ISEA *Sus* and *Sus scrofa*. Altogether, this analysis validates the existence of gene-flow between wildboar and ISEA *Sus* as well as between pygmy hog and wildboar. Yet it also suggests that this admixture could have been bidirectional in which case the south Chinese wildboar population forms the best proxy for these events.

To coalesce the known demographic implications by far, we further fitted various fitted various gene flow models, which are based on our priori assumption of *Suidae* systematic, to a phylogenetic scenario using G-phocs. We separated admixture branches into each of the *Sus scrofa* populations, to better account for

their variable levels of basal admixture. (Supplementary Fig 2.22, Supplementary Note). In support of our D-statistic findings, high probability of gene flow between the common ancestor of wild boar and pygmy hog was inferred. In our full model, signatures of admixture in the ISEA population were also examined and significant gene flow between the ISEA *Sus* and Asian wild boar was found in agreement with previous analyses (Frantz et al., 2013). Furthermore, a model assuming a basal ghost population was applied and confirmed a post-speciation gene flow between the common ancestor of wild boar and the ghost population (Supplementary data 4, Supplementary Fig 2.23, Supplementary Note).

Testing for incomplete lineage sorting

It is well known that incomplete ancestral linage sorting can inflate ad-mixture signals. Therefore, we used two methods to test for incomplete lineage sorting (ILS) in our data. First, we calculated calculate the maximum length of a shared haplotype by pygmy hog and wild boar due to ILS (probability < 0.005). With the deep divergence split between pygmy hog and wild boar (at least 4.2 My), the estimated length of a shared haplotype due to ILS is extremely small (<688bp) and significantly different from the window size used in our sliding-window D-statistic analysis (100kb) Furthermore, in our Saguaro analysis, we also filter out all the segment having alternative topology with a length <= 688 bp.

Secondly, we also follow the approach described in Huerta-Sánchez's(Huerta-Sánchez et al., 2014) to assess the probability of ILS. We simulated 10000 loci with length of 100kb under the model described in Supplementary Note, and calculated D-statistics with the same quadruplets we used in the sliding-window analysis. All results from the simulations resulted in P < 0.001 against ILS for all comparisons (Supplementary Fig 2.25). Thus, we conclude that it is unlikely that ILS have contributed significantly to our observed introgression signal.

Identification and functional analysis of introgressed genes

The different introgression signals that we observe in *Sus scrofa* could have played an important role in its successful expansion. We therefore accessed the functional annotation of the genes overlapping introgressed regions. However, given our low coverage unphased genomic datasets and limited sample size, our ability to reveal ultra-short introgression segments broken down by long-term recombination is limited. With this in mind, we undertook a functional annotation analysis for candidate introgressed genes. For the introgressed pygmy hog genes in wild boars (384 genes), enrichment for GO terms related to the sensory perception of taste, olfactory pathways and participating in glycolysis and fatty acid metabolism was

observed (Supplementary Fig 2.26, Supplementary data 5). This finding is in agreement with the knowledge that smelling, taste and energy metabolism pathway do have specific roles in adaptive capacity to environment (Chevin et al., 2010; Paudel et al., 2013; Kishida et al., 2015). However, it should be noted that especially olfactory genes are prone to copy number variation making them consistent enriched in such analysis. The Ghost introgression genes (104 genes) are related to more broad GO terms including neurogenesis, immune response and TCA cycle (Supplementary data 5). Of the Ghost introgressed genes, the POR gene is of most interest as it is involved in Vitamin D metabolism(Larson-Meyer et al., 2017), which could potentially boost the fitness during the expansion of wild boar from sun-belt region (Southern Asia) to short daylength region (Northern Asia and European).

2.3 Discussion

Our analyses reveal the phylogeny and diversification times of the *Suidae* family (Fig 2.2). The demographic analysis suggests at least three independent events of inter-species gene flow during *Suidae* evolution – the most notable from an ancient and now extinct linage (Fig 2.3 and Supplementary Fig 2.23). Combined, these results allow us to dramatically refine the evolutionary history of the *Suidae* family. After the Sub-Saharan suids evolved during the late Miocene (~10.2 Mya, Fig 2.1d), the divergence between pygmy hog and *Sus* took place around the Miocene/Pliocene boundary (~6.1Mya, Fig 2.1e), followed by the emergence of ISEA *Sus* and wild boar (Fig. 1f). At around 1 Mya, populations of wild boar from Asia started to spread and reach Europe around 0.8 Mya (Fig 2.1g&h). During this migration, wild boar colonized Eurasia and efficiently replaced all but one of the local species along the way (Fistani, 1996; Guérin and Faure, 1997), with pygmy hog as the only survivor. Moreover, during this the expansion, despite long divergences (approximately 2 My between wild boar and pygmy hog, even longer for the ghost linage), wild boar hybridized with both pygmy hogs and an extinct, more divergent, *Suinae* species (Fig 2.1h). The frequent climatic fluctuations during the Pleistocene led to alternating warm and cold periods (ices ages) (Guo et al., 2001; Hansen et al., 2013), which likely resulted in multiple rounds of north-south directed migration (Seebacher and Post, 2015). While expanding instantly, wild boar had greater chances of encountering and temporal co-existing with local species, enabling possible interspecies hybridization. Although our knowledge of the impact of admixture on the fitness of expanding populations is still limited, it is likely that

changes in the genetic architecture that arise from admixture will generate heterosis that could boost adaptation to local niches (i.e. high grassland for pygmy hog, high altitude for ghost lineage).

Here, we have shown that an effective replacement of species is accompanied by consistently absorption of part of gene pool of the local related species. This suggests that admixture may play a role as an evolutionary biological driving force in successful range expansion and provides pertinent evolutionary hypothesis on the model of massive species replacement. With the booming development of paleogenomics technology, several case studies have directly verified ancient gene-flow between genomes of extinct species and extent recipient species (Huerta-Sánchez et al., 2014; Barlow et al., 2018). Future studies, where the genome of fauna from early/middle Pleistocene remain is retrieved, will probably further refine the *Sus scrofa* expansion from Asia to Europe. Overall, the demographic history of pig species not only demonstrates how explosive and invasive range expansion can be, but also reminds us of the ubiquity of inter-species hybridization during speciation.

2.4 Methods

Sampling, genome sequencing, alignment and SNP calling

The pygmy hog used for this research consists of three individuals sampled from the wild and three individuals from captivity. Whole genome Illumina PE 100 bp re-sequencing was performed at SciGenom Laboratories in Chennai, India on these six pygmy hog samples. The *Babyrousa babyrussa* was sampled from Copenhagen Zoo. Libraries of ~300 bp fragments were prepared using Illumina paired-end kits (Illumina, San Diego, CA) and 100bp paired-end sequenced with Illumina HiSeq. A selection of published genome from other *Suinae* species was included (Supplementary data 1). All these samples were also sequenced with the Illumina sequence technology. The whole genome sequencing data were trimmed using sickle (https://github.com/najoshi/sickle) with default parameters. The trimmed reads were aligned to the Sscrofa 11.1 reference genome using the Burrows-Wheeler Aligner (BWA 0.7.5a) (Li and Durbin, 2009). Local re-alignment was performed using GATK v3.6.0 RealignmentTargetCreator and IndelRealigner and variants were called using GATK UnifiedGenotyper (McKenna et al., 2010), with the –stand_call_conf option set to 50, the –stand_emit_conf option set to 20, and the -dcov option set to 200. Variants with a read-depth between 0.5 and 2.0 times of the average sample genome coverage were selected and stored in variant calling format. We identified the sex of all individuals by calculating the ratio of read depth

on X chromosome and the autosomes. For the individuals whose molecular sex are male, we filtered out variants in the non-PAR regions which are heterozygous and with a coverage larger than the average read-depth in autosomes. We do not have any explicit pedigree for the *Babyrousa Babyrussa* sample. To avoid potential biases caused by recent interbreeding, we did decide to use warthog as the outgroup in all the analyses related to introgression. Pygmy hog samples were collected within the Pygmy Hog Conservation Programme in Assam, India in accordance with ethical and legal regulations in India. The *Babyrousa babyrussa* was sampled from a dead individual in Copenhagen Zoo in accordance with ethical and legal regulations in Denmark (The Animal Experimentation Act LBK no. 253, March 8th 2013). This study was ethically approved by the European Research Council under the European Community's 256 Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n° 249894".

Phylogenetic analysis

Phylogenetic trees on autosomes were construct based on the maximum likelihood (ML) method as implemented in RAxML 8.2.3 (Stamatakis, 2014) using the best fitting model of substitutions, identified by jModelTest2 (Darriba et al., 2012) on 100 random subsets of 1 Mbp. In order to eliminate possible bias stemming from alignment and genotyping errors, we only used autosome one-to-one orthologous gene coding sequences (CDS) (Frantz, 2015) between pig and cow for this analysis. A list of One-to-one orthologous genes (between cow and pig) and coordinates of corresponding one-to-one CDS region were extracted from ENSEMBL with biomart (Kinsella et al., 2011). Finally, we got 486203 CDS regions from 18313 genes. The total number of SNPs in the one-to-one gene regions was 2571419. We used both supermatrix and supertree techniques (Delsuc et al., 2005), using *Babyrousa babyrussa* as an outgroup. In the supermatrix approach, the concatenated CDS alignment was analyzed under best fitting substitution model (GTR+Γ+I) with 100 bootstrap replications. In the supertree approach, pig-to-cow orthologous genes with CDS alignments longer than 1000 bp were used. Individual gene trees were inferred separately under GTR+Γ+I substitution model with 100 rapid bootstraps. All gene trees with an average bootstrap value above 40 were combined into a consensus tree using the software ASTRAL-II (Mirarab and Warnow, 2015). To assess support for particular clades in the supertree analysis, we calculated concordance factors in DensiTree (Bouckaert, 2010).

RAxML 8.2.3 was used to reconstructed ML phylogenetic trees on the whole X chromosome and on the two regions which have anomalous phylogenetic relationship in the SAGUARO analysis. For mitochondria DNA analysis, we used a

Bayesian Markov Chain Monte Carlo simulation (MCMC) to estimate the most likely phylogenetic trees with MrBayes 3.2.3 (Ronquist et al., 2012), using the best fitting model of substitutions, identified by JModelTest2. The length of the MCMC was set to 10,000,000. The parameter estimates and consensus trees resulting from 10 MrBayes runs were recorded and compared. The best supported phylogenetic consensus tree was summarized with TRACER v1.6 (http://tree.bio.ed.ac.uk/software/tracer/) discarding the first 10% as burn-in. All trees were depicted using the software FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Molecular clock analyses

We estimated divergence times using an approximate likelihood method as implemented in MCMCtree (Yang, 2007), with an independent relaxed-clock and birth-death sampling. To overcome difficulties arising from computational efficiency and admixture, we only used coding sequences (CDS) with pig-to-cow ortholog filter. We fitted an GTR + Γ4 model to each genomic bin and estimated a mean mutation rate by fitting a strict clock to each fragment setting a root age at 20Mya, which represent the earliest *Tayasuidae* fossil. This mean rate was used to adjust the prior on the mutation rate (rgene) modelled by a gamma distribution as G (1,241). The BDS and sigma2 values were set at 7 5 1 and G (1,10) respectively. We ran 100,000 (25% burn in) MCMC samples for fossil calibration reported previously. We used a float prior and a maximum bound age, with a scale parameter of c=2. Allowing the MCMC to explore a wide range of time for the divergence between *Babyrussa* and *Suinae* and calibrate the time later than the split of 'new world' peccaries (tU=2 [20 Mya], p=0.1, c=2). For MRCA of sub-Saharan African suid and *Sus*, we used the same fossil calibration as in Frantz et al. 2013 (tL=0.55 [5.5 Mya], p=0.9, c=0.5). For MRCA of *Sus*, we used a minimum bound at 2 Ma [tL=0.2 [2 Mya], p=0.1, c=0.5] to represents the earliest appearance of *Sus* in the fossil record of Island Southeast Asia.

Detecting gene flow among *Suidae*

We integrated Patterson's D-statistic to examine the phylogenetic distribution of derived alleles at loci that display either an ABBA or BABA allelic configuration across the chromosomes among *Suidae* using warthog as an outgroup. For admixture estimation, we assigned 18 autosomes and selected a block size of 5Mb to calculate the standard errors on D-statistics using Admixtools (Patterson et al., 2012). We also identified candidate introgression loci using D-statistic and fd in slide window (Rosenzweig et al., 2016a). To avoid D returning inflated values in

small genomic regions (Martin et al., 2015), we set the window size as 100Kb and summarized the results in Venn diagrams.

SAGUARO

Phylogenetic relationships of genomic regions may differ from the species tree due to incomplete linage sorting and introgression. To test whether this is the case in our analysis, and to detect breakpoints between genomic segments supporting different local topologies, we used the machine-learning approach implemented in SAGUARO (Zamani et al., 2013). We first ran SAGUARO with 6 representative individuals for the overview of whole genome (See Supplementary data 3). Then, to better estimate length of phylogenetic incongruent regions, we performed the same approach but using the quadruplets in sliding-window D-statistic analysis. We constrained SAGUARO to use only nucleotide positions with no missing data and ran with 20 iterations and 500 neurons.

Fitting models of population history

We used qpGraph (Patterson et al., 2012) to fit admixture graphs to six populations representing European wild boar, North Chinese wild boar, South Chinese wild boar, ISEA, pygmy hog and warthog as the outgroup. We filter the dataset using following criteria: SNPs with at least 10Kb distance from one another, no more than 10% missing data. This resulted in 361,837 SNPs. To explore the space of all possible admixture graphs, we used a heuristic search algorithm first described in Ní Leathlobhair et al. 2018 (Leathlobhair et al., 2018) (code available at https://github.com/ekirving/qpbrute). Given an outgroup with which to root the graph, a stepwise addition order algorithm was used for adding leaf nodes to the graph. At each step, insertion of a new node was tested at all branches of the graph, except the outgroup branch. Where a node could not be inserted without producing f4 outliers (i.e. $|Z| >= 3$) then all possible admixture combinations were also attempted. If a node could not be inserted via either approach, that sub-graph was discarded. If the node was successfully inserted, the remaining nodes were recursively inserted into that graph. All possible starting node orders were attempted to ensure full coverage of the graph space. We fitted 2,444 unique admixture graphs for these 6 populations and recorded the 37 graphs that left no f4 outliers (i.e. $|Z| < 3$). We then used the MCMC algorithm implemented in the ADMIXTUREGRAPH R package (Leppälä et al., 2017) to compute the marginal likelihood of these 37 models and their Bayes Factors (BF). We ran two independent replications, each with 2 million iterations, 5 heated chains, a burn in of 50%, and no thinning. Convergence was assessed by calculating the potential

scale reduction factor (PSRF) for the model likelihoods using the CODA R package (Best and Cowles, 1997). We found one particularly well supported model (1d9676e) which, when compared to all others, had K < 119 (Supplementary Fig. 20-21). We also note that among the remaining models there are small differences in the admixture topologies, however, most models support gene-flow between ISEA *Sus* and Chinese pigs, as well as between the pygmy hog and basal wild boar (Supplementary Fig 2.21).

We further carried out our demographic analysis is based on the G-PhoCS, were applied to 10,000 loci of 1 kb of length chosen via a series of filter to obtain putatively neutral loci. We filtered out exons of protein coding genes and 10 kilobases (kb) flanking them on each side, as well as conserved noncoding elements (CNEs) and 100 bases on each side of these elements. We selected 1 kb loci located at least 30 kb apart. We identified a collection of 11,274 loci that followed these criteria, and random selected 10,000 loci for the G-PhoCS. Multiple sequence alignments for these loci were extracted using sequence data from the all individual genomes. We conditioned inference on the population phylogeny based upon the neighbor-joining tree constructed with MEGA (Kumar et al., 2016) on the basis of the IBS distance matrix data of neutral loci used in G-PhoCS analysis calculated by PLINK 1.9 (Chang et al., 2015) (Supplementary Fig. 24). The prior distributions over model parameters was defined by a product of Gamma distributions with $\alpha = 1$ and $\beta = 10{,}000$ for population size and divergence time scaled by mutation rate, and $\alpha = 0.002$ and $\beta = 0.00001$ for the migration rates. Markov Chain was run for 100,000 burn-in iterations, after which parameter values were sampled for 200,000 iterations every 10 iterations, resulting in a total of 20,001 samples from the approximate posterior. Convergence was inspected manually for each run (effective sample size (ESS) for all parameters >200). We converted probabilities into rates using the formula $p=1-e^{-m}$ (where p is the probability of gene flow and m is the total migration rate) (VonHoldt et al., 2016). We checked for convergence between runs using Tracer v1.7 (Rambaut et al., 2018).

Finally we also used Treemix v1.12 (Pickrell and Pritchard, 2012) to test models of possible admixture for Babyrussa, warthog, pygmy hog, ISEA, European pigs, Northern China pigs and Southern China pigs. Windows with 500 consecutive SNPs were used to account for the non-independence of SNPs located in close vicinity. Migrations from m0 to m10 were tested, with five replicates per m to assess consistency.

Probability of introgression fragment from shared ancestral lineage

Based on the equation described in Huerta-Sánchez et al (Huerta-Sánchez et al., 2014), we calculate the probability of a haplotype shared by pygmy hog and wild boar as a result of incomplete ancestral lineage sorting. Briefly, let k be the introgressed haplotype length of the two species' branches since divergence. The expected length of a shared ancestral sequence is $L = 1/(r \times t)$, where r is the recombination rate and t is branch lengths of pygmy hog and wild boar since divergence. The probability of a length of at least k is 1-GammaCDF(k, shape=2, L=1/L), where GammaCDF is the Gamma distribution function. The lower estimate of 4.2 My of the Sus – pygmy hog was used as branch length and an assumed generation time of 5 years. The recombination rate was set to 0.8 cM/Mb (Tortereau et al., 2012).

Another approach to assess probability of ILS is comparing D-statistics between populations under simulations of demographic model. However, it requires a very detailed and precise demographic model to obtain a better assessment. The historical demographic information of pygmy hog and the ancestral population of Suinae species are still deficient. Here, we can only fit in a simplified model. Inaccuracy of the effective ancestral population size and bottleneck event may lead to over-/underestimation of ILS. With this in mind, we compared D-statistics with the same quadruplet as we used in the sliding-window steps under simulations of a simple demographic model with no gene flow. (see Supplementary Note for details). All simulations resulted in $P < 0.001$ against ILS for all comparisons (see Supplementary Fig 2.25 and Supplementary Note).

Functional annotation of genes in introgressed regions

We applied a functional annotation analysis using PANTHER v.11 (Westbury et al., 2018) on the candidate introgressed genes. Genes from the pygmy hog/Sus scrofa introgression and the Sus scrofa/ghost lineage introgression were analyzed separately. Gene-enrichment analyses were performed using clusterProfiler(Yu et al., 2012). False discovery rate (FDR) was performed to adjust P-values using the Benjamini and Hochberg method. A P-value of <0.05 was used as the cut-off criterion.

Data Availability

The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information files, or from the corresponding author upon request.

Accession codes

Raw reads of pygmy hog and *Babyrousa babyrussa* have been deposited in the European Nucleotide Archive (ENA) under accession PRJEB30129. Sequences for the sequenced *Sus scrofa* and other species have been deposited on the EBI Sequence Read Archive under accession number ERP001813.

Acknowledgements

L.Liu has received financial support from the China Scholarship Council (Grant No. 201707720055). This project was financially supported by a European Research Council grant (ERC-2009-AdG: 249894). Sample of *Babyrousa babyrussa* was kindly provided by Prof. Dr. M Fredholm, University of Copenhagen.

Contributions

M.A.M.G., O.M. and L.L. designed the study. M.A.M.G., O.M., and H.-J.M. initially conceived the project; G.N provided the pygmy hog samples; L.L analyzed the data; Y.L.L and L.L. performed the phylogenetic analyses; E.K.I.-P preformed the qpgraph analyses; L.L., M.B., H.-J.M. and O.M. discussed the results; L.L. wrote the manuscript; M.B., H.-J.M., L.A.F.F., M.A.M.G. and O.M provided valuable suggestion and comments to improve the manuscript.

Description of supplementary Material

For a compact layout, in this thesis I did not include all supplementary material. I presented Supplementary Figures which may help the reader. For sake of coherence, I kept the original number of Supplementary Figures and tables.
Complete supplementary material and supplementary note are available at: https://www.nature.com/articles/s41467-019-10017-2.

Additional file

Correction to: Nature Communications
Published online 2019.
In the original Article, we analyzed 38 genomes from pygmy hog and related suid species. Our analysis identified a signal of introgression between *Sus scrofa* and pygmy hog but also reinforced the idea that there was gene flow between *Sus scrofa* and an extinct (ghost) *Suidae* linage (Ai et al., 2015). In our original analysis, however, for the D-statistics equation in Admixtools (Patterson et al., 2012), we mistakenly interpreted as $D = \frac{ABBA-BABA}{ABBA+BABA}$, while it was in fact $D = \frac{BABA-ABBA}{ABBA+BABA}$. This led to an inversion of the direction of admixture in the original version of the

article. To remain consistent with the rest of the analyses in the article, we rectified the formula of D-statistics as $D = \frac{ABBA-BABA}{ABBA+BABA}$ and updated Fig. 2.2b (***in this thesis we used the corrected figure**). The results after correcting the formula indicate that in fact there is an excess of shared derived alleles between the pygmy hog and Island of Southeast Asian (ISEA) *Sus* (ABBA) and not between the pygmy hog and *Sus scrofa* (BABA). The excess of ABBA can be the result of admixture either between pygmy hog and ISEA *Sus* or between *Sus scrofa* and the archaic ghost linage or both. Methods to detect hybridization such as Patterson's D, fd or Twisst (Green et al., 2010; Martin et al., 2015; Martin and Van Belleghem, 2017), however, are inadequate to distinguish gene-flow events between P1 and P3 (Fig.2, orange arrow) from those between P2 and an archaic ghost lineage basal to P3 (Fig.2, blue arrow).

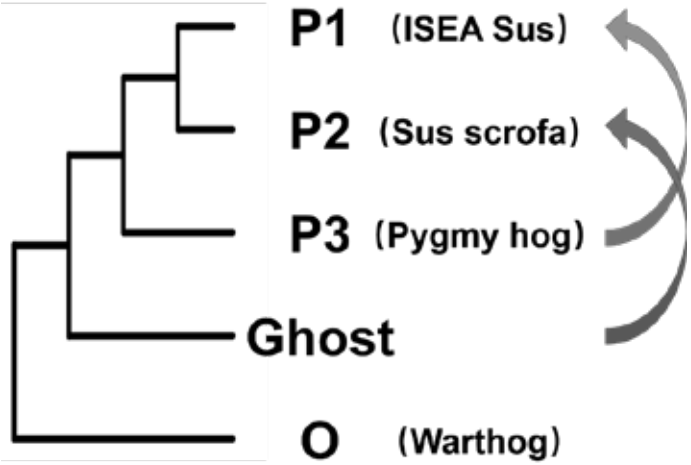


Fig. 2 Schematic representation of the potential gene-flow scenarios discussed in the text. Two-way arrows pointing the two hybridized species.

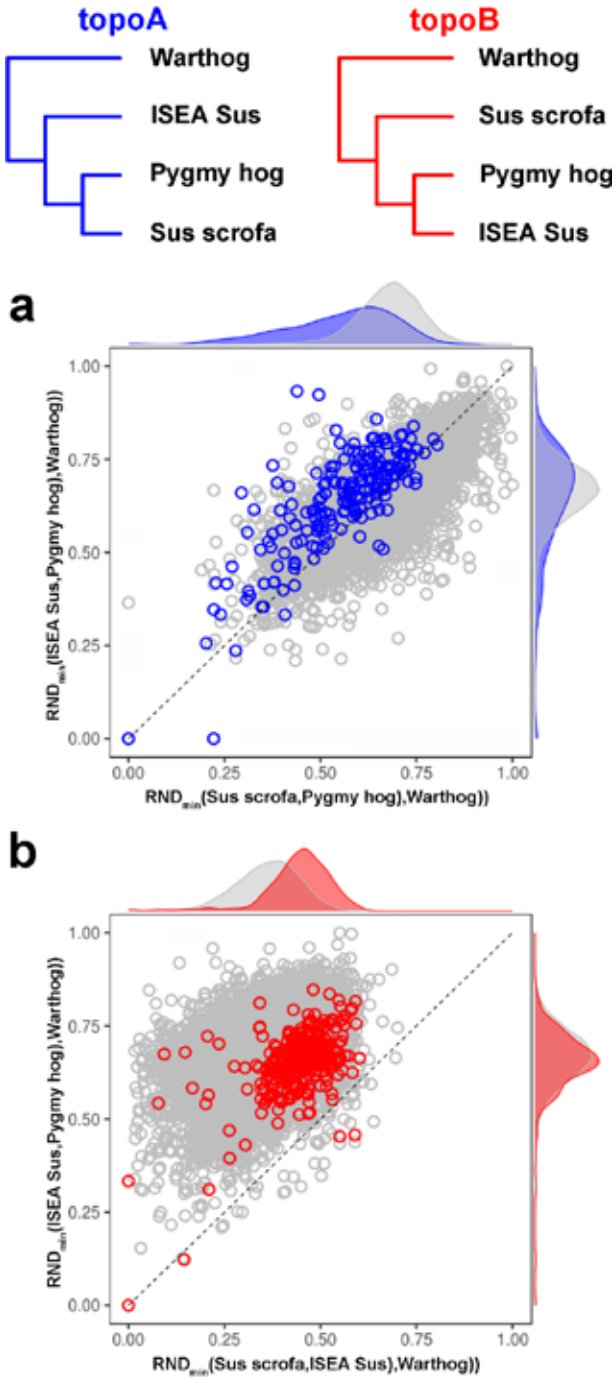
To address this issue an alternative method was used, based on relative nucleotide distance, as implemented in RNDmin (Rosenzweig et al., 2016) to calculate the relative node depth between each taxon. To do so, we first used BEAGLE v4.1(Browning and Browning, 2007) to perform haplotype phasing on genotypes of all individuals (excluding *Babyrousa babyrussa, Potomochoerus larvatus and Potomochoerus porcus*) with default parameters. Then we calculated RNDmin in 100kb sliding windows along autosomes with less than 50% missing sites using PopGenome(Pfeifer et al., 2014). Comparisons were carried out between pygmy

hog, ISEA *Sus* and *Sus scrofa*, using the genome of a warthog as an outgroup. This shows that, for the genomic regions supporting topoA (Fig. 3a), the distribution of RNDmin computed between the pygmy hog and *Sus scrofa* is significantly lower (16.2%, Welch's t-test p-value < 1.8e-14) than RNDmin between pygmy hog and ISEA *Sus*, indicating the first two species share more similarities. We also note that this pattern is more apparent in regions supporting topoA than in the overall genomic background (grey dots in Fig. 3a). Altogether, this indicates that the average genomic distance between the pygmy hog and *Sus scrofa* is smaller than between the pygmy hog and the ISEA *Sus* clade, which supports the existence of gene-flow between pygmy hog and *Sus scrofa*.

If the result of our D-statistics were affected by an admixture between pygmy hog and ISEA *Sus*, we would expect that, in regions that display topoB (FIG. 3b), the RMDmin computed between the pygmy hog and ISEA *Sus* should be lower than in the overall genomic background. Our analysis, however, shows that this is not the case. In fact, we found these distributions to be very similar (Fig. 3b), especially when compared to the result shown in Fig. 3a. In addition, we found the distance between ISEA *Sus* and *Sus scrofa* to be higher (23.2%, Welch's t-test p-value < 2.2e-16) than the overall genomic background in the regions that show topoB, suggesting that topoB was caused by the admixture with an unsampled taxon.

Altogether, our analyses indicate that the findings and interpretations presented in the original article are correct. Firstly, there was no admixture between ISEA *Sus* and pygmy hog. Instead, the combined results of the D-statistics and RMDmin are more consistent with a ghost admixture between a distantly related taxon and *Sus scrofa*, and this inflated the number of shared derived alleles between ISEA *Sus* and the pygmy hog. Secondly, the shorter distance between *Sus scrofa* and pygmy hog suggests that there was also admixture between these two species.

Fig. 3 Phylogenetic tree indicated the alternative topologies (topoA and topoB) to the main species tree. Scatter and density plots below show the distribution of RND_{min} between different comparisons. a) Blue circles represent RND_{min} distribution within windows supporting topoA. b) Red circles represent RND_{min} distribution within windows supporting topoB. Grey circles represent RND_{min} distribution of the 100kb windows among all autosomes. Populations used to calculate RND_{min} is shown on axis lables. Warthog was used as outgroup. (figure on next page)


topoA
Warthog
ISEA Sus
Pygmy hog
Sus scrofa
topoB
Warthog
Sus scrofa
Pygmy hog
ISEA Sus
a
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
RND_min(ISEA Sus,Pygmy hog),Warthog))
RND_min(Sus scrofa,Pygmy hog),Warthog))
b
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
RND_min(ISEA Sus,Pygmy hog),Warthog))
RND_min(Sus scrofa,ISEA Sus),Warthog))

References

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* 47, 217–225. doi:10.1038/ng.3199.

Andrews, P. (1988). Cainozoic Paleontological Sites in Western Kenya - Pickford,M. *J. Hum. Evol.* 17, 273.

Barlow, A., Cahill, J. A., Hartmann, S., Theunert, C., Xenikoudakis, G., Fortes, G. G., et al. (2018). Partial genomic survival of cave bears in living brown bears. *Nat. Ecol. Evol.* doi:10.1038/s41559-018-0654-8.

Best, N. G., and Cowles, M. K. (1997). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output. *MRC Biostat. Unit, Cambridge Univ.* %6, 7–11. Available at: http://oro.open.ac.uk/22547/.

Bouckaert, R. R. (2010). DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* 26, 1372–1373. doi:10.1093/bioinformatics/btq110.

Breeding, C., Of, R., and Endangered, C. CRITICALLY ENDANGERED PYGMY HOG (Porcula salvania).

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi:10.1086/521987.

Chan, K. M. A., and Levin, S. A. (2005). Leaky prezygotic isolation and porous genomes: Rapid introgression of maternally inherited DNA. *Evolution (N. Y).* 59, 720–729. doi:10.1111/j.0014-3820.2005.tb01748.x.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi:10.1186/s13742-015-0047-8.

Chevin, L.-M., Lande, R., and Mace, G. M. (2010). Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *PLoS Biol.* 8, e1000357. doi:10.1371/journal.pbio.1000357.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). JModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* 9, 772. doi:10.1038/nmeth.2109.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375. doi:10.1038/nrg1603.

Fernández, A. I., Muñoz, M., Alves, E., Folch, J. M. arí., Noguera, J. L. ui., Enciso, M. P. ére., et al. (2014). Recombination of the porcine X chromosome: a high density linkage map. *BMC Genet.* 15, 148. doi:10.1186/s12863-014-0148-x.

Fistani, A. B. (1996). Sus scrofa priscus (Goldfuss, de Serres) (Mammalia, Artiodactyla, Suidae) from the Middle Pleistocene layers of Gajtan 1 site, southeast of Shkoder (North Albania). *Ann. Paläontologie* 82, 177–229.

Frantz, L. A. F. (2015). Speciation and domestication in Suiformes: a genomic perspective.

Frantz, L. A. F., Madsen, O., Megens, H. J., Groenen, M. A. M., and Lohse, K. (2014). Testing models of speciation from genome sequences: Divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Mol. Ecol.* 23, 5566–5574. doi:10.1111/mec.12958.

Frantz, L. A. F., Rudzinski, A., Nugraha, A. M. S., Evin, A., Burton, J., Hulme-Beaman, A., et al. (2018). Synchronous diversification of sulawesi's iconic artiodactyls driven by recent geological events. *Proc. R. Soc. B Biol. Sci.* 285, 20172566. doi:10.1098/rspb.2017.2566.

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

Frantz, L., Meijaard, E., Gongora, J., Haile, J., Groenen, M. A. M., and Larson, G. (2016). The Evolution of Suidae. *Annu. Rev. Anim. Biosci.* 4, 61–85. doi:10.1146/annurev-animal-021815-111155.

Funk, S. M., Verma, S. K., Larson, G., Prasad, K., Singh, L., Narayan, G., et al. (2007). The pygmy hog is a unique genus: 19th century taxonomists got it right first time round. *Mol. Phylogenet. Evol.* 45, 427–436. doi:10.1016/j.ympev.2007.08.007.

Geraads, D., Spassov, N., and Garevski, R. (2008a). New specimens of <I>Propotamochoerus</I> (Suidae, Mammalia) from the late Miocene of the Balkans.

Neues Jahrb. für Geol. und Paläontologie - Abhandlungen 248, 103–113. doi:10.1127/0077-7749/2008/0248-0103.

Geraads, D., Spassov, N., and Garevski, R. (2008b). New specimens of <I>Propotamochoerus</I> (Suidae, Mammalia) from the late Miocene of the Balkans. *Neues Jahrb. für Geol. und Paläontologie - Abhandlungen* 248, 103–113. doi:10.1127/0077-7749/2008/0248-0103.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi:10.1126/science.1188021.

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi:10.1038/nature11622.

Groves, C. (1981). *Ancestor for the pigs: taxonomy and phylogeny of the genus Sus*. Dept. of Prehistory, Research School of Pacific Studies, Australian National University.

Guérin, C., and Faure, M. (1997). The wild boar (Sus scrofa priscus) from the post-Villafranchian lower Pleistocene of Untermassfeld. *Das Pleistozän von Untermassfeld bei Meiningen* 1, 375–384.

Guo, Z. T., Peng, S. Z., Hao, Q. Z., Biscaye, P. E., and Liu, T. S. (2001). Origin of the miocene - Pliocene Red-Earth formation at Xifeng in northern China and implications for paleoenvironments. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 170, 11–26. doi:10.1016/S0031-0182(01)00235-8.

HAN, and D.-F. (1975). Quaternary mammalian fossils from Bijiashan, Luizhou, Guangxi. *Vertebr. Palasiat.* 13, 250–256. Available at: https://ci.nii.ac.jp/naid/20000758019/ [Accessed January 23, 2018].

HAN, and D.-F. (1987). Artiodactyla fossils from Liucheng Gigantopithecus cave in Guangxi. *Mem. Inst. Vertebr. Palaeontol. Palaeoanthropology, Acad. Sin.* 18, 135–208. Available at: https://ci.nii.ac.jp/naid/20000758018/ [Accessed January 23, 2018].

Hansen, J., Sato, M., Russell, G., and Kharecha, P. (2013). Climate sensitivity, sea level and atmospheric carbon dioxide. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 371, 20120294. doi:10.1098/rsta.2012.0294.

Hardjasasmita, H. S. (1987). Taxonomy and phylogeny of the Suidae (Mammalia) in Indonesia. *Scr. Geol.* 85, 1–68. Available at: http://www.narcis.nl/publication/RecordID/oai:naturalis.nl:317394 [Accessed May 17, 2018].

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. doi:10.1038/nature13408.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, bar030–bar030. doi:10.1093/database/bar030.

Kishida, T., Thewissen, J., Hayakawa, T., Imai, H., and Agata, K. (2015). Aquatic adaptation and the evolution of smell and taste in whales. doi:10.1186/s40851-014-0002-z.

Kolbe, J. J., Larson, A., Losos, J. B., and de Queiroz, K. (2008). Admixture determines genetic diversity and population differentiation in the biological invasion of a lizard species. *Biol. Lett.* 4, 434–7. doi:10.1098/rsbl.2008.0205.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi:10.1093/molbev/msw054.

Larson-Meyer, D. E., Ingold, B. C., Fensterseifer, S. R., Austin, K. J., Wechsler, P. J., Hollis, B. W., et al. (2017). Sun exposure in pigs increases the vitamin D nutritional quality of pork. *PLoS One* 12, e0187877. doi:10.1371/journal.pone.0187877.

Lawson Handley, L.-J., Estoup, A., Evans, D. M., Thomas, C. E., Lombaert, E., Facon, B., et al. (2011). Ecological genetics of invasive alien species. *BioControl* 56, 409–428. doi:10.1007/s10526-011-9386-2.

Leathlobhair, M. N., Perri, A. R., Irving-Pease, E. K., Witt, K. E., Linderholm, A., Haile, J., et al. (2018). The evolutionary history of dogs in the Americas. *Science (80-.).* 361, 81–85. doi:10.1126/science.aao4776.

Leppälä, K., Nielsen, S. V., and Mailund, T. (2017). Admixturegraph: An R package for admixture graph manipulation and fitting. *Bioinformatics* 33, 1738–1740. doi:10.1093/bioinformatics/btx048.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.
Ma, J., Iannuccelli, N., Duan, Y., Huang, W., Guo, B., Riquet, J., et al. (2010). Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. *BMC Genomics* 11, 13. doi:10.1186/1471-2164-11-159.
Made, J. van der (1999). Biometrical trends in the Tetraconodontinae, a subfamily of pigs. *Trans. R. Soc. Edinb. Earth Sci.* 89, 199–225. doi:10.1017/S0263593300007136.
Martin, S. H., Davey, J. W., and Jiggins, C. D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32, 244–257. doi:10.1093/molbev/msu269.
Martin, S. H., and Van Belleghem, S. M. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics* 206, 429–438. doi:10.1534/genetics.116.194720.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110.
Mirarab, S., and Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52. doi:10.1093/bioinformatics/btv234.
Munz, E. D. (2017). Psychotherapie in der Psychiatrie. *Nervenheilkunde* 36, 800–805. doi:10.1007/s13398-014-0173-7.2.
Nakagome, S., Pecon-Slattery, J., and Masuda, R. (2008). Unequal rates of Y chromosome gene divergence during speciation of the family Ursidae. *Mol. Biol. Evol.* 25, 1344–1356. doi:10.1093/molbev/msn086.
Nuijten, R. J. M., Bosse, M., Crooijmans, R. P. M. A., Madsen, O., Schaftenaar, W., Ryder, O. A., et al. (2016). The Use of Genomics in Conservation Management of the Endangered Visayan Warty Pig (Sus cebifrons). *Int. J. Genomics* 2016, 1–9. doi:10.1155/2016/5613862.
Oliver, W. L. R. (2001). Taxonomy and conservation of Asian wild pigs. *Asian Wild Pig News* 1, 3–5.
Orliac, M. J., Pierre-Olivier, A., and Ducrocq, S. (2010). Phylogenetic relationships of the Suidae (Mammalia, Cetartiodactyla): New insights on the relationships within Suoidea. *Zool. Scr.* 39, 315–330. doi:10.1111/j.1463-6409.2010.00431.x.
Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi:10.1534/genetics.112.145037.
Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A. F., Bosse, M., Bastiaansen, J. W. M., et al. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* 14, 449. doi:10.1186/1471-2164-14-449.
Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31, 1929–1936. doi:10.1093/molbev/msu136.
PHCP (2008). Conservation Strategy and Action Plan for Pygmy Hog in Assam EcoSystems-India.
Pickford Senut, B., Hadoto, D., M. (1993). *Geology and Palaeobiology of the Albertine Rift Valley, Uganda - Zaire - Volume I: Geology*. Cifeg.
Pickford, M. (1988). Revision of the Miocene Suidae of the Indian subcontinent. *Muenchner Geowissenschaftliche Abhandlungen R. A Geol. und Palaeontol.* 12, 1–92. Available at: https://scholar.google.com/scholar?q=pickford+sanitheriidae&btnG=&hl=en&as_sdt=0%2C3#1%5Cn%3CGo to ISI%3E://ZOOREC:ZOOR12600006940 [Accessed September 21, 2017].
Pickford, M. (2013). Suids from the Pleistocene of Naungkwe Taung, Kayin State, Myanmar. *Paleontol. Res.* 16, 307–317. doi:10.2517/1342-8144-16.4.307.
Pickrell, J. K., and Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* 8, e1002967. doi:10.1371/journal.pgen.1002967.
Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 00, 1–3. doi:10.1093/sysbio/syy032.
Rhymer, J. M., and Simberloff, D. (1996). EXTINCTION BY HYBRIDIZATION AND INTROGRESSION. *Annu. Rev. Ecol. Syst.* 27, 83–109. doi:10.1146/annurev.ecolsys.27.1.83.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi:10.1093/sysbio/sys029.

Ropiquet, A., and Hassanin, A. (2006). Hybrid origin of the Pliocene ancestor of wild goats. *Mol. Phylogenet. Evol.* 41, 395–404. doi:10.1016/j.ympev.2006.05.033.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., and Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Mol. Ecol.* 25, 2387–2397. doi:10.1111/mec.13610.

Roth, J., and Wagner, J. A. (1854). *Die fossilen Knochenüberreste von Pikermi in Griechenland: Gemeinschaftlich bestimmt u. beschrieben nach d. Materialien, welche durch die von dem Erstgenannten im Winter 1852/3 dortselbst vorgenommenen Ausgrabungen erlangt wurden*. Verlag d. Akad.

Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* 46, 919–925. doi:10.1038/ng.3015.

Seebacher, F., and Post, E. (2015). Climate change impacts on animal migration. *Clim. Chang. Responses* 2, 5. doi:10.1186/s40665-015-0013-9.

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., et al. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* 561, 113–116. doi:10.1038/s41586-018-0455-x.

Smith, J., and Kronforst, M. R. (2013). Do Heliconius butterfly species exchange mimicry alleles? *Biol. Lett.* 9, 20130503–20130503. doi:10.1098/rsbl.2013.0503.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., 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. doi:10.1186/1471-2164-13-586.

van der Made, J., Morales, J., and Montoya, P. (2006). Late Miocene turnover in the Spanish mammal record in relation to palaeoclimate and the Messinian Salinity Crisis. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 238, 228–246. doi:10.1016/j.palaeo.2006.03.030.

von Koenigswald, G. H. R. (1963). Fossil Pygmy Suidae from Java and China. *Proceedings, Ser. B* 66, 192–197.

VonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., et al. (2016). Whole-genome sequence analysis shows that two endemic species of North American Wolf are admixtures of the coyote and gray Wolf. *Sci. Adv.* 2, e1501714–e1501714. doi:10.1126/sciadv.1501714.

Westbury, M. V, Hartmann, S., Barlow, A., Wiesel, I., Leo, V., Welch, R., et al. (2018). Extended and Continuous Decline in Effective Population Size Results in Low Genomic Diversity in the World's Rarest Hyena Species, the Brown Hyena. *Mol. Biol. Evol.* 35, 1225–1237. doi:10.1093/molbev/msy037.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088.

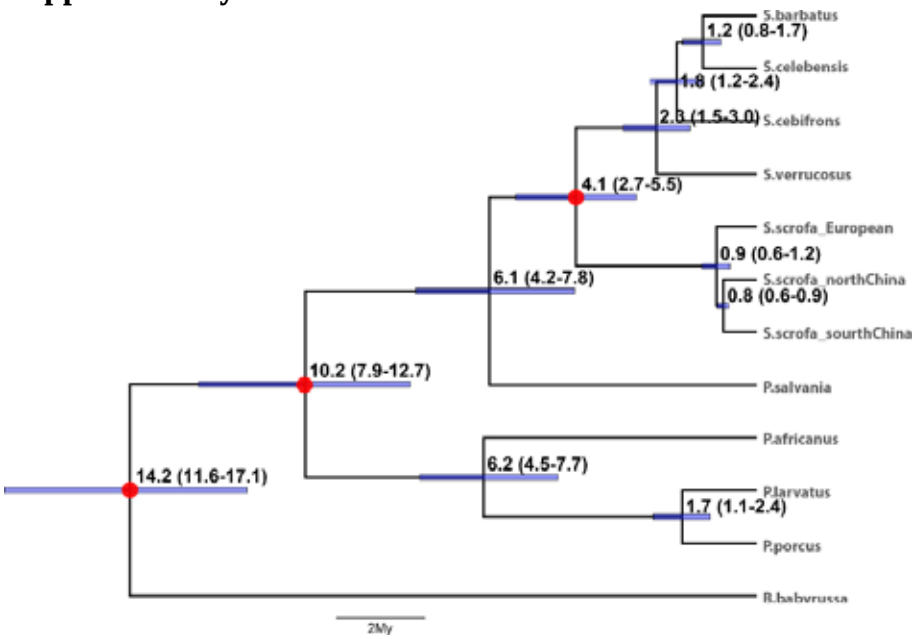
Yannic, G., Dubey, S., Hausser, J., and Basset, P. (2010). Additional data for nuclear DNA give new insights into the phylogenetic position of Sorex granarius within the Sorex araneus group. *Mol. Phylogenet. Evol.* 57, 1062–1071. doi:10.1016/j.ympev.2010.09.015.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118.

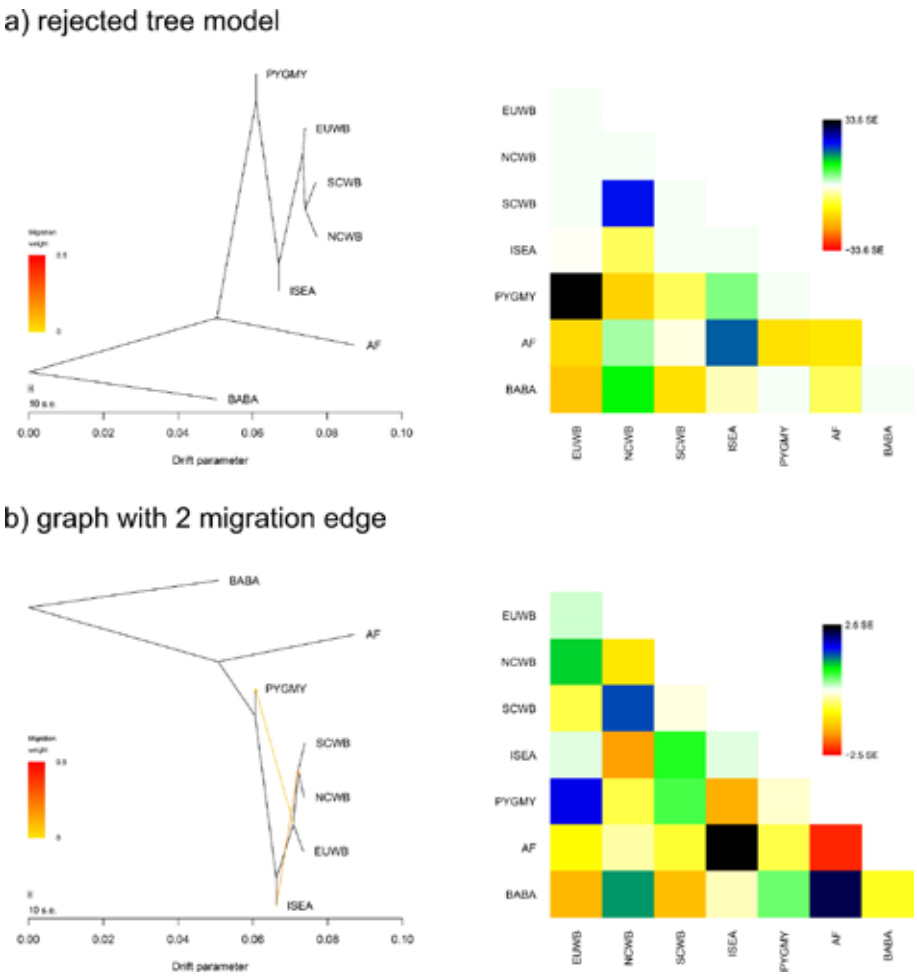
Zachos, J., Pagani, H., Sloan, L., Thomas, E., and Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science (80-.).* 292, 686–693. doi:10.1126/science.1059412.

Zamani, N., Russell, P., Lantz, H., Hoeppner, M. P., Meadows, J. R. S., Vijay, N., et al. (2013). Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics* 14, 347. doi:10.1186/1471-2164-14-347.

Supplementary material



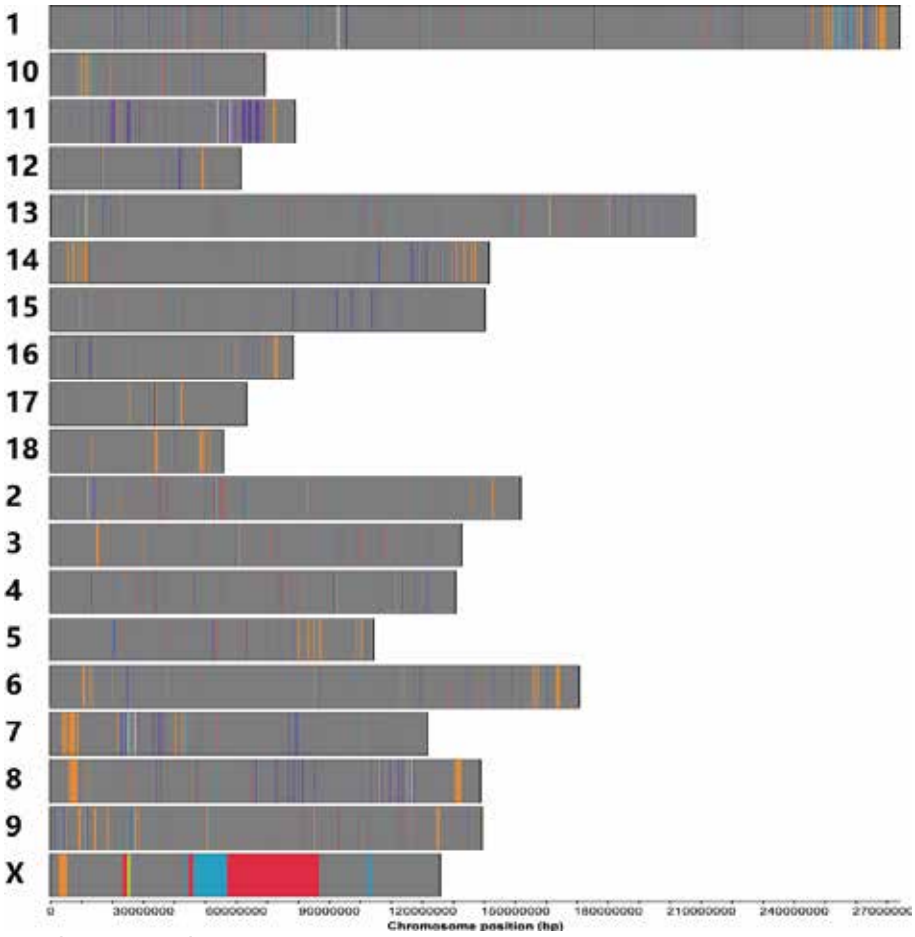
Supplementary Figure 2.4
Phylogenetic time tree
Species tree with mean posterior age (in millions of years). Blue bars indicated 95% confidence intervals. Red dots indicated the calibration points. We used a float prior and a maximum bound age, with a scale parameter of c=2. For the root divergence, we set the prior to (tU=2 [20 Mya], p=0.1, c=2). For MRCA of *Suinae* and *Sus*, we used the same fossil calibration as in Frantz et al. 2013 (tL=0.55 [5.5 Mya], p=0.9, c=0.5 and tL=0.2 [2 Mya], p=0.1, c=0.5, respectively).



Supplementary Fig 2.5

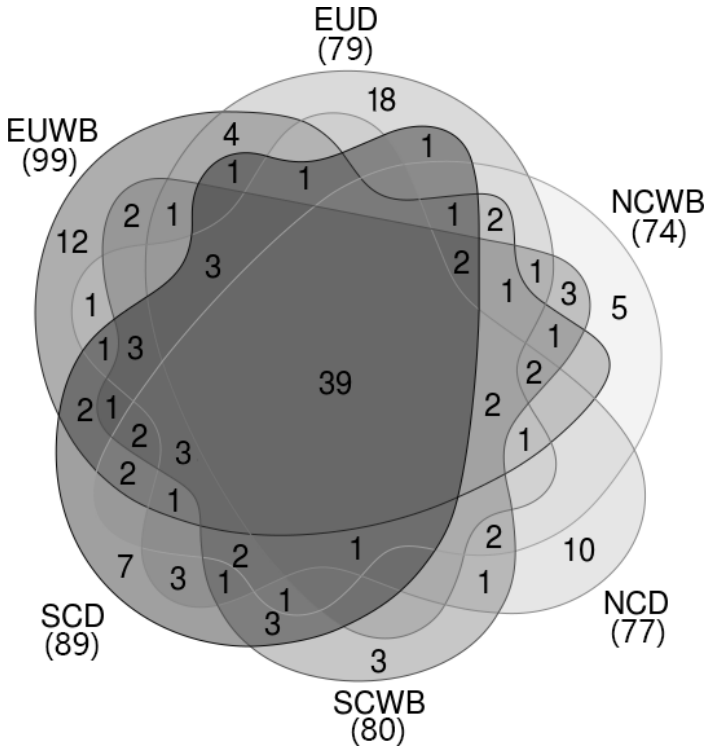
Admixture graph inferred using Treemix

a) A simple tree-like model without admixture fits the data poorly, as can be seen from the matrix of residuals between empirical and modelled allele frequency covariance on the right. b) The optimal placement of two admixture event are from the common ancestor of Eurasian wild boar to pygmy hog, as well as from ISEA Sus to common ancestor of Asian wild boar.



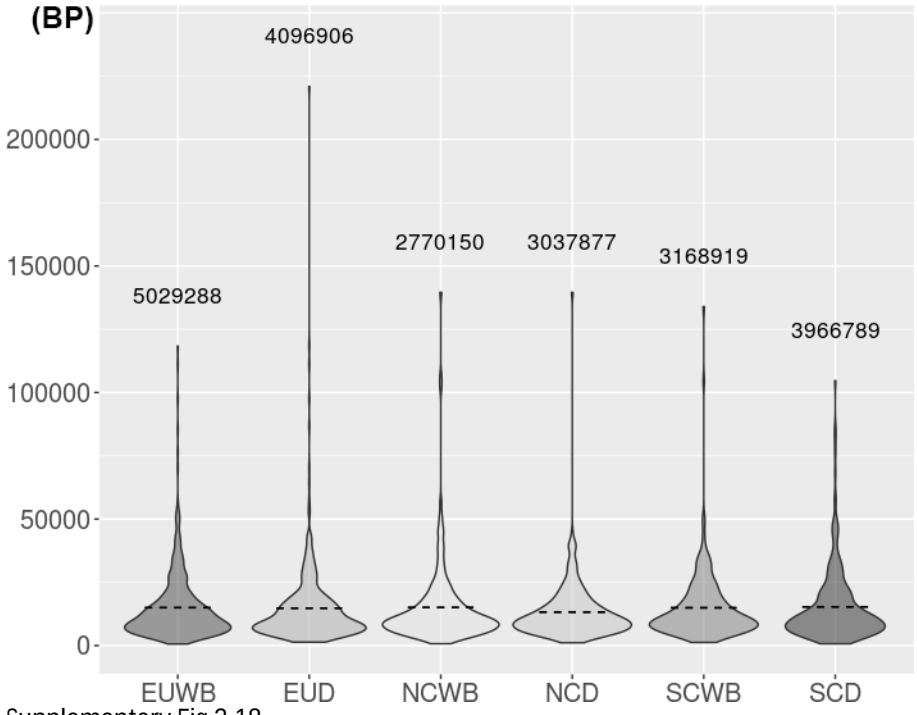
Supplementary Fig 2.9

Distribution of chromosomal segments supporting the 12 most frequent tree topologies assigned by SAGUARO to segments along the autosomes and the X chromosome. The numbers given next to phylogenies indicate chromosome IDs. See Supplementary data 3 for further details.



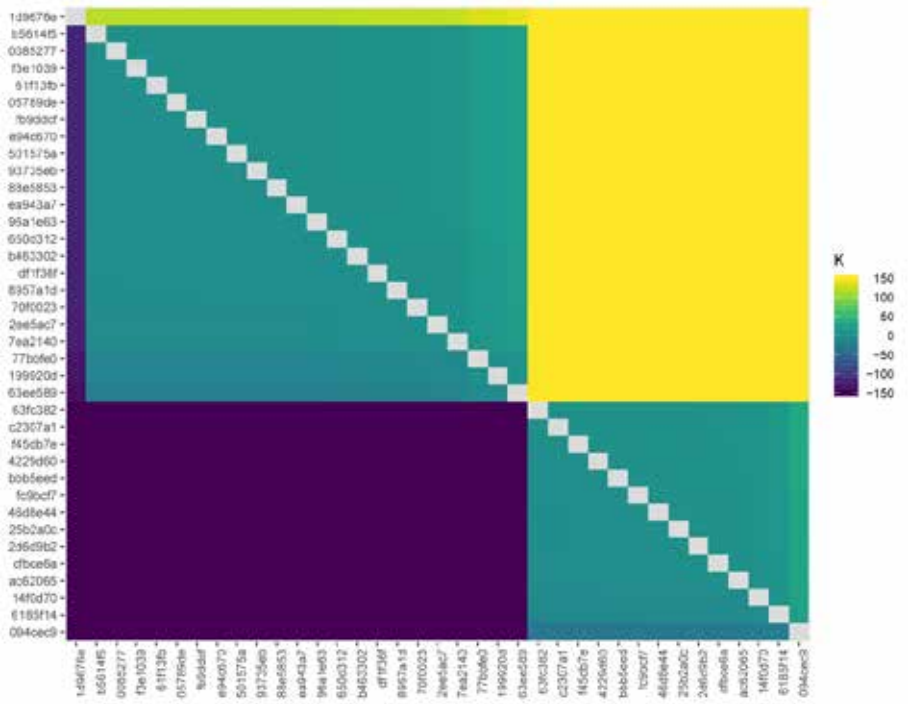
Supplementary Fig 2.17

The Venn diagram shows shared autosomal segments with 'ghost' introgression signal in sliding-window ABBA-BABA analysis among different swine populations. For European pigs, we calculated the average D-value for the clear 'ghost' introgression region on X chromosome and any autosomal windows with D-value lower than this will be regarded as introgression region. Numbers in parentheses refer to the total amount of introgression windows in each population. SCWB = South Chinese wild boar, SCD = South Chinese domesticated pig, NCWB = North Chinese wild boar, NCD = North Chinese domesticated pig, EUWB = European wild boar, EUD = European domesticated pig.



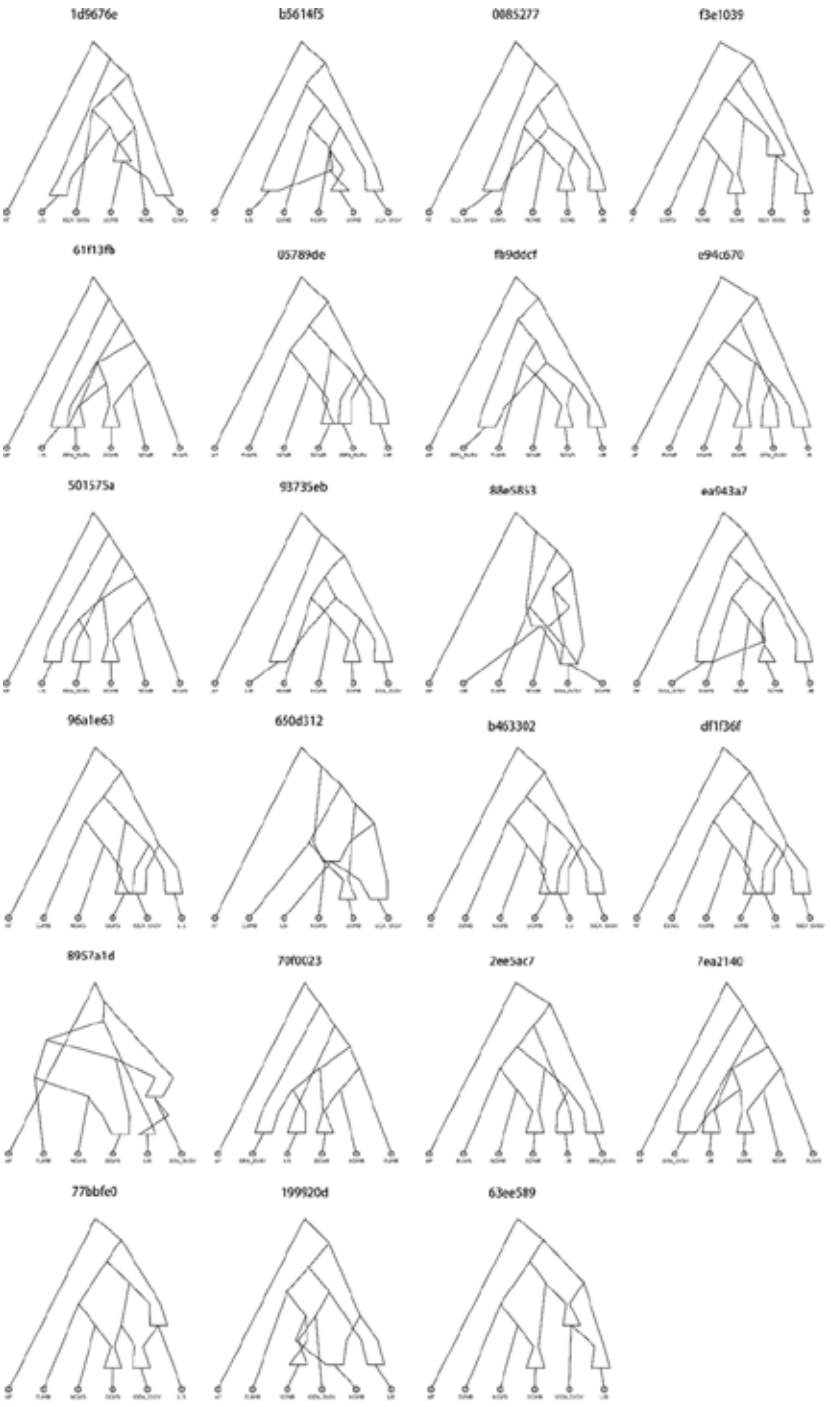
Supplementary Fig 2.18

Violin plot representing the length of autosomal 'ghost' introgression region in different population inferred by Saguaro. Numbers on top of each violin are the total length of introgression region. Dash lines indicated the average length. SCWB = South Chinese wild boar, SCD = South Chinese domesticated pig, NCWB = North Chinese wild boar, NCD = North Chinese domesticated pig, EUWB = European wild boar, EUD = European domesticated pig.

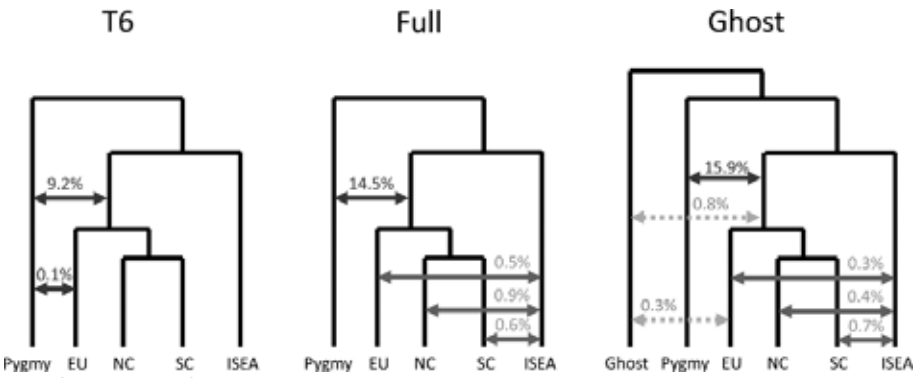


Supplementary Fig 2.20

Heatmap showing the comparation of the Bayes factors for all model pairs. |K|>3 is considered as the significant threshold. Coordinates of the scale indicated the internal code for each model.


1d9676e
b5614f5
0085277
f3e1039
61f13fb
057894e
fb9ddcf
c94c670
501575a
93735eb
88e5853
ea943a7
96a1e63
650d312
b463382
df1f36f
8957a1d
70f0023
2ee5ac7
7ea2140
77bbfe0
199920d
63ee589

Supplementary Fig 2.21
23 most well supported models (based on K) used for ADMIXTUREGRAPH. Strings on top of each diagram indicated the internal code for each model. (AF= *Phacochoerus africanus*; LIB=pygmy hog; ISEA_SVSV=*Sus verrucosus*; EUWB=European wildboar; NCWB=North Chinese wildboar; SCWB=South Chinese wildboar)



Supplementary Fig 2.23
Migration bands are shown in blue, red and gray with associated values indicating estimates of total migration rates, which equal the probability that a lineage will migrate through the band during the time period when the two populations co-occur. Thig figure only show a subset of models tested and only significant migration signals were delineated (See Supplementary data 4 for further details). G-Phocs result is consistent with our admixture analysis between pygmy hog and *Sus scrofa*. In our model with ghost population, European pigs appeared to have more introgression from ghost population than other pigs, which supports the hypothesis that European pig used to hybrid with other species. Notably, G-Phocs always estimated a migration signal from Sus scrofa to other species. A possible reason is that Sus scrofa population remain the biggest effective population size and genetic diversity, which may overweight ILS. ISEA=Island of South Asia pigs, SC = South Chinese pig, NC = North Chinese pig, EU = European pig.

3

Genetic consequences of long-term small effective population size in the critically endangered pygmy hog

Langqing Liu[1*], Mirte Bosse[1], Hendrik-Jan Megens[1], Manon de Visser[1], Martien AM Groenen[1], Ole Madsen[1*]

[1] Animal Breeding and Genomics, Wageningen University & Research, The Netherlands

Evolutionary Application
doi:10.1111/eva.13150

Abstract

Increasing human disturbance and climate change have a major impact on habitat integrity and size, with far-reaching consequences for wild fauna and flora. Specifically, population decline and habitat fragmentation result in small, isolated populations. To what extend different endangered species can cope with small population size is still largely unknown. Studies on the genomic landscape of these species can shed light on past demographic dynamics and current genetic load, thereby also providing guidance for conservation programmes. The pygmy hog (*Porcula salvania*) is the smallest and rarest wild pig in the world, with current estimation of only a few hundred living in the wild. Here, we analyzed whole genome sequencing data of six pygmy hogs, three from the wild and three from a captive population, along with 30 pigs representing six other Suidae. First, we show that the pygmy hog had a very small population size with low genetic diversity over the course of the past ~1 million years. One indication of historical small effective population size is the absence of mitochondrial variation in the six sequenced individuals. Second, we evaluated the impact of historical demography. Runs of homozygosity (ROH) analysis suggests that the pygmy hog population has gone through past but not recent inbreeding. Also, the long-term, extremely small population size may have led to the accumulation of harmful mutations suggesting that the accumulation of deleterious mutations is exceeding the purifying selection. Thus, care has to be taken in the conservation programme to avoid or minimize the potential for inbreeding depression, and guard against environmental changes in the future.

Key words: conservation genomics, inbreeding, deleterious variants, population genomics

3.1 Introduction

During the last glacial maximum, the ranges of most temperate species shifted and shrunk as temperatures decreased (Davis and Shaw, 2001). During the Holocene, human populations expanded rapidly, and negatively affected biotic recoveries and natural range expansions through both hunting and land clearing (Ellis, 2015). Thus, the combined effects of climate changes and human activities have reduced population sizes of many species throughout the world to a critically small size over the past 10,000 years (Miraldo et al., 2016; Pimm and Raven, 2017).

Small, fragmented, and isolated populations lead to reduced genetic variation and increased inbreeding and genetic drift (Lynch et al., 1995, 2016). Inbreeding can have a negative effect on population viability through inbreeding depression, which is a consequence of an increase of harmful mutations in the homozygous state of inbred individuals (Pekkala et al., 2012, 2014; Kardos et al., 2017). In extremely small populations, genetic drift tends to prevail over natural selection, limiting the potential for purifying selection against deleterious variation, and even allowing deleterious variants to increase in frequency (Funk et al., 2016; Lynch et al., 2016). Importantly, low variety of genetic variation is expected to reduce the opportunities for selection and to limit adaptive potential in populations that experience rapid environmental changes, e.g. new diseases and climate fluctuation (Piertney and Oliver, 2006; Hamilton and Miller, 2016).

Studies on demographic history and erosion of genomic variation of endangered populations, can show the impact of losing genomic diversity and accumulation of genetic load. For instance, in the endangered Cheetah (*Acinonyx jubatus*) population, long term decline and subsequent bottlenecks have resulted in excessive deleterious mutations, reducing reproductive success (Merola, 1994; Dobrynin et al., 2015). However, not all populations with low genetic diversity suffer from inbreeding depression. Similar patterns of long-term decline are apparent in the genomes of island foxes, which resulted in extensive runs of homozygosity and increased genetic load. Yet, the lack of apparent phenotypical defects suggests that deleterious variants were purged from the island fox population in parallel with further adaptation to the local environment (Robinson et al., 2016, 2018). It is, therefore, important to understand demographic history as well as temporal changes in mutational load in small, fragmented populations in order to predict the impact of inbreeding and increase the chances of long-term population persistence.

The pygmy hog (*Porcula salvania*) is the smallest and the rarest wild suid in the world, and so far known as the sole living representative of the genus *Porcula*. The pygmy hog has been classified as a critically endangered species by the

International Union of Conservation of Nature (IUCN) since 2008. The pygmy hog is confined to the tall grass savanna of the Himalayan foothills. Since the early 20th century, human settlement and agriculture led to accelerated fragmentation and loss of pygmy hog habitat (Peet et al., 1999). The pygmy hog was believed to be extinct in most of its natural range in the Terai and Duars region (Oliver and Deb Roy, 1993) until they were rediscovered in 1971. Currently, only one viable wild population remains, in Manas National Park, northern Assam, India. Considering its critical status and the unique habitat it lives in, a recovery programme for this species, the Pygmy Hog Conservation Programme (PHCP), was initiated in 1995 (PHCP, 2008). Starting with six wild caught hogs, the breeding programme exceeded early expectations. The captive population is now around 80 (Leus, 2017). Although the PHCP has benefited from several decades of planned breeding and pedigree management, so far there has been no information on the genetic diversity in the individuals that were used to establish the breeding programme. This information is essential to inform the breeding programme to prevent inbreeding issues.

It is still largely unknown whether the small population size has experienced purifying selection of harmful mutations and whether current inbreeding leads to inbreeding depression. To infer the demographic history, and eventual inbreeding concerns, we studied whole genome data of six pygmy hogs: three from the wild and three from the breeding programme. By combining the pygmy hog information with 30 pigs belonging to six other old-world pig species (Sup Table 3.1), we interpreted our findings in the context of these other pig species, whose demographic history has been well studied. For example, the critically endangered, Javan warty pig (*Sus verrucosus*), which is highly inbred due to recent zoo management (Semiadi and Meijaard, 2006). And even far more widespread species, such as the European wild boar (*Sus scrofa*), have experienced profound population bottlenecks, due to glaciations and, historically, hunting and habitat loss (Groenen et al., 2012).

In this study we aim at using a comparative genomics approach to infer past population dynamics and assess the consequences of severe population decline. Our results provide a detailed genomic estimation of the pygmy hog's population history, genomic diversity, inbreeding status and genetic load. These results provide a strong foundation in evaluating the conservation status of the pygmy hog and highlighting the importance of genomic monitoring in population management of pygmy hogs and other endangered species, both *in situ* and *ex situ* conservation programmes.

3.2 Materials and methods

Whole-genome resequencing and variant calling and filtering

The pygmy hog samples used for this research are derived from three wild and three captive individuals. On these samples, whole genome Illumina PE 100 bp re-sequencing was performed at SciGenom Laboratories in Chennai, India. A selection of other *Suidae* species was included (Sup Table 3.1). All these samples were also sequenced with the Illumina sequence technology. The whole genome sequencing data were trimmed using sickle (Version 1.33, https://github.com/najoshi/sickle) with default parameters. The trimmed reads were aligned to the Sscrofa 11.1 reference genome. Since there are multiple closely related species to the reference species, we used the unique alignment option of MOSAIK aligner (Version 2.2.30) (Lee et al., 2014) to increase mapping accuracy (Pightling et al., 2014). Local re-alignment was performed using GATK (Version 3.7) RealignmentTargetCreator and IndelRealigner and variants were called using GATK UnifiedGenotyper (McKenna et al., 2010), with the –stand_call_conf option set to 50, the –stand_emit_conf option set to 20, and the -dcov option set to 200. Variants with a read-depth between 0.5 and 2.0 times of the average sample genome coverage were selected and stored in variant calling format (Sup Table 3.1).

Mitochondrial genome assembly and analysis

As no pygmy hog mitochondrial sequence was available, we reconstructed one, using the short-read data from the high-coverage individual (Sup Table 3.1). We assembled the mitochondrial genome through iterative mapping using MITObim v1.8 (Hahn et al., 2013) on 100 million trimmed and merged reads, subsampled using seqtk (version 1.3 r106), https://github.com/lh3/seqtk). Mitochondrial reconstruction was performed in three independent runs using three different starting bait reference sequences. The references included the domestic pig (AF034253.1), common warthog (DQ409327.1), and cattle (AY526085.1). We implemented MITObim using default parameters apart from mismatch value where we used zero. We resolved the circularity of mitochondrial DNA using the published control region sequences (Funk et al., 2007). All three independent MITObim assembly runs produced identical pygmy hog mitochondrial sequences, providing strong evidence that our reconstructed mitochondrial genome is correct. The reconstructed mitochondrial genome served as a reference sequence for subsequent mitochondrial DNA mapping analyses. We mapped the trimmed and merged reads from our 6 pygmy hogs to the reconstructed reference sequence using BWA-mem (version 0.7.15) (Li and Durbin, 2009) using default parameters and parsed the mapped files using Samtools (version 0.1.19-44428cd) (Li et al.,

2009). Local re-alignment was performed using GATK RealignmentTargetCreator and IndelRealigner and variants were called using GATK UnifiedGenotyper (McKenna et al., 2010), with the –stand_call_conf option set to 50, the –stand_emit_conf option set to 20. The consensus sequences were constructed using ANGSD (version 0.929) (Korneliussen et al., 2014). Mitochondrial genome sequence was aligned and analyzed using MEGA7 (Kumar et al., 2009).

Genetic diversity

Nucleotide diversity was calculated for bins of 10 kbp over the entire genome within each individual, following the description in Bosse et al. (2012b) . Nucleotide diversity was represented by "SNPbin". ''SNPbin'' is the SNP count per 10 kbp window, corrected for the number of bases within that bin that was not covered after the read-depth filtering, so that the eventual SNP count per bin (SNPbin) is proportional to 10,000 covered bases. SNP count is the total number of SNPs counted in a bin of 10 kbp. We assessed genetic diversity by calculating heterozygosity for each SNPbin, here defined as the number of heterozygous genotypes divided by the number of called sites within a single individual. Heterozygosity was calculated for the entire autosomal genome and in 100 kb sliding windows with a 10 kb step size. Windows with more than 20% of the sites failing the quality filters, or with fewer than 20 kb of confidently called sequence were excluded. Peaks of heterozygosity within a genome were defined as windows with heterozygosity greater than two standard deviations above the mean, based on the genome-wide distribution of per-window heterozygosity. Overlapping windows of high heterozygosity were merged using BEDTools (version 2.28.0) (Quinlan, 2014).

Runs of homozygosity (ROH) analysis

For homozygosity analysis, we calculated the runs of homozygosity (ROH) to estimate autozygosity for the sequenced individual. ROH for an individual were calculated based on the following criteria specified in Bosse et al. (2012b) and using the python script specified in Bortoluzzi et al. (2019). This included the number of SNPs, in a window size of 10 Kb, counted below 0.25 times the average whole-genome SNP count; and the homozygous stretches contained at least 10 consecutive windows which showed a total SNP average lower than the genomic average. Sufficiently covered windows with 0.5–2 times the individual average depth was considered. The relaxed threshold for individual windows were used within a homozygous stretch to avoid local assembly or alignment errors, which was done by allowing for maximum twice the genomic average SNP count, and the

average SNP count within the candidate ROH to not exceed 1/4 the genomic average. The inbreeding coefficient derived from ROH genomic coverage (FROH) was calculated by dividing total ROH length per individual by total genome length across all autosomes (~2.4Gb) for each individual.

Variant Annotation

All variants were annotated using Variant Effect Predictor (ensembl-vep version 91.1) (Gentleman, 2015), with "--species sus_scrofa --fork 4 --canonical --stats_text --gene_phenotype --numbers --domains --symbol --buffer_size 100000 --offline --force_overwrite --vcf --sift b". Functional significance of amino acid substitutions was predicted using SIFT (Kumar et al., 2009). Putative deleterious mutation was further evaluated by pig Combined Annotation Dependent Depletion (pCADD) (Groß et al., 2020).

Functional, pathway and interaction enrichment analysis

ClusterProfiler (version 3.6.0) (Yu et al., 2012) was applied to perform GO analysis (including cellular composition, molecular function and biological process terms) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. False discovery rate (FDR) was performed to adjust P-values using the Benjamini and Hochberg method. A P-value of <0.05 was used as the cut-off criterion.

PSMC analysis

To derive an estimate of the historic effective population size, a Pairwise Sequential Markovian Coalescence (PSMC version 0.6.4-r49) model was used (Li and Durbin, 2011). This software uses the time to most recent common ancestor of a diploid genome (determined by looking at the density of heterozygotes) to estimate the effective population size (Ne) in the past. The individual whole genome consensus sequence, called by SAMtools (Li et al., 2009), was used as an input for this analysis. We used a generation time of five years (in concordance with the studbook files that showed a generation time for the captive population of 4.5 years) and a default mutation rate / generation of 2.5×10^{-8}. The following parameters were used: Tmax = 20; n = 64 ('4+50*1+4+6').

Forward simulations of genetic variation in pygmy hog population

To evaluate the demographic parameters that could lead to a purging of genetic load, we performed forward-time simulations as described in Robinson, et al (Robinson et al., 2018). We simulated neutral and deleterious variation under a

constant population size, involving the establishment of a small population (N = 100, 200, 300, 500 or 1000 individuals) derived from a large ancestral population (N = 10000 individuals). Based on the PSMC results, the pygmy hog kept a low and stable population size from 100~200kya. We assumed the generation time to be 5 years. The small populations were randomly sampled from the ancestral population and kept at constant size for 20,000 or 40,000 generations. Each simulated individual consisted of a diploid 2Mb genome, consisting of 2000 ''genes'' carried on 18 chromosomes proportional to chromosome lengths in the pig genome. A mutation rate of $2{,}5 \times 10^{-8}$ and recombination rate was set to 0.8 cM/Mb (Tortereau et al., 2012). Each model was run for 20,000 and 40,000 generations following a 100,000 generation burn-in period. 100 replicates were performed for each dominance value and each population size. The average number of alleles and the average number of homozygous alleles carried by each individual were calculated for deleterious ($s < 0$) and neutral mutations ($s = 0$). Deleterious mutations were grouped as strongly ($s < -0.01$), moderately ($-0.01 < s < -0.001$), and weakly deleterious ($-0.001 < s < 0$). One-way ANOVA and Tukey HSD post hoc tests were used to evaluate significant differences in the number of total alleles and the number of homozygous alleles between different models.

3.3 Result

Relatedness between pygmy hog samples

According to pedigree information provided by the breeding programme (Figure 3.1, Sup Table 3.2), the three captive individuals were representatives of the third and fourth generations of the captive population. Two of the captive individuals (PYGMY2 and PYGMY3) are maternal half-sibs. Assuming that all wild founders are not from the same family, no related mating caused by the breeding scheme was observed within the breeding programme. Among all six pygmy hog samples, only a fraction, around 10%, of these variants was specific to either wild or captive individuals, and no significant difference between the number of SNPs between wild and captive animals was observed.

Due to being maternally inherited and lack of recombination, variation in mitochondrial genomes can provide unique insight into population structure. We assembled the complete mitochondrial genome from the wild caught individual with the highest read depth (Sup Table 3.1). Next, we mapped the reads from the six individuals to the assembled mitochondrial sequence to assess the mitochondrial variation in our sequenced pygmy hogs. We observed that all six pygmy hogs carried identical mitochondrial genomes.

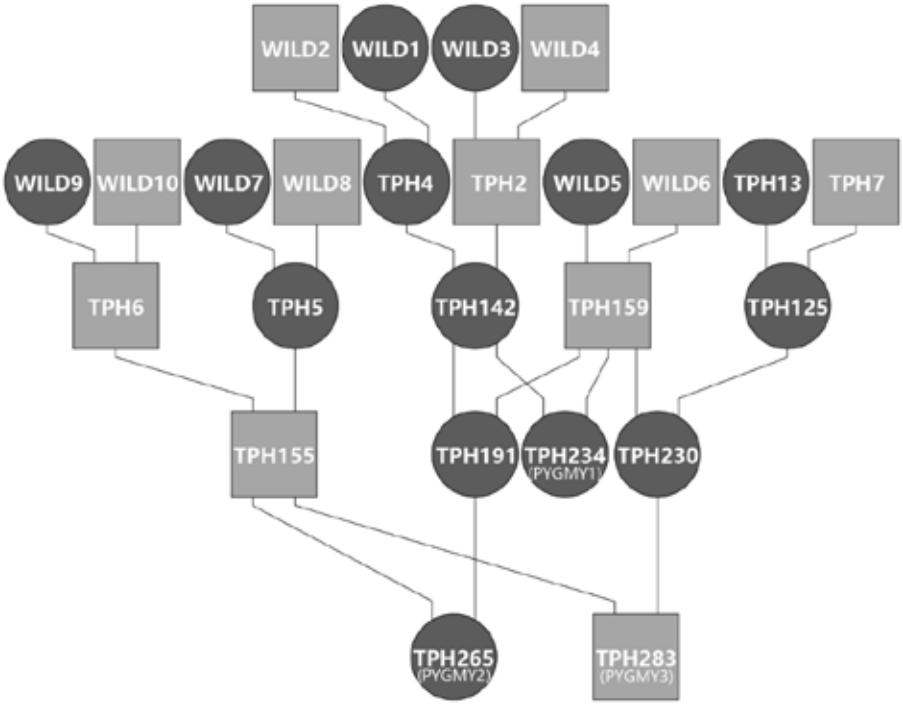


Figure 3.1 Studbook information. Part of the pedigree of the captive pygmy hog populations in the wild and the captive program reconstructed from the studbook files. Founder individuals were indicated as "WILD". PYGMY1 (born 2007, sampled 2014); PYGMY2 (born 2008, sampled 2014); PYGMY3 (born 2009, sampled 2014).

Genome-wide diversity, inbreeding and demographic history

We compared the genome-wide autosomal nucleotide diversity between the pygmy hog and the other *Sus* species. Overall, genome-wide nucleotide diversity (π) of the pygmy hog is much lower (3.33 ± 1.36) than for all the other *Suidae* species, which are less threatened (15.15 ± 10.30). This number is even lower than what is observed in *Sus verrucosus* (4.82 ± 5.03) or European wild boar (8.42 ± 7.50) (Figure 3.2).

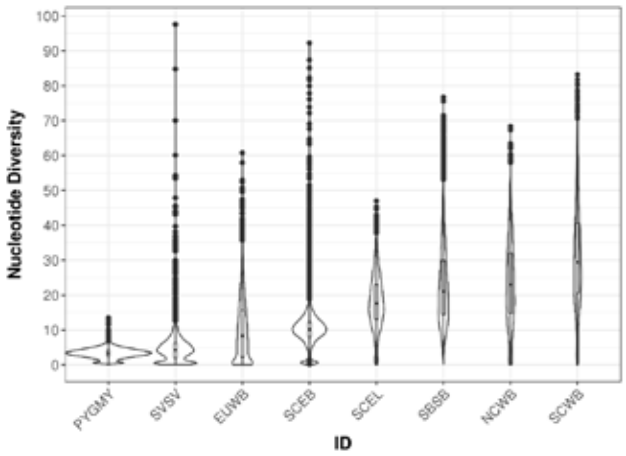


Figure 3.2 Nucleotide Diversity (*10^{-4}bp) in the sampled populations. (PYGMY = pygmy hog, SBSB = Sus barbatus, SCEB = Sus cebrifons, SCEL = Sus celebensis, SVSV = Sus verrucosus, EUWB = European wild boar, NCWB = Northern China wild boar, SCWB = Southern China wild boar)

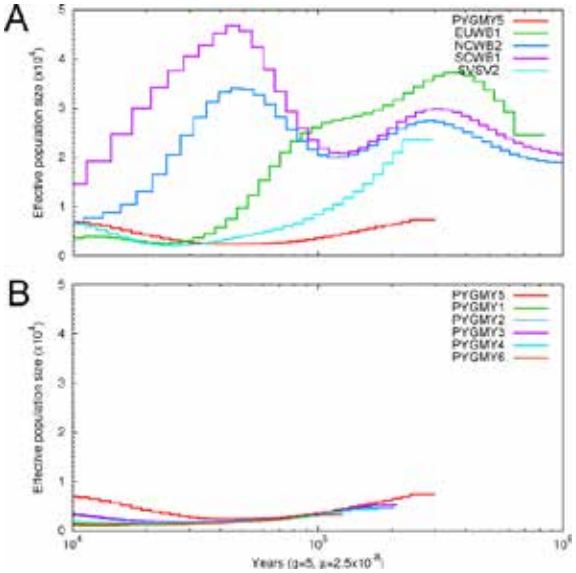


Figure 3.3 Demographic history of pygmy hogs compared to other pig species. Demographic history was inferred using a hidden Markov model (HMM) approach as implemented in pairwise sequentially Markovian coalescence (PSMC). (A). Comparison between pygmy hog and other Sus species. (B). Historical effective population size for all pygmy hog individuals. (PYGMY = pygmy hog, SBSB = Sus barbatus, SVSV = Sus verrucosus, EUWB = European wild boar, NCWB = Northern China wild boar, SCWB = Southern China wild boar, detailed sample abbreviations see Sup table 3.1.)

The mean heterozygosity per 10kb window across the autosomal genome showed that the pygmy hog has very low levels of autosomal heterozygosity (Sup Table 3.3). The distribution of heterozygous peaks in pygmy hogs shows that 87.4% are shared by all individuals (Sup Figure 3.1a). These conserved peaks of heterozygosity are strongly enriched for olfactory receptor (OR) genes (Sup Figure 3.1b). It is well known that OR gene repertoires evolve rapidly through gene duplication, pseudogenization, and loss in other pig species and mammals (Paudel et al., 2013). It is likely that a large fraction of these conserved heterozygosity peaks is ambiguous and caused by copy number variation of OR gene families. We therefore excluded regions with OR genes, which results in a lower heterozygosity distribution (Sup Figure 3.1c). The 30 genes within the remaining diverse hotspots are mainly related to energy metabolism processes and immune response (Sup Figure 3.1d and Sup Table 3.4).

We investigated the historical changes of effective population size within pygmy hogs and compared them with other *Suidae* species. The results of the PSMC analysis revealed a persistent low effective population size smaller than ~500 from 100,000 up to 10,000 years ago (Figure 3.3).

ROH were separated into four size classes. Among the *Suidae* species, the pygmy hog has an intermediate ROH coverage (Figure 3.4, Sup Figure 3.2). On average we found that the captive pygmy hogs have 408 ± 190 ROH with a total coverage of 17.8 ± 4.1% (means ± SDs, equal to 422 ± 101 Mb) and that the wild pygmy hogs contain 420 ± 142 ROH with a total coverage of 23.2 ± 2.9% (576 ± 74 Mb). This average is higher for pygmy hogs than for most Island of Southeast Asia (ISEA) *Sus* species (6.3 ± 1.3%, 157 ± 32 Mb). Compared to the highly inbred *Sus verrucosus* individual (48.9%, 1217 Mb), or European wild boars (56.0 ± 0.4%, 1388 ± 11 Mb), the proportion of ROH in pygmy hog genomes is significantly lower. In most pygmy hog individuals, the largest proportion of the genome was covered by short ROH (size ranges of 0-1 Mb and 1-5 Mb). Notably, one pygmy hog (PYGMY2) has significantly more long ROH than the other five individuals (t-test, p-value=0.04475).

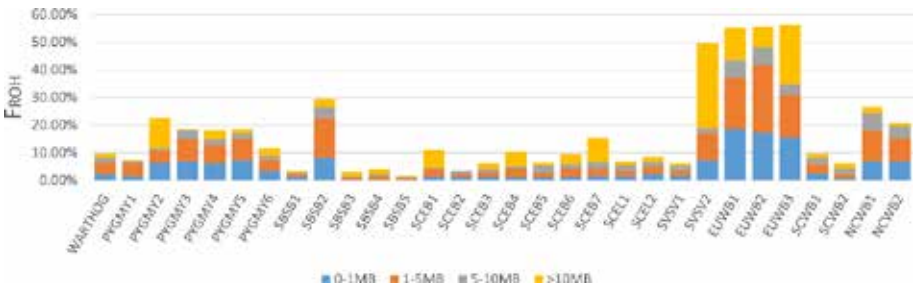


Figure 3.4 Proportion of the genome covered by ROH (F_{ROH}). ROH are divided into four categories ranging from a relatively 'small' (0.2 – 1 Mb) size category to a relatively 'large' (>10Mb) size category. WARTHOG = Phacochoerus africanus, PYGMY = pygmy hog, SBSB = Sus barbatus, SCEB = Sus cebrifons, SCEL = Sus celebensis, SVSV = Sus verrucosus, EUWB = European wild boar, NCWB = Northern China wild boar, SCWB = Southern China wild boar, detailed sample abbreviations see Sup table 3.1.)

Analysis of genetic load in pygmy hog genomes

We functionally annotated the variations found in our data. Variations which are predicted to result in frameshift, stop-gained and deleterious missense translation were selected. Compared with other species, pygmy hog harbors the highest number of frameshifts, stop-gained and missense variants (Figure 3.5), and the overwhelming majority of these variants is fixed within pygmy hogs (Sup Figure 3.3). The functional implication underlying the putative deleterious variants in the pygmy hog genomes was further investigated. First, to avoid uncertainty caused by alignment ambiguities, variants located in OR genes were excluded (see above). Next, to assess potential genetic load in the pygmy hog, we further extracted pygmy hog specific frameshift, stop-gained and missense mutations. Overall, 5972 frameshift mutations, 389 stop-gained mutations and 1772 deleterious missense mutations were observed to be pygmy-hog specific. To assess the potential impact of these functional variants, we used pig Combined Annotation Dependent Depletion (pCADD) scores to evaluate the predicted impact of stop gained and missense mutations (Groß et al., 2020). The pCADD scores are derived from a supervised classification that integrates multiple annotations, including conservation score (e.g. PhyloP, PhastCons and GERP), transcriptomic and epigenomic parameters (e.g. RNA-seq and ChIP-seq). Pygmy hogs appear to have more high impact mutations compared to the common warthog (*Phacochoerus africanus*), European wild boar, and Javan warty pig (Sup Figure 3.4). An enrichment of frameshift, stop-gained and missense mutations in the N- and C-terminal end of the affected genes can be observed (Sup Figure 3.5a). We further

predicted the distribution of high-impact variation within protein domains. In the terminal region of proteins, we found a relatively larger proportion of variants located in protein domains (Sup Figure 3.5b). Functional analysis did not reveal a significant gene ontology enrichment for frameshift and stop-gained mutations, or for missense mutations, however, significant enrichment for genes involved in immunity and hemostasis was found (Sup Figure 3.5b and Sup Table 3.5-3.7).

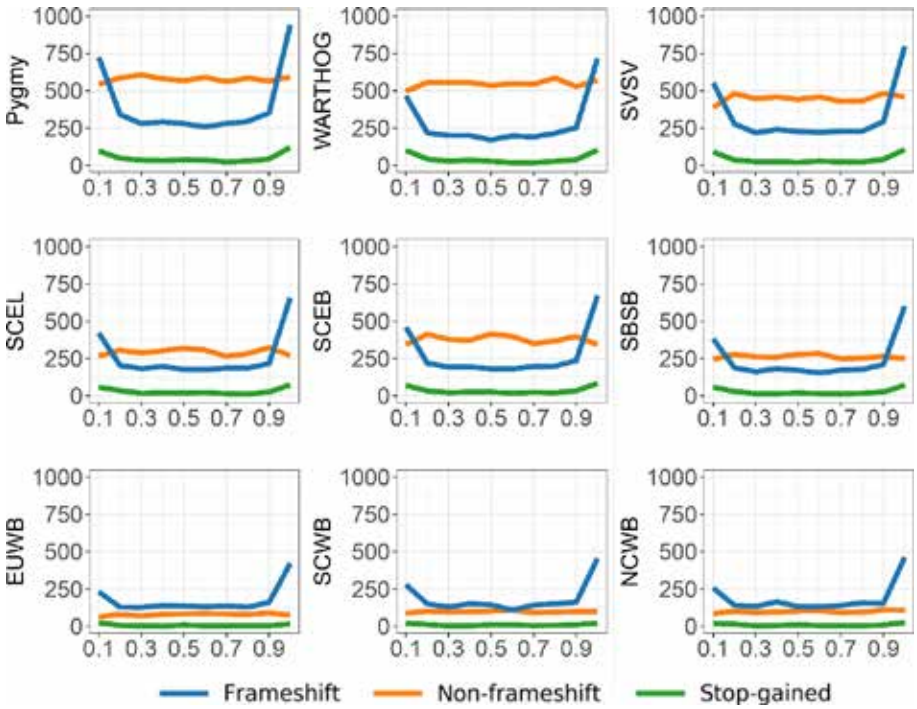


Figure 3.5 Relative position in the protein for frameshift, non-frameshift, and stop-gained variants in the different suid population. The x-axis displays the relative position of amino acid along the protein. The y-axis displays the average amount of variants within populations. WARTHOG = Phacochoerus africanus, PYGMY = pygmy hog, SBSB = Sus barbatus, SCEB = Sus cebrifons, SCEL = Sus celebensis, SVSV = Sus verrucosus, EUWB = European wild boar, NCWB = Northern China wild boar, SCWB = Southern China wild boar, detailed sample abbreviations see Sup table 3.1.)

Purging of deleterious alleles will occur naturally, as inbreeding increases the frequency of homozygotes where recessive effects are exposed to selection. However, whether the persistence of a small population size over long periods of time can be attributed to continued purging of harmful recessive mutations has not

been studied. To investigate this, we conducted forward-in-time simulations from 100k years ago onward with different consistent effective population sizes. Our results indicated that the population size is the key factor that influences the genetic load (Sup Figure 3.6). Numbers of strongly and moderately deleterious alleles were predicted to be remarkably increased in the current populations following the reduction of population size. The total number of deleterious alleles per individual in the current population relative to the ancestral population varied according to selection and dominance coefficients. Although we can still observe the elimination of harmful mutations in large populations, in small population, selection against deleterious alleles was weakened dramatically and the accumulation of additive deleterious alleles became more severe. Under additive regime scenario, with a population size smaller than 1000, current genomes always contained more deleterious alleles per individual than in the ancestral genomes. Under recessive regime scenario, when population size is smaller than 100, deleterious alleles per individual will excess the ancestral genomes. (Sup Figure 3.6a&b). Moreover, all harmful mutations, including high impact mutations tended to be homozygous, which is consistent with the actual pygmy hog genomes. In sum, these findings suggest the limitation of purging of high impact alleles in the historically persistent small population of pygmy hogs.

3.4 Discussion

This study offers insight into the historic demography and current genetic conservation status of the critically endangered pygmy hog. The continued low population size for the past one hundred thousand years, the very low genetic diversity, and the accumulation of potentially harmful mutations, are supporting the endangered conservation status of this species. Being a small and isolated population, the pygmy hog has low genomic diversity and heterozygosity compared to other pig species. Although having low genetic variation similar to other critically endangered species, such as the Cheetah and the Tasmanian devil, the pygmy hog genome possesses a relative low level of ROH compared to the mentioned endangered species. In the meantime, unlike the island foxes, the effective size of pygmy hog populations have been so small for a very long time that effective purging of harmful mutations is likely impossible. This makes the pygmy hog an interesting model for studying the survival of small populations.

Demographic analyses of the pygmy hog revealed a persistent low effective population size with fewer than ~500 animals over the past one hundred thousand years to ten thousand years. These results are consistent with paleontological

evidence where all fossil finds of pygmy hogs outside the Assam region were from ~1 Mya (Pickford, 2013). This suggests that the pygmy hog used to have a broader distribution range and then started contracting already during the Pleistocene. Phytogeographic analysis shows that the type of grasslands currently found at the southern foothills of the Himalayas, were far more widespread across parts of South Asia during the Pliocene and early Pleistocene (Dowsett et al., 1994; Dennell, 2011). According to our PSMC analysis, the pygmy hog was not noticeably affected by the Last Glacial Maximum (LGM), which, in contrast, had a huge effect on effective population size in the Eurasian wild boar (Groenen et al., 2012). Palaeoclimatologists have hypothesized that the southern flank of the Himalayas during the LGM harbored a range of climatological refuges (Singh et al., 2010). This would continue to provide a suitable habitat, allowing pygmy hog to continue to have a constant, local, population size.

After persisting a long period of low effective population size, the current pygmy hog population is harboring more deleterious mutations, or precisely high impact mutations, than other *Suidae* species. Notably, reference bias can influence variant calling by missing alternative alleles or by wrongly calling heterozygous sites as homozygous for the reference allele (Ros-Freixedes et al., 2018). This effect increases with the genetic distance toward the reference genome (*Sus scrofa*) (Liu et al., 2019). Although we do expect some bias in this estimation of high impact mutations in pygmy hogs, caused by the genetic distance to the reference genome, the pygmy hog does harbor more high impact mutations than the warthog (Figure 5, Figure S4). Since the African warthog is even more distantly related to *Sus scrofa*, distance to the reference genome alone cannot explain the high frequency of high impact mutations in the pygmy hog. Therefore, the pygmy hog appears to have a dramatically increased rate of accumulation of high impact mutations.

In pygmy hog genomes, high impact mutations show a pattern of historical purifying selection, since most of them are located at the N- and C-terminal end of genes. However, even within the two tails of proteins, which generally contain less functional domains, there are still abundant mutations that may influence the function of the protein. The gene set enrichment analysis clearly shows that certain GO terms are strongly associated with pygmy hog-specific missense mutations. These GO terms are mostly related to the immune response and blood coagulation pathways. In the meantime, selection against deleterious recessive alleles is less efficient when population size is small (García-Dorado, 2012). A previous study suggested that the minimum effective population size to avoid severe inbreeding depression in the short term is $Ne \approx 70$ (Caballero et al., 2017). This is consistent

with our simulations, which show an elevation of deleterious mutations in small populations. Moreover, the majority of the deleterious mutations is in the homozygous state, suggesting that the accumulation of deleterious mutations is exceeding the purging effect. The overall ROH coverage in pygmy hog indicated a low level of inbreeding. As a result, recessive deleterious mutations had less chance to become exposed before they became fixed in the population. Such dynamic relationship between inbreeding and purging has thus far not been observed in other endangered populations illuminating the importance of species-specific genetic analysis for predicting and enhancing population persistence. The results underly the assumption that the predicted variants in pygmy hog are predominantly harmful, or greatly affect gene function. High impact mutations can also show selective advantage by genetic hitchhiking in regions under selection, sometimes even boosting the fitness in specific lineage due to the local adaptation (Bosse, 2019; Bosse et al., 2019). With the limited information we have, it is difficult to assess the actual effect of specific alleles, some of which potentially could be related to shaping species characteristic like e.g. behavioral traits and adaptation to a specific habitat. However, since the accumulation of deleterious mutations is well in excess of a purging effect, we believe that the majority of these predicted high impact variants have a negative effect on fitness.

The current pygmy hog population exhibits low nucleotide diversity and heterozygosity compared to other pig species, which conforms to its critically endangered status. Comparing the mitochondrial genomes of three wild-caught pygmy hogs and three captive individuals, we find that there is no variation within the analyzed samples. Although the small sampling size, six individuals in this case, may lead to sampling biases of maternal linage. The wild individuals and founders of the captative population were independently sampled in 2014 and 1996, respectively, reducing the possibility of sampling biases to a certain extent. This indicates a very low mitochondrial DNA diversity and a potential maternal bottleneck happened before the establishment of the captive population. Severe unbalanced sex-ratio is often observed in species on the edge of extinction (Bessa-Gomes et al., 2004; Allentoft et al., 2010; Pečnerová et al., 2017). The same situation may have happened to pygmy hogs in the 60s, when they almost disappeared from the wild.

The long-term small population size and potential historical bottleneck lead to reduction in genetic diversity, which further limits the ability to adapt to environmental changes. By comparison, the close relative to the pygmy hog, *Sus scrofa*, is widely distributed over Eurasia continent, whereas the pygmy hog is highly specialized, only living in the tall grass savannah. A conservation programme

was used to transfer pygmy hogs to the Zurich and London Zoos in 1998 and 1876 respectively (Oliver and Roy, 1993), but both failed. The dependence on a specific ecological niche and a reduced adaptability to environmental changes could be the consequence of the reduced standing genetic variation and accumulation of genetic load. Thus, maximum efforts should be made to protect the fragile high-grassland ecosystem.

While genome-wide allelic diversity may be low, the pygmy hog does not show extreme ROH coverage. Specifically, long ROH are rare, compared to for instance the sequenced male Javan warty pig (SVSV01M01) or European wild boars, which are known to have gone through series of recent population bottlenecks (Groenen et al., 2012; Nuijten et al., 2016). Between wild and captive pygmy hog populations, there is no significant difference in the total length of ROH. The overall ROH landscape in pygmy hog indicates very little recent inbreeding. The observed ROH were possibly caused by an ancient bottleneck followed by a gradual breakdown of ancient long ROH (Speed and Balding, 2015). Notably, although the pedigree information does not indicate closely related mating, one of the pygmy hogs (PYGMY2) has significantly more long ROH than the other individuals. Thus, the founders of the maternal and the paternal lineage of PYGMY2 seem to share more relatedness than the founders of other two captive individuals. This result is a warning that the arbitrary assumption in conservation practices that the wild capture founders are genetically unrelated is not always valid.

Considering the genetic diversity and inbreeding level, the initial founders of the PHCP were sufficiently representatives of the wild population. However, the captive individuals in this study represent the third and fourth generation of the breeding programme. The observation that the most recent generation showed a significant decrease in individual heterozygosity indicates that drift effects likely are becoming prominent after more than five generations (Purohit et al., 2019). Since there is no other existing wild population and pygmy hog is the only member in its genus, ‘genetic rescue’ is not feasible for the pygmy hog population. Fortunately, signatures of very recent inbreeding, compared to some of the other endangered pig populations, are relatively mild. Furthermore, no noticeable morphological changes have been reported (i.e., length, weight, external appearance) (Narayan et al., 1999; Deka et al., 2009). Other additional phenotypic traits have not yet been examined within this population and therefore it is not known if the low levels of genetic diversity are impacting population fitness. Considering the recent decline of genetic diversity in captive pygmy hogs (Purohit et al., 2019), genetic defects may become apparent due to the recessive deleterious alleles being homozygous. Recent studies have shown the potential of using genomics information to monitor

deleterious mutations in breeding programme (Charlier et al., 2016; Derks et al., 2018, 2019). A genomic method to measure the kinship in captivity is pressingly needed for the PHCP to prevent close relatives from mating and to estimate individual genetic load to guide the artificial selection against harmful mutations causing genetic defects. To preserve the evolutionary potential of the pygmy hog population, it is essential to enlarge the extant population and prevent a severe decline of population size due to disease outbreaks or anthropogenic threats.

3.5 Conclusion

In conclusion, long-term persistence of extremely small population size can lead to an increase in genetic load. Although species management through breeding programmes can prevent the occurrence and expression of harmful alleles, genetic diversity cannot be boosted by human intervention, but only by natural mutation and introgression with closely-related species. Monitoring the individual heterozygosity of subsequent generations is, hence, crucial for maintaining the genetic diversity in captive pygmy hogs and to inform future conservation breeding decisions.

Data and software availability

The authors declare that all data and software supporting the findings of this study are available within the article and its Supplementary Information files, or from the corresponding author upon request. Raw reads of all samples used in this study have been deposited in the European Nucleotide Archive (ENA) under accession ERP001813, ERP112560 and ERP118195.

Description of supplementary Material

For a compact layout, in this thesis I did not include all supplementary material. I presented Supplementary Figures which may help the reader. For sake of coherence, I kept the original number of Supplementary Figures and tables.
Complete supplementary material are available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.13150

Reference:

Allentoft, M. E., Bunce, M., Scofield, R. P., Hale, M. L., and Holdaway, R. N. (2010). Highly skewed sex ratios and biased fossil deposition of moa: ancient DNA provides new insight on New Zealand's extinct megafauna. *Quat. Sci. Rev.* 29, 753–762. doi:10.1016/j.quascirev.2009.11.022.

Bessa-Gomes, C., Legendre, S., and Clobert, J. (2004). Allee effects, mating systems and the extinction risk in populations with two sexes. *Ecol. Lett.* 7, 802–812. doi:10.1111/j.1461-0248.2004.00632.x.

Bortoluzzi, C., Bosse, M., Derks, M. F. L., Crooijmans, R. P. M. A., Groenen, M. A. M., and Megens, H. (2019). The type of bottleneck matters: Insights into the deleterious variation landscape of small managed populations. *Evol. Appl.*, eva.12872. doi:10.1111/eva.12872.

Bosse, M. (2019). No "doom" in chicken domestication? *PLoS Genet.* 15. doi:10.1371/journal.pgen.1008089.

Bosse, M., Megens, H., Derks, M. F. L., Cara, Á. M. R., and Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. *Evol. Appl.* 12, 6–17. doi:10.1111/eva.12691.

Bosse, M., Megens, H. J., Madsen, O., Paudel, Y., Frantz, L. A. F., Schook, L. B., et al. (2012). Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. *PLoS Genet.* 8, e1003100. doi:10.1371/journal.pgen.1003100.

Caballero, A., Bravo, I., and Wang, J. (2017). Inbreeding load and purging: Implications for the short-term survival and the conservation management of small populations. *Heredity (Edinb).* 118, 177–185. doi:10.1038/hdy.2016.80.

Charlier, C., Li, W., Harland, C., Littlejohn, M., Coppieters, W., Creagh, F., et al. (2016). NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res.* 26, 1333–1341. doi:10.1101/gr.207076.116.

Davis, M. B., and Shaw, R. G. (2001). Range shifts and adaptive responses to Quaternary climate change. *Science* 292, 673–9. doi:10.1126/science.292.5517.673.

Deka, P. J., Narayan, G., Oliver, W. L. R., and Fa, J. E. (2009). Reintroduced pygmy hogs (Porcula salvania) thrive a year after release- more hogs released in Sonai Rupai Wildlife Sanctuary, Assam, India. *Suiform Sound.* 9, 23–28. Available at: https://s3.amazonaws.com/academia.edu.documents/41175913/0deec5240b90120f08000000.pdf20160115-19908-2xk60k.pdf?response-content-disposition=inline%3Bfilename%3DEmerging_infectious_diseases_swine_flu.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential.

Dennell, R. W. (2011). "The Colonization of 'Savannahstan': Issues of Timing(s) and Patterns of Dispersal Across Asia in the Late Pliocene and Early Pleistocene," in *Vertebrate Paleobiology and Paleoanthropology* (Springer), 7–30. doi:10.1007/978-90-481-9094-2_2.

Derks, M. F. L., Gjuvsland, A. B., Bosse, M., Lopes, M. S., Van Son, M., Harlizius, B., et al. (2019). Loss of function mutations in essential genes cause embryonic lethality in pigs. *PLoS Genet.* 15, e1008055. doi:10.1371/journal.pgen.1008055.

Derks, M. F. L., Megens, H. J., Bosse, M., Visscher, J., Peeters, K., Bink, M. C. A. M., et al. (2018). A survey of functional genomic variation in domesticated chickens. *Genet. Sel. Evol.* 50, 17. doi:10.1186/s12711-018-0390-1.

Dobrynin, P., Liu, S., Tamazian, G., Xiong, Z., Yurchenko, A. A., Krasheninnikova, K., et al. (2015). Genomic legacy of the African cheetah, Acinonyx jubatus. *Genome Biol.* 16, 277. doi:10.1186/s13059-015-0837-4.

Dowsett, H., Thompson, R., Barron, J., Cronin, T., Fleming, F., Ishman, S., et al. (1994). Joint investigations of the Middle Pliocene climate I: PRISM paleoenvironmental reconstructions. *Glob. Planet. Change* 9, 169–195. doi:10.1016/0921-8181(94)90015-9.

Ellis, E. C. (2015). Ecology in an anthropogenic biosphere. *Ecol. Monogr.* 85, 287–331. doi:10.1890/14-2274.1@10.1002/(ISSN)1557-7015.MONOGRAPHSCENTENNIALPAPERS.

Funk, S. M., Verma, S. K., Larson, G., Prasad, K., Singh, L., Narayan, G., et al. (2007). The pygmy hog is a unique genus: 19th century taxonomists got it right first time round. *Mol. Phylogenet. Evol.* 45, 427–436. doi:10.1016/j.ympev.2007.08.007.

Funk, W. C., Lovich, R. E., Hohenlohe, P. A., Hofman, C. A., Morrison, S. A., Sillett, T. S., et al. (2016). Adaptive divergence despite strong genetic drift: genomic analysis of the evolutionary mechanisms causing genetic differentiation in the island fox (*Urocyon littoralis*). *Mol. Ecol.* 25, 2176–2194. doi:10.1111/mec.13605.

García-Dorado, A. (2012). Understanding and predicting the fitness decline of shrunk populations: Inbreeding, purging, mutation, and standard selection. *Genetics* 190, 1461–1476. doi:10.1534/genetics.111.135541.

Gentleman, R. I. and R. (2015). R : A Language for Data Analysis and Graphics. *Comput. Graph. Stat.* 5, 299–314. Available at: http://www.jstor.org/stable/pdf/1390807.pdf?acceptTC=true.

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi:10.1038/nature11622.

Groß, C., Derks, M., Megens, H.-J., Bosse, M., Groenen, M. A. M., Reinders, M., et al. (2020). pCADD: SNV prioritisation in Sus scrofa. *Genet. Sel. Evol.* 52, 4. doi:10.1186/s12711-020-0528-9.

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41, e129–e129. doi:10.1093/nar/gkt371.

Hamilton, J. A., and Miller, J. M. (2016). Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conserv. Biol.* 30, 33–41. doi:10.1111/cobi.12574.

Kardos, M., Qvarnström, A., and Ellegren, H. (2017). Inferring Individual Inbreeding and Demographic History from Segments of Identity by Descent in Ficedula Flycatcher Genome Sequences. *Genetics* 205, 1319–1334. doi:10.1534/genetics.116.198861.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15, 356. doi:10.1186/s12859-014-0356-4.

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

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., and Marth, G. T. (2014). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *PLoS One* 9, e90581. doi:10.1371/journal.pone.0090581.

Leus, K. (2017). "Ex-situ conservation of wild pigs and peccaries: Roles, status, management successes and challenges," in *Ecology, Conservation and Management of Wild Pigs and Peccaries* (Cambridge University Press), 420–436. doi:10.1017/9781316941232.039.

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

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496. doi:10.1038/nature10231.

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

Liu, L., Bosse, M., Megens, H.-J., Frantz, L. A. F., Lee, Y.-L., Irving-Pease, E. K., et al. (2019). Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* 10, 1992. doi:10.1038/s41467-019-10017-2.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., et al. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* 17, 704–714. doi:10.1038/nrg.2016.104.

Lynch, M., Conery, J., and Burger, R. (1995). Mutation Accumulation and the Extinction of Small Populations. *Am. Nat.* 146, 489–518. doi:10.1086/285812.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110.

Merola, M. (1994). A Reassessment of Homozygosity and the Case for Inbreeding Depression in the Cheetah, Acinonyx jubatus: Implications for Conservation. *Conserv. Biol.* 8, 961–971. doi:10.1046/j.1523-1739.1994.08040961.x.

Miraldo, A., Li, S., Borregaard, M. K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., et al. (2016). An Anthropocene map of genetic diversity. *Science* 353, 1532–1535. doi:10.1126/science.aaf4381.

Narayan, G., Deka, P. J., Chakraborty, A., and Oliver, W. L. R. (1999). *Increase in the captive population of pygmy hogs Sus salvanius: Health problems and husbandry*. Available at: https://www.scopus.com/inward/record.url?eid=2-s2.0-0442311627&partnerID=40&md5=f97a55d6b90ee1c50d4759c9f5ea2cb5 [Accessed May 29, 2019].

Nuijten, R. J. M., Bosse, M., Crooijmans, R. P. M. A., Madsen, O., Schaftenaar, W., Ryder, O. A., et al. (2016). The Use of Genomics in Conservation Management of the Endangered Visayan Warty Pig (*Sus cebifrons*). *Int. J. Genomics* 2016, 1–9. doi:10.1155/2016/5613862.

Oliver, W. L. R., and Deb Roy, S. (1993). The pygmy hog (Sus salvanius). *Pigs, Peccaries Hippos status Surv. Conserv. action plan. IUCN, Gland*, 121–129.

Oliver, W. L. R., and Roy, S. D. (1993). The pygmy hog. *Pigs, Peccaries, Hippos Status Surv. Conserv. Action Plan* 19, 121.

Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A. F., Bosse, M., Bastiaansen, J. W. M., et al. (2013). Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* 14, 449. doi:10.1186/1471-2164-14-449.

Pečnerová, P., Díez-del-Molino, D., Dussex, N., Feuerborn, T., von Seth, J., van der Plicht, J., et al. (2017). Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth. *Curr. Biol.* 27, 3505-3510.e3. doi:10.1016/j.cub.2017.09.064.

Peet, N. B., Watkinson, A. R., Bell, D. J., and Sharma, U. R. (1999). The conservation management of Imperata cylindrica grassland in Nepal with fire and cutting: An experimental approach. *J. Appl. Ecol.* 36, 374–387. doi:10.1046/j.1365-2664.1999.00405.x.

Pekkala, N., Emily Knott, K., Kotiaho, J. S., and Puurtinen, M. (2012). Inbreeding rate modifies the dynamics of genetic load in small populations. *Ecol. Evol.* 2, 1791–1804. doi:10.1002/ece3.293.

Pekkala, N., Knott, K. E., Kotiaho, J. S., Nissinen, K., and Puurtinen, M. (2014). The effect of inbreeding rate on fitness, inbreeding depression and heterosis over a range of inbreeding coefficients. *Evol.*

Appl. 7, 1107–1119. doi:10.1111/eva.12145.

PHCP (2008). Conservation Strategy and Action Plan for Pygmy Hog in Assam EcoSystems-India.

Pickford, M. (2013). Suids from the Pleistocene of Naungkwe Taung, Kayin State, Myanmar. *Paleontol. Res.* 16, 307–317. doi:10.2517/1342-8144-16.4.307.

Piertney, S. B., and Oliver, M. K. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb).* 96, 7–21. doi:10.1038/sj.hdy.6800724.

Pightling, A. W., Petronella, N., and Pagotto, F. (2014). Choice of Reference Sequence and Assembler for Alignment of Listeria monocytogenes Short-Read Sequence Data Greatly Influences Rates of Error in SNP Analyses. *PLoS One* 9, e104579. doi:10.1371/journal.pone.0104579.

Pimm, S. L., and Raven, P. H. (2017). The Fate of the World's Plants. *Trends Ecol. Evol.* 32, 317–320. doi:10.1016/J.TREE.2017.02.014.

Purohit, D., Ram, M. S., Pandey, V. K., Pravalika, S., Deka, P. J., Narayan, G., et al. (2019). Cross-specific markers reveal retention of genetic diversity in captive-bred pygmy hog, a critically endangered suid. *Conserv. Genet. Resour.*, 1–5. doi:10.1007/s12686-019-01091-1.

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* 47, 11.12.1-11.12.34. doi:10.1002/0471250953.bi1112s47.

Robinson, J. A., Brown, C., Kim, B. Y., Lohmueller, K. E., and Wayne, R. K. (2018). Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. *Curr. Biol.* 28, 3487-3494.e4. doi:10.1016/J.CUB.2018.08.066.

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., vonHoldt, B. M., Marsden, C. D., et al. (2016). Genomic Flatlining in the Endangered Island Fox. *Curr. Biol.* 26, 1183–1189. doi:10.1016/J.CUB.2016.02.062.

Ros-Freixedes, R., Battagin, M., Johnsson, M., Gorjanc, G., Mileham, A. J., Rounsley, S. D., et al. (2018). Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genet. Sel. Evol.* 50. doi:10.1186/s12711-018-0436-4.

Semiadi, G., and Meijaard, E. (2006). Declining populations of the Javan warty pig Sus verrucosus . *Oryx* 40, 50. doi:10.1017/S003060530600007X.

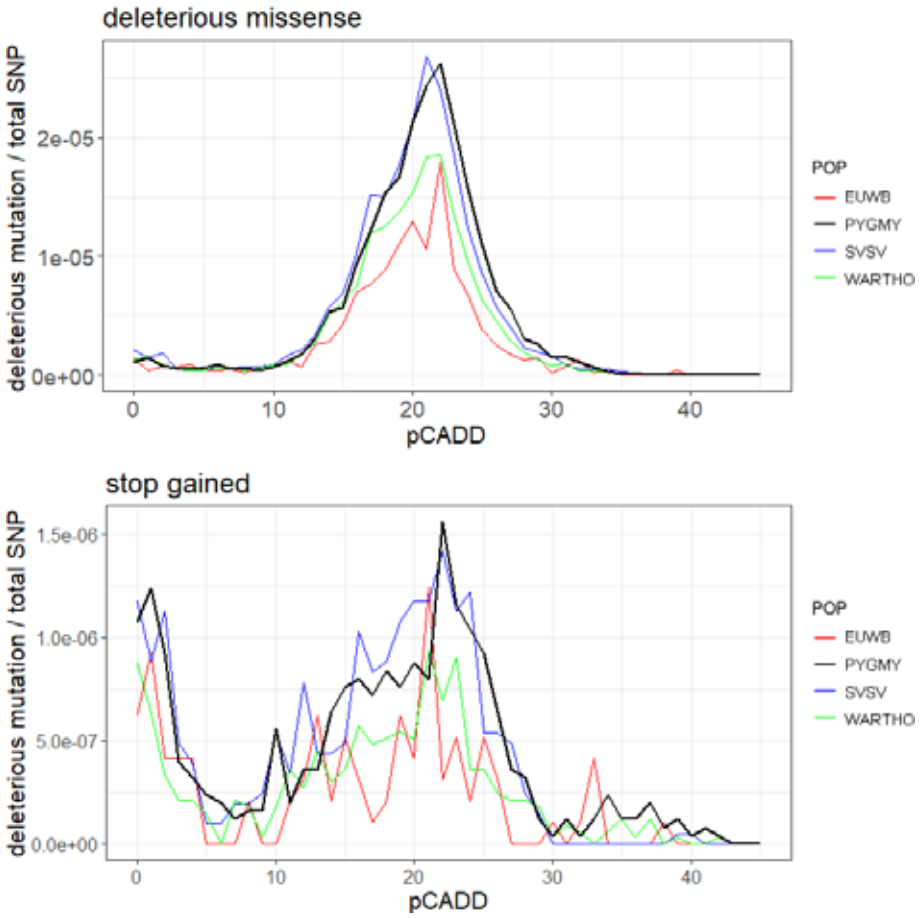
Singh, S. P., Singh, V., and Skutsch, M. (2010). Rapid warming in the Himalayas: Ecosystem responses and development options. *Clim. Dev.* 2, 221–232. doi:10.3763/cdev.2010.0048.

Speed, D., and Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nat. Rev. Genet.* 16, 33–44. doi:10.1038/nrg3821.

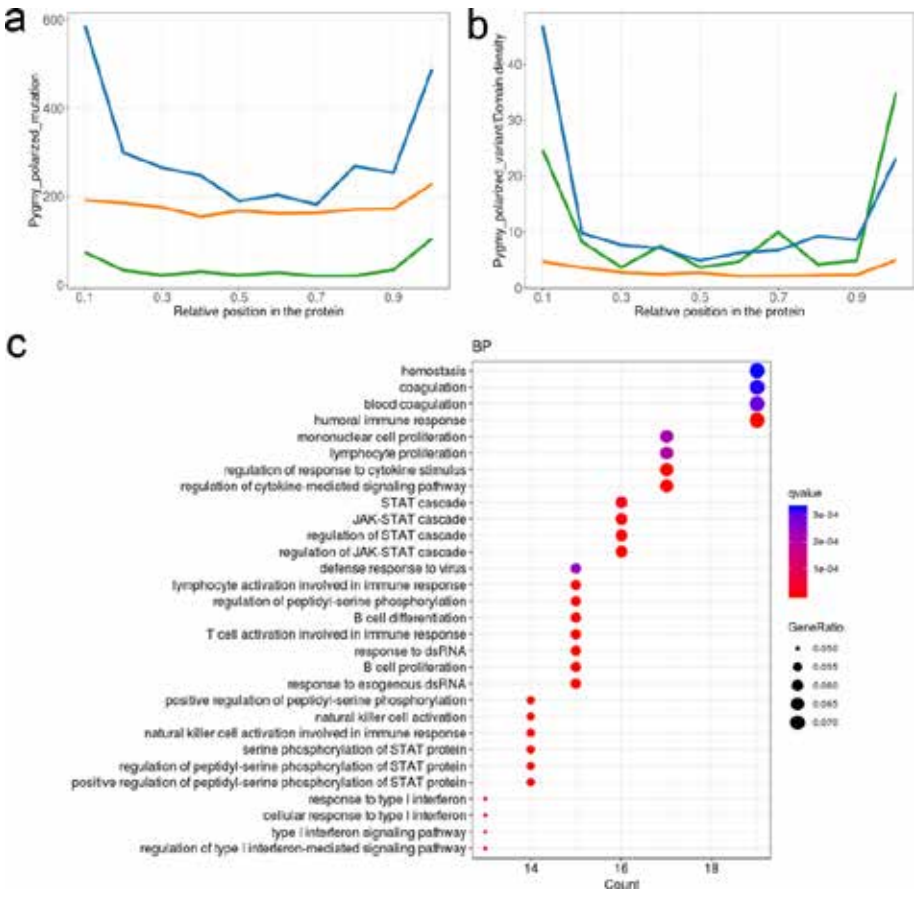
Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., et al. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* 13, 586. doi:10.1186/1471-2164-13-586.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118.

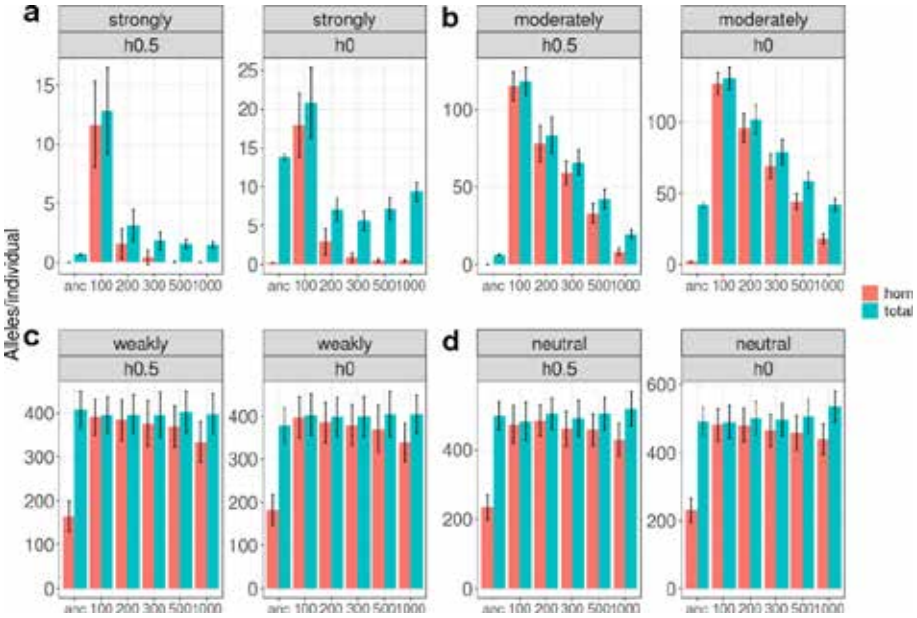
Supplementary material



Sup Fig. 3.4 Distribution of pig CADD score (pCADD) of deleterious missense and stop gained mutation.



Sup Fig. 3.5 a) Relative position in the protein for pygmy hog specific frameshift, non-frameshift, and stop-gained variants. The x-axis displays the relative position of amino acid along the protein. b) Distribution for pygmy hog specific frameshift, non-frameshift, and stop-gained variants related to protein domain. The y-axis displays counts of variant located in a protein functional domain. The x-axis displays the relative position of amino acid along the protein normalized by the distribution of protein domain. Information of functional domain was downloaded from https://pfam.xfam.org/. c) Functional enrichment result for gene set harbored missense mutations. The number of horizontal axis is the number of enriched genes. The bubble size indicates the ratio of genes in each term or pathway, and different colours correspond to different adjusted p-values. The p-values are adjusted by Benjamini-Hochberg method.



Sup Fig. 3.6 Forward simulations of genetic variation in pygmy hog population. Charts shows the total number of derived alleles and the number of homozygous derived alleles per individual (Y-axis) by dominance and selection strength. Bar height represents the mean across all 50 replicates. Selection strengths are set as four categories: a) strongly deleterious, b) moderately deleterious, c) weakly deleterious, and d) neutral. Simulations include a mixture of neutral and deleterious alleles in which deleterious alleles are either entirely additive (additive regime, h0.5) or entirely recessive (recessive regime, h0). Anc is the ancestral population and the numbers (100-1000) represent the population size used in the simulations.

4

Genome assembly of Visayan warty pig (*Sus cebifrons*) provides insight into genome evolution of *Sus* during speciation

Langqing Liu[1], Hendrik-Jan Megens[1], Richard PMA Crooijmans[1], Mirte Bosse[1], Qitong Huang[1,2], Linda van Sonsbeek[3], Martien AM Groenen[1], Ole Madsen[1]

[1] Animal Breeding and Genomics, Wageningen University & Research, The Netherlands; [2] Center for Animal Genomics, Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518124, China; [3]Rotterdam Zoo, 3041 JG Rotterdam, Netherlands

Manuscript in preparation

Abstract

Environmental induced speciation is common in animal and plant, but the genomic consequences of speciation remain poorly understood. An excellent model for understanding rapid evolution is provided by the *Sus* genus, which diverged relatively recent and lacks post-zygotic isolation. Here, we present a high-quality reference genome of the Visayan warty pig, which is specialized to tropical island environment. Comparing the genome sequence and chromatin contact map between Visayan warty pig (*Sus cebifrons*) and domestic pig, we characterized the dynamic of chromosomal structure evolution during *Sus* speciation. We further investigated the different signatures of adaptive selection and domestication in Visayan warty pig and domestic pig. Moreover, we studied the evolution of olfactory and gustatory genes and showed the genetic basis of species-specific sensation.

Key words: *de novo* genome assembly, speciation, chromosomal rearrangement, adaptation, domestication

4.1 Introduction

The process of speciation is often initiated by changes in the environment, where exposure to new environmental selective forces directs populations toward new evolutionary trajectories (Templeton, 2008; Coyne and Orr, 2009). In animals, behavioral modifications are sometimes associated with ecological differences during an initial colonization event, or sudden environmental fluctuation (Hoy, 1990; Gottlieb, 2002; Rundle, Howard D. and Boughman, 2015). But no matter which forces are at work during speciation, for a successful response to altered conditions there must be sufficient genetic variation for the population to reach a new adaptive norm through natural selection.

The genus *Sus*, pigs and hogs, comprises of at least seven morphologically and genetically well-defined species. Except for *Sus scrofa*, that inhabit the Eurasian continent, all other *Sus* species are restricted on Islands of South East Asia (ISEA). Recent findings showed that these species diverged during the late Pliocene (4 - 2.5 Mya), as a result of isolation on different islands of ISEA due to climate fluctuation (Frantz et al., 2013; Liu et al., 2019). One of the representative species of ISEA *Sus*, the Visayan warty pig (*Sus cebifrons*) has been assigned the highest threat status, Critically Endangered, by the IUCN Red List. Visayan warty pig is named after the islands it lives on, and the three pairs of fleshy warts on the face of boars. Living in social groups, this species requires dense forested areas, on primary and secondary forests. Visayan warty pigs are omnivores and eat a wide variety of forest food.

Contrary to the endemic nature of the ISEA *Sus* species, such as the Visayan warty pig, the wild boar (*Sus scrofa*) has a widespread geographical Eurasian distribution. The widespread distribution has recently been suggested to have been facilitated by admixture events with local pig species during the expansion from Asia to Europe. Meanwhile, *Sus scrofa* have also been highly interacting with human during the process of domestication. Having been domesticated 9,000 to 10,000 years ago independently in East Anatolia and in China (Larson et al., 2005, 2010; Frantz et al., 2015), pigs have experienced long-term artificial selection for various traits including production, fertility, and disease resistance,

Since the draft reference genome of domesticated pig, published in 2012 (Groenen et al., 2012), the number of *Sus scrofa* genomes available has increased dramatically during the last couple of years (Fang et al., 2012; Li et al., 2013, 2017; Vamathevan et al., 2013). However, to date, no reference genome has been reported for the other members of *Suidae* family. Here we provide a high-quality draft Visayan warty pig genome sequence. The complete assembled sequences of two closely related pig species genomes provides an opportunity to further understand the dynamic nature of mammalian chromosomes over a relative short

evolutionary scale. The highly phenotypic and ecological divergency of these two sister species also provides an excellent comparative framework to study the genomic consequences of distinct evolutionary and demographic history.

4.2 Result

Genome sequencing, assembly and quality evaluation

For the genome assembly, 10X genomics, Hi-C and mate pair sequences were employed. The size of the Visayan warty pig genome was estimated at 2.25 Gb (Sup Fig. 4.1). The genome sequence was constructed by a hybrid assembly approach (see Material and Methods), which finally yielded a contig N50 size of 159.6 Kb, a scaffold N50 size of 141.8 Mb and an assembled genome size of 2.48 Gb, comprising of 1585 scaffolds (Table 4.1, Sup Fig. 4.2, Sup Table. 4.1). The assembly of the mitochondrial genome sequence resulted in a 16.56 kb assembly with a cyclic structure (Sup Fig. 4.3). A schematic representation of some characteristics of the genome is shown in Fig. 1a.

A total of 99.73% of the short sequence reads covered 99.41% of the genome assembly map. The percentage of homologous SNPs reflects the accuracy of the genome assembly. 201,798 SNPs are in homozygous state (0.008% of the whole genome), indicating that the genome assembly shows high quality at the single-base level. The Visayan warty pig assembly was also assessed for completeness using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015). The BUSCO analysis gave 95.7% of the BUSCOs genes as complete, suggesting a high completeness of the assembly (Sup Table. 4.2).

Table 4.1 Assembly and annotation statistics of the draft genome of Visayan warty pig

Assembly features	
Total length (bp)	2,459,328,531
Scaffold count	1,585
Longest scaffold (bp)	289,860,557
N50 of scaffold (bp)	141,782,568
N50 of contig (bp)	159,795
GC ratio (%)	41.78
Genome annotation	
Number of putative coding genes	21,153
Average gene model length (bp)	583,09.93
Average CDS length (bp)	1,640.66
Average gene exon length (bp)	220.89
Average exon number per gene	11.56
Average gene intron length (bp)	4,565.05
Total size of Tes (bp)	1,079,874,097
TEs in genome (%)	43.91

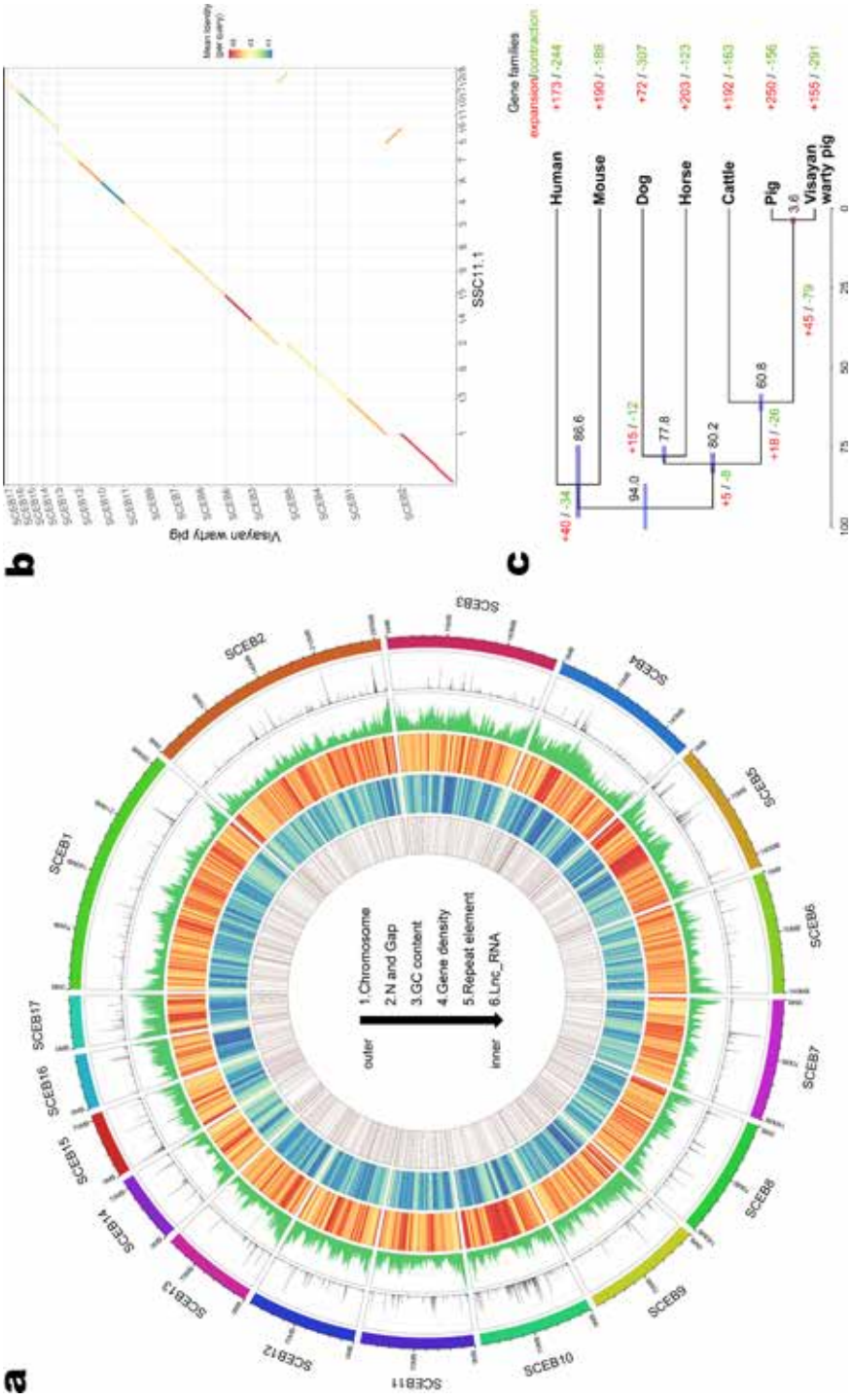

a
SCEB1
SCEB2
SCEB3
SCEB4
SCEB5
SCEB6
SCEB7
SCEB8
SCEB9
SCEB10
SCEB11
SCEB12
SCEB13
SCEB14
SCEB15
SCEB16
SCEB17
1.Chromosome
2.N and Gap
3.GC content
4.Gene density
5.Repeat element
6.Lnc_RNA
outer
inner
b
Visayan warty pig
SSC11.1
c
Gene families
expansion/contraction
Human
Mouse
Dog
Horse
Cattle
Pig
Visayan warty pig
+173 / -244
+190 / -186
+72 / -307
+203 / -123
+192 / -163
+250 / -156
+155 / -291
+40 / -34
+15 / -12
+5 / -8
+18 / -26
+45 / -79
86.6
94.0
77.8
80.2
60.8
3.6
100
75
50
25
0

Fig. 4.1 Overview of the assembly characteristics of the Visayan warty pig genome. a) Circular diagram depicting the characteristics of the Visayan warty pig genome. The tracks from outer to inner circles are indicated in the legend (density of coding gene are shown in track 4). b) Whole genome alignment between Visayan warty pig and Duroc pig (Sus scrofa11.1). c) Divergence times and history of orthologous gene families. Numbers on the nodes represent divergence times, with the error range shown in blue bar. The numbers of gene families that expanded (red) or contracted (green) in each lineage after speciation are shown on the corresponding branch.

Contiguity and Chromosomal rearrangements

The chromosome-level Visayan warty pig assembly allows comparisons to the assembled Duroc pig genome (Sus scrofa11.1) (Kurtz et al., 2004) to assess the genomic dynamics between these closely related species. We assessed the scaffold orderings of these two genomes through whole-genome sequence alignments (Fig. 4.1b). Our Visayan warty pig assembly reveals a high degree of co-linearity with the Duroc pig PacBio assembly. Beside the strong overall collinear relationship between the two genomes, we could confirm previous described chromosomal rearrangements between the two species (Sup Fig. 4.4). The karyotype of Visayan warty pig (2n = 34) seems to contain the centric fusions SSC13/16 and SSC14/18, which hereby gave rise to the largest and third largest Visayan warty pig chromosomes respectively. Chromosome interactions of Visayan warty pig and Duroc pig was assess from Hi-C data. The Hi-C interaction maps between homologous chromosome pairs show obvious similarities (Sup. Fig 4.5) and the fused chromosomes in Visayan warty pig retained the same chromosome interaction patterns as the corresponding Duroc pig homologous chromosomes (Fig. 4.2). The inter-chromosomal interactions in the fused chromosome SCEB1 (SSC13/16) are mainly restricted within the two chromosome arms, while SCEB3 (SSC14/18) shows more trans-interactions. We further quantified the similarity of interaction matrixes between Visayan warty pig and Duroc pig using Phylo-HMRF (Yang et al., 2019). Around 23.5% of the total interactions are categorized as highly or medium conserved (Fig. 4.2, Sup Fig. 4.6, see Material and Methods).

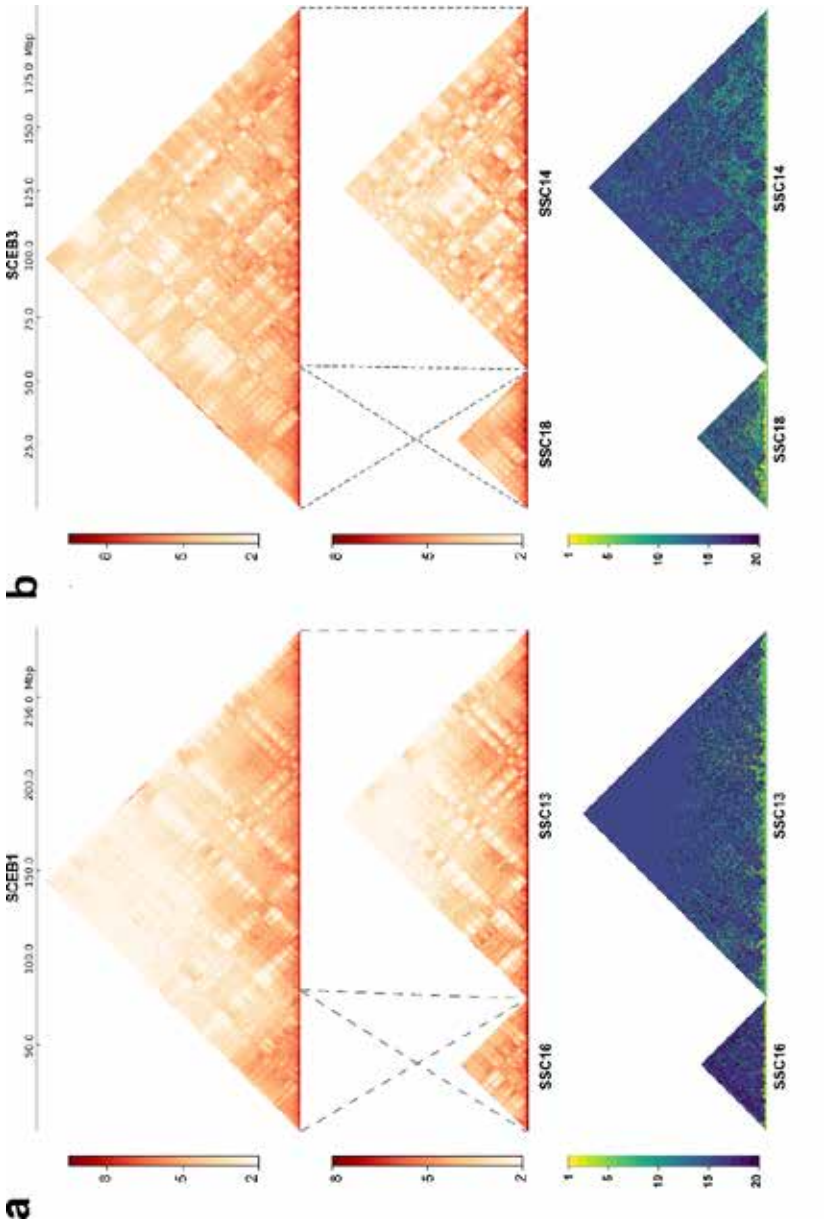


Fig 4.2. Chromosome interactions in homologous chromosome pairs. Hi-C contact map of a) SCEB1 and b) SCEB3, with signal scale displayed at the left. Dash lines highlight the homologous region between Visayan warty pig (top panel) and Duroc pig (middle panel). The darker color in the Hi-C contact map represents higher contact frequency. Bottom panel shows cross-species Hi-C contact frequency states identified by PhyloHMRF. The darker color represents less conserve state between two species.

Repetitive sequences, telomeres and centromeres

The numbers of different repeat classes and the divergence distribution of these repetitive elements were identified in Visayan warty pig genome (Sup Table. 4.3 and Sup Fig. 4.7 respectively). In total 43.9% of the Visayan warty pig genome consists of repetitive elements. The main repetitive transposable elements (TEs) were long interspersed nuclear elements (LINEs) (~511 Mb, covering 20.8% of genome).

Putative telomeres were identified at the proximal ends of Visayan warty pig chromosome assemblies of SCEB1-4, SCEB6, SCEB8, SCEB9 and SCEB11-16 (Sup Fig. 8, Sup Table. 4.4) and putative centromeres were identified in the Visayan warty pig chromosome for SCEB1-11, SCEB13, SCEB14 and SCEB16 (Sup Fig. 4.8, Sup Table. 4.5). Chromosome SCEB7 and SCEB14 both harbored two centromeric repeat regions, which is consist with observations for the Duroc genome (SSC8 and SSC11). However, in SCEB6 we only identified one centromeric region, while there are two in the homologous SSC15. Chromosomes SCEB1-11, SCEB13 and SCEB14 are metacentric, whilst chromosome SCEB17 is acrocentric.

Genome comparison between *Sus cebifrons* and *Sus scrofa*

Genome variations, such as insertions, deletions, inversions and duplications, are important sources of the genetic diversity that shapes phenotypic variations. We identified 52,827 deletions and 59,299 insertions (INDELs) with a length greater than 100 bp using *Sus scrofa* as the reference. The average lengths of the deletions (374 bp) and insertions (342 bp) were very similar (Sup Fig. 4.9), and >70% of the INDELs were overlapping with TEs. The TE distribution patterns for insertions and deletions are distinct. Short interspersed nuclear elements (SINEs) have the highest proportion associated with INDELs (38.3% for insertion, 30.6% for deletion). Long interspersed nuclear elements (LINEs) have the second highest proportion of deletions accounting for 27.0%, whereas simple sequence repeats (SSRs) have the largest number of insertions representing approximately 23.9% (Sup Fig. 4.10, Sup Table. 4.6). In addition, we identified 115 large inversions (INVs) with length longer than 100 kb, of which two, on SCEB4 (SSC6) and SCEB9 (SSC3), were longer than 1 Mb (Sup Fig. 4.5). We investigated the relationship between these major inversions and the density of the transposal elements. We found that the breakpoints of these two INVs were associated with low TE content, which may have induced these inversion events (Sup Fig. 4.11-4.12).

Genome annotation of the Visayan warty pig

RNA-seq data from the Visayan warty pig and the proteomes of six other mammalian species were used for the structural and functional gene annotations. By using an integrative annotation approach (see Material and Methods) we identified 21,153 protein-coding genes. Among the 21,153 protein-coding genes, 20,750 (99%) were functionally annotated with GO terms (Sup Table. 4.7). Gene annotation of mitochondria was performed using MitoZ (Meng et al., 2019) and 13 protein-coding genes as well as 21 tRNA genes were annotated (Sup Table. 4.8).

Of the predicted protein-coding genes, 15,175 genes (90%) are shared with Duroc pig, and 14,643 genes (80%) are shared with human. We compared the gene models of Visayan warty pig with other mammalian species. The structural feature of genes is highly conserved within the mammalian linages, and Visayan warty pig had very similar gene features with Duroc pig (Sup Fig. 4.13, Sup Table. 4.7).

In the Visayan warty pig genome, we predicted 21,467 non-coding RNA transcripts, including 11,767 Long noncoding RNAs (lncRNAs). Based on miRBase (http://www.mirbase.org) hairpin sequence alignments, 1037 miRNA loci were identified. On the basis of alignment evidence from Ensembl noncoding RNA sequences (release 95), approximately 20,515 non-coding RNA genes in Visayan warty pig had a reciprocal best hit (RBH) or second-best hit against Duroc pig (Sup Table. 4.9).

We estimated the number of expanded and contracted gene families among the six mammalian genomes (Fig 4.1c, Sup Table. 4.12-4.14). In the Visayan warty pig, Duroc pig and the ancestral *Sus* genus branch the most prominent gene family involved in gene expansion related to olfactory function.

Phylogenetic and divergence-time analysis

We obtained a high-confidence orthologous gene set of Visayan warty pig together with pig, cattle, horse, dog, mouse and human. From 18,586 gene family clusters we obtained 10,057 confidence single-copy gene families (Sup Fig. 4.15) which was used for phylogenetics analysis. The phylogenetic time tree (Sup Fig. 4.15) indicated that Visayan warty pig and Duroc pig diverged ~3.6 million years ago (Fig. 4.1c, Sup Fig. 4.16) and that the evolutionary rate in Visayan warty pig was ~0.82 × 10−9, which was significantly lower than that Duroc pig (t test, $P < 0.01$) (Sup Fig. 4.17).

Selective constraints on functional elements

Conservation of DNA sequences across distantly related species reflects functional constraints. We first generated multiple genome alignments from Visayan warty

pig, Duroc pig, cattle, horse, dog, mouse and human genome. Among the aligned mammalian genomes, we predicted 16.6 million highly conserved elements (HCEs) at a resolution of 20 bp or greater, spanning 7.5% of the Visayan warty pig genome. Functional annotations revealed that ~3.5% of these HCEs are associated with protein-coding genes, whereas the majority of the remaining HCEs are located in intron and intergenic regions (Sup Fig. 4.18). We further subset the *Sus* genus specific HCEs (SHCEs). Noncoding HCEs may play important roles in the regulation of gene expression. We compared the transcription factor binding sites found in the ENCODE project with these SHCEs. In total, 37.9% (49,846) of the SHCEs contain known transcription factor binding sites (Sup Table. 4.10), which are significantly associated with transcription factor functioning in temporal and spatial of cell differentiation (Sup Table. 4.11). To investigate evolutionary constraints on genes, we identified positively selected genes (PSGs) and rapidly evolving genes (REGs) for the 10,057 one-to-one orthologs. Functional enrichment analyses of the PSGs and REGs in Duroc pig and the ancestral *Sus* genus branch both exhibit enrichment in immune response and energy metabolism functions (Sup Table. 4.15-4.22). Specifically, we observed a series of PSGs and REGs in Visayan warty pig involved in regulation of growth (Sup Fig 4.19, Sup Table. 4.23-4.26). In addition, we identified PSGs in Visayan warty pig related to adaptive thermogenesis and chromatin organization.

Genomic variations related to species characteristics

In mammals, the chemosensory system mediates many behaviors such as locating food, mating and finding shelter sites. *Suidae* species do not have particularly good vision. Thus, they heavily utilize their olfactory and gustatory sensation (Sutherland-Smith, 2015). For example, olfactory receptor (OR) gene duplications are evident in the genomes of domesticated pig, which potentially enhances the adaptations to novel habitats (Truong Nguyen et al., 2012; Paudel et al., 2015). Having distinct evolutionary history, habitat and diet, Visayan warty pig and Duroc pig represent a prefect pair to identify genomic variations correlated with particular species characteristics.

Evolution of olfaction

In the Visayan warty pig genome, we identified 1,163 OR genes of which 921 are functional ORs. This number of functional ORs is approximately equivalent to the number found in Duroc pig (1,213 ORs, 1,089 functional ORs) and is more than most of other mammals (Fig. 4.3a, Sup table. 4.27). Based on the OR families we estimated the dynamics of ORs and found OR expansion at the ancestral branch of

Sus genus, Visayan warty pig and Duroc pig (Sup Fig. 4.20). This suggest that an expanded OR repertoire became important already in the *Sus* ancestor and we therefore looked in detail into specific OR function in Visayan warty pig. The diversity of the OR gene repertoire in the Visayan warty pig is represented by different families. Similar to Duroc pig (Truong Nguyen et al., 2012), the identified Visayan warty pig OR genes could be clustered into 2 classes, 17 families and further grouped into 428 subfamilies, based on phylogenetic analyses and their sequence similarity (Fig. 4.3b).

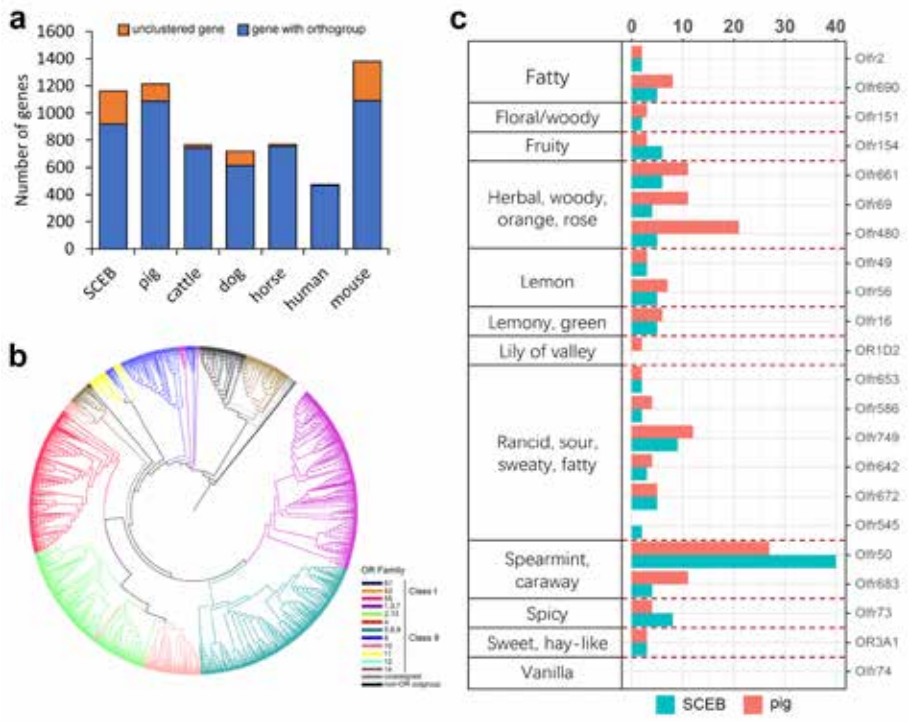


Fig.4.3 Evolution of Olfactory gene family. a). Shared and specific olfactory receptor (OR) genes in Visayan warty pig and six other mammalian species. b) Phylogenetic tree constructed with OR protein sequences of Visayan warty pig showing classification of ORs in Visayan warty pig. The different family members are colored according the colors shown in the legend of panel b. c) Potential association between olfactory receptor gene and odorant recognition.

To identify potential target specificity of pig OR genes in odor perception, we compared the amino acid sequences of translated Visayan warty pig OR genes and Duroc pig OR genes to those of other species with annotated information on

odorant specificity. Using Nguyen et al.'s methods (Truong Nguyen et al., 2012), two sequences were considered to have the same odorant recognition if their identity reached more than 60%. Of the 921 Visayan warty pig functional OR genes, 121 ORs matched to an OR with known odorant (Fig. 4.3c). We found that expanded ORs in Visayan warty pig are mainly involved in spearmint, caraway, spicy and fruit odor detection, while Duroc pig is more enriched in herbal, woody, orange, rose and fatty odor detection. Additionally, we calculated the amino acid diversity of each OR gene families. In general, OR genes in Duroc pig show higher diversity than those in Visayan warty pig (Sup table. 4.28).

Evolution of gustation

Taste perception affects mainly diet preference and consequently food/feed intake. The mammalian gustatory system usually discriminates five major basic taste classes: salty, sour, sweet, umami and bitter. In addition, as fat is an essential ingredient in feed, fat taste is one of the taste types with high potential implications relevant to the swine industry. We did a systematic study on taste transduction related genes in *Sus* (Fig. 4.4a, Sup table. 4.29).

For the 35 well characterized gustatory genes (Danilova et al., 1999; Glaser et al., 2001; Huang et al., 2006; Ishimaru et al., 2006; Bachmanov and Beauchamp, 2007; Horio et al., 2011; Roura et al., 2013; Li and Zhang, 2014), we carried out a comprehensive evolutionary analysis. Among all the taste receptor genes, only *TAS2R42* and *TAS2R39* had significantly elevated dN/dS ratios in the Visayan warty pig. The cluster of genes sensing amino acids showed a high proportion of members under relaxed selection in both Visayan warty pig and Duroc pig (Sup table. 4.29). All fatty acid sensors in Duroc pig are under relaxed selection, while none of them shows the same pattern in Visayan warty pig.

PKD1L3 and *PKD2L1* genes participate in reception of sourness. In the duroc pig annotation, *PKD1L3* is pseudogenized and identified as a long non-coding RNA (*LOC102158108*). In the Visayan warty pig genome, the putative PKD1L3 is overlapping with DHODH and is identified as an antisense long non-coding RNA (Fig. 4.4b).

Pigs have a limited ability to taste NaCl (Danilova et al., 1999; Hellekant and Danilova, 1999). Na+ taste reception involves the selective epithelial amiloride-sensitive sodium channel, *ENaC*. In most mammals, there are four *ENaC* channel subunits, α, β, γ, and δ. By comparing the ENaC genes among mammals, we found a *Sus* genus specific insertion (S420_Y421insNYG) in the *ENaC* channel subunit β (*SCNN1B*) (Fig. 4.4c). In a second candidate NaCl receptor, vanilloid receptor subtype 1 (*TRPV1*), we also found a *Sus* specific deletion (T662del in Visayan warty

pig and S661_R665del in Duroc pig, using human *TRPV1* as reference, Sub Fig. 4.21). These two amino acid variants are located in hydrophobic regions on the protein surface (Sub Fig. 4.22).

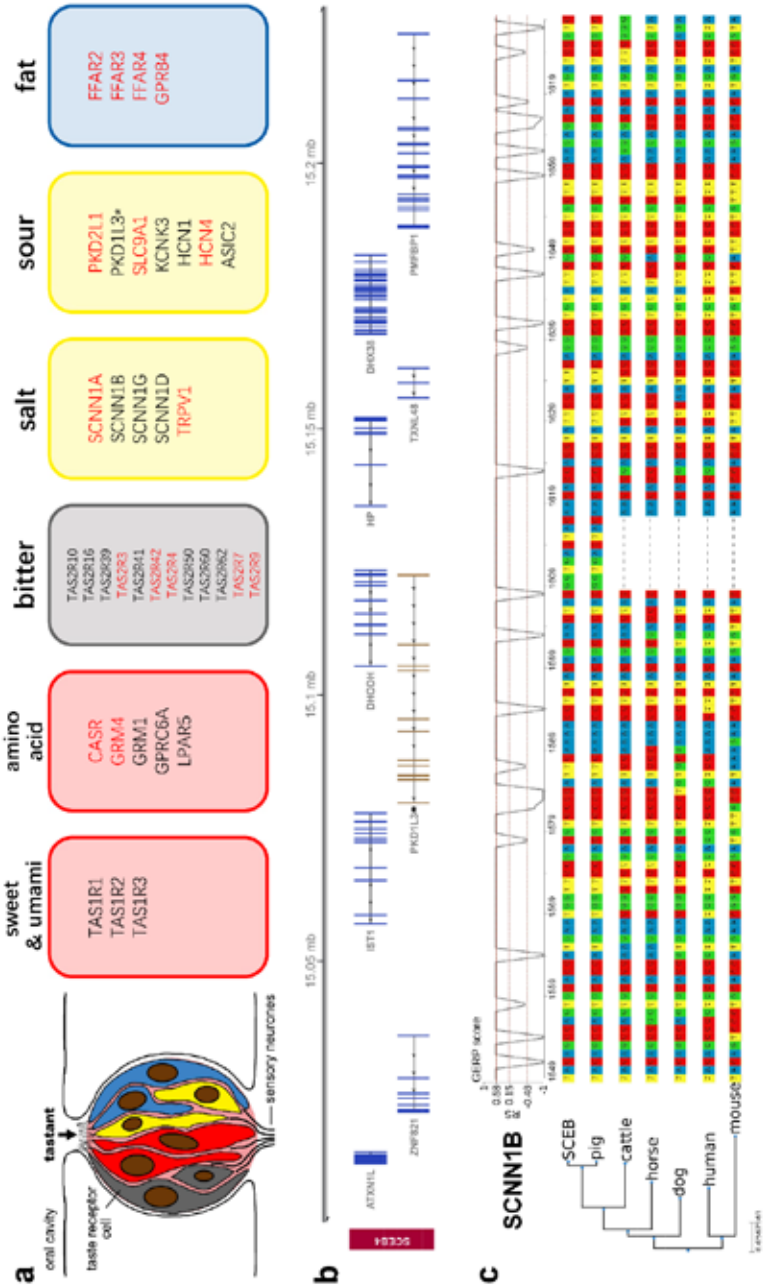


Fig.4.4 Genomic features related to taste transduction. a) Diagram of the taste bud cell and related taste receptor genes. The red font indicates different selection pressure present in Visayan warty pig and Duroc pig. Black Asterisk indicates a putative pseudogene. b) Genome track for Visayan warty pig. PKD1L3 is nested with DHODH. Functional genes are shown in blue, while the putative pseudogene is shown in brown. c) Multispecies DNA alignment of SCNN1B, showing the region with *Sus* genus specific insertion. GERP scores are shown on top of the alignment. Dashed lines represent upper quartile, average and lower quartile respectively.

4.3 Discussion

With the decrease in sequencing costs and improvement in genome assembly algorithms, it is expected that many more commercial pig and closely related pig species genome sequences will be published at a striking pace. In this study, we have produced a hybrid assembly approach for the Visayan warty pig genome using 10x Chromium linked-reads to generate long contigs. These contigs were scaffolded using Hi-C chromatin contact map and jump reads. This pipeline enabled us to produce a highly contiguous genome, with all chromosomes reconstructed as single scaffolds.

Evolution of *Suidae* genomes

On the basis of a comparison of the genome assembly of Visayan warty pig with that of Duroc pig, both genome sequences displayed considerable continuity with comparable scaffold N50 lengths. The characteristics of short read sequencing technology constrains its ability to resolve low-complexity repeats (Minoche et al., 2011). This explains the relatively low contig N50 (159.6kb) and incomplete telomeres & centromeres observed in the Visayan warty pig genome assembly. However, the contig N50 is less likely to affect Hi-C based genome scaffolding (Zhang et al., 2019). Thus, the clustering, ordering and orientation of our final assembly is likely reliable. We also compared basic genome structural features, including genes lengths, coding regions, and non-coding regions of the Visayan warty pig genome with the Duroc pig, all of which showed a striking level of similarity. Interestingly, we found that the breakpoints of inverted homologous region between Visayan warty pig and Duroc pig are highly associated with low TE content. It was hypothesized that TEs can induced chromosomal rearrangements. TE contents are observed to be enriched in chromosome regions proximal to the inversion breakpoints in plants, invertebrates and vertebrates (Fedoroff, 1988; Lim and Simmons, 1994; Bennetzen, 2005; Bourque, 2009; Konkel and Batzer, 2010;

Parisod et al., 2010). However, to date, most of these studies were done within one species, and no unified characterization of the mechanisms governing the inversion process has been addressed. Our cross-species comparison implied that the different role of TEs in the inversion process may reflect differences in evolutionary timescales.

Using fossil calibrations, we estimated the divergence between Visayan warty pig and Duroc pig at 4.57 million to 2.71 million years ago (early Pliocene), which is reflected in the high collinearity between them. In addition, we observed a chromosome number difference when comparing these two species (Visayan warty pig 2n=34, Duroc pig 2n=38), which has been reported previously using G-banding and painting probes (Musilova et al., 2010). A systematic survey of the karyotype of members of the *Suidae* family indicated that the centric fusion in the Visayan warty pig genome homologous to *Sus scrofa* chromosomes (SSC) 13/16 is more likely to be the ancestral karyotype of *Suidae* (Sup Fig 4.23). Pygmy hog (*Porcula salvania*) and other *Sus* species are all carrying separated SSC13/SSC16 homologous chromosome (Bosma et al., 1983, 1991; Rothschild and Ruvinsky, 2011). However, in the more ancestral linages, *Babyrousa* and Sub-Saharan Suids (MELANDER and HANSEN‐MELANDER, 1980; Bosma et al., 1996; Thomsen et al., 1996; Musilova et al., 2010), SSC13/SSC16 are fused. We hypothesize that the ancestral chromosome SSC13/SSC16 split in the common ancestor of *Porcula* and *Sus*, while the centric fusion independently occurred in Visayan warty pig. The other centric fusion, homologous to SSC14/18, is only found in Visayan warty pig. Notably, chromosome fusion/fission seems to happen quite frequently among the *Suidae* linage. For example, the karyotype of Sub-Saharan suids (2n = 34) differs from that of the domestic pig by the presence of 2 fusion chromosomes homologous to SSC13/16 and 15/17 (Musilova et al., 2010), and some wild boars living in Western Europe carry a single centric fusion between SSC15 and SSC17 (Raudsepp and Chowdhary, 2011). Moreover, fertile hybrid off-spring between *Sus scrofa* and Visayan warty pig, as well as other ISEA *Sus* (Blouch and Groves, 1990; IUCN, 1993; Ruvinsky et al., 2011), have been recorded. Genome analyses also reveal that *Sus scrofa* exhibited extensive gene flow with other *Sus* species, and even with its sister genus Porcula, during range expansion (Frantz et al., 2013, 2014; Liu et al., 2019). This indicates a limited post-zygotic reproductive isolation despite frequent chromosomal re-arrangements. In meiotic prophase, homologous chromosomes must be organized into compacted loop arrays in order to identify one another, pair along their length and physically link to ensure their accurate recombination and segregation in the meiosis I division (Keeney et al., 2014; Patel et al., 2019). The comparison of chromosome interaction maps suggests that, in the short divergence time between

Visayan warty pig and Duroc pig (~3 Mya), the chromosome structure remains the same after chromosome fusion. This may explain the loose reproduction barrier between *Sus* species. Even with different chromosome numbers, the synapsis during meiosis is probably not interrupted, leading to the absence of post-zygotic reproductive isolation.

Selective constraints on functional elements and genes

With the newly generated Visayan warty pig genome assembly, we were able to conduct evolutionary genomic analyses aimed at revealing genomic variations correlated with particular features in the *Sus* genus. Noncoding HCEs play an important role in the regulation of gene expression, which among others has been reported to direct gene expression involved in establishing of the anatomical structures in the embryo (Pennacchio et al., 2006). We found *Sus* genus specific noncoding HCEs to be significantly associated with transcription factors functioning in temporal and spatial cell differentiation, for example at the *PAX9* gene. *PAX9* regulates squamous cell differentiation in the esophageal epithelium, which has been reported to be associated with rumen development in pecoran (Doane and Elemento, 2017; Chen et al., 2019). Although *Artiodactyla* species are highly diverse in morphology, most of the families have multichambered stomach, except *Suidae* which is the only monogastric *Artiodactyla* (Langer, 1974). The *Sus* genus specific noncoding HCE associated with *PAX9* could potentially be related to the monogastric trait, corresponding to the omnivorous diet (Miller and Ullrey, 1987; Robeson et al., 2018).

Functional enrichment analyses of the PSGs and REGs in the ancestral *Sus* genus branch both exhibit enrichment in immune response to various pathogens and energy metabolism functions. Although fossil evidence cannot resolve the origin and dispersion route of *Sus* species, it has been well supported that during Pliocene this genus effectively colonized most of the Eurasia continent and ISEA (Azzaroli, 1992; van der Made et al., 2006; Frantz et al., 2013; Pickford, 2013). The genomic changes associated with olfactory function, immune system and metabolism may relate to the wide distribution and effective local adaptation of the ancestral *Sus* population.

Comparing the genome of Visayan warty pig with domestic pigs further showed species specific characteristics and the impact of domestication. In the genome of the Visayan warty pig, we observed a series of PSGs and REGs involved in regulation of growth. Like many island species, the Visayan warty pig is relatively small in size, with an adult weight of 20-40 kg (Rabor, 1977; Oliver et al., 1993; Clauss et al., 2008). Accordingly, adult weight of Duroc pig shows a up to ten-fold

difference (250-300 kb) to that of Visayan warty pig (Evans et al., 1946; Comission, 2010). We also identified PSGs in Visayan warty pig related to thermoregulation. The geographical distribution of Visayan warty pig has always been restricted to the Visayan islands which have a typical tropical rainforest climate. This region does not experience significant seasonal changes and keeps a high annual mean temperature (Villafuerte et al., 2020). The genomic changes associated with body size and body temperature could reflect the adaptation to a tropical island environment. In addition, we identified six PSGs related to DNA ligation and DNA repair in Visayan warty pig: *BRCA1*, *RIF1*, *TIMELESS*, *DNA2*, *BABAM2*, *XRCC6*, which may have related to the described chromosome rearrangements.

Domestication and subsequent long-term selective breeding have profoundly changed the anatomical, physiological, and behavioral characteristics of livestock with clear selective genomic signatures. We found a functional enrichment of the PSGs and REGs in Duroc pig exhibiting enrichment in neural functions, immune response, energy metabolism functions, growth and muscle development, which is consistent with previous reports of genes related to pig domestication. A typical altered trait of domesticated animals is reduction of alertness and sensitivity to environmental changes. To fit the demand of meat production, selection in pig breeding programs has been vigorous for factors such as meat yield and growth rate and adaptation to high-calorie diets and minimal activity. Nonetheless, we would like to note that without the genome information of the wild boar, we cannot pinpoint when did the selection constraints occur. Future studies, where genome assembly of wild boar and, ideally, genome on early pig domestication are available, will probably further reveal the genetic basis underlying the evolution of the pig.

Evolution of olfactory and gustatory sensation

As two of the basic senses of mammals, olfaction and taste play a very important role in daily life. These two types of chemical sensors are important for recognizing environmental conditions, and have a positive coactivation (integration) in flavor perception (Murphy and Cain, 1980; Veldhuizen et al., 2009).

Olfactory receptors are largely responsible for odor perception and detection of chemical cues, facilitating the differentiation of tens of thousands of unique odorants. This makes olfaction an important physiological function crucial for the success of animals because of their role in recognizing suitable food, mates, offspring, territories, and the presence of predators or prey. We Identified 921 and 1089 functional ORs in Visayan warty pig and Duroc pig, respectively, which are higher than for most mammals, but lower than e.g. mouse (1135 ORs). The mouse

has a large number of OR genes, which probably reflects its nocturnal lifestyle, complex living conditions (Cuesta et al. 2009) and the ability to identify various odors from complex environments. Olfactory gene families were expanded in the ancestral *Sus* branch, and have been independently accumulating in two *Sus* species linages. Functional categories of expanded ORs in Visayan warty pig are different from Duroc pig, which suggests that environmental adaptation of the two species are orchestrated by the olfactory system. For example, we found expanded ORs involved in the identification of fatty odor in Duroc pig. Wild *Suiformes* are foraging animals with a diet consisting roughly 90% of plant-derived foods such as fruits, roots, leaves and grasses. The relative amount of dietary fats is usually low (less than 10%) (Leus, 1994; Ballari and Barrios-García, 2014; Souron et al., 2015), while in the swine industry additional fat is added to feed for a better performance (Lewis and Lee Southern, 2000). Fat related OR duplications could be the consequence of adaptation to a high-fat diet. Moreover, diversification of genes within each OR subfamily in Visayan warty pig is lower than in Duroc pig. Visayan warty pig is specialized to the tropical island habitat in Southeast Asia. In contrast, the source population of the domesticated pigs, the wild boar (*Sus scrofa*) has an enormously wide distribution across the globe (also include ISEA). This requires wild boar to heavily utilize their olfactory system to acclimate the complex and various environment. We hypothesis that the high diversity of OR genes already existed in the pig genome before domestication and has been retained in current domesticated pig genome.

Taste receptors play a fundamental role in survival through the identification of dietary nutrients or potentially toxic compounds. Thus, dietary adaptation through taste sensory mechanisms is emerging as a major evolutionary force. *TAS2R42* and *TAS2R39* in Visayan warty pig are under positive selection. *TAS2Rs* are part of the sensory mechanism to identify potentially toxic compounds and elicits bitter taste (Bachmanov and Beauchamp, 2007). This positive selection can be explained by adaptation to avoid undesirable substances, e.g. plant and insect toxins. Accordingly, relaxed selection constraint can lead to trait reduction or loss (Lahti et al., 2009; Bely, 2010; Wicke et al., 2016). Relaxed selection of fat sensory genes in Duroc pig indicated a weakened fatty taste sensibility, which may be due to the selective breeding history of high-fat diet tolerance.

It was reported that pigs have a limited ability to taste NaCl (Danilova et al., 1999; Hellekant and Danilova, 1999). Compared to other mammalian species, we identified two *Sus* genus specific INDELs in *SCNN1B* and *TRPV1*. Those INDELs are not located in the transmembrane region, and are both in-frame mutations, which may not affect the structural integrity of the proteins. However, the multispecies

alignments show that around those INDELs sequences are relatively conserved. Also, by modelling the protein structure, we hypothesize that the conformation of the hydrophobic surface has changed. As channel proteins, those amino acid mutations could potentially influence the combination and transport of ions, which further reduced the nerve response to salty gustatory stimulate in *Sus* species.

Notably, we observed pseudogenization of *PKD1L3* in both Visayan warty pig and Duroc pig, which overlaps with *DHODH*. *PKD1L3* presents in most mammalian species and does not overlap with *DHODH*. *PKD1L3* and *PKD2L1* genes participate in reception of sourness. *PKD1L3* and *PKD2L1* are co-expressed, and interact through their transmembrane domains (Ishimaru et al., 2006). The resulting heterodimer (*PKD1L3*/*PKD2L1*) forms a unique hydrogen gated channel for acid stimulation (Ishii et al., 2012). Under *in vitro* conditions, *PKD2L1* is only activated by acid when it is co-expressed with *PKD1L3* (Ishimaru et al., 2006). However, *in vivo* studies have shown that *PKD2L1* expressing cells response to acid when lacking *PKD1L3* expression (Huang et al., 2006; Chang et al., 2010). Nerve pulse recordings during taste stimulation indicate that acids give a distinct taste to the pig (Hellekant and Danilova, 1999; Glaser et al., 2001). This suggests that an unknown compensatory effect for H+ response may exist. Or instead of generating sense of sourness, *PKD1L3* absent individuals might still respond to acid simply by chemical stimulation. Further research is needed to answer this unresolved question.

4.4 Conclusion

Our comprehensive evolutionary and comparative genome analyses provided insight into the distinct evolutionary scenarios occurring during recent species divergence. Comparing the chromosome-level assembly of Visayan warty pig with domestic Duroc pig has revealed the potential biological mechanism of frequent post-divergence hybridization among *Suidae*. We identified numerous genetic variations correlated with natural selection for local adaptation and artificial selection during domestication, shedding new light on the origin of species-specific characteristics. This study provides valuable genomic resources as well as insights into not only the evolution and diversification of Suiform but also our understanding of mammalian biology.

4.5 Material and Methods

Sample collection and sequencing

DNA of a female Visayan warty pig was obtained from tissue samples at the Rotterdam Zoo. Tissues for RNA-seq were dissected from seven organs (liver, spleen, lymph, muscle, lung, frontal lobe, olfactory bulb).

10× Genomics library: Visayan warty pig's genomic DNA was size selected for fragments >40 kb on a BluePippin instrument (Sage Sciences, Beverly, MA, USA) and Illumina sequencing libraries were constructed using the 10× Genomics Chromium Controller instrument with the Genome Reagents Kit v2 chemistry (10× Genomics, Pleasanton, CA, USA) according to the manufacturer's recommendations. The resulting Illumina library was sequenced on a NextSeq500 using a High Output Kit v2 for paired-end, 2 × 151 bp run (Illumina, San Diego CA, USA).

Hi-C library: Hi-C library was generated by Dovetail Genomics as described by Lieberman-Aiden et al (2009). Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from protein. Purified DNA was treated to remove biotin that was not linked to ligated fragments. The DNA was then sheared to a mean fragment size of ~350 bp, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina Hi-seq 4000 sequencing platform (2x151bp).

Long jumping libraries: Briefly, DNA was extracted from the same Visayan warty pig tissues, and used to create two mate-pair libraries, with insert sizes of 10 and 20 kb respectively. The mate-pair libraries were prepared with Illumina's "Nextera Mate-Pairs Sample Prep kit" followed by the "TruSeq DNA Sample Prep kit" and sequenced on Illumina HiSeq 2500 platform.

RNA-seq: High-quality RNA (1 µg) was used to generate TruSeq Stranded RNA-seq libraries (TruSeq Stranded RNA Sample Preparation Kit, Illumina, San Diego, CA, USA) by the HTS lab (University of Illinois, Urbana, IL, USA) following standard protocols and sequenced on an Illumina HiSeq 2000 platform.

Genome assembly

Based on linked-reads that were sequenced with a standard Illumina system, we used the Supernova assembler (version 1.2.2, 10X Genomics Inc., Pleasanton, CA, USA) with the parameters (run --maxreads=1,500,000,000; mkoutput --style=pseudohap2 --minsize=500) to construct contigs and scaffolds of the Visayan warty pig.

Chromosome assembly using Hi-C

The input *de novo* assembly, Chicago library reads, and Dovetail Hi-C library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam et al., 2016). An iterative analysis was conducted. First, Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper. The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a default threshold. After aligning and scaffolding Chicago data, Dovetail Hi-C library sequences were aligned and scaffolded following the same method. After scaffolding, shotgun sequences were used to close gaps between contigs.

Reference guided and long jumping reads scaffolding

Since the high relatedness between Visayan warty pig and Duroc pig (*Sus scrofa*) (Frantz et al., 2013), the HiRise assembly was subjected to a secondary assembly with AlignGraph (Bao et al., 2014) using the Sus scrofa11.1 reference genome (Warr et al., 2019) as guidance. Briefly, in the guided secondary assembly with AlignGraph the de-novo generated scaffolds are aligned to the reference and the long jumping reads (10kb and 20kb insert size) are mapped to the assembled scaffolds and to the reference. This results in a paired-end multi-positional de Bruijn graph from which the scaffolds are extended if possible.

Mitochondrial genome assembly

We randomly selected 4 Gb PE reads from resequencing library to assemble mitochondrial genome sequence using MitoZ, resulting in 16.52 kb assembly with a cyclic structure (Sup Fig. 4).

Evaluation of genome assembly

The Illumina short reads from the 10X libraries were used for kmer frequency analyses of genomes. A k-mer count analysis was done using Jellyfish v2.3.0 (Marçais and Kingsford, 2011) on the Illumina data. From the paired end reads, only the first read was truncated to 100 bp to avoid the lower quality part of the read. The second read was omitted from this analysis to avoid counting overlapping k-mers. A k-mer size k=21 was used. After converting the k-mer counts into a histogram format, this file was analyzed using the Genomescope (Vurture et al., 2017).

Evaluation of genome completeness with BUSCO

BUSCO v3.1.0 (Simão et al., 2015) was used to assess the genome completeness by estimating the percentage of expected single copy conserved orthologs captured in the Visayan warty pig assembly, according to the mammalia_odb9 BUSCO set.

Cross-species Hi-C processing

Hi-C data for pig were downloaded from FR-AgENCODE (http://www.fragencode.org/results.html) (Foissac et al., 2019). Each mate of the Hi-C sequencing read pairs was aligned separately using bwa mem 0.7.5a (Li and Durbin, 2009) with parameters '-E50 -L0'. We mapped approximately 7.31E+07 and 2.26E+07 Hi-C contact pairs for Visayan warty pig and Duroc pig genome, respectively. Next, we aligned the Hi-C contact pairs of Visayan warty pig to the pig genome. We mapped the aligned loci of the two ends of a contact pair in the Hi-C data of the Visayan warty pig from the original genome assembly to the pig genome with reciprocal mapping using the tool liftOver (Hinrichs, 2006) based on the whole genome alignment. We used HiCExplorer v3.0 (Ramírez et al., 2018) to create the contact matrices and extract the Hi-C contact maps of each species at the resolution of 50kb. Normalization was performed using Knight-Ruiz matrix balancing (Knight and Ruiz, 2013). We applied Phylo-HMRF (Yang et al., 2019) to the multi-species Hi-C data to predict 20 hidden states. For all the autosomes and chromosome X in the Duroc pig genome, we run Phylo-HMRF jointly on the multiple synteny blocks of the chromosomes and identified possible different evolutionary patterns of the Hi-C contact frequencies across species in a genome-wide manner. We further categorized the 20 estimated hidden states based on their conservation status.

Identification of repetitive elements

We identified repetitive elements by a combination of homology alignment and *de novo* searches, as follows. We used RepeatMasker (Tarailo-Graovac and Chen, 2009) with Repbase (v.16.10) (Jurka et al., 2005) to scan for sequences homologous to annotated repeat sequences in published databases and then used RepeatModeler (Jurka et al., 2005) (http://www.repeatmasker.org/RepeatModeler.html) with the default parameters to predict *de novo* TEs. We combined the repeat sequences identified by both methods as the final annotated repeat set. We integrated the overlapping TEs and removed those with scores less than 200. Subsequently, tandem repeats were annotated using Tandem Repeat Finder (Benson, 1999) (version 4.07b; with following settings: 2 7 7 80 10 50 2000 -d -h).

Gene prediction

Gene prediction was conducted with the BRAKER v2.1.2 (Hoff et al., 2019) pipeline by integrating *ab initio*, transcriptome-based and protein homology-based evidence. *Ab initio* gene predictions were performed with Augustus v3.3.2 (Stanke et al., 2006) and GeneMark-ET v4.33 (Lomsadze et al., 2014). Before starting the annotation process, we used the mammalia_odb9 BUSCO set to train Augustus. For prediction based on the RNA-seq data, all RNA reads were aligned to the Visayan warty pig genome by HISAT (Kim et al., 2015). The results were used to identify candidate exon regions, donor and acceptor sites. Protein sequences of all mammalian species were downloaded from UniProt (Bateman et al., 2015) as protein homology-based evidence.

Single exon genes (SEG) were identified as mono-exonic genes containing full-length ORFs, as specified by the gene prediction pipeline. The same approach was applied to the human, mouse, dog, horse, cattle and pig genomes. Lists of non-redundant SEG from all six genomes were obtained and aligned to the Visayan warty pig genome using TBLASTN (Camacho et al., 2009) (E-value cutoff: 1E-5). We aligned homologous genomic sequences against matching proteins to define gene models using GeneWise v2.4.1 (Birney and Durbin, 2000). Gene annotation of mitochondria was performed using MitoZ software.

EVidenceModeler v1.1.1 (Haas et al., 2008) was used to integrate the genes predicted by BRAKER and SEG approaches and generate a consensus gene set.

Functional annotation

We annotated the functions of the protein-coding genes using BLASTP (E-value cutoff: 1E-5) (Camacho et al., 2009) based on entries in the Swiss-Prot and TrEMBL databases (Bateman et al., 2015). The motifs and domains were annotated using

InterProScan v5.35 (Jones et al., 2014) by searching against publicly available databases, including ProDom, PRINTS, Pfam, SMRT, PANTHER and PROSITE. The Gene Ontology (GO) IDs for each gene were assigned according to the corresponding InterPro entry.

Non-coding gene element prediction

We also predicted gene structures of tRNAs, rRNAs and other non-coding RNAs. A total of 37,019 tRNAs were predicted using t-RNAscan-SE v2.0 (Lowe and Eddy, 1996). Because rRNA genes are highly evolutionarily conserved, we choose human rRNA sequence as references and then predicted rRNA genes using Blast tool with default parameters. Small nuclear and nucleolar RNAs were annotated using the Infernal v1.1.2 (Nawrocki and Eddy, 2013). Long non-coding RNAs were detected using FEELnc (https://github.com/tderrien/FEELnc) with default parameters and the output transcripts of the HISAT alignments.

Whole genome alignments and structural variation

The Nucmer program from Mummer v3.23 (Haas et al., 2008) was used with -maxgap 50 and -breaklen 400 for comparing the genomes between Visayan warty pig to Duroc pig to obtain the synteny blocks. Structural variations between Visayan warty pig and Duroc pig were called using svmu as described by Chakraborty et al (2019)

For multiple genome alignments from Visayan warty pig, Duroc pig, cattle, horse, dog, mouse and human, we first aligned all genomes to the human genome using LAST v980 (Kiełbasa et al., 2011), Then, we used MULTIZ v11.2 (Blanchette et al., 2004) to merge the pairwise alignments into multiple genome alignments using the human genome as the reference. Approximately 241 Mb syntenic sequence are shared by the seven mammalian species.

Gene family, phylogenetic analyses and divergence time calibration

We used the Orthofinder v2.3.3 to identify gene families/clusters (Emms and Kelly, 2019). Genome sequences and annotations for Duroc pig, cattle, horse, dog, mouse and human were downloaded from Ensembl (https://www.ensembl.org). The longest proteins of each gene were aligned to one and another. The species-specific gene families were determined according to the presence or absence of genes for a given species. The shared family expansion and contraction analysis were conducted with CAFÉ v3.1 (De Bie et al., 2006).

Phylogenetic relationships were resolved based on the maximum-likelihood (ML) method as implemented in RAxML v8.2.3 (Stamatakis, 2014) using the best-fitting

model of substitutions, identified by jModelTest2 (Darriba et al., 2012). 10,057 high-quality 1:1 single copy orthologous genes were used.

We estimated divergence times using an approximate likelihood method as implemented in MCMCtree (Yang, 2007), with an independent relaxed clock and birth-death sampling. We used a float prior and a maximum bound age, with a scale parameter of c=2. For the root divergence, we set the prior to (tU=2 [100 Mya], p=0.1, c=2). For MRCA of *Sus*, we used the same fossil calibration as in Frantz et al. (2013) (tL=0.2 [2 Mya], p=0.1, c=0.5).

DNA evolutionary rates in each species, the common ancestor of *Sus* and *Artiodactyla* were calculated using r8s (Sanderson, 2003) based on the whole genome alignments

Identification of highly conserved elements (HCEs)

PHAST v1.5 program package (Hubisz et al., 2011) was used to identify highly conserved elements (HCEs) based on the multiple-genome sequence alignments. Firstly, we extracted the 4-fold degenerate sites from the multiple genome alignments generated above and used phyloFit (in PHAST v1.5 package) to estimate a neutral phylogenetic model (also considered as the nonconserved model in PhastCons). Then we ran PhastCons (in PHAST v1.5 package) to estimate conserved and nonconserved models and predicted conserved elements and conservation scores based on the conserved and nonconserved models. We identified 1,162,748 HCEs, covering 137.2 Mb of the human genome. Further, we identified Sus genus specific HCEs (SHCEs), which are conserved in Sus species, but not in other mammalian species. We used phyloP (in PHAST v1.5 package, with the parameters: phyloP --method LRT --mode CONACC) under the non-conserved model. Finally, we identified 49,846 SHCEs, covering 13.9 Mb.

Non-coding conserved regions usually contain regulatory elements, i.e. transcription factor binding sites (TFBSes). We downloaded the human TFBS data from the UCSC data set generated by the ENCODE project (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/encRegTfbsClustered/encRegTfbsClusteredWithCells.hg38.bed.gz) to compare the conservation degrees of SHCEs. We found that of 4,379,357 human TFBSes used in analysis, 71,908 (1.6%) was conserved in *Sus*.

Identification of selective constraint in genes

Positively selected genes (PSGs)

The strength of positive selection on each codon of each orthologous gene along a specific targeted lineage of a phylogenetic tree was estimated with the branch site

model using the Codeml program of the PAML package v4.9. To determine to what degree these codon sequences, along the targeted lineage fit the branch site model (including positive selection) better than the one containing neutral selection or negative selection. An alternative branch site model (Model = 2, NSsites = 2 and Fix = 0) and a null branch site model (Model = 2, NSsites = 2, Fix = 1 and Fix ω = 1) was combined to calculate log-likelihood values for each model using likelihood ratio tests. The log-likelihood values generated were used to assess the model fit, using the Chi-square test with one degree of freedom. Genes with a p value less than 0.05 were treated as candidates under positive selection.

Rapidly evolved genes (REGs)

Branch model in Codeml was used to identify evolutionary rate among orthologous genes. To determine whether these codon sequences along the targeted lineage are evolving under a different evolutionary rate than other lineages, an alternative branch site model (Model = 2) and a null branch site model (Model = 0) were combined to calculate log-likelihood values for each model using likelihood ratio tests. Genes with a p value less than 0.05 and a higher ω value for the foreground than the background branches were considered as evolving with a faster rate.

Relaxed selected genes

RELAX (Wertheim et al., 2015) analyses were also performed to evaluate whether selective constraints are stronger in foreground branches compared with background branches. The analyses were performed using the RELAX tool on datamonkey server (test.datamonkey.org/relax/). An input file that contains the codon sequence alignment and the RAxML tree was provided, and the foreground branches (test branches) and background branches (reference branches) were then indicated before running the analysis. A likelihood ratio test was also performed to evaluate if selection varied between test and reference branches. A significant result (P value < 0.05) indicates a stronger level of selection if K is greater than 1 or a weaker selection or relaxed constraint if K is less than 1.

Functional, pathway and interaction enrichment analysis

KOBAS (version 3.0) (Wu et al., 2006) was applied to perform GO analysis (including cellular composition, molecular function and biological process terms), Reactome pathway and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. False discovery rate was performed to adjust p values using the Benjamini and Hochberg method. An adjusted p value of < 0.05 was used as the cutoff criterion.

Interspecies conservation of taste receptor genes

We used genomic evolutionary rate profiling (GERP++) (Davydov et al., 2010) to estimate the interspecies conservation of taste receptor genes. GERP++ provides conservation scores through the quantification of position-specific constraint in multiple-species alignments. We calculated and attributed the GERP scores for each nucleotide site.

Protein structure modelling

The three dimensional homology models of proteins were obtained using Swiss-Model (https://swissmodel.expasy.org/) (Waterhouse et al., 2018) to search for templates and build the models. The predicted protein structures were estimated based on target–template alignment using ProMod3 on SWISS-MODEL server (https://swissmodel.expasy.org/). Models were built based on the target–template alignment using ProMod3. The global and per-residue model quality was assessed using the QMEAN scoring function.

Ethics statement

DNA of Visayan warty pig was obtained from tissue samples collected by veterinarians at the Rotterdam Zoo, The Netherlands, according to national legislation. Tissue samples obtained are derived from an animal culled within the wildlife management program of the Rotterdam Zoo.

Description of supplementary Material

For a compact layout, in this thesis I did not include all supplementary material. I presented Supplementary Figures which may help the reader. For sake of coherence, I kept the original number of Supplementary Figures and tables.
Complete supplementary material are available at the Open Science Framework repository: https://osf.io/6yznw.

References

Azzaroli, A. (1992). Suids of the early villafranchian of villafranca d'asti and china. *Rend. Lincei* 3, 109–124. doi:10.1007/BF03002969.

Bachmanov, A. A., and Beauchamp, G. K. (2007). Taste Receptor Genes. *Annu. Rev. Nutr.* 27, 389–414. doi:10.1146/annurev.nutr.26.061505.111329.

Ballari, S. A., and Barrios-García, M. N. (2014). A review of wild boar Sus scrofa diet and factors affecting food selection in native and introduced ranges. *Mamm. Rev.* 44, 124–134. doi:10.1111/mam.12015.

Bao, E., Jiang, T., and Girke, T. (2014). AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics* 30, i319–i328. doi:10.1093/bioinformatics/btu291.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., et al. (2015). UniProt: A hub for protein information. *Nucleic Acids Res.* 43, D204–D212. doi:10.1093/nar/gku989.

Bely, A. E. (2010). Evolutionary loss of animal regeneration: Pattern and process. in *Integrative and Comparative Biology* (Oxford Academic), 515–527. doi:10.1093/icb/icq118.

Bennetzen, J. L. (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr. Opin. Genet. Dev.* 15, 621–627. doi:10.1016/j.gde.2005.09.010.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi:10.1093/nar/27.2.573.

Birney, E., and Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res.* 10, 547–548. doi:10.1101/gr.10.4.547.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smith, A. F. A., Roskin, K. M., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715. doi:10.1101/gr.1933104.

Blouch, R. A., and Groves, C. P. (1990). Naturally occurring suid hybrid in Java. *Z. Säugetierkd.* 55, 270–275. Available at: http://www.biodiversitylibrary.org/ [Accessed April 23, 2020].

Bosma, A. A., De Haan, N. A., Blouch, R. A., and Macdonald, A. A. (1991). Comparative cytogenetic studies in Sus verrucosus, Sus celebensis and Sus scrofa vittatus (Suidae, Mammalia). *Genetica* 83, 189–194. doi:10.1007/BF00126224.

Bosma, A. A., Mellink, C. H. M., Yerle, M., and Zijlstra, C. (1996). Chromosome homology between the domestic pig and the babirusa (family Suidae) elucidated with the use of porcine painting probes. *Cytogenet. Genome Res.* 75, 32–35. doi:10.1159/000134452.

Bosma, A. A., Oliver, W. L. R., and Macdonald, A. A. (1983). The karyotype, including G- and C-banding patterns, of the pigmy hog Sus (Porcula) salvanius (Suidae, Mammalia). doi:10.1007/BF00123219.

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.* 19, 607–612. doi:10.1016/j.gde.2009.10.013.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421.

Chakraborty, M., Emerson, J. J., Macdonald, S. J., and Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* 10, 1–11. doi:10.1038/s41467-019-12884-1.

Chang, R. B., Waters, H., and Liman, E. R. (2010). A proton current drives action potentials in genetically identified sour taste cells. *Proc. Natl. Acad. Sci. U. S. A.* 107, 22320–22325. doi:10.1073/pnas.1013664107.

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., et al. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science (80-.).* 364. doi:10.1126/science.aav6202.

Clauss, M., Nijboer, J., Loermans, J. H. M., Roth, T., Van Der Kuilen, J., and Beynen, A. C. (2008). Comparative digestion studies in wild suids at Rotterdam Zoo. *Zoo Biol.* 27, 305–319. doi:10.1002/zoo.20191.

Comission, E. (2010). Farmer's Hand Book on Pig Production Food and Agriculture Organization of the United Nations.

Coyne, J. a, and Orr, H. A. (2009). Speciation: A Catalogue and Critique of Species Concepts. *Philos. Biol. an Anthol.*, 272–292. Available at: http://books.google.com/books?id=tBxGpaV-ocsC.

Danilova, V., Roberts, T., and Hellekant, G. (1999). Responses of single taste fibers and whole chorda tympani and glossopharyngeal nerve in the domestic pig, Sus scrofa. doi:10.1093/chemse/24.3.301.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). JModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* 9, 772. doi:10.1038/nmeth.2109.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6. doi:10.1371/journal.pcbi.1001025.

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi:10.1093/bioinformatics/btl097.

Doane, A. S., and Elemento, O. (2017). Regulatory elements in molecular networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 9, e1374. doi:10.1002/wsbm.1374.

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi:10.1186/s13059-019-1832-y.

Evans, B. R., Evans, G. G., and others (1946). The story of Durocs, the truly American breed of swine. *story Durocs, truly Am. breed swine.*

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

Fedoroff, N. V. (1988). *The discovery and characterization of transposable elements. The collected papers of Barbara McClintock*. Garland Pub. doi:10.1016/0092-8674(88)90481-3.

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., et al. (2019). Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol.* 17, 108. doi:10.1186/s12915-019-0726-5.

Frantz, L. A. F., Madsen, O., Megens, H. J., Groenen, M. A. M., and Lohse, K. (2014). Testing models of speciation from genome sequences: Divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Mol. Ecol.* 23, 5566–5574. doi:10.1111/mec.12958.

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

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H. J., Cagan, A., Bosse, M., et al. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* 47, 1141–1148. doi:10.1038/ng.3394.

Glaser, D., Wanner, M., Tinti, J. M., and Nofre, C. (2001). Pig Responses to Taste Stimuli. doi:10.1007/978-1-4615-0671-3_58.

Gottlieb, G. (2002). Developmental-behavioral initiation of evolutionary change. *Psychol. Rev.* 109, 211–218. doi:10.1037/0033-295X.109.2.211.

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi:10.1038/nature11622.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9. doi:10.1186/gb-2008-9-1-r7.

Hellekant, G., and Danilova, V. (1999). Taste in domestic pig, Sus scrofa. *J. Anim. Physiol. Anim. Nutr. (Berl).* 82, 8–24. doi:10.1046/j.1439-0396.1999.00206.x.

Hinrichs, A. S. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 34, D590–D598. doi:10.1093/nar/gkj144.

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). "Whole-genome annotation with BRAKER," in *Methods in Molecular Biology* (Humana, New York, NY), 65–95. doi:10.1007/978-1-4939-9173-0_5.

Horio, N., Yoshida, R., Yasumatsu, K., Yanagawa, Y., Ishimaru, Y., Matsunami, H., et al. (2011). Sour taste responses in mice lacking pkd channels. *PLoS One* 6. doi:10.1371/journal.pone.0020007.

Hoy, R. R. (1990). Evolutionary innovation in behavior and speciation: Opportunities for behavioral neuroethology. *Brain. Behav. Evol.* 36, 141–153. doi:10.1159/000115303.

Huang, A. L., Chen, X., Hoon, M. A., Chandrashekar, J., Guo, W., Tränkner, D., et al. (2006). The cells and logic for mammalian sour taste detection. *Nature* 442, 934–938. doi:10.1038/nature05084.

Hubisz, M. J., Pollard, K. S., and Siepel, A. (2011). Phastand Rphast: Phylogenetic analysis with space/time models. *Brief. Bioinform.* 12, 41–51. doi:10.1093/bib/bbq072.

Ishii, S., Kurokawa, A., Kishi, M., Yamagami, K., Okada, S., Ishimaru, Y., et al. (2012). The response of PKD1L3/PKD2L1 to acid stimuli is inhibited by capsaicin and its pungent analogs. *FEBS J.* 279, 1857–1870. doi:10.1111/j.1742-4658.2012.08566.x.

Ishimaru, Y., Inada, H., Kubota, M., Zhuang, H., Tominaga, M., and Matsunami, H. (2006). Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12569–12574. doi:10.1073/pnas.0602702103.

IUCN (1993). Status Survey and Conservation Action Plan: Pigs, Peccaries, and Hippos. Available at: https://portals.iucn.org/library/sites/library/files/documents/1993-055.pdf [Accessed April 23, 2020].

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi:10.1159/000084979.

Keeney, S., Lange, J., and Mohibullah, N. (2014). Self-Organization of Meiotic Recombination Initiation: General Principles and Molecular Pathways. *Annu. Rev. Genet.* 48, 187–214. doi:10.1146/annurev-genet-120213-092304.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi:10.1101/gr.113985.110.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317.

Knight, P. A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* 33, 1029–1047. doi:10.1093/imanum/drs019.

Konkel, M. K., and Batzer, M. A. (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. in *Seminars in Cancer Biology*, 211–221. doi:10.1016/j.semcancer.2010.03.001.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. doi:10.1186/gb-2004-5-2-r12.

Lahti, D. C., Johnson, N. A., Ajie, B. C., Otto, S. P., Hendry, A. P., Blumstein, D. T., et al. (2009). Relaxed selection in the wild. *Trends Ecol. Evol.* 24, 487–496. doi:10.1016/j.tree.2009.03.010.

Langer, P. (1974). Stomach evolution in the artiodactyla. *Mammalia* 38, 295–314. doi:10.1515/mamm.1974.38.2.295.

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science (80-.).* 307, 1618–1621. doi:10.1126/science.1106927.

Larson, G., Liu, R., Zhao, X., Yuan, J., Fuller, D., Barton, L., et al. (2010). Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proc. Natl. Acad. Sci. U. S. A.* 107, 7686–7691. doi:10.1073/pnas.0912264107.

Leus, K. (1994). Foraging behaviour, food selection and diet digestion of Babyrousa babyrussa (Suidae, Mammalia). The University of Edinburgh. College of Medicine and Veterinary Medicine. Royal (Dick) Veterinary School.

Lewis, A. J., and Lee Southern, L. (2000). *Swine nutrition*. CRC press doi:10.1201/9781420041842.

Li, D., and Zhang, J. (2014). Diet shapes the evolution of the vertebrate bitter taste receptor gene repertoire. *Mol. Biol. Evol.* 31, 303–309. doi:10.1093/molbev/mst219.

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

Li, M., Chen, L., Tian, S., Lin, Y., Tang, Q., Zhou, X., et al. (2017). Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* 27, 865–874. doi:10.1101/gr.207456.116.

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

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–93. doi:10.1126/science.1181369.

Lim, J. K., and Simmons, M. J. (1994). Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster. *BioEssays* 16, 269–275. doi:10.1002/bies.950160410.

Liu, L., Bosse, M., Megens, H.-J., Frantz, L. A. F., Lee, Y.-L., Irving-Pease, E. K., et al. (2019). Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* 10, 1992. doi:10.1038/s41467-019-10017-2.

Lomsadze, A., Burns, P. D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 42, e119–e119.

doi:10.1093/nar/gku557.
Lowe, T. M., and Eddy, S. R. (1996). TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi:10.1093/nar/25.5.0955.
Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi:10.1093/bioinformatics/btr011.
MELANDER, Y., and HANSEN-MELANDER, E. (1980). Chromosome studies in African wild pigs (Suidae, Mammalia). *Hereditas* 92, 283–289. doi:10.1111/j.1601-5223.1980.tb01709.x.
Meng, G., Li, Y., Yang, C., and Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res.* 47, 63. doi:10.1093/nar/gkz173.
Miller, E. R., and Ullrey, D. E. (1987). THE PIG AS A MODEL FOR HUMAN NUTRITION. Available at: www.annualreviews.org [Accessed April 26, 2020].
Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* 12, R112. doi:10.1186/gb-2011-12-11-r112.
Murphy, C., and Cain, W. S. (1980). Taste and olfaction: Independence vs interaction. *Physiol. Behav.* 24, 601–605. doi:10.1016/0031-9384(80)90257-7.
Musilova, P., Kubickova, S., Hornak, M., Cernohorska, H., Vahala, J., and Rubes, J. (2010). Different Fusion Configurations of Evolutionarily Conserved Segments in Karyotypes of <i>Potamochoerus porcus</i> and <i>Phacochoerus africanus</i> *Cytogenet. Genome Res.* 129, 305–309. doi:10.1159/000314954.
Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi:10.1093/bioinformatics/btt509.
Oliver, W. L. R., Cox, R., and Groves, C. (1993). The Philippine Warty Pigs (Sus philippensis and S. cebifrons). *Pigs, Peccaries, and Hippos*, 145–155.
Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., et al. (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.* 186, 37–45. doi:10.1111/j.1469-8137.2009.03096.x.
Patel, L., Kang, R., Rosenberg, S. C., Qiu, Y., Raviram, R., Chee, S., et al. (2019). Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat. Struct. Mol. Biol.* 26, 164–174. doi:10.1038/s41594-019-0187-0.
Paudel, Y., Madsen, O., Megens, H. J., Frantz, L. A. F., Bosse, M., Crooijmans, R. P. M. A., et al. (2015). Copy number variation in the speciation of pigs: A possible prominent role for olfactory receptors. *BMC Genomics* 16. doi:10.1186/s12864-015-1449-9.
Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502. doi:10.1038/nature05295.
Pickford, M. (2013). Suids from the Pleistocene of Naungkwe Taung, Kayin State, Myanmar. *Paleontol. Res.* 16, 307–317. doi:10.2517/1342-8144-16.4.307.
Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., et al. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 26, 342–350. doi:10.1101/gr.193474.115.
Rabor, D. S. (1977). *Philippine birds and mammals.* UP Science Education Center.
Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., et al. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* 9, 1–15. doi:10.1038/s41467-017-02525-w.
Raudsepp, T., and Chowdhary, B. P. (2011). Cytogenetics and chromosome maps. *Genet. Pig Second Ed.*, 134–178. doi:10.1079/9781845937560.0134.
Robeson, M. S., Khanipov, K., Golovko, G., Wisely, S. M., White, M. D., Bodenchuck, M., et al. (2018). Assessing the utility of metabarcoding for diet analyses of the omnivorous wild pig (*Sus scrofa*). *Ecol. Evol.* 8, 185–196. doi:10.1002/ece3.3638.
Rothschild, M. F., and Ruvinsky, A. (2011). *The genetics of the pig: Second edition*. CABI.
Roura, E., Baldwin, M. W., and Klasing, K. C. (2013). The avian taste system: Potential implications in poultry nutrition. *Anim. Feed Sci. Technol.* 180, 1–9. doi:10.1016/j.anifeedsci.2012.11.001.
Rundle, Howard D. and Boughman, J. W., and HOWARD D. RUNDLE AND JANETTE W. BOUGHMAN

(2015). Behavioral Ecology and Speciation. *Evol. Behav. Ecol.*, 471–487.
Ruvinsky, A., Rothschild, M. F., Larson, G., and Gongora, J. (2011). Systematics and evolution of the pig. doi:10.1079/9781845937560.0001.
Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. doi:10.1093/bioinformatics/19.2.301.
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.
Souron, A., Merceron, G., Blondel, C., Brunetière, N., Colyn, M., Hofman-Kamińska, E., et al. (2015). Three-dimensional dental microwear texture analysis and diet in extant Suidae (Mammalia: Cetartiodactyla). *Mammalia* 79, 279–291. doi:10.1515/mammalia-2014-0023.
Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.
Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi:10.1093/nar/gkl200.
Sutherland-Smith, M. (2015). "Suidae and Tayassuidae (Wild Pigs, Peccaries)," in *Fowler's Zoo and Wild Animal Medicine, Volume 8* (Elsevier), 568–584. doi:10.1016/b978-1-4557-7397-8.00058-x.
Tarailo-Graovac, M., and Chen, N. (2009). "Using RepeatMasker to identify repetitive elements in genomic sequences," in *Current Protocols in Bioinformatics* (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 4.10.1-4.10.14. doi:10.1002/0471250953.bi0410s25.
Templeton, A. R. (2008). The reality and importance of founder speciation in evolution. *BioEssays* 30, 470–479. doi:10.1002/bies.20745.
Thomsen, P. D., Høyheim, B., and Christensen, K. (1996). Recent fusion events during evolution of pig chromosomes 3 and 6 identified by comparison with the babirusa karyotype. *Cytogenet. Genome Res.* 73, 203–208. doi:10.1159/000134339.
Truong Nguyen, D., Lee, K., Choi, H., Choi, M., Thong Le, M., Song, N., et al. (2012). The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. doi:10.1186/1471-2164-13-584.
Vamathevan, J. J., Hall, M. D., Hasan, S., Woollard, P. M., Xu, M., Yang, Y., et al. (2013). Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. *Toxicol. Appl. Pharmacol.* 270, 149–157. doi:10.1016/j.taap.2013.04.007.
van der Made, J., Morales, J., and Montoya, P. (2006). Late Miocene turnover in the Spanish mammal record in relation to palaeoclimate and the Messinian Salinity Crisis. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 238, 228–246. doi:10.1016/j.palaeo.2006.03.030.
Veldhuizen, M. G., Shepard, T. G., Wang, M. F., and Marks, L. E. (2009). Coactivation of gustatory and olfactory signals in flavor perception. *Chem. Senses* 35, 121–133. doi:10.1093/chemse/bjp089.
Villafuerte, M. Q., Macadam, I., Daron, J., Katzfey, J., Cinco, T. A., Ares, E. D., et al. (2020). Projected changes in rainfall and temperature over the Philippines from multiple dynamical downscaling models. *Int. J. Climatol.* 40, 1784–1804. doi:10.1002/joc.6301.
Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi:10.1093/bioinformatics/btx153.
Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2019). An improved pig reference genome sequence to enable pig genetics and genomics research. *bioRxiv*, 668921. doi:10.1101/668921.
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi:10.1093/nar/gky427.
Wertheim, J. O., Murrell, B., Smith, M. D., Pond, S. L. K., and Scheffler, K. (2015). RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* 32, 820–832. doi:10.1093/molbev/msu400.
Wicke, S., Müller, K. F., DePamphilis, C. W., Quandt, D., Bellot, S., and Schneeweiss, G. M. (2016). Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants.

Proc. Natl. Acad. Sci. U. S. A. 113, 9045–9050. doi:10.1073/pnas.1607576113.

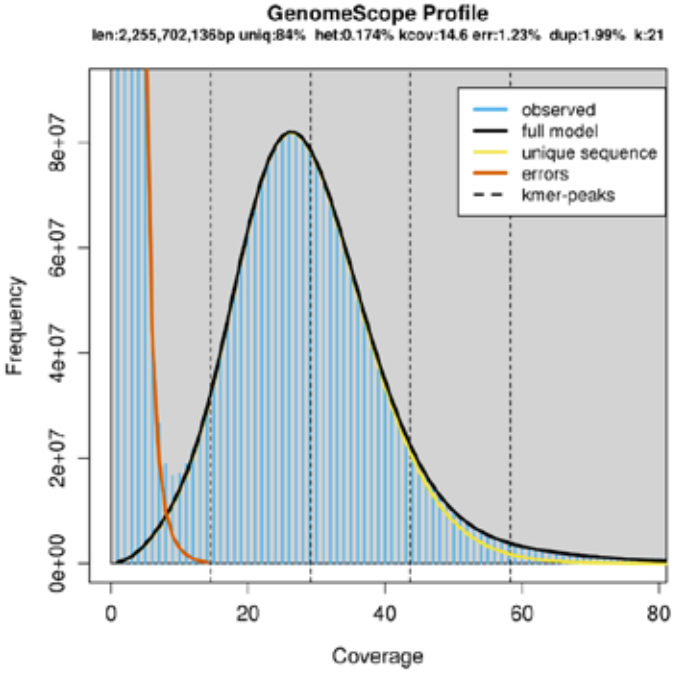
Wu, J., Mao, X., Cai, T., Luo, J., and Wei, L. (2006). KOBAS server: A web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* 34. doi:10.1093/nar/gkl167.

Yang, Y., Zhang, Y., Ren, B., Dixon, J. R., and Ma, J. (2019). Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF. *Cell Syst.* 8, 494-505.e14. doi:10.1016/j.cels.2019.05.011.

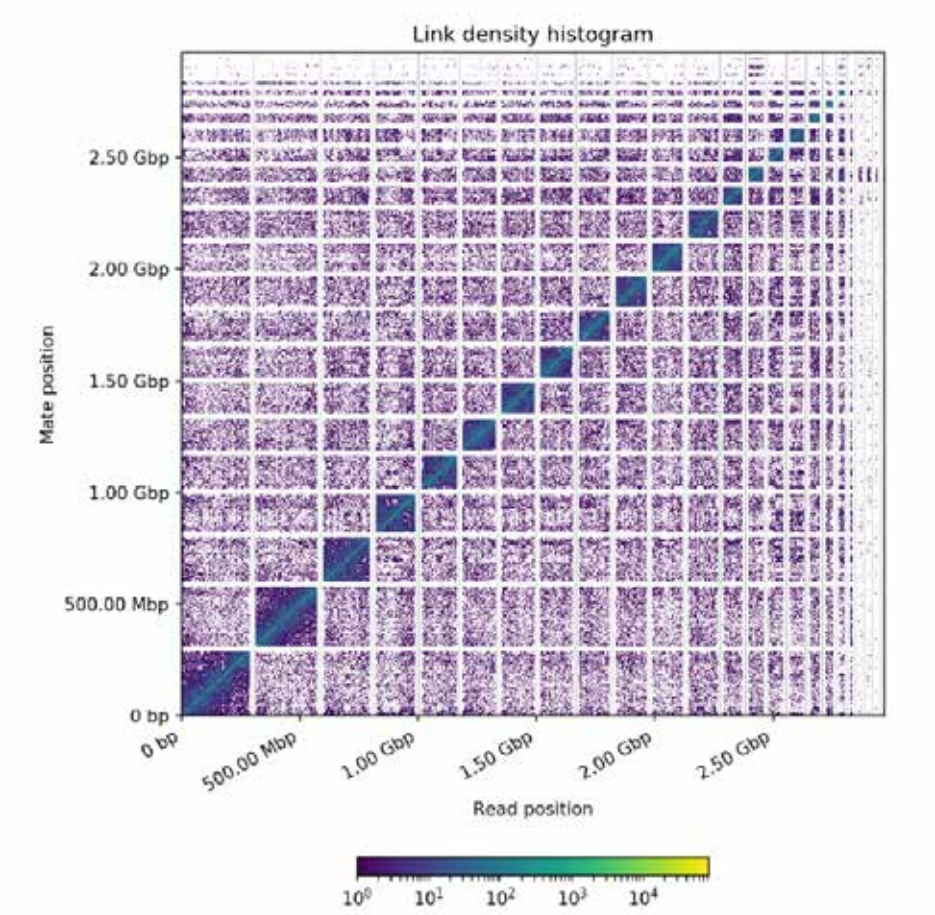
Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088.

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* 5, 833–845. doi:10.1038/s41477-019-0487-8.

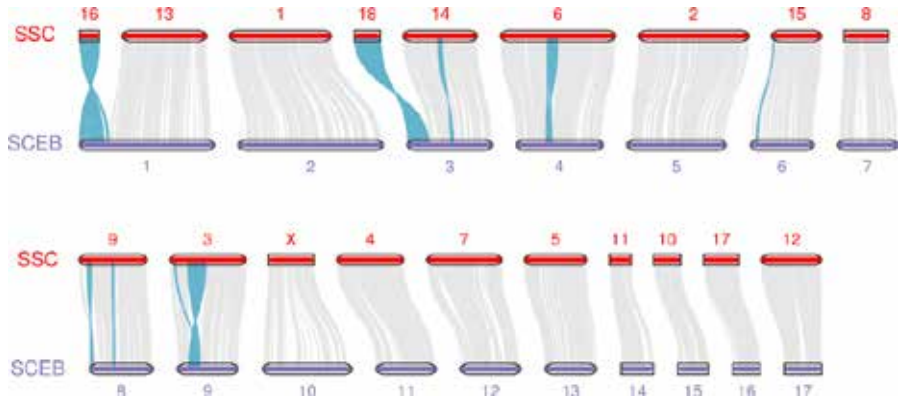
Supplementary Material



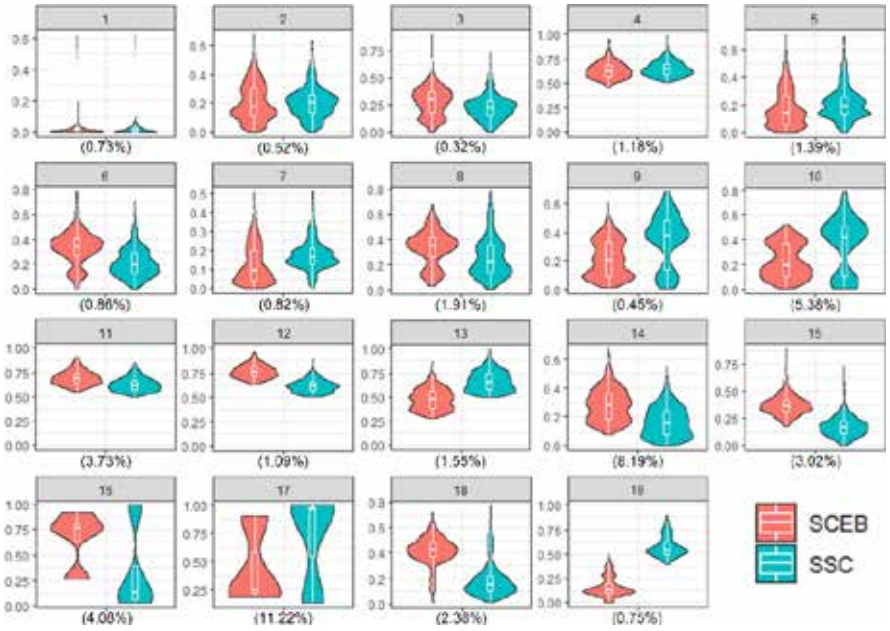
Sup Fig. 4.1 Genomes size estimations. Genomes size estimations for Visayan warty pig generated by Genomescope, providing a k-mer analysis (k=21) by Jellyfsh using filtered data to estimate haploid genome size, heterozygosity and coverage.



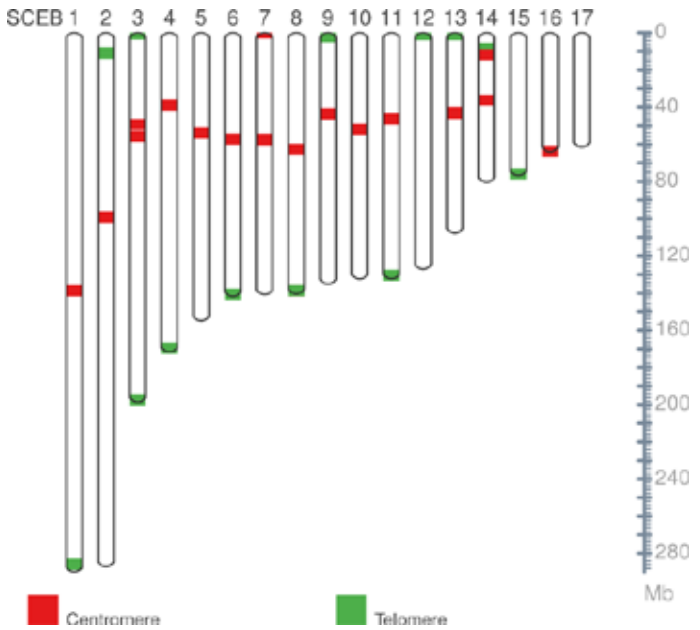
Sup Fig. 4.2 Hi-C interactions with 100-kb resolutions. The color of each square gives the number of read pairs within that bin. Strong interactions are indicated in yellow and weak interactions are indicated in blue. White vertical and black horizontal lines have been added to show the borders between scaffolds. Scaffolds less than 1 Mb are excluded.



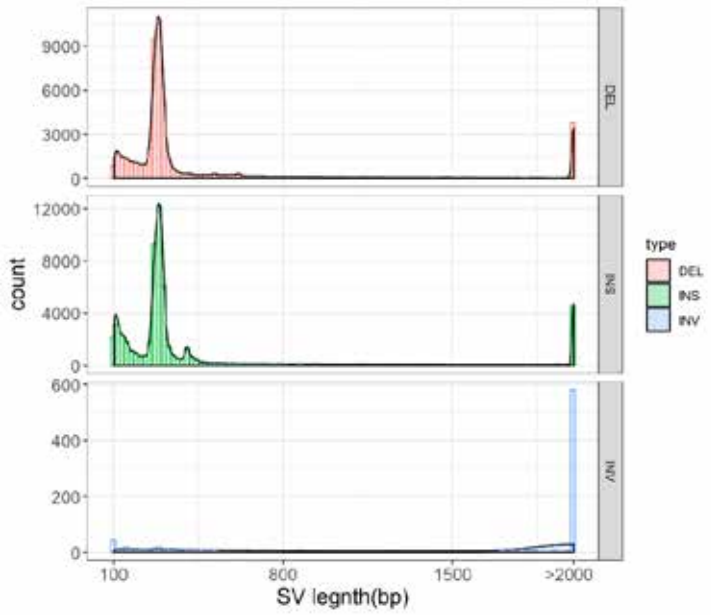
Sup Fig. 4.4 Synteny between Visayan warty pig and Duroc pig. Maps of the 17 Visayan warty pig chromosomes (purple) and of the 19 Duroc chromosomes (red) based on positions of 18,175 orthologue pairs demonstrate highly conserved synteny. Inversions are shown in blue.



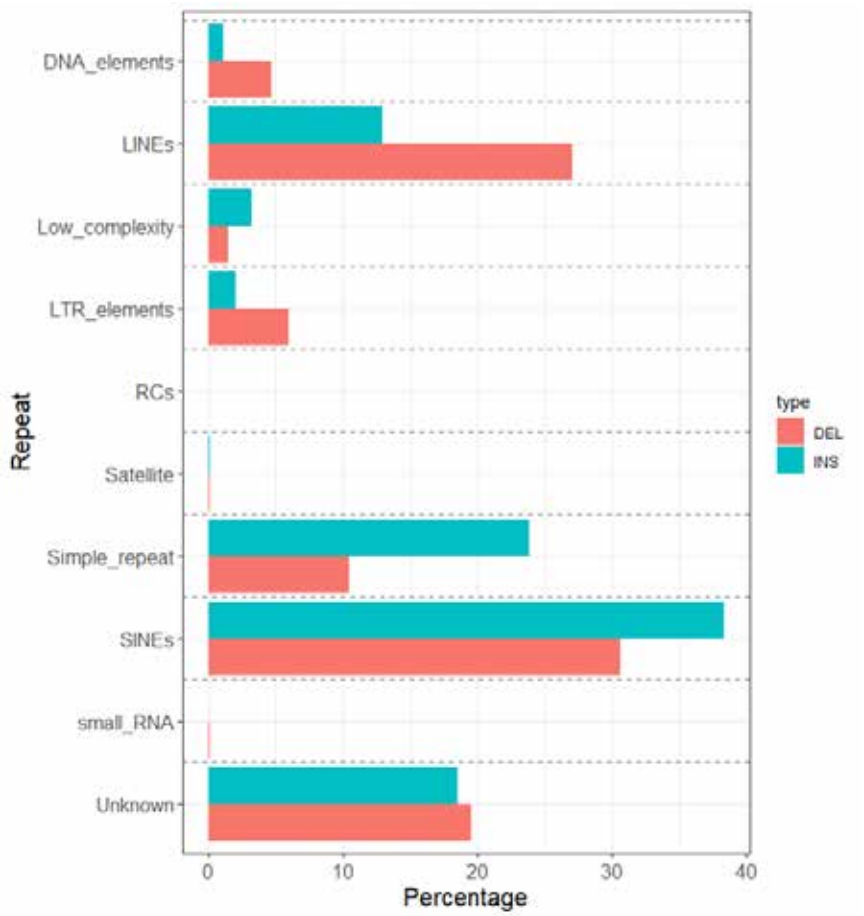
Sup Fig. 4.6 Hi-C evolutionary patterns identified by phylo-HMRF in all the major synteny blocks on all chromosomes between Visayan warty pig and Duroc pig. The boxplot shows the normalized cross-species Hi-C contact frequency distributions in the corresponding state. Percentage of each state among the genome is shown below each panel. State 20 represent the missing data point, which contains ~50.3% of Hi-C signal.



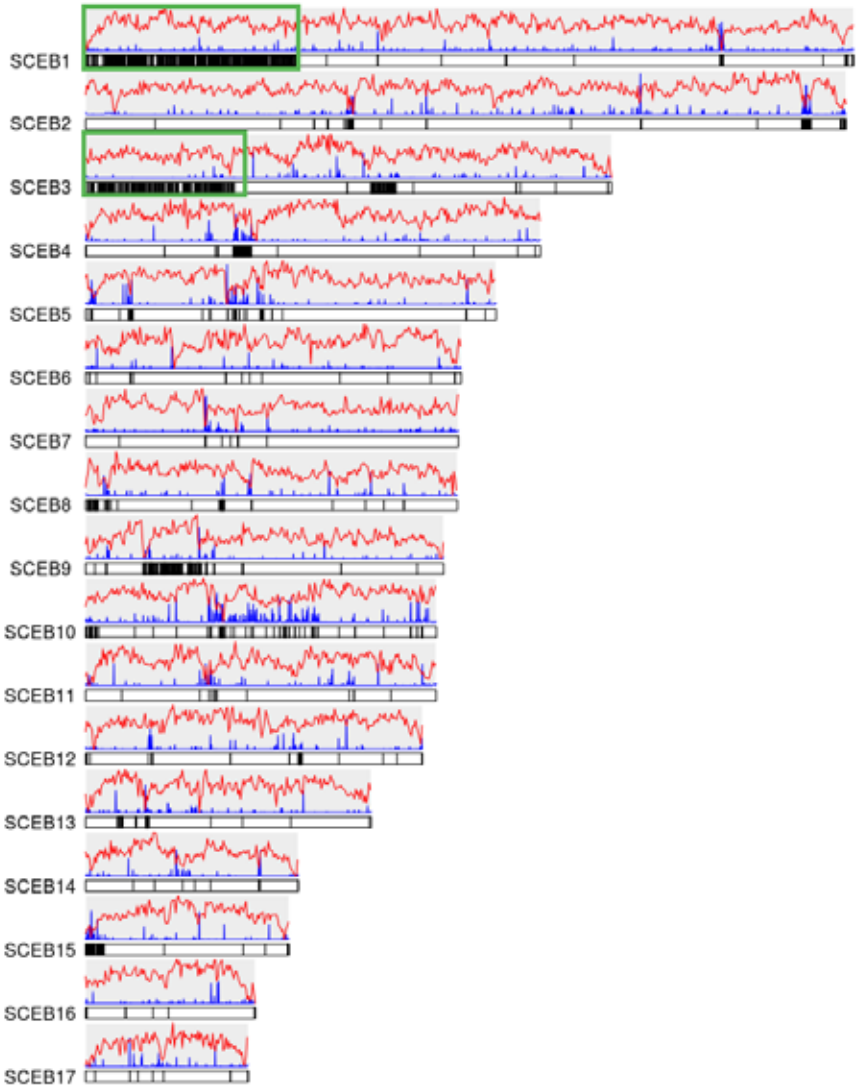
Sup Fig. 4.8 Predicted telomere and centromere locations in the Visayan warty pig assembly.



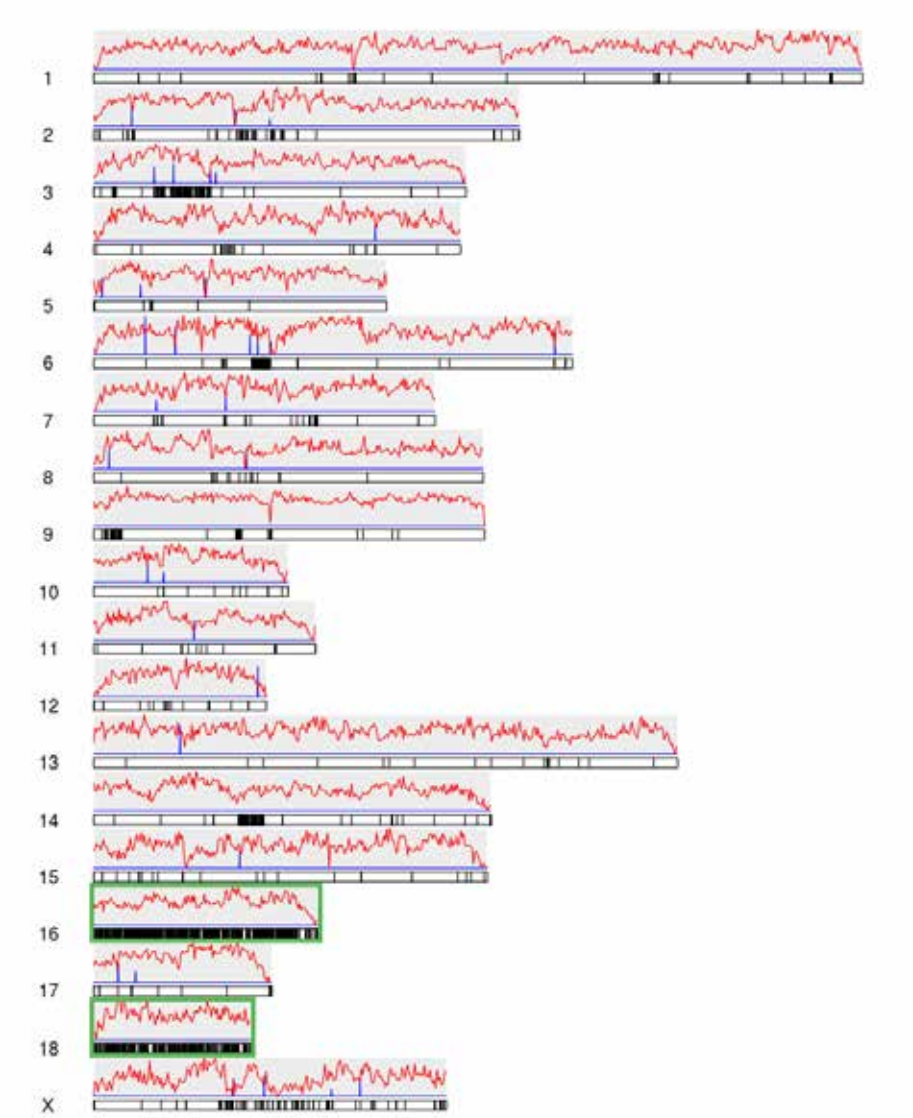
Sup Fig. 4.9 Length distribution of structural variations. Structural variation (deletion/insertion/inversion) was detected using Sus scrofa11.1 assembly as reference. Structural variations larger than 2000 bp were collapsed into '>2000' catalog.



Sup Fig. 4.10 Classification of repeat elements associated with insertions/deletions. A repeat element was defined as being associated with an indel if it overlaps with the structural variation.



Sup Fig. 4.11 Chromosome diagrams of Visayan warty pig assembly. Inversions were detected using Sus scrofa11.1 assembly as reference and shown by black bar. Inverted regions caused by chromosome fusion are indicated by green boxes. Red lines above each chromosome indicate density of Transposable elements (Z-transformed). Blue lines indicate density of Ns and gaps (Z-transformed).



Sup Fig. 4.12 Chromosome diagrams of Sus scrofa11.1 assembly. Inversions were detected using Visayan warty pig assembly as reference and showed in black bar. Inverted regions caused by chromosome fusion are indicated by green boxes. Red lines above each chromosome indicate density of Transposable elements (Z-transformed). Blue lines indicate density of Ns and gaps (Z-transformed).

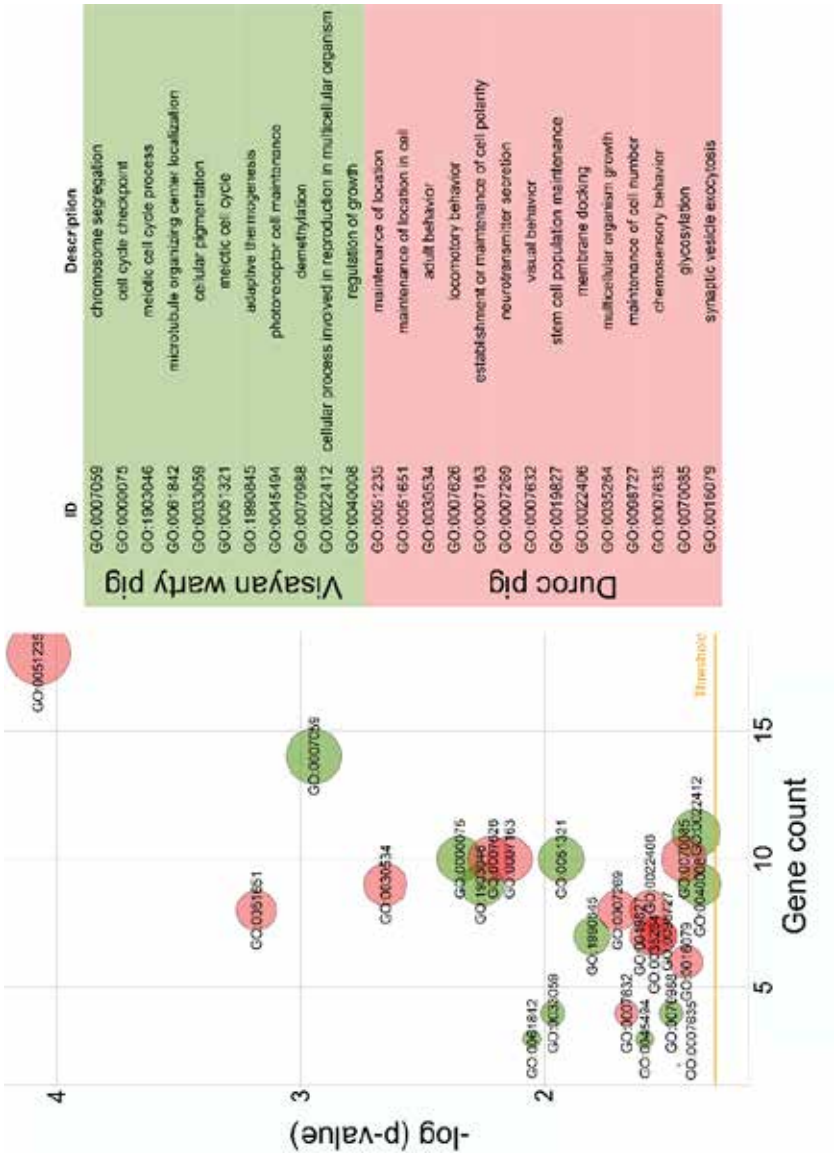
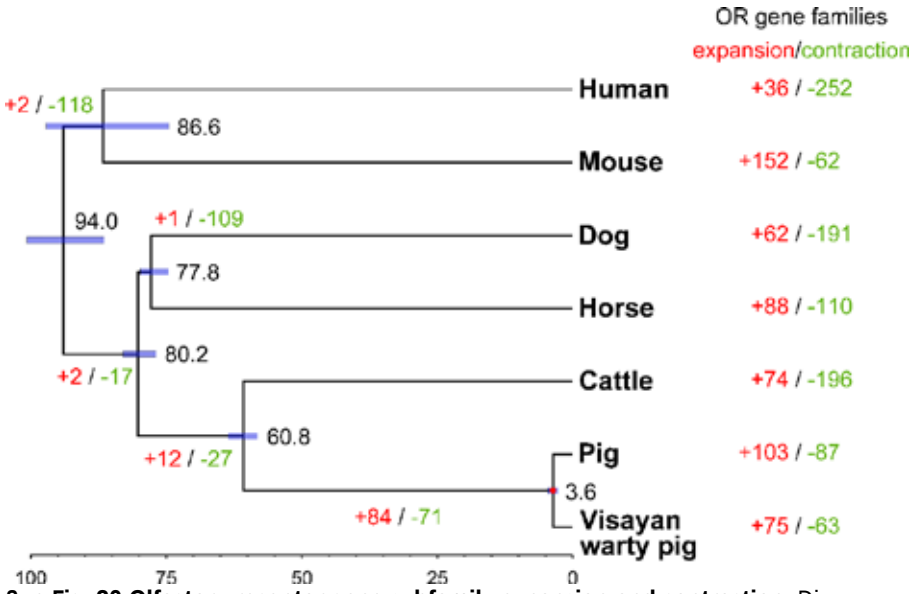
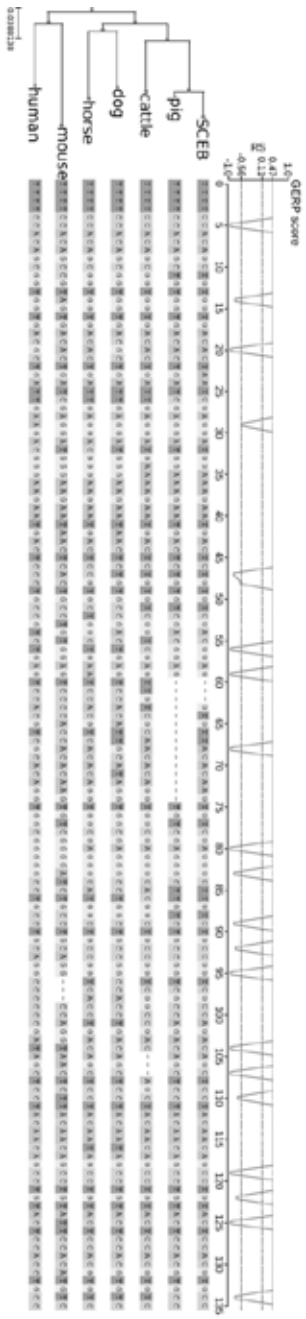


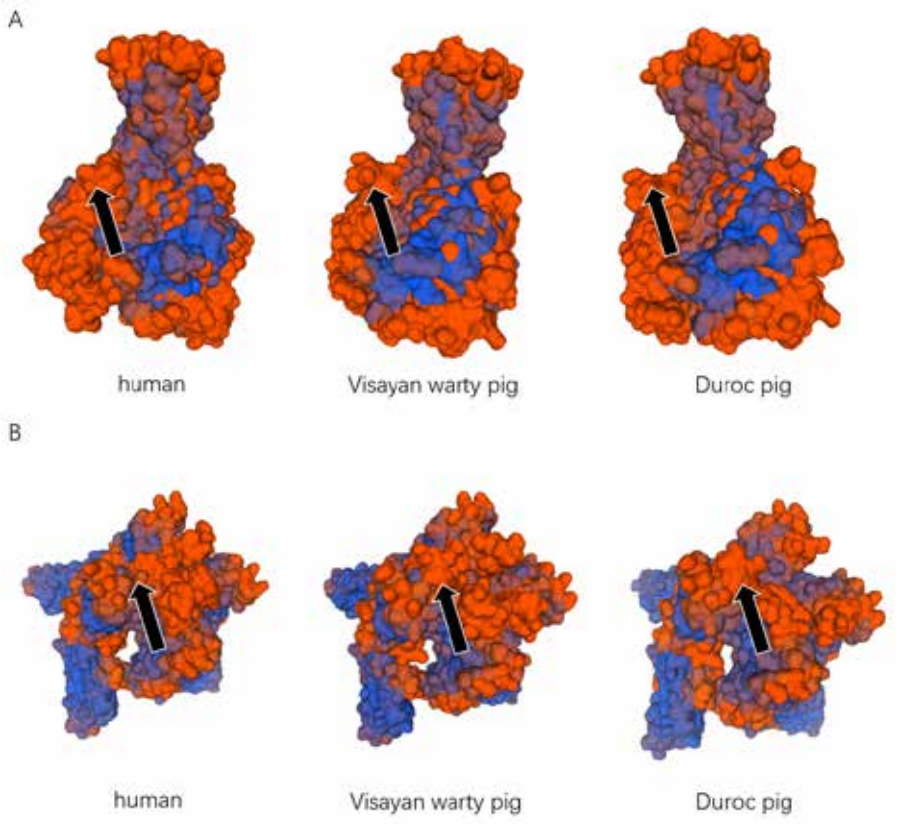
Fig. 4.19 Evolution of genes and gene families in Sus. Enriched GO terms for the positively selected genes in Visayan warty pig and Duroc pig. The level 3 GO terms were used for visualization. Detailed enrichment results can be found in Sup Table 20 and 24.



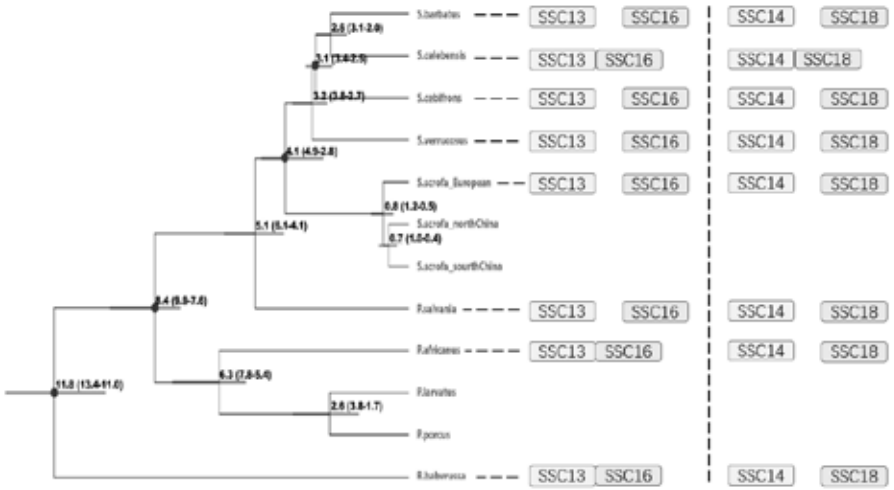
Sup Fig. 20 Olfactory receptor gene subfamily expansion and contraction. Divergence times and history of orthologous olfactory receptor (OR) gene families. Numbers on the nodes represent divergence times, with the error range shown by blue bar. The numbers of OR gene families that expanded (red) or contracted (green) in each lineage after speciation are shown on the corresponding branch.



Sup Fig. 4.21 Multispecies DNA alignment of TRPV1, showing the region with Sus genus specific insertion. GERP scores are shown on top of the alignment. Dashed lines represent upper quartile, average and lower quartile respectively.



Sup Fig. 22 Protein 3D structures estimated by SWISS-MODEL. Computed molecular surfaces of A) SCNN1B and B) TRPV1. Sus genus specific INDEL regions are highlighted by black arrows. Electrostatic potential of modelled surface residues is showed in color. Red indicates a more negative electric potential, whereas blue indicates a more positive electric potential.



Sup Fig. 4.23 Karyotype diagrams of *Suidae* species. Phylogenetic tree on the left is adapted from Chapter 2. Karyotypes of chromosome which are homologous to SSC13, SSC16, SSC14 and SSC18 are shown on the right. Rounded rectangles indicate chromosomes.

5

Sus reference-guided genome assemblies provide a high-resolution view of diversification, reticulation and adaptation during pig evolution

Langqing Liu[1], Hendrik-Jan Megens[1], Martien AM Groenen[1], Ole Madsen[1]

[1] Animal Breeding and Genomics, Wageningen University & Research, The Netherlands

Manuscript under preparation

Abstract

The diverse habitats and unique domestication history have made *Sus scrofa* and its related species attractive models for evolutionary biology. We assembled the genomes of three *Sus* species and one closely related outgroup. Along with published *Sus* genome sequences, we present a near complete collection of *Suinae* genomes. We conducted comparative genomic analyses to investigate the genetic mechanisms underlying *Sus* speciation. Our tests with complex models of speciation revealed that the extensive inter-species gene flow is concomitant with past climatic fluctuations. By investigating the switch of evolutionary rate, we identified candidate genes likely to be associated with wild boar range expansion and the pig domestication process. The data and results presented in this study provide a refined scenario of *Sus* evolution.

Key words: reference-guided assembly, introgression, positive selection, purifying selection

5.1 Introduction

Investigation of the molecular patterns and mechanisms underlying biological diversification remains central in answering unresolved questions about taxonomic diversity on Earth. Species diversity is shaped by a variety of different, possibly interacting, processes including different modes of speciation, divergent adaptation, adaptive radiation, mass extinction, species replacement, and introgression from distant lineages. By comparing the patterns of genomic difference and similarity between members in species complex will inform the study of species diversification. The biota of the island of Southeast Asia (ISEA) is regarded as the world's most diverse but at the same time most imperiled region (Wallace, 1855). Large-scale climatic fluctuations beginning in the early Pliocene changed the landscape and further stimulated intermittent allopatric and parapatric speciation events, which resulted in high species richness. The biodiversity hotspot found in ISEA is hosting at least seven morphologically defined species of pig of the genus *Sus*. Aside from *Sus scrofa*, which is distributed across most of the world, all other species of the genus *Sus* are restricted to ISEA. *Sus scrofa* has undergone an extremely rapid radiation during the Pleistocene. Due to the lack of postzygotic reproductive barrier in *Sus*, this range expansion was accompanied by a consistent gene flow with local related species (Frantz et al., 2013; Liu et al., 2019). *Sus scrofa* later was domesticated independently in Asia and Europe some 10,000 years ago, followed by a strong and continuous artificial selection for meat industrial production in the last 200 years. Thus, the genus *Sus* provides an excellent model to study many of the species diversification processes mentioned above.

So far, there are two chromosome level genome assemblies of *Sus* species available, i.e., *Sus scrofa* (Warr et al., 2020) and *Sus cebifrons* (Chapter 4). For three other *Sus* species (*Sus verrucosus*, *Sus celebensis* and *Sus barbatus*) and the outgroup species *Porcula salvania* (Pygmy hog), also medium coverage resequencing data is available, which was used for evolutionary studies based on mapping against the *Sus scrofa* reference genome (see Fig 5.1 for the geographical distribution). Using a reference genome from a single species may potentially bias the results of cross species comparison (Günther and Nettelblad, 2019). In this study, we present reference-guided genome assemblies of the three *Sus* species and the outgroup species *Porcula salvania* (Sup Table 5.1). With these reference-guided assemblies combined with the two chromosomal *Sus* assemblies, which represent nearly all species of the *Sus* lineage, we test the genetic mechanisms underlying *Sus* speciation using comparative genomic and molecular evolution analyses. This study demonstrates the power of directly using genome sequences to address evolutionary and functional genomic questions, in particular by using genome wide scanning for

selections signal, and assessing impact of reticulation speciation diversification process.

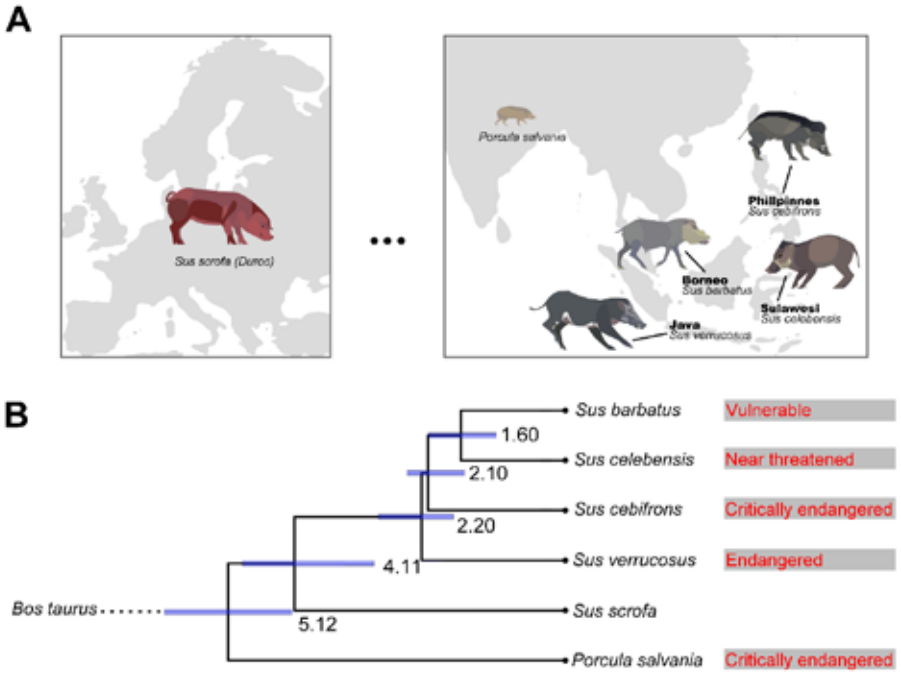


Fig 5.1 Geographic distribution, phylogenetic relationships and divergence times of *Suinae* species. A) A map of Europe and Southern Asia depicting the modern distributions of *Suinae* species (Note: distribution of wild boar is covering the whole Eurasia). B) Phylogenetic relationships among *Suinae* species inferred from nuclear DNA. Node labels show age in millions of years and 95% confidence interval. Tip labels show the taxon and the conservation status according to IUCN red list (https://www.iucnredlist.org/). (animal illustrations credit to Sheila McCabe, https://faunalfrontier.com/)

5.2 Results

Validation of the Reference-Based Assembly Strategy

We adapted the reference-guided assembly approach from Lischer and Shimizu (2017). This approach first maps reads against the reference genome to identify conserved regions. In the following steps, reads with no similarity to the related genome are combined with the conserved regions and further integrated into scaffolds. (Sup Fig 5.1).

To validate the reference-guided approach, we took ~25X resequencing data from the same individual which was used to generate the complete *de novo* assembly of

Sus cebifrons and performed a reference-guided assembly with the *Sus scrofa* as reference. This reference-guided assembly resulted in a total size of 2.34 Gb, with a contig N50 of 78 kb. We mapped the contigs from the reference-guided assembly to the complete *de novo* assembly. In total, 5.94% of the genome was misassembled (translocations, relocations, inversions or duplication) in the reference-guided assembly (Sup Table 5.2). The completeness of the obtained reference-guided assembly was also assessed with BUSCO (Table 5.1, Sup Table 5.3), which showed that 92.30% of BUSCO genes was retrieved from the reference-guided assembly compared to 95.70% BUSCO genes for the complete *de novo* assembly. However, the reference-guided assembly shows more fragmented gene models (5.00%, while 2.40% for the complete *de novo* assembly). To confirm that the reference-guided assembly method resulted in the correct reconstruction of gene sequences, we compared the BUSCO genes in the reference-guided assembly and the complete *de novo* assembly. In 4,104 genes, covering ~22.8 Mb, only 31 mismatches were observed. Overall, our evaluation shows the accuracy and robustness of this reference-guided assembly pipeline.

Table 5.1 Genome assembly Statistics for the two *de novo* assemblies and five reference-guided assemblies.

	De Novo Assembly		Reference-Guided Assembly				
	S.scrofa	S.cebifrons	S.cebifrons	S. verrucosus	S. celebensis	S. barbatus	P. salvania
Total length (Gb)	2.50	2.46	2.34	2.44	2.42	2.42	2.38
Contig N50 (Kb)	48231	159	78	66	95	81	30
Scaffolds count	613	1585	13006	17932	15215	18414	17928
GC%	41.87	41.79	41.68	41.79	41.82	41.81	41.79
N%	1.19	0.80	0.81	0.78	0.76	0.84	0.88
Complete BUSCO%	93.84	95.72	94.55	92.64	93.81	94.04	92.63

Reference-guided assemblies, quality assessment, and annotations

For *Sus verrucosus*, *Sus celebensis* and *Sus barbatus*, paired-end sequencing reads were assembled using the *Sus cebifrons* genome as reference since *Sus cebifons* is more related to these species than *Sus scrofa*. *Porcula salvania* has a similar evolutionary distance to all the *Sus*. Considering the similar karyotype between *Porcula salvania* and *Sus scrofa* (Bosma et al., 1983) and the higher continuity of the *Sus scrofa* reference genome, we used the Sus scrofa genome as reference to guide the assembly of *Porcula salvania*. The summary statistics of the genome assemblies are shown in Table 5.1. The completeness of the obtained assemblies was assessed with BUSCO (Sup Table 5.3). The reference-guided assemblies resulted in 91%-93%

of the mammalian core genes being either completely or partially represented in the assembly of the different species.

We annotated the genome using ab initio gene prediction and homology-based searching. Combining these two methods, we predicted 18,408 functional annotated genes in *Sus verrucosus*, 19,661 genes in *Sus celebensis*, 18,854 genes in *Sus barbatus*, and 19,689 genes in *Porcula salvania*, respectively (Sup Table 5.4). Species-specific gene duplications or losses due to the misassembly were examined by checking the gene distribution of gene coverage (Sup Fig 5.3). The coverage distributions are approximating a normal distribution (Shapiro-Wilk normality test, p-value < 0.05).

Gene orthology and divergence time

Gene orthology inference was performed on proteomes of the 4 reference-guided assemblies, *Sus scrofa*, *Sus cebifrons*, and *Bos taurus*. This resulted in 16,543 orthogroups, of which 10,177 were single-copy among the used ungulates. We constructed a phylogenetic tree and determined the time of speciation (Fig 5.1). We estimated that *Porcula salvania* and the *Sus* lineage diverged about 5.1 million years ago (Mya). The divergence between *Sus scrofa* and ISEA *Sus* occurred ~4.1 Mya. For the insular *Sus* lineages, *Sus verrucosus* on Java first diversified ~2.2 Mya, followed by the emergence of *Sus cebifrons* on the Philippines ~2.1 Mya. Finally, the divergence between *Sus celebensis* and *Sus barbatus* occurred ~1.6 Mya.

Whole genome synteny

We obtained the synteny blocks among the 6 *Suinae* species by aligning their genome sequences to the pig reference genome. All the newly assembled species showed an over 90% alignment ratio to the pig genome (Sup Fig 5.2, Sup Table 5.5). The whole genome alignments were transformed into a multiple sequence alignment format using the pig genome as reference. Approximately 2.2 Gb syntenic sequences are shared by all the *Suinae* genomes, with an average length of 14 kb. Considering the potential assembly error in the low complexity regions, we further subset the alignment with less than 10% repetitive content. This resulted in a final 1.5 Gb aligned sequences representing 62% of the pig genome.

Hybridization history of *Sus* genus

Whole genome alignments were used to infer the gene flow among *Sus* species. In order to differentiate between incomplete lineage sorting and gene-flow, we conducted admixture analysis (using D-statistics) and reconstructed local genealogy for synteny blocks longer than 10kb. Overall, we found strong evidence of admixture

between *Sus scrofa* and *Sus verrucosus*, while there is no significant gene flow between *Sus scrofa* and other ISEA *Sus* (Fig 5.2, Sup Table 5.6). Furthermore, we estimated the admixture fraction from *Sus verrucosus* to *Sus scrofa* to be 1.4% (Sup Fig 5.4, Sup Table 5.7). Besides the signal for admixture between *Sus verrucosus* and *Sus scrofa*, the D-statistics supports that the admixture of *Sus verrucosus* with *Sus barbatus* is similar to the admixture with *Sus celebensis* (using *Sus cebifrons* as non introgressed; D=0.12; D=0.11). Furthermore, the 53% overlap of the admixture regions between *Sus celebensis* and *Sus barbatus* suggests an ancient admixture between *Sus verrucosus* and the common ancestor of *Sus celebensis* and *Sus barbatus*. We further investigated the genes located in the introgressed regions and their functions. In total, 481 genes are located in the introgressed fragments between *Sus scrofa* and *Sus verrucosus* (Sup Fig 5.4, Sup Table 5.8). We found enrichment of gene ontology terms involved in Sialic acid transport and Fatty acid beta-oxidation (Sup Table 5.9). Meanwhile, 2271 genes were found that are related to the admixture between *Sus verrucosus* and the common ancestor of *Sus celebensis* and *Sus barbatus.* (Sup Fig 5.5, Sup Table 5.10). We found significant functional enrichment for nervous system development (Sup Table 5.11).

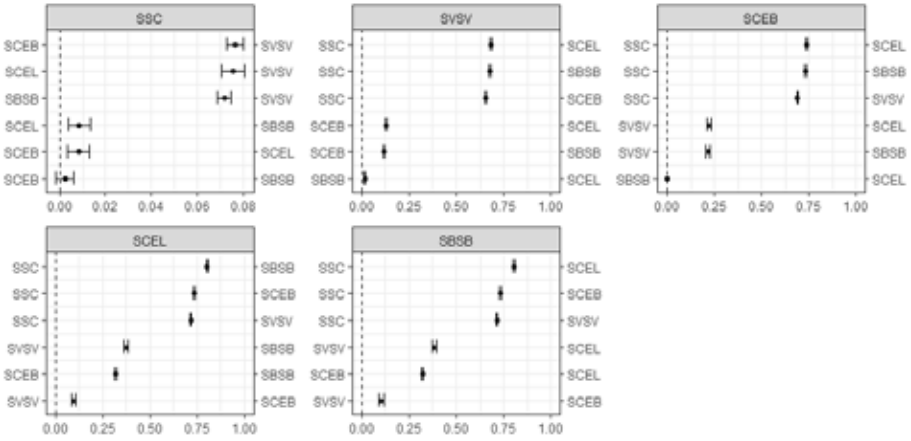


Fig 5.2 Admixture between *Suinae* species. Diagrams depicting the excess derived allele sharing when comparing sister taxa and outgroups (pygmy hog). We computed D statistics of the form D (X, Y, SSC, PYGMY). Error bars correspond to standard errors. Red dots indicate that the D statistic is significantly different from 0 (|Z|>=3), such that the *Sus scrofa* genome shares more derived alleles with the genome on the right (Y) than the left (X).

Selection changes on Single-Copy Genes

We assessed the direction and the strength of changes of selection across functional genes by dN/dS and the selection strength parameter K through branch-site models (Yang, 2007; Wertheim et al., 2015). We tested for positive selection and purifying selection at the ancestral *Sus* linage, the *Sus scrofa* linage, and the basis of the ISEA *Sus* clade for 10,177 single-copy orthologous genes, with cattle orthologs as outgroup. We used the likelihood ratio (LRT) to identify significant positive or purifying selected genes. These genes were further analyzed to assess how the selection pattern changed along the phylogeny. A total of 42 genes was found to have evolved under positive selection at the ancestral *Sus* and *Sus scrofa* lineage, but with a switch to purifying selection in the ISEA *Sus* clade (Sup Table 5.12). In the meantime, a total of 122 genes was found to have evolved under purifying selection at the ancestral *Sus* and ISEA *Sus* clade but with a switch to positive selection in the *Sus scrofa* linage (Sup Table 5.13).

5.3 Discussion

The knowledge of the genomic mechanisms underlying speciation and adpatation is still scarce, and this is particularly true when studying domesticated animals. Traditional methods lack the ability to distinguish between natural selection on ancestral variants and artificial selection during and after domestication. In this study, we obtained genomic sequences for three *Sus* species and one closely related outgroup, the pygmy hog, which were combined with the previously assembled genome of *Sus scrofa* and *Sus cebifrons*. These genomes are valuable resources, which in this study were further exploited for advancing our understanding of the genomic pattern observed in adaptive radiation and domestication.

Reference guided assembly for pig species

The Mammalian gene content is found to be evolutionary stable for ~100 My (Zhao and Schranz, 2019). Moreover, almost complete synteny and large blocks of collinearity were also observed between *Sus scrofa* and *Sus cebifrons* (chapter 4, this thesis). The divergence time among *Sus* species is ~4 Mya, while the split between pygmy hog and *Sus* occurred ~5 Mya. Thus, given the short divergence time, non-conserved synteny and noncollinearity were therefore not expected to be a primary concern here.

We verified the quality the referenced-guided assemblies, which show high completeness, especially in the genic regions. However, we should keep in mind that the reference-guided assemblies may still miss characteristics that are specific to a

particular species but not found in the reference used. For instance, species-specific gene duplications or losses may be missed if exclusively comparing the assembled genomes. These features can for most part be identified by taking advantage of the gene coverage, similar as is done for copy-number variation detection. In our study, the distributions of the gene coverage were in general normally distributed, with the mean centered on the expected average coverage. This suggests that there are little species-specific duplications or losses in the reference-guided assemblies (Yoon et al., 2009). Moreover, species-specific features are in any case out of the scope of the current study, as we specifically investigated sequences common to all pig species.

Reference bias

In standard next-generation genome sequencing, DNA is fragmented and sequenced. The sequenced reads are subsequently aligned to a linear reference genome of the same or a closely related species. Because a reference genome normally comprises a single sequence from one individual, it does not in full capture the genomic diversity within the reference species, and even less to other closely related species. This results in a reference bias since reads with high similarity to the reference have a higher mapping rate and mapping quality compared to the non-reference reads. Incorrect read mapping in turn leads to false negative or false positive variant calls, which can be a single nucleotide polymorphism, insertion, deletion, or structural variation. This systematic bias is, in general, towards the reference allele, resulting in higher genomic similarity between the reference species and the other species. One of the aims of this study was to directly compare the genome sequences from different species. For this analyses, homologous genome segments extracted from the multiple whole genome alignment were used. This potentially avoids the bias towards the reference.

Speciation with gene flow in *Sus*

Geological evidence suggests that during most of the late Pliocene to early Pleistocene four major land bridges provided access from Asia into Wallacea, the geologically mobile region between Sundaland and Sahul-land (Tjia, 2006). The land bridges later became deep sea passages due to the rise of the sea level and tectonic subsidence. The temporal connection between the insular island and continent allowed animal migration. It appears that the late Pliocene to early Pleistocene was a time when three faunal dispersals in Southeast Asia occurred. One of these took place from Sumatra and Java along the Lesser Sundas to Timor; the other two were from Borneo to Sulawesi and from Borneo to the Philippines (Groves, 1985). The Philippines appear never to have a land bridge to the Asian mainland (Heaney, 1985,

1986), but it is possible that western Sulawesi was still attached structurally to Sundaland during the Pliocene (Hall, 1996), although not necessarily by continuous dry land.

Our admixture analysis revealed the inter-specific gene-flow to be highly correlated with the geological accessibility between two species. The mainland Southeast Asia is frequently connected with Java, and we found clear evidence of admixture from *Sus scrofa* to *Sus verrucosus* (see Figure 5.1 for the species distribution). Similarly, the fluctuation of the sea level not only induced the speciation in Sunda-shelf and the Philippines (Frantz et al., 2013, 2014), but also allowed post-speciation gene-flow between geologically isolated species. We observed gene-flow signals from *Sus verrucosus* to *Sus celebensis* and to *Sus barbatus*, and this admixture may have taken place before the divergence of *Sus celebensis* and *Sus barbatus*. Due to the much lower geographic accessibility of the Philippines, in all comparisons, *Sus cebifrons* did not show any significant gene-flow signal. Our results are mostly consistent with previous studies (Frantz et al., 2013, 2014), except for the similar admixture fraction from *Sus verrucosus* into *Sus celebensis* and *Sus barbatus*. This can be explained by the different individuals we used in this study. The different admixture fraction between individuals led us to the hypothesis that there is a more recent admixture from *Sus verrucosus* into *Sus celebensis* than into *Sus barbatus*, and that the introgressed fragments have not got fixed within the population. Moreover, we did not find the previous reported minor gene flow between *Sus scrofa*, *Sus celebensis* and *Sus barbatus*. We note that using *Sus scrofa* as reference, reference bias will force resequencing samples to appear more similar to *Sus scrofa* than they actually are, which may have led to false positive gene flow signals between *Sus scrofa* and resequencing samples. In this study we eliminated reference bias by working on pure assemblies generated from the sequencing data in our experiment. The results illustrate that, in admixture analyses on re-sequenced data aligned against a reference genome, concerns about reference bias should be taken into consideration. Thus, further investigation is needed to shed more light on the effects of reference bias in admixture analyses.

Candidate genes involved in *Sus scrofa* adaptive radiation and domestication

Evolutionary mechanisms that may result in the appearance of new advantageous traits (e.g., boosting viability or adapting to environmental changes) include positive selection on protein-coding genes, where frequency of mutations altering the function of genes are raised in the population because they are favorable. By contrast, purifying selection is the mechanism preventing the fixation of deleterious

mutations, as those mutations are detrimental for the organism. Therefore, under a continuously changing environment, the appearance of an advantageous trait by positive selection in an ancestral species may be followed by a switch in the selective pressure, with this trait undergoing purifying selection in the descendant species (the gene is still functional but not showing additional advantage). Examples of this scenario with a pattern of positive selection in internal branches of a phylogeny, followed by a shift to purifying selection are found in both animals and plants (Perry et al., 2012; Zhao et al., 2013; Schwerdt et al., 2015; Daub et al., 2017; Popejoy et al., 2020). The near complete genome collection of *Sus* species enabled us to reconstruct the ancestral state of the *Sus* linage and estimate the changes of selection pressure across the phylogeny.

Based on the archaeological and paleoclimatological evidence, ancestral *Sus* and present-day *Sus scrofa* shared a similar geographic distribution (van der Made et al., 2006; Frantz et al., 2016). The global temperature and atmospheric carbon dioxide level of 5 Mya also resemble current conditions (Hansen et al., 2013). Genes involved in continental habitat adaptation may have been under the similar selection pressure in ancestral *Sus* and *Sus scrofa*. Meanwhile, as the ISEA *Sus* are specialized to tropical island habitats, continental adaptive genes may have experienced purifying selection constraint. A total of 42 genes were found to have evolved under positive selection in the ancestral *Sus* and *Sus scrofa* lineages but showing a switch to purifying selection in the ISEA *Sus* clade (Sup. Table 5.12). For example, we identified the *ESR1* gene, which in domestic pigs is associated with litter size (Muñoz et al., 2007). Wild boar is known to have the biggest litter size within the *Suidae* family. Both ancestral *Sus* and *Sus scrofa* have colonized the whole Eurasian continent. These finding suggest that this high fecundity could have played a role in the successful range expansion of *Sus scrofa*. During the range expansion of *Sus scrofa* from Southern Asia to the North, the most immediate challenge would have been the different temperature. Mammals will adjust blood flow to acclimate to extreme climate conditions (Jansky and Hart, 1968; Stocks et al., 2004; Horwitz, 2011). We identified genes affecting sinus rhythm (*FXP2*), along with genes involved artery development and embryonic heart tube development (*JCAD* and *ZNF335*), which are associated with the function of cardiovascular system. We also found the *CHMP1A* gene which previously has been associated with tanning ability (Nan et al., 2009; Visconti et al., 2018). This gene is also related to the tropical adaptation in cattle (Makina et al., 2015) and could potentially have boosted the fitness during the expansion of wild boar from a sun-belt region (Southern Asia) to short day-length regions (Northern Asia and Europe).

Domestication is often associated with morphological and or behavioral changes induced by human mediated involuntary or voluntary selection that results in direct control over breeding to improve traits that are beneficial for humans (Zeder, 1982). A total of 122 genes were found to have evolved under purifying selection at the ancestral *Sus* and ISEA *Sus* clade, but which showed a switch to positive selection solely in the *Sus scrofa* linage (Sup.Table 5.13). These genes may reflect the unique evolutionary process during *Sus scrofa* domestication. Most significantly, we found genes involved in cognitive ability (*ATF7IP*, *BRWD1*, *C20orf96*, *PEF1*, *PTPRO*, *ZBBX*) and chronotype (*ARID2*, *PABPC1L*, *TNRC6B*). This observation is consistent with the hypothesis that the primary selective pressure during the initial stages of domestication is on behavior, and in particular for tameness (lack of fearful or aggressive responses to human caretakers). Such a reduction in acute fear and long-term stress is a prerequisite to successful breeding in captivity (Künzl and Sachser, 1999; Price, 1999; Gariépy et al., 2001; Schütz et al., 2001). We observed a gene related to brain region volumes (*NUAK1*). This finding is in agreement with the wildly observed fact that domesticated animals show reduction in total brain size and specific brain regions compared to their wild ancestor (Kruska, 1987, 1996, 2005; Zeder, 2012). Another typical change associated with domestication in mammals is morphological change of teeth. We also found genes associated with orofacial development (*ABCA4*) and dental crown development (*CA12* and *ITPKB*). Due to agriculturalization and domestication, the diet of pigs shifted to household scraps primarily with rich starch content (Weber and Price 2015), which requires less mastication. Pigs' jawbones eventually experienced a shrinkage from the lack of frequent, forceful use (Evin et al., 2015). By the same token, domesticated pigs tend to have a high linear enamel hypoplasia (LEH) occurrence rate (Magnell and Richard Carter and Magnell O., 2007; Hu et al., 2009). We also found genes related to energy metabolism and storage (*ATP2A3*, *MRPL19*, *PARP8*, *SLC12A8*). This is in line with the breeding goal of domesticated pig. To meet the demand of meat production, domesticated pig was selected to adapt to the high energy diets and inactivity. Interestingly, there are several genes affecting lung function (*BACH2*, *ITGA1*, *MN1*, *MTERF3*, *NR5A2*, *SCAF8*). Extensive evidence exists for the different relative weight of lungs between wild boar and domesticated pig (ZHAI and Wang, 2001; Gun et al., 2009; Razmaite et al., 2009). There is even one extreme case reporting different number of lobes (Cabral et al., 2001). This may be a consequence of adaptation to the captive living environment and restriction of movement.

In addition to the above-mentioned genes, we found genes in adaptive radiation category and domestication category, which shared similar functions. They are

regulating immune response, growth and development, and most importantly, fat metabolism. These genes and their associated traits are also very important for both wild and captive living. However, in the absence of phenotypic records, it is difficult or even impossible, to predict the actual phenotypic consequences of the selection on those genes.

5. 4 Conclusion

This study highlights the value of using comparative genomic approaches to examine evolutionary genomics patterns in a near complete phylogenetic framework. We provided insights into the distinct evolutionary scenarios occurring under adaptive radiation and under artificial selection during domestication. Reference-guided assemblies with medium-coverage libraries hold strong potential to reveal interesting genomic characteristics and features before high-quality genome assemblies become available. Such preliminary studies will be useful to guide genome sampling projects across the tree of life.

5.5 Materials and Methods

Reference-guided *de novo* assembly pipeline

The samples used in this study were chosen from previously described and published data in order to ensure that each was representative of their respective species.

We adapted the reference-guided *de novo* assembly pipeline described by Lischer and Shimizu (2017). In brief, it first identifies the homologous regions between target and reference genome, then *de novo* assembles the homologous regions separately. The assembled contigs are further merged into scaffolded with the guide of reference genome. We modified the assembly approach to be aware of divergent and repetitive regions, and integrated an iterative gap-closure and error correction to the final scaffolds. These modifications enable multithreading for most of the steps in the pipeline to reduce runtime. Scripts used in this pipeline were deposited in https://github.com/llq0325/ref-guided-assembly.

Evaluation of the reference-guided *de novo* assemblies

The completeness of all reference-guided genome assemblies was investigated with BUSCO (v2, mammalian dataset) (Simão et al., 2015). For each species, we calculated the proportion of the sequence covered by mapped reads and average read depth with bedtools v2.2 (Quinlan, 2014).

Genome Annotation

We identified repetitive elements by a combination of homology alignment and *de novo* searches. We used RepeatMasker (Tarailo-Graovac and Chen, 2009) with Repbase (v.16.10) (Jurka et al., 2005) to scan for sequences homologous to annotated repeat sequences in published databases and then used RepeatModeler (Jurka et al., 2005) (http://www.repeatmasker.org/RepeatModeler.html) with default parameters.

Gene annotation was performed using BRAKER v2 (Hoff et al., 2019) with the *ab initio* gene prediction and homology-based searching with proteome from *Sus scrofa* and *Sus cebifrons* as protein homology evidence. EVidenceModeler v1.1.1 (Haas et al., 2008) was used to integrate the genes predicted by BRAKER and generate a consensus gene set.

Whole genome alignments

The Nucmer program from Mummer v3.23 (Haas et al., 2008) was used with -maxgap 50 and -breaklen 400 for comparing the genomes between the pig reference genome and genomes of other species to obtain the synteny blocks.

To generate multiple genome alignments, we first aligned all genomes to the pig reference genome using LAST v980 (Kiełbasa et al., 2011), Then, we used MULTIZ v11.2 (Blanchette et al., 2004) to merge the pairwise alignments into multiple genome alignments using the pig genome as the reference. Approximately 1.7 Gb syntenic sequence are shared by the six pig species.

Admixture analyses

D-statsitics were calculated using the python script ABBABABAwindows.py (github.com/simonhmartin/genomics_general), and standard errors were calculated employing a weighted block jackknife with a lock size of 5 Mbp. We used the pygmy hog genome as an outgroup. To quantify genealogical relationships among taxa, neighbour-joining phylogenies were inferred for windows of 10 kb using PHYLIP v3.6 (Retief, 2000).

Orthology Inference

We used Orthofinder v2.3.3 to identify gene families/clusters (Emms and Kelly, 2019). Genome sequences and annotations for pig and cattle are download from Ensembl (https://www.ensembl.org). The longest proteins of each gene were aligned to one another. The species-specific gene families were determined according to the presence or absence of genes for a given species.

Positively selected genes

The strength of positive selection on each codon of each orthologous gene along a specific targeted lineage of a phylogenetic tree was estimated with the branch site model using Codeml program of the PAML package v4.9. We infer the fitness of the codon sequences along the targeted lineage. An alternative branch site model (Model = 2, NSsites = 2 and Fix = 0) and a null branch site model (Model = 2, NSsites = 2, Fix = 1 and Fix ω = 1) were combined to calculate log-likelihood values for each model using likelihood ratio tests. The log-likelihood values generated were used to assess the model fit, using the Chi-square test with one degree of freedom. Genes with p value less than 0.05 were treated as candidates that under positive selection.

Relaxed selected genes

RELAX (HyPhy v2.3) (Wertheim et al., 2015) analyses were also performed to evaluate whether selective constraints are stronger in foreground branches compared with background branches. An input file that contains the codon sequence alignment and the phylogenetic tree was reconstructed using RAxML v8.2.9, and the foreground branches (test branches) and background branches (reference branches) were then indicated before running. A likelihood ratio test was also performed to evaluate if selection varied between test and reference branches. A significant result (P value < 0.05) indicates a stronger level of selection if K is greater than 1 or a weaker selection or relaxed constraint if K is less than 1.

Functional, pathway and interaction enrichment analysis

KOBAS (version 3.0) (Wu et al., 2006) was applied to perform GO analyses (including cellular composition, molecular function and biological process terms), Reactome pathway and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses. False discovery rate was performed to adjust p values using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). An adjusted p value of < 0.05 was used as the cutoff criterion.

Description of supplementary Material

For a compact layout, in this thesis I did not include all supplementary material. I present Supplementary Figures which may help the reader. For sake of coherence, I kept the original number of Supplementary Figures and tables.

Complete supplementary material are available at the Open Science Framework repository: https://osf.io/wbrhe.

Reference:

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smith, A. F. A., Roskin, K. M., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715. doi:10.1101/gr.1933104.

Bosma, A. A., Oliver, W. L. R., and Macdonald, A. A. (1983). The karyotype, including G- and C-banding patterns, of the pigmy hog Sus (Porcula) salvanius (Suidae, Mammalia). doi:10.1007/BF00123219.

Cabral, V. P., Oliveira, F. S., Machado, M. R. F., Ribeiro, A. A. C. M., and Orsi, A. M. (2001). Study of lobation and vascularization of the lungs of wild boar (Sus scrofa). *Anat. Histol. Embryol.* 30, 205–209. doi:10.1046/j.1439-0264.2001.00315.x.

Daub, J. T., Moretti, S., Davydov, I. I., Excoffier, L., and Robinson-Rechavi, M. (2017). Detection of pathways affected by positive selection in primate lineages ancestral to humans. *Mol. Biol. Evol.* 34, 1391–1402. doi:10.1093/molbev/msx083.

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi:10.1186/s13059-019-1832-y.

Evin, A., Dobney, K., Schafberg, R., Owen, J., Strand Vidarsdottir, U., Larson, G., et al. (2015). Phenotype and animal domestication: A study of dental variation between domestic, wild, captive, hybrid and insular Sus scrofa. *BMC Evol. Biol.* 15, 1–16. doi:10.1186/s12862-014-0269-x.

Frantz, L. A. F., Madsen, O., Megens, H. J., Groenen, M. A. M., and Lohse, K. (2014). Testing models of speciation from genome sequences: Divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. *Mol. Ecol.* 23, 5566–5574. doi:10.1111/mec.12958.

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

Frantz, L., Meijaard, E., Gongora, J., Haile, J., Groenen, M. A. M., and Larson, G. (2016). The Evolution of Suidae. *Annu. Rev. Anim. Biosci.* 4, 61–85. doi:10.1146/annurev-animal-021815-111155.

Gariépy, J. L., Bauer, D. J., and Cairns, R. B. (2001). Selective breeding for differential aggression in mice provides evidence for heterochrony in social behaviours. *Anim. Behav.* 61, 933–947. doi:10.1006/anbe.2000.1700.

Groves, C. P. (1985). Plio-Pleistocene mammals in island Southeast Asia. in *Modern Quaternary research in Southeast Asia. Vol. 9*, 43–54.

Gun, S., Ma, Y., and Wang, G. (2009). Measurements of major internal organs and digestibility of wild pigs (Sus scrofa) captured from Ziwulin Mountains, Gansu. *Acta Theriol. Sin.* 29, 96–100.

Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genet.* 15, e1008302. doi:10.1371/journal.pgen.1008302.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9. doi:10.1186/gb-2008-9-1-r7.

Hall, R. (1996). Reconstructing Cenozoic SE Asia. *Geol. Soc. Spec. Publ.* 106, 153–184. doi:10.1144/GSL.SP.1996.106.01.11.

Hansen, J., Sato, M., Russell, G., and Kharecha, P. (2013). Climate sensitivity, sea level and atmospheric

carbon dioxide. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 371, 20120294. doi:10.1098/rsta.2012.0294.

Heaney, L. R. (1985). Zoogeographic evidence for middle and late Pleistocene land bridges to the Philippine Islands. *Mod. Quat. Res. Southeast Asia. Vol. 9* 9, 127–143.

Heaney, L. R. (1986). Biogeography of mammals in SE Asia: estimates of rates of colonization, extinction and speciation. *Biol. J. Linn. Soc.* 28, 127–165. doi:10.1111/j.1095-8312.1986.tb01752.x.

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). "Whole-genome annotation with BRAKER," in *Methods in Molecular Biology* (Humana, New York, NY), 65–95. doi:10.1007/978-1-4939-9173-0_5.

Horwitz, B. A. (2011). Homeostatic Responses to Acute Cold Exposure: Thermogenic Responses in Birds and Mammals. *Compr. Physiol.*, 359–377. doi:10.1002/cphy.cp040116.

Hu, Y. W., Luan, F. S., Wang, S. G., Wang, C. S., and Richards, M. P. (2009). Preliminary attempt to distinguish the domesticated pigs from wild boars by the methods of carbon and nitrogen stable isotope analysis. *Sci. China, Ser. D Earth Sci.* 52, 85–92. doi:10.1007/s11430-008-0151-z.

Jansky, L., and Hart, J. S. (1968). Cardiac output and organ blood flow in warm- and cold-acclimated rats exposed to cold. *Can. J. Physiol. Pharmacol.* 46, 653–659. doi:10.1139/y68-096.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi:10.1159/000084979.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi:10.1101/gr.113985.110.

Kruska, D. (1987). How fast can total brain size change in mammals? *J. Hirnforsch.* 28, 59–70.

Kruska, D. (1996). The effect of domestication on brain size and composition in the mink (Mustela vison). *J. Zool.* 239, 645–661. doi:10.1111/j.1469-7998.1996.tb05468.x.

Kruska, D. C. T. (2005). On the evolutionary significance of encephalization in some eutherian mammals: Effects of adaptive radiation, domestication, and feralization. *Brain. Behav. Evol.* 65, 73–108. doi:10.1159/000082979.

Künzl, C., and Sachser, N. (1999). The behavioral endocrinology of domestication: A comparison between the domestic guinea pig (Cavia aperea f. porcellus) and its wild ancestor, the cavy (Cavia aperea). *Horm. Behav.* 35, 28–37. doi:10.1006/hbeh.1998.1493.

Lischer, H. E. L., and Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics* 18, 474. doi:10.1186/s12859-017-1911-6.

Liu, L., Bosse, M., Megens, H.-J., Frantz, L. A. F., Lee, Y.-L., Irving-Pease, E. K., et al. (2019). Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* 10, 1992. doi:10.1038/s41467-019-10017-2.

Magnell and Richard Carter, O., and Magnell O., C. R. (2007). THE CHRONOLOGY OF TOOTH DEVELOPMENT IN WILD BOAR – A GUIDE TO AGE DETERMINATION OF LINEAR ENAMEL\rHYPOPLASIA IN PREHISTORIC AND MEDIEVAL PIGS . *Vet. Ir Zootech.* 40, 43–48.

Makina, S. O., Muchadeyi, F. C., Van Marle-Köster, E., Taylor, J. F., Makgahlela, M. L., and Maiwashe, A. (2015). Genome-wide scan for selection signatures in six cattle breeds in South Africa. *Genet. Sel. Evol.* 47, 1–14. doi:10.1186/s12711-015-0173-x.

Muñoz, G., Ovilo, C., Estellé, J., Silió, L., Fernández, A., and Rodriguez, C. (2007). Association with litter size of new polymorphisms on ESR1 and ESR2 genes in a Chinese-European pig line. *Genet. Sel. Evol.*

39, 195. doi:10.1186/1297-9686-39-2-195.

Nan, H., Kraft, P., Qureshi, A. A., Guo, Q., Chen, C., Hankinson, S. E., et al. (2009). Genome-wide association study of tanning phenotype in a population of european ancestry. *J. Invest. Dermatol.* 129, 2250–2257. doi:10.1038/jid.2009.62.

Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., et al. (2012). Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* 22, 602–610. doi:10.1101/gr.130468.111.

Popejoy, A., Domanska, D., and Thomas, J. (2020). Genome-Wide Search for Candidate Drivers of Adaptation Reveals Genes Enriched for Shifts in Purifying Selection (SPurS). *bioRxiv*, 2020.01.11.902759. doi:10.1101/2020.01.11.902759.

Price, E. O. (1999). Behavioral development in animals undergoing domestication. *Appl. Anim. Behav. Sci.* 65, 245–271. doi:10.1016/S0168-1591(99)00087-8.

Quinlan, A. R. (2014). "BEDTools: The Swiss-Army tool for genome feature analysis," in *Current Protocols in Bioinformatics* (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 11.12.1-11.12.34. doi:10.1002/0471250953.bi1112s47.

Razmaite, V., Kerziene, S., and Jatkauskiene, V. (2009). Body and carcass measurements and organ weights of Lithuanian indigenous pigs and their wild boar hybrids.

Retief, J. D. (2000). "Phylogenetic analysis using PHYLIP," in *Bioinformatics methods and protocols* (Springer), 243–258.

Schütz, K. E., Forkman, B., and Jensen, P. (2001). Domestication effects on foraging strategy, social behaviour and different fear responses: A comparison between the red junglefowl (Gallus gallus) and a modern layer strain. *Appl. Anim. Behav. Sci.* 74, 1–14. doi:10.1016/S0168-1591(01)00156-3.

Schwerdt, J. G., Mackenzie, K., Wright, F., Oehme, D., Wagner, J. M., Harvey, A. J., et al. (2015). Evolutionary dynamics of the cellulose synthase gene superfamily in grasses. *Plant Physiol.* 168, 968–983. doi:10.1104/pp.15.00140.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.

Stocks, J. M., Taylor, N. A. S., Tipton, M. J., and Greenleaf, J. E. (2004). Human Physiological Responses to Cold Exposure. *Aviat. Sp. Environ. Med.* 75, 444–457.

Tarailo-Graovac, M., and Chen, N. (2009). "Using RepeatMasker to identify repetitive elements in genomic sequences," in *Current Protocols in Bioinformatics* (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 4.10.1-4.10.14. doi:10.1002/0471250953.bi0410s25.

Tjia, H. D. (2006). Geological evidence for Quaternary land bridges in insular Southeast Asia. *Archeol. Indones. Perspect. RP Soejono's Festschift. Indones. Inst. Sci.*, 71–78.

van der Made, J., Morales, J., and Montoya, P. (2006). Late Miocene turnover in the Spanish mammal record in relation to palaeoclimate and the Messinian Salinity Crisis. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 238, 228–246. doi:10.1016/j.palaeo.2006.03.030.

Visconti, A., Duffy, D. L., Liu, F., Zhu, G., Wu, W., Chen, Y., et al. (2018). Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure. *Nat. Commun.* 9, 1–7. doi:10.1038/s41467-018-04086-y.

Wallace, A. R. (1855). On the law which has regulated the introduction of new species. *Ann. Mag. Nat. Hist.* 16, 184–196. doi:10.1080/037454809495509.

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference

genome sequence to enable pig genetics and genomics research. *Gigascience* 9, 1–14. doi:10.1093/gigascience/giaa051.

Wertheim, J. O., Murrell, B., Smith, M. D., Pond, S. L. K., and Scheffler, K. (2015). RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* 32, 820–832. doi:10.1093/molbev/msu400.

Wu, J., Mao, X., Cai, T., Luo, J., and Wei, L. (2006). KOBAS server: A web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* 34. doi:10.1093/nar/gkl167.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi:10.1101/gr.092981.109.

Zeder, M. A. (1982). The Domestication of Animals. *Rev. Anthropol.* 9, 321–327. doi:10.1080/00988157.1982.9977605.

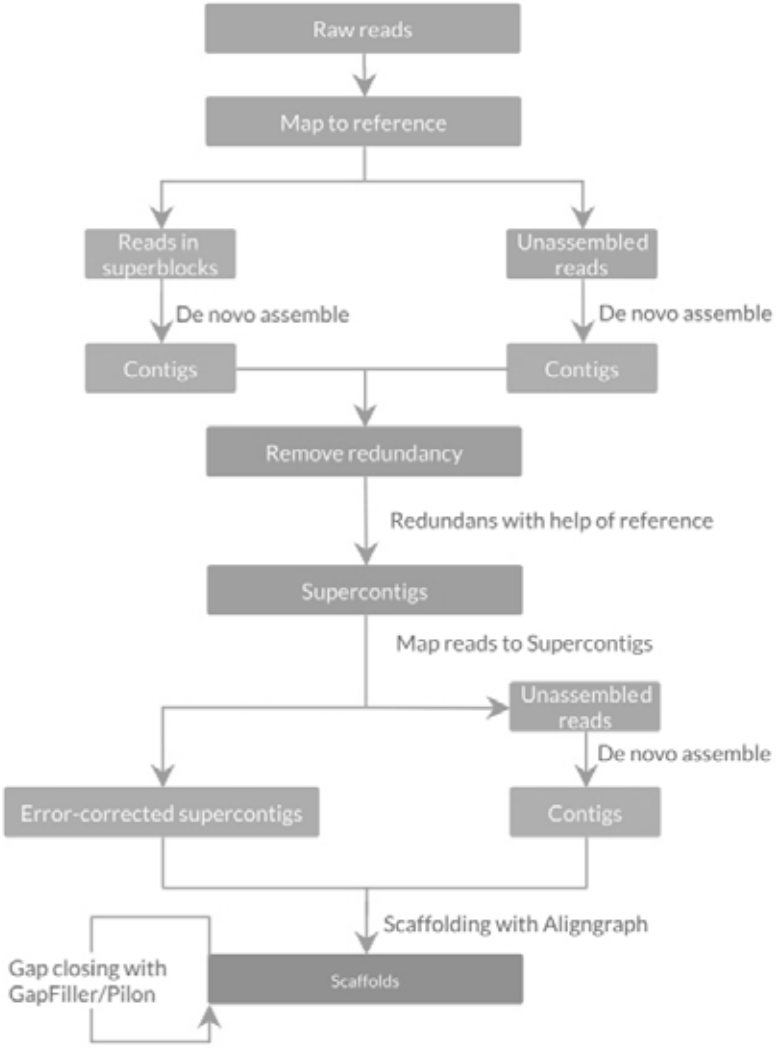
Zeder, M. A. (2012). Pathways to animal domestication. *Biodivers. Agric. Domest. Evol. Sustain.*, 227–259. doi:10.1017/CBO9781139019514.013.

ZHAI, R. G., and Wang, X. G. (2001). Analysis on slaughter performance of wild crossbred pigs. *Res. Agric. Mod.* 22, 171–173.

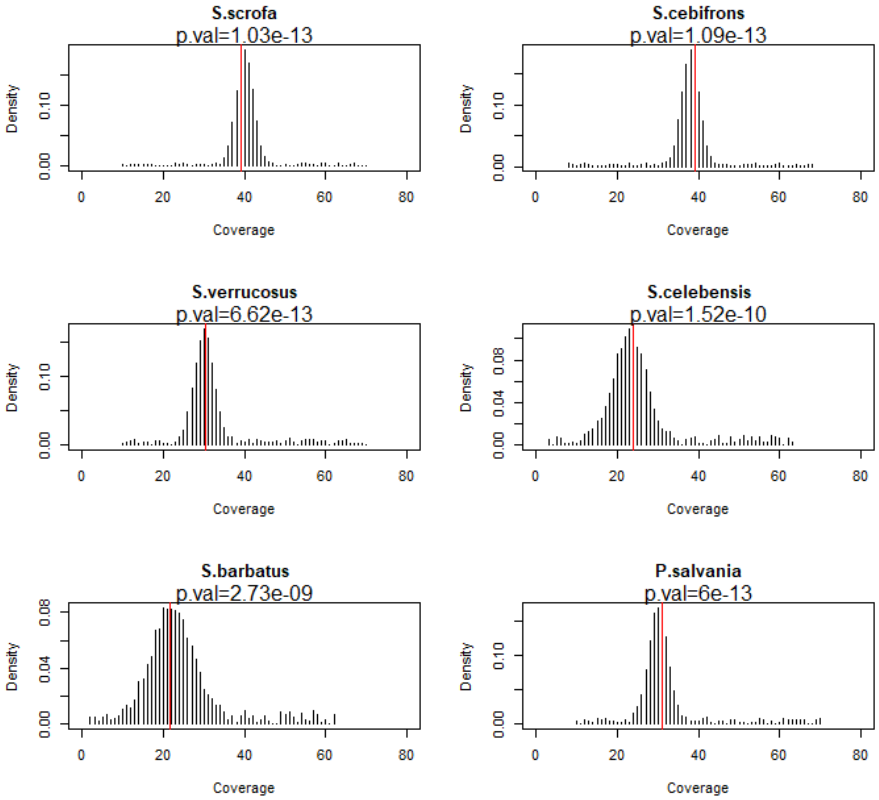
Zhao, M., Du, J., Lin, F., Tong, C., Yu, J., Huang, S., et al. (2013). Shifts in the evolutionary rate and intensity of purifying selection between two Brassica genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. *Plant J.* 76, 211–222. doi:10.1111/tpj.12291.

Zhao, T., and Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2165–2174. doi:10.1073/pnas.1801757116.

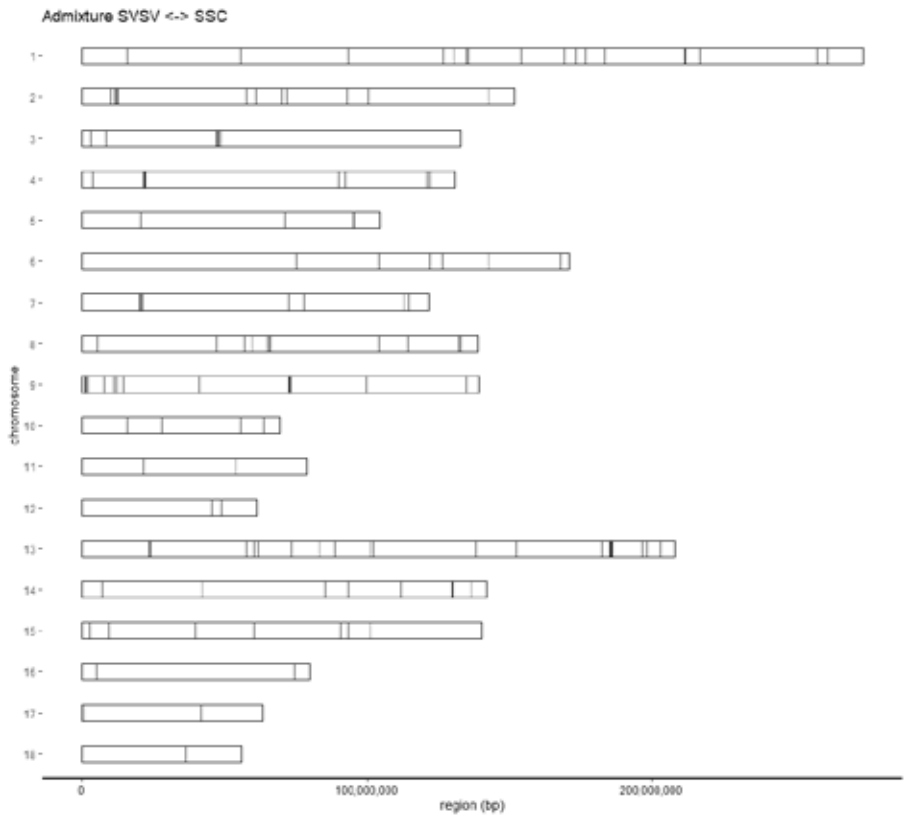
Supplementary Martial



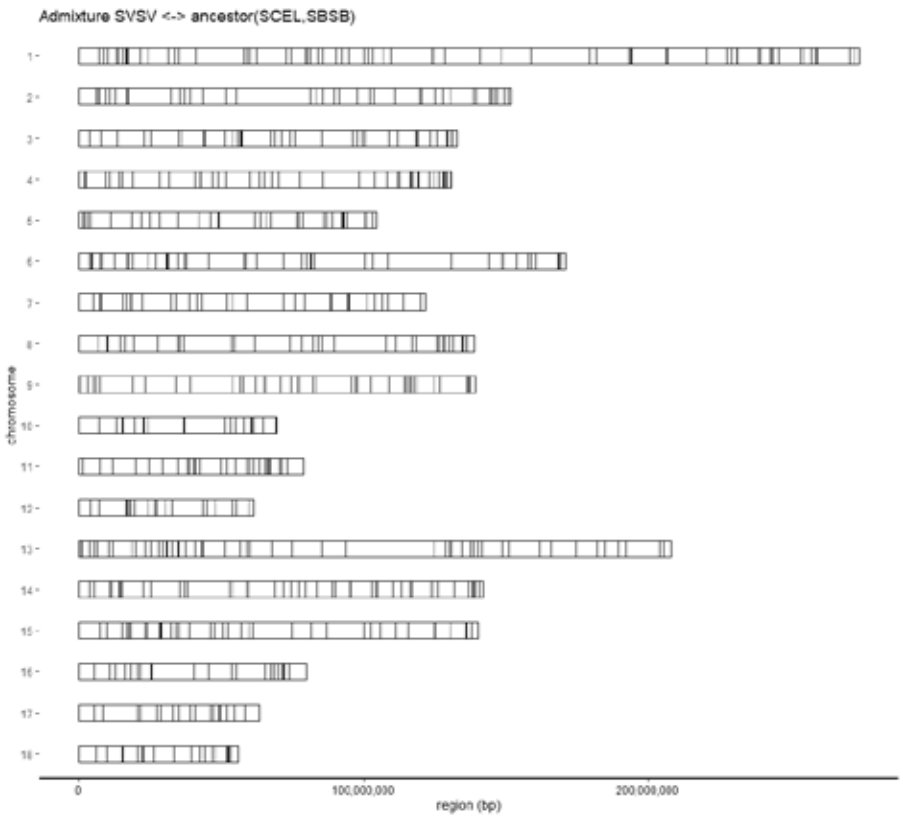
Sup Fig 5.1 Schematic flowchart of the reference-guided assembly pipeline. Reads and their alignments are shown in blue. Regions of constant coverage were definded as blocks. Adjacent blocks were combined into superblocks until they reached a minimal length of 10 kb. Superblocks were defined in an overlapping fashion, such that blocks could belong to several superblocks. All reads of a superblock were assembled with reads that had not been aligned. Resulting contigs were merged into a nonredundant set of supercontigs. Short read alignments against the supercontigs allowed for error correction and scaffolding. Short read alignments against the scaffolds enabled a final quality assessment and filtering.



Sup Fig 5.3 Gene average depth for all species. Number of reads mapping to the gene was calculated with bedtools. Red line represents the average sequence coverage in the whole genome. The normality of the distribution of read depth was test with Shapiro–Wilk test.



Sup Fig 5.4 Distribution of chromosomal segments supporting admixture between *Sus verrucosus* and *Sus scrofa*. Black rectangle represents the introgressed segment.



Sup Fig 5.5 Distribution of chromosomal segments supporting admixture between *Sus verrucosus* and ancestor of *Sus celebensis and Sus barbatus*. Black rectangle represents the introgressed segment.

6

General discussion

6.1 Introduction

Developments in sequencing technology have revolutionized our understanding of evolutionary biology, and have led to the concept of evolutionary genomics. Continuous efforts to sequence additional species have generated a comprehensive dataset for biologists to study the evolution of the diversity of life on earth, to explore the underlying biological mechanisms and to aid conservation of ecosystems. In this thesis I systematically described the evolutionary history of the *Suidae* species and examined complex speciation scenarios with hybridization. The critically endangered pygmy hog was used as a model species to reconstruct the demographic trajectory and evaluate corresponding genetic consequences of it. With the genome sequences of *Sus*, I characterized the evolution of genome architecture underlying the rapid speciation and adaptation. I also demonstrated the importance of a fine-structured reference genome for the interpretation of selective signatures. In the following sections, I attempt to generalize these findings to provide answers to the fundamental questions of contemporary interest in evolutionary genomics.

6.2 A resource for the pig research community

While our findings represent a significant contribution to the understanding of genome evolution in *Suidae*, one of the major outcomes of this work is a resource that has been created for the pig research community. For long periods of time, studies on pigs were mainly focused on the domesticated pig and the pig research community lacked the resources to place pig into a broader evolutionary context. In recent years, the Animal Breeding and Genomics group at Wageningen University has sequenced hundreds of genomes of domestic and wild *Suidae*, significantly expanding available genome resources. In Chapter 2 of this thesis, whole genome sequencing data of pygmy hog is presented. With this, we are approaching the complete phylogenetic framework of *Suidae* species. Moreover, in Chapter 4, I presented a chromosome level genome assembly and annotation of *Sus cebifrons*. Our results suggest that having a highly contiguous genome sequence is a great addition to large scale genome resequencing at low to medium coverage, especially for non-coding regulatory elements and chromosome structure.

So far, several species and sub-species of extant Suidae still have not got their genome sequenced, namely, *Babyrousa bolabatuensis* (Bola Batu babirusa), *Babyrousa celebensis* (North Sulawesi babirusa), *Babyrousa togeanensis* (Togian babirusa), *Hylochoerus meinertzhageni* (giant forest hog), *Phacochoerus aethiopicus* (desert warthog), *Sus ahoenobarbus* (Palawan bearded pig) and *Sus oliveri* (Mindoro

warty pig). Due to the lack of sufficient morphological differentiation, the taxonomy of these species remains controversial. The genus *Babyrousa* was first described as a monotypic taxon (Groves, 1980). Later, it was recommended treating the subspecies under genus *Babyrousa* as separate species based on morphological classification (Meijaard and Groves, 2002). Likewise, *Sus ahoenobarbus* (Palawan bearded pig) and *Sus oliveri* (Mindoro warty pig) used to be subspecies of *Sus barbatus* (Bornean bearded pig) and *Sus philippensis* (Philippine warty pig), respectively. Based on phylogeny inferred from partial mitochondrial DNA (Lucchini et al., 2005) both subspecies have been upgraded to separate species. Sequencing the genomes of all species of the *Suidae* family can provide a full resolution of their phylogeny and resolve the taxonomy question. In the meantime, there are also great merits for generating chromosome-scale assemblies for all *Suidae* species. As with the rational of on-going ambitious genome projects, like the Earth BioGenome Project (EBP) (Lewin et al., 2018), reference-quality genomes across the phylogenetic spectrum serve as backbone of understanding evolutionary and other biological processes, such as speciation, extinction and adaptation. This knowledge creates new foundation to provide solutions for preserving biodiversity and ecosystems, and finally save mankind from itself.

6.3 Resolving complex speciation models

Phylogenetic reconstruction has now advanced into the era of phylogenomics that takes full use of sequence information from whole genomes. This provides great power in resolving difficult taxonomy problems. However, processes that generate gene tree discordance may impede species tree reconstruction. Gene flow among populations and species is one of the major evolutionary processes that can generate gene tree discordance. The traditional mode of species divergence assuming a model of strict allopatry is now being amended with many empirical examples of divergence accompanied by gene flow, or allelic introgression across species boundaries (see General introduction). Another natural evolutionary process responsible for gene tree discordance across the entire tree of life is incomplete lineage sorting (ILS). ILS can be defined as a genealogy different from the species phylogeny, which arise due to the retention of ancestral polymorphisms in different species or populations.

In Chapter 2, to effectively accommodate ILS, I used ASTRAL to reconstruct *Suidae* phylogeny. ASTRAL incorporates a multispecies coalescent (MSC) model, which uses a simple heuristic that provides estimates that are statistically consistent when gene trees arise under the MSC (Mirarab and Warnow, 2015). It computes branch lengths in coalescent units and measures branch support by calling local posterior probability.

However, jointly considering ILS and gene flow remains a great challenge. The MSC method will fit a single neutral coalescent model to each branch in a species phylogeny. As to the coalescent model, it relies on the assumption that speciation is a series of isolation events unaccompanied by gene flow. Failing to account for gene flow during species tree estimation surely impacts demographic parameter estimation, such as divergence time, population size and migration rate. Recently, Flouri et al., (2020) proposed a full likelihood method for the joint estimation of population demographic parameters. It extended the MSC to incorporate cross-species gene flow. However, currently the demographic parameters of this model are fixed, with the number and directions of the introgression events specified by the user. The future research in the field of species tree reconstruction should therefore include developing more sophisticated frameworks for hypothesis testing using the MSC and developing network coalescent models to test finer models with variable gene flow events. This will finally allow us to reconstruct the detailed evolutionary history and understand genetic diversity across time and space.

6.4 Fitting complex speciation models to genome sequences

There are several statistical methods available for identifying introgression from genome-wide data. The most well-known is D-statistic (also known as Patterson's D or ABBA-BABA test), a statistical quantification of SNP patterns to detect hybridization between taxa (Patterson et al., 2012). It measures the excess sharing of derived alleles between each of two populations in a pair (ingroup populations) and an outgroup population, labeled as "ABBA" and "BABA" (Fig 6.1). D-statistic is based on a phylogenetic argument that, due to incomplete lineage sorting, ABBA and BABA counts are expected to be similar under a null hypothesis of no gene flow. An excess of either pattern represents a signal that can be used to detect introgression. Identifying specific genes or segments of the genome that are introgressed is more challenging, because ILS and gene flow produce very similar patterns of shared genetic diversity. Inferences regarding specific regions must instead rely on demographic models that include assumptions about parameters such as divergence time, effective population size (Ne), and recombination rate. The extent of ILS (and the resulting gene tree discordance) is correlated with the effective population size and negatively correlated with the time between speciation events (Maddison, 1997). Also, in the presence of selection, Ne is negatively correlated with the recombination rate (Pease and Hahn, 2013). Taken together, Huerta-Sánchez et al., (2014) provide a statistical framework to test whether the lengths of putative introgressed haplotypes exceeded the length of segments resulting from ILS. For the

more recent admixture, it is still challenging to rule out ancestral polymorphisms from true introgression. For this reason, ancient sequence data have proven useful. If historical samples are available, comparing the ancient and modern genomes would be an efficient way of excluding shared ancestral polymorphisms. It allows inference of ancestry of specific genes or genomic regions that an admixed individual derives from an ancestral population. For example, with the first genome sequencing data from a Neanderthal individual (Green et al., 2010), researchers directly investigated the relative contribution of Neanderthals to individual human genomes using D-statistic. Further investigations has shown that in modern human genomes more than 95% of polymorphic derived alleles shared with Neanderthals are due to incomplete lineage sorting, and thus less than 5% of shared alleles are informative about introgression (Lin et al., 2015).

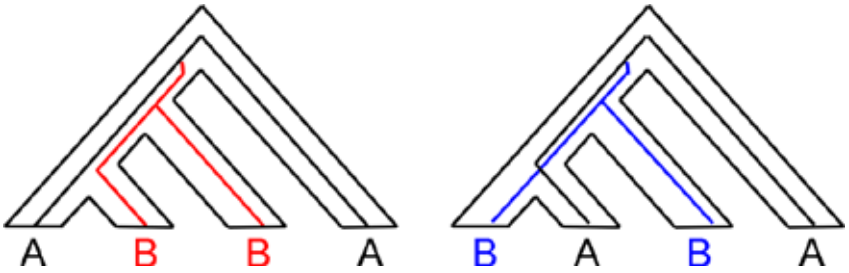


Fig 6.1 The ABBA-BABA test is used to detect an excess of one pattern of discordance relative to the other in four taxon phylogenies. An equal number of incongruent ABBA and BABA allele patterns are expected under ILS alone. A significant excess of ABBA or BABA allele patterns is consistent with a history of introgression.

Four taxon tests (e.g. D-statistic, F-statistic, fd) can estimate the introgression probability and internal branch lengths in coalescent units on the species tree. However, with a null hypothesis of symmetric genealogies, these methods often do not have an implicit assumption about the direction of gene-flow. In Chapter 2, I show that such statistics can be difficult to interpret and can lead to biased inferences. Admixture involving unsampled or extinct lineages can result in complex site patterns and might influence the results of the D-statistics (Durand et al., 2011), so called ghost introgression. Direct genetic traces of ghost introgression are detected as highly divergent haplotypes in present day genomes. These divergent haplotypes can be either in mitochondrial DNA (mtDNA) or in the nuclear genome. Deep mitochondrial divergence along with shallow overall nuclear genomic differentiation may indicate mitochondrial replacement/capture through genetic introgression (Mao et al., 2013; Rakotoarivelo et al., 2019; Zhang et al., 2019).

However, for the nuclear genome, it is more complicated. Intense episodes of continuous gene flow and recombination will shuffle the introgression pattern along the genome. In Chapter2, I propose that combining a relative distance method (RNDmin, Rosenzweig et al., 2016) and a four taxon test (D-statistic) can successfully resolve the genealogy discordance caused by a ghost introgression, which can also be applied to other similar studies. Model-based approaches can also address complex speciation hypothesis testing. Large-scale simulation based on demographic inferences drawn from whole genome sequences (e.g., G-PhoCS, MOMI, admixturegraph (Gronau et al., 2011; Leppälä et al., 2017; Kamm et al., 2018)) or site frequency spectrum (Excoffier et al., 2013; Martin and Amos, 2020), have shown their power of detecting and quantifying ghost introgression. However, the number of coalescence-events one can include in a model grows quadratically with the number of lineages (individuals) (Oulu, 2003; Weber, 2008), limiting the application in practice, especially when one can only sequence one or two individuals per population.

6.5 Hybridization and range expansion

It has become clear that speciation with gene flow is more of a rule than an exception. Much of our knowledge about the role of hybridization in evolution comes from studies of secondary contacts during Quaternary environmental changes (Hewitt, 2011). Climate changes are expected to shape the ecological niche and biodiversity distribution. For example, rising temperatures can result in animals and plants migrating northward or uphill to escape the heat and vice versa. Generally, the distribution of species is limited by the ability of populations to adapt to environmental pressures at the range margins (Bridle and Vines, 2007). Populations at the range edge are directly facing the severe selection pressure and the danger of extinction. They can avoid becoming extinct by the arise of alleles for adaptive traits to the new environment. However, studies further predict that the source of local adaptation at a range edge does not always result from new mutations. Compared with the rapid climate fluctuations, the waiting time until favorable mutations emerge is too long (Orr and Unckless, 2008; Bijlsma and Loeschcke, 2012). An alternative source of adaptive allelic variants at the range edge is hybridization, interbreeding of distinct evolutionary populations or species. Interspecific admixture can provide genetic variation that allows populations to adapt to selective pressures, either through the transfer of alleles for key traits that are already adapted to the new environment, or through an increase in overall genetic diversity which can then facilitate adaptation (Stelkens et al., 2014; Hamilton and Miller, 2016). Ecological

conditions that lead to hybridization and establishment of hybrid populations also tend to lead to range expansion. Range expansion is by definition a non-equilibrium process, because spatial distribution of individuals within a population is based on the individual fitness. Once the adaptive allele is introduced by hybridization, individuals with higher adaptability to the new environment will aggregate at the range edge, thus further accelerating the establishment of populations in newly colonized areas. Moreover, hybridization can lead to replacement of the resident species by the species invasion or the breakdown of species boundaries (Huxel, 1999). In Chapter 2, the range expansion of *Sus scrofa* during the Plio-Pleistocene (2-1Mya) represents an illustrative example of this phenomenon. During the expansion from south to north Asia, *Sus scrofa* not only replaced many local species, it also "absorbed" the genetic material from closely related species' genomes (ISEA *Sus*, pygmy hog and an unknown *Suidae*). Some genes harbored within the introgressed genomic fragments are further detected to be associated with altitude adaptation, which may potentially boost the viability of pigs during northward migration. These results highlight that admixture events can involve multiple species and that hybridization plays a role as an important evolutionary force in range expansion. Ironically, in the light of global climate change and human-mediated species invasion, hybridization might be a factor that contributes to the collapse of populations and resulting in biodiversity loss. Thus, hybridization is not just a topic of academic interest but an unavoidable issue of conservation of biodiversity in the wild.

6.6 Demographic history inferred from genomes

Uncovering the temporal fluctuations of population size and understanding corresponding consequences provide information about how species reacted to past events such as glacial cycles, climate change or human intervention. Such prior information can be used directly in species conservation (Frankham et al., 2002). To make decisions regarding a conservation programme, it is important to understand the level of inbreeding in the population, which also depends on the population size and its demographic history. Traditionally, inbreeding levels was estimated using pedigree or low density genetic markers (e.g., microsatellite, SNP array) (Kardos et al., 2016). Recently, whole-genome sequencing has enabled a transition from focusing on genome-wide estimates of inbreeding to examining patterns of inbreeding across the genome, which means literally observing inbreeding across the genome. In Chapter 3, I used whole genome sequencing data to characterize the genetic diversity, demographic history, and level of inbreeding of the critically endangered pygmy hog. The pairwise sequentially Markovian coalescent (PSMC)

analysis shows that pygmy hog exhibited persistent low effective population size smaller than ~500 from 100,000 to 10,000 years ago. However, due to the homogeneous distribution of heterozygous sites in the pygmy hog genome, PSMC cannot make reliable estimates of demography more recently than 10,000 years ago. A way to infer recent demography is by investigating runs of homozygosity (ROH). In pygmy hog an absence of long ROH indicates that no recent inbreeding happened in the pygmy hog population. Our study highlights the power of combining several methods to reconstruct a complete and continuous demographic history.

6.7 Evolution of genome architecture

6.7.1 Highly contiguous genome assemblies

The development of genome sequencing technologies and assembly tools has rapidly advanced with genome projects. Since the Illumina short read sequencing technique was introduced, the substantial decrease in sequencing costs and efforts has enabled genome sequencing project for non-model species. The read length for Illumina sequencers is typically 150bp to 250bp, with less than 1% error rate. To assemble genomes with lengths of billions of base pairs, such reads are too short in length to obtain high quality assemblies. This is mainly due to the limited power of the short reads for resolving tandem repeats, GC-rich regions and duplicated segments in the genome. To overcome these limitations, complementary approaches based on the short read platform were developed. One example is 10X genomics linked read sequencing technology (Zheng et al., 2016). In 10X genomics sequencing, DNA is partitioned and dispersed to barcoded oligonucleotides, which are then sequenced using a short-read sequencer. The resulting reads are then assembled using the long-range linkage information from barcodes into long DNA molecules. However, extreme repetitive sequences can still introduce ambiguity and prevent complete reconstruction of a genome. Some of the newer chromosome conformation capture techniques like Hi-C provide genome-wide contact maps between loci within chromosomes (Lieberman-Aiden et al., 2009). Because the intrachromosomal interaction probability rapidly decays with increasing genomic distance and is much higher than interchromosomal interaction, Hi-C data can be used to order and orient contigs, even producing scaffolds which can span complete chromosomes (Zheng et al., 2016). In Chapter 4, I show that by only using sequencing data from short read platforms (10X and Hi-C), one can generate a highly contiguous genome assembly. Further, in Chapter 5, using the available genome sequences as reference, I applied a reference guided assembly approach to generate genome assemblies with reasonable quality using previous published medium coverage

resequencing data. These studies provide a good example of initiating a new genome project and fully utilizing the available genome resources.

De novo genome assemblies are rarely considered complete due to technical limitations during the sequencing and assembly process. Thus, constant improvements in sequencing technologies and assembly algorithms make genome building an iterative process. Examples are the human reference genome which is currently version 38 (GRCh38) and pig which is version 11 (SSC11.1, Warr et al., 2020). For the genome sequence of *Sus cebifrons* presented in this thesis, there is also considerable room for improvement, especially considering the contig N50. In recent years, two platforms have acquired major roles in improving *de novo* genome sequencing of larger eukaryotic animals, namely PacBio single molecule real time sequencing and Oxford Nanopore Technology sequencing. Both techniques produce long reads of up to several thousand kilobases. The long read length can easily bridge the repetitive genome regions and results in high contiguous scaffolds. Currently, sequencing accuracy is limited by the technology and chemicals used in long read sequencing. To eliminate sequencing errors, long read technologies require high depth sequencing when used for *de novo* genome assembly. Lower cost, and lower sequencing errors of Illumina short read still provide a solid alternative to long read sequencing. Thus, new technologies require more time to be adapted and improved. Nevertheless, genome projects could immensely benefit from a hybrid approach. Combining sequencing data obtained from short and long read technologies has shown to substantially increase assembly quality, especially for complex genome structures. Most recently, The Telomere-to-Telomere (T2T) consortium used almost all the available sequencing technics (including 120x coverage of Oxford Nanopore, 70x PacBio CLR, 30x PacBio HiFi, 50x 10X Genomics, as well as BioNano DLS and Arima Genomics Hi-C) to sequence the CHM13hTERT human cell line and assembled a complete human genome with only 5 known gaps within the rDNA gene cluster (Logsdon et al., 2020; Miga et al., 2020). Moreover, groups, affiliated with Earth BioGenome Project (EBP), such as the Vertebrate Genomes Project (VGP) and the 10,000 Plants(10KP), have made significant progress in building up standardized pipelines to generate high-quality and complete references (Cheng et al., 2018; Rhie et al., 2020). Thus, there is no doubt that this represents an exciting area of future research. The continuous improvement of these cutting-edge technologies and bioinformatics tools will allow for sequencing every single base in the genome and ultimately every representative genome in the Tree of life.

6.7.2 Chromosomal rearrangement

As the most conspicuous genetic variation, chromosome rearrangement has been a topic of interests for many decades. From chromosome banding to whole genome sequencing, higher data density allowed for a much more detailed analyses of chromosomes. In theory, comparing *de novo* assembly of the query genome and the reference genome allows detection of all forms of chromosomal rearrangement present in the query genome. In Chapter 4, I compared the genome assemblies of two closely related species, *Sus scrofa* and *Sus cebifrons*. Three major questions were addressed: two known chromosome fusion events, conservation of syntenic blocks between species and the disruption of syntenic blocks (via chromosomal inversion events).

By directly comparing the two genome sequences, I observed the previous reported chromosome fusions (Musilova et al., 2010), that led to the different karyotypes of these two species. It is well known that different *Sus* species can produce fertile hybrid offspring. With the genome sequence and Hi-C chromatin interaction information, I confirmed that *Sus scrofa* and *Sus cebifrons* share very similar chromatin conformation even in chromosomes involved in fusion events. A successful meiosis requires accurate homologous chromosome paring, recombination and segregation by recognizing the compacted loop arrays of the homologs (Keeney et al., 2014; Patel et al., 2019). This is the basis of the hypothesis that chromosome structure is one of the determining factors of post-zygotic isolation. A recent study used mitochondrial genetic divergence as proxy to predict the reproductive compatibility between species (Allen et al., 2020). Similar to that, based on the mechanism of meiosis, I propose that the similarity of chromatin conformation can be a more direct proxy to infer hybrid fertility. Thus, in the future, comparative analysis of chromatin conformation based on Hi-C should be explored to evaluate the role of chromosome topology in the speciation process.

The genome organization reveals a macro perspective of the evolutionary history. Mammalian genomes are overall highly syntenic (Zhao and Schranz, 2019). Comparison of syntenic blocks between *Sus scrofa* and *Sus cebifrons* genome assemblies confirms that ~95% of the genetic elements are localized in the same order. Within a short evolutionary time scale (~4 Million years), paracentric inversions serve as the dominant mechanism for shuffling the order of genes along a chromosome. Transposable elements (TEs) have long been considered as activate components of the genomes, inducing structural rearrangements by chromosome breakage, thereby inactivating or duplicating genes and adding or removing regulatory regions (Slotkin and Martienssen, 2007; Akagi et al., 2013). The presence of paralogous copies of TEs in a genome can provide the substrate for aberrant

transposition and ectopic recombination leading to novel structural rearrangements and genomic plasticity. To date, most of the conclusions are drawn based on studies within species. In Chapter 4, I examined TEs distribution in the genomes of two closly related species in order to better understand the genomic changes around large inversions. The results, that TE contents are lower in chromosome regions proximal to the inversion breakpoints in both *Sus scrofa* and *Sus cebifrons* genome, cannot be explained by current theory (Fig 6.2). The dramatic decrease of TE content can overlap with the inverted segment or be located outside the inverted region. Moreover, this pattern is usually only present at one side of the inversion breakpoints. The cross-species comparison implied that the different effects of TEs in inversion may be explained by the different evolutionary timescales. We hypothesize that the lowered TE content may prevent secondary chromosomal rearrangement and maintain the inverted allele frequency. Thus, chromosomal inversions between species have generally reached fixation in both populations. Due to the time constrains of my PhD, I wasn't able to test our hypothesis in broader taxonomic groups. Further theoretical and empirical work is therefore needed to evaluate these hypotheses.

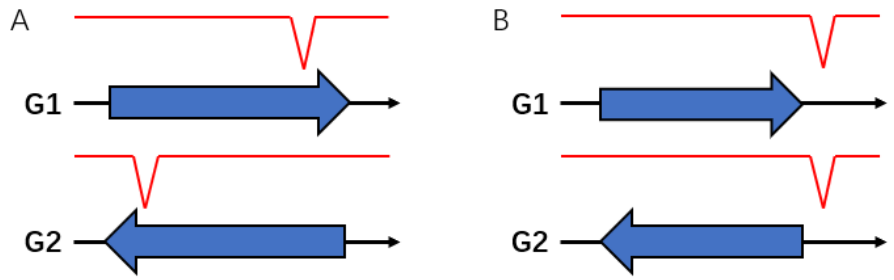


Fig 6.2 Example of chromosomal inversion and associated TE distribution. G1 and G2 represent the query and target genome. Red line above each genome represent the TE content. The blue arrows represent the inverted segments.

6.7.3 Adaptation genomics

Another major theme of this thesis is to discover genomic determinants of phenotypic differences between species, which is important to understand nature's fascinating phenotypic diversity. In Chapter 4, I introduced a molecular evolutionary analysis using the *Sus scrofa* and the *Sus cebifrons* genomes, which are phenotypically different and living in different natural environments. Our analyses detected candidate genes that show differences in the coding regions. The functional annotation of these candidate genes offers potential insights for the observed

environmental adaptation and phenotypic divergence. Notably, the current pig reference genome was made from a domesticated pig (a Duroc). Since domesticated pig has gone through a complex evolutionary history before domestication and after the split from other *Sus* species, it is important to distinguish the genetic traces left by different evolutionary dynamics (selection during domestication and natural selection) in the modern pig genome. In Chapter 5, having a near complete collection of *Sus* species, I was able to reconstruct the ancestral codons sequence of *Sus*. This enabled me to infer the selection constrain in each of the *Sus* species lineages, as well as in the ancestral branch of *Sus*. I specifically tested for long-term shifts in selection pressure associated with changes in ecology and function. The analyses revealed a subset of positively selected genes in the pig genome categorized into two categories: adaptive radiation or artificial selection, which correspond to range expansion and domestication events, respectively. A general remark concerning this work is the importance of carefully contrasting any observations suggestive of positive selection between (or among) different partitions of a phylogeny. Proper prior hypothesis about the location or timing of the divergence point should help identify historically important selection events and uncover mechanisms of evolution. Ideally, in the future, when more genome sequences of wild boar are available, this approach can better help us to understand the selection constrain in the pig genome at different times of evolution.

In this thesis, I mainly discuss the coding sequences of genes involved in adaptation and domestication, although it is now well established that evolutionary changes in gene expression are widespread, both within and between species (Kleinjan and Van Heyningen, 2005; Wray, 2007; Stern and Orgogozo, 2008). Regulatory sequence changes do not alter the protein structure, rather influence phenotypes by controlling the spatial and temporal distribution, as well as the quantity of protein produced. SNPs located within epigenetic switches, e.g., promoters and enhancers, may underlie inter-species differences in gene expression levels and also underlie phenotypic adaptation and speciation (Elena and Lenski, 2003). The speed of genetic adaptation depends on the rate of emergence of fitness mutations as well as the strength of selection pressure. The rate of epigenetic variation is usually much faster than genetic mutation (Rando and Verstrepen, 2007). Under rapid environmental change, epigenetic variation can adjust phenotypes instantaneously (i.e. during development) without modifying the DNA sequence, thereby quickly providing adaptive plasticity. Examples of epigenetic switches of gene regulation include transcription factor activity, DNA methylation, chromatin modification and chromatin accessibility (Henikoff and Greally, 2016). Epigenetic switches can emerge through positive feedback loops in biochemical or genetic networks. Epigenetic

variations contained in germline cells are heritable from parents to offspring (Jablonka et al., 1992; Grossniklaus et al., 2013). Collectively, DNA regulatory elements and epigenetic components constitute a gene regulatory network, hence playing a central role in organisms' responses to environment. Further integrative studies using multi-omics data (e.g., genomics, transcriptomics, epigenomics) should improve the understanding of adaption mechanisms in specific environmental dynamics.

6.8 Methodological challenges in the post genomic era

6.8.1 Reference bias

Resequencing is currently the most commonly used DNA sequencing approach. Resequencing data analyses often begin with aligning reads to a reference genome, with the reference represented as a linear string of bases. Single linear haploid references such as the human GRCh38 assembly or the pig SSC11.1 assembly used in this thesis originated from a single individual or a mix of several individuals. Thus, the underrepresentation of the entire genetic variation of population leads to reference bias, a tendency to miss-align or incorrectly align reads containing non-reference alleles. This can ultimately lead to confounding scientific results, especially for analyses related to hypervariable regions (Brandt et al., 2015) or allele-specific effects (Degner et al., 2009). Generally, reference bias can influence variant calling by missing alternative alleles or by wrongly calling heterozygous sites as homozygous for the reference allele which is known to influence estimates of heterozygosity and allele frequencies. It impacts any comparative study where reads are mapped to a divergent reference. For example, reference bias can lead to the underestimation of rates of molecular evolution or the overestimation of phylogenetic discordance due to stochastic genealogical processes (i.e., incomplete lineage sorting) or hybridization (Günther and Nettelblad, 2019).

Efforts have been made to address reference bias, which can roughly be classified into three directions. First, based on the linear reference, some studies suggest replacing the widely used linear reference with a customized reference that incorporates sample-specific variation (Stevenson et al., 2013; Buchkovich et al., 2015; Sarver et al., 2017). This can increase alignment and genotyping accuracy. However, this linear reference-based framework will always be limited with respect to larger-scale structural variation (e.g., chromosomal translocations and inversions). Thus, this approach is most useful for generating comparative evolutionary genomic data sets of orthologous loci that can be used for phylogenetic and population genomic inferences. Second, alternative mapping strategies such as mapping against

genome graphs could eliminate reference bias already at the mapping step (Paten et al., 2017; Garrison et al., 2018; Rakocevic et al., 2019). Genome graphs include the reference genome together with genetic variation and polymorphic haplotypes and can not only increase the number of aligned reads and resolve haplotypes, but also allow a better representation of the diversity among population groups (Fig 6.2). A typical application of graph reference is for "pangenome", which has been generated for human, pig, goat and many other extensively studied species (Eizenga et al., 2020). Although graph genome tools could eventually transform how whole-genome sequences are analyzed, current implementations have important limitations (Marschall et al., 2018). The accuracy of graph genomes highly relies on the prior identification of structural variations and complex genomic rearrangements, which generally need to be detected by high quality sequencing. This leads to the third approach, using an unbiased *de novo* assembly to make inferences on individual genomes. Single molecule and synthetic long reads are advantageous for structural variation calling because they can overcome repetitive regions and span structural variations entirely. For example, a recent study using a combination of long-read technologies reported a sevenfold increase in structural variations compared with the number of structural variations using standard short-read whole-genome sequencing. In Chapter 4 and 5, I show the possibility to directly compare genome sequences from different species. With that, I successfully inferred structural variations, as well as gene flow and selection signals. However, building an alignment-based framework for analysis of mammalian genomes is a huge computational task. Until we break through this bottleneck, graph genome approaches embracing high quality *de novo* genome assemblies may be the best solution to harness the full potential of genome data.

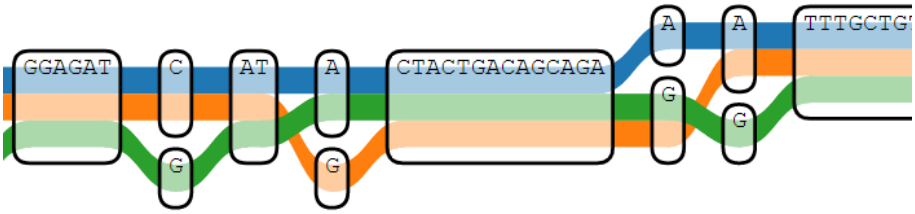


Fig 6.2 Example of a fragment of variation graph. Colored paths represent genomes which traverse sequences (nodes). This visualization was rendered using the SequenceTubeMap (https://github.com/vgteam/sequenceTubeMap).

6.8.2 Sequence comparison methods

Reconstructing phylogenies of DNA and amino acid sequences is of fundamental importance in biological research, particularly in molecular biology and genomics. It is the key step in molecular evolutionary analyses, e.g., gene function and regulatory region prediction. Traditionally, phylogenetic and phylogeographic inferences are based on nucleotide substitutions. Indels (insertion and deletion) and structural variations in molecular sequence alignments are removed prior to the analysis by coding them as missing data or by deleting whole columns that contain gaps. Moreover, mutation rates and patterns of indels and structural variations are less well studied than those associated with single nucleotide or amino acid substitutions. With the increasing number of whole genome sequences available, the rate of indel and structural variation events observed in whole genome comparisons is too high to ignore, especially for across species comparison. Several methods have been developed for including indel events. Thorne et al (1991) developed the first probabilistic model (the TKF91 model), describing the evolution of indels on a phylogenetic tree as a Poisson process (Bouchard-Cŏté and Jordan, 2013). The TKF91 model was further extended to allow longer indels (the TKF92 model, Thorne et al., 1992). The TKF91 and TKF92 models can be represented as Hidden Markov Models (HMM, Thorne et al., 1992). Recently developed pair-HMMs can closely approximate the features of the indel model in various length, assuming geometrically distributed indel lengths, and without assuming reversibility (De Maio, 2020). Lastly, a wide array of alignment-free approaches to perform sequence comparison have been developed. Currently, the most widely used alignment-free algorithms are based on k-mer counts (Reinert et al., 2009; Schwende and Pham, 2014). K-mer methods work by projecting sequences into a feature space of k-mer counts, which is further transformed into numerical values (e.g., k-mer frequencies) that can be used to calculate sequence distances. Alignment-free algorithms are rapidly extending the range of applications and answering previously intractable questions in phylogenomics (Fan et al., 2015), horizontal gene transfer (Bernard et al., 2016), population genetics (Haubold et al., 2013) and evolution of regulatory sequences (Zhao et al., 2016). It is conceivable that molecular evolution studies in the future will abandon alignment altogether and uncover the evolution trajectory for every single base.

References

Akagi, K., Li, J., and Symer, D. E. (2013). How do mammalian transposons induce genetic variation? A conceptual framework: The age, structure, allele frequency, and genome context of transposable elements may define their wide-ranging biological impacts. *BioEssays* 35, 397–407. doi:10.1002/bies.201200133.

Allen, R., Ryan, H., Davis, B. W., King, C., Frantz, L., Irving-Pease, E., et al. (2020). A mitochondrial genetic divergence proxy predicts the reproductive compatibility of mammalian hybrids. *Proc. R. Soc. B Biol. Sci.* 287, 20200690. doi:10.1098/rspb.2020.0690.

Bernard, G., Chan, C. X., and Ragan, M. A. (2016). Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci. Rep.* 6, 1–12. doi:10.1038/srep28970.

Bijlsma, R., and Loeschcke, V. (2012). Genetic erosion impedes adaptive responses to stressful environments. *Evol. Appl.* 5, 117–129. doi:10.1111/j.1752-4571.2011.00214.x.

Bouchard-Cŏté, A., and Jordan, M. I. (2013). Evolutionary inference via the Poisson Indel Process. *Proc. Natl. Acad. Sci. U. S. A.* 110, 1160–1166. doi:10.1073/pnas.1220450110.

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 Genes, Genomes, Genet.* 5, 931–941. doi:10.1534/g3.114.015784.

Bridle, J. R., and Vines, T. H. (2007). Limits to evolution at range margins: when and why does adaptation fail? *Trends Ecol. Evol.* 22, 140–147. doi:10.1016/j.tree.2006.11.002.

Buchkovich, M. L., Eklund, K., Duan, Q., Li, Y., Mohlke, K. L., and Furey, T. S. (2015). Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. *BMC Med. Genomics* 8, 1–15. doi:10.1186/s12920-015-0117-x.

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P. M., et al. (2018). 10KP: A phylodiverse genome sequencing plan. *Gigascience* 7, 1–9. doi:10.1093/gigascience/giy013.

De Maio, N. (2020). The Cumulative Indel Model: Fast and Accurate Statistical Evolutionary Alignment. *Syst. Biol.* 0, 1–22. doi:10.1093/sysbio/syaa050.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., et al. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25, 3207–3212. doi:10.1093/bioinformatics/btp579.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28, 2239–2252. doi:10.1093/molbev/msr048.

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., et al. (2020). Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* 21, 139–162. doi:10.1146/annurev-genom-120219-080406.

Elena, S. F., and Lenski, R. E. (2003). Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* 4, 457–469. doi:10.1038/nrg1088.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet.* 9, 1003905. doi:10.1371/journal.pgen.1003905.

Fan, H., Ives, A. R., Surget-Groba, Y., and Cannon, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* 16, 522. doi:10.1186/s12864-015-1647-5.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2020). A bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* 37, 1211–1223. doi:10.1093/molbev/msz296.

Frankham, R., Ballou, S. E. J. D., Briscoe, D. A., and Ballou, J. D. (2002). *Introduction to conservation genetics*. Cambridge university press.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* 36, 875–881. doi:10.1038/nbt.4227.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi:10.1126/science.1188021.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* 43, 1031–1035.

doi:10.1038/ng.937.
Grossniklaus, U., Kelly, B., Ferguson-Smith, A. C., Pembrey, M., and Lindquist, S. (2013). Transgenerational epigenetic inheritance: How important is it? *Nat. Rev. Genet.* 14, 228–235. doi:10.1038/nrg3435.
Groves, C. P. (1980). Notes on the systematics of Babyrousa (Artiodactyla, Suidae). *Zool. Meded.* 55, 29–46.
Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genet.* 15, e1008302. doi:10.1371/journal.pgen.1008302.
Hamilton, J. A., and Miller, J. M. (2016). Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conserv. Biol.* 30, 33–41. doi:10.1111/cobi.12574.
Haubold, B., Krause, L., Horn, T., Pfaffelhuber, P., and Hancock, J. (2013). An alignment-free test for recombination. *Bioinformatics* 29, 3121–3127. doi:10.1093/bioinformatics/btt550.
Henikoff, S., and Greally, J. M. (2016). Epigenetics, cellular memory and gene regulation. *Curr. Biol.* 26, R644–R648. doi:10.1016/j.cub.2016.06.011.
Hewitt, G. M. (2011). Quaternary phylogeography: the roots of hybrid zones. *Genetica* 139, 617–638. doi:10.1007/s10709-011-9547-3.
Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. doi:10.1038/nature13408.
Huxel, G. R. (1999). Rapid displacement of native species by invasive species: Effects of hybridization. *Biol. Conserv.* 89, 143–152. doi:10.1016/S0006-3207(98)00153-0.
Jablonka, E., Lachmann, M., and Lamb, M. J. (1992). Evidence, mechanisms and models for the inheritance of acquired characters. *J. Theor. Biol.* 158, 245–268. doi:10.1016/S0022-5193(05)80722-2.
Kamm, J. A., Terhorst, J., Durbin, R., and Song, Y. S. (2018). Efficiently inferring the demographic history of many populations with allele count data. *bioRxiv*, 287268. doi:10.1101/287268.
Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., and Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evol. Appl.* 9, 1205–1218. doi:10.1111/eva.12414.
Keeney, S., Lange, J., and Mohibullah, N. (2014). Self-Organization of Meiotic Recombination Initiation: General Principles and Molecular Pathways. *Annu. Rev. Genet.* 48, 187–214. doi:10.1146/annurev-genet-120213-092304.
Kleinjan, D. A., and Van Heyningen, V. (2005). Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* 76, 8–32. doi:10.1086/426833.
Leppälä, K., Nielsen, S. V., and Mailund, T. (2017). Admixturegraph: An R package for admixture graph manipulation and fitting. *Bioinformatics* 33, 1738–1740. doi:10.1093/bioinformatics/btx048.
Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333. doi:10.1073/pnas.1720115115.
Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* 326, 289–293. doi:10.1126/science.1181369.
Lin, Y. L., Pavlidis, P., Karakoc, E., Ajay, J., and Gokcumen, O. (2015). The evolution and functional impact of human deletion variants shared with archaic hominin genomes. *Mol. Biol. Evol.* 32, 1008–1019. doi:10.1093/molbev/msu405.
Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., et al. (2020). The structure, function, and evolution of a complete human chromosome 8. *bioRxiv*, 2020.09.08.285395. doi:10.1101/2020.09.08.285395.
Lucchini, V., Meijaard, E., Diong, C. H., Groves, C. P., and Randi, E. (2005). New phylogenetic perspectives among species of South-east Asian wild pig (Sus sp.) based on mtDNA sequences and morphometric data. *J. Zool.* 266, 25–35. doi:10.1017/S0952836905006588.
Maddison, W. P. (1997). Gene trees in species trees. Maddison doi:10.1093/sysbio/46.3.523.
Mao, X., He, G., Hua, P., Jones, G., Zhang, S., and Rossiter, S. J. (2013). Historical introgression and the persistence of ghost alleles in the intermediate horseshoe bat (*Rhinolophus affinis*). *Mol. Ecol.* 22, 1035–1050. doi:10.1111/mec.12154.

Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., et al. (2018). Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.* 19, 118–135. doi:10.1093/bib/bbw089.

Martin, S. H., and Amos, W. (2020). Signatures of introgression across the allele frequency spectrum. *bioRxiv*, 1–23. doi:10.1101/2020.07.06.189043.

Meijaard, E., and Groves, C. (2002). Upgrading three subspecies of babirusa (Babyrousa sp.) to full species level. *Asian Wild Pig News* 2, 33–39.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585, 79–84. doi:10.1038/s41586-020-2547-7.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52. doi:10.1093/bioinformatics/btv234.

Musilova, P., Kubickova, S., Hornak, M., Cernohorska, H., Vahala, J., and Rubes, J. (2010). Different Fusion Configurations of Evolutionarily Conserved Segments in Karyotypes of <i>Potamochoerus porcus</i> and <i>Phacochoerus africanus</i> *Cytogenet. Genome Res.* 129, 305–309. doi:10.1159/000314954.

Orr, H. A., and Unckless, R. L. (2008). Population extinction and the genetics of adaptation. *Am. Nat.* 172, 160–169. doi:10.1086/589460.

Oulu (2003). *Gene Genealogies, Variation and Evolution A primer in coalescent theory*. Oxford University Press, USA.

Patel, L., Kang, R., Rosenberg, S. C., Qiu, Y., Raviram, R., Chee, S., et al. (2019). Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat. Struct. Mol. Biol.* 26, 164–174. doi:10.1038/s41594-019-0187-0.

Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res.* 27, 665–676. doi:10.1101/gr.214155.116.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093. doi:10.1534/genetics.112.145037.

Pease, J. B., and Hahn, M. W. (2013). More Accurate Phylogenies Inferred From Low-Recombination Regions In The Presence Of Incomplete Lineage Sorting. *Evolution (N. Y).* 67, 2376–2384. doi:10.1111/evo.12118.

Rakocevic, G., Semenyuk, V., Lee, W. P., Spencer, J., Browning, J., Johnson, I. J., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* 51, 354–362. doi:10.1038/s41588-018-0316-4.

Rakotoarivelo, A. R., O'donoghue, P., Bruford, M. W., and Moodley, Y. (2019). An ancient hybridization event reconciles mito-nuclear discordance among spiral-horned antelopes. *J. Mammal.* 100, 1144–1155. doi:10.1093/jmammal/gyz089.

Rando, O. J., and Verstrepen, K. J. (2007). Timescales of Genetic and Epigenetic Inheritance. *Cell* 128, 655–668. doi:10.1016/j.cell.2007.01.023.

Reinert, G., Chew, D., Sun, F., and Waterman, M. S. (2009). Alignment-free sequence comparison (I): Statistics and power. *J. Comput. Biol.* 16, 1615–1634. doi:10.1089/cmb.2009.0198.

Rhie, A., Mccarthy +2, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., et al. (2020). Towards complete and error-free genome assemblies of all vertebrate species. *Constantina Theofanopoulou* 52, 65. doi:10.1101/2020.05.22.110833.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., and Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Mol. Ecol.* 25, 2387–2397. doi:10.1111/mec.13610.

Sarver, B. A. J., Keeble, S., Cosart, T., Tucker, P. K., Dean, M. D., and Good, J. M. (2017). Phylogenomic insights into mouse evolution using a pseudoreference approach. *Genome Biol. Evol.* 9, 726–739. doi:10.1093/gbe/evx034.

Schwende, I., and Pham, T. D. (2014). Pattern recognition and probabilistic measures in alignment-free sequence analysis. *Brief. Bioinform.* 15, 354–368. doi:10.1093/bib/bbt070.

Slotkin, R. K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285. doi:10.1038/nrg2072.

Stelkens, R. B., Brockhurst, M. A., Hurst, G. D. D., and Greig, D. (2014). Hybridization facilitates evolutionary rescue. *Evol. Appl.* 7, 1209–1217. doi:10.1111/eva.12214.

Stern, D. L., and Orgogozo, V. (2008). The loci of evolution: How predictable is genetic evolution? *Evolution (N. Y).* 62, 2155–2177. doi:10.1111/j.1558-5646.2008.00450.x.

Stevenson, K. R., Coolon, J. D., and Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* 14, 536. doi:10.1186/1471-2164-14-536.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33, 114–124. doi:10.1007/BF02193625.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.* 34, 3–16. doi:10.1007/BF00163848.

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* 9, 1–14. doi:10.1093/gigascience/giaa051.

Weber, F. (2008). The coalescent model.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216. doi:10.1038/nrg2063.

Zhang, D., Tang, L., Cheng, Y., Hao, Y., Xiong, Y., Song, G., et al. (2019). "ghost Introgression" As a Cause of Deep Mitochondrial Divergence in a Bird Species Complex. *Mol. Biol. Evol.* 36, 2375–2386. doi:10.1093/molbev/msz170.

Zhao, J., Song, X., and Wang, K. (2016). LncScore: Alignment-free identification of long noncoding RNA from assembled novel transcripts. *Sci. Rep.* 6, 1–12. doi:10.1038/srep34838.

Zhao, T., and Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2165–2174. doi:10.1073/pnas.1801757116.

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311. doi:10.1038/nbt.3432.

Summary

Summary

Understanding the origin of species and biodiversity is one of the fundamental objectives in evolutionary biology. Continuous efforts to sequence additional species have generated a comprehensive dataset for biologists to explore the underlying biological mechanisms contributing to evolution and to aid conservation of ecosystems. The aim of my thesis is to utilize genome scale data to provide new insight into the evolutionary process and mechanisms in *Suidae* evolution. In this work, I provided a comprehensive view of the evolutionary history of pig species, spanning from species origin to current time. Furthermore, the refined comparative genomic framework of *Suidae* species contributes to our understanding of the effects of complex speciation and hybridization from an evolution genomics perspective.

In **Chapter 2** we sequenced and analyzed the genomes of the highly endangered pygmy hog (*Porcula salvania*). Phylogenetic reconstructions using whole-genome data strongly support pygmy hog as a distinct genus separated from other *Suinae* species, whereby the controversial taxonomic classification is resolved. Time of divergence estimation suggested that pygmy hog initially emerged during the Miocene/Pliocene boundary some 5.1 million years ago. Moreover, admixture analyses revealed at least two independent events of inter-species gene flow during wild boar range expansion across Eurasia. Despite the large phylogenetic divergence, wild boar interbred with pygmy hog and a now-extinct species which exhibit a deep phylogenetic placement between pygmy hog and warthog. Our analyses highlight the important role of admixture as evolutionary biological driving force in successful range expansion and species replacement.

In **Chapter 3** we provided an in-depth analysis of the formation of the current pygmy hog population and demonstrated consequences of the historical demography. Using whole genome sequencing data of six individual pygmy hogs, our demographic analysis revealed that pygmy hog has remained at small population sizes with low genetic diversity since ~1 Mya. The absence of mitochondrial variation in the six sequenced individuals suggested a historical maternal bottleneck. Runs of homozygosity (ROH) analysis showed that pygmy hog does not comprise high ROH coverage and long ROH, indicating very little recent inbreeding. Also, our genome wide scan and simulation of harmful mutations suggested that the long-term, extremely small population size may have constrained the purifying effect and led to the accumulation of genetic load.

In **Chapter 4** we generated a *Sus cebifrons* (the Visayan warty pig) genome assembly using linked-read sequencing and chromosome conformation capture techniques. The assembly is on chromosome level, which consisting of 17

chromosomes and yielding a genome size of 2.48 Gb and scaffold N50 of 141.8 Mb. The alignment of the *Sus cebifrons* assembly and *Sus scrofa* assembly (Duroc - Sscrofa11.1) revealed a high degree of collinearity, but also chromosome fission and fusion. The comparison of chromosome interaction maps suggested that, due to the short divergence time between Visayan warty pig and Duroc pig (~3 Mya), the chromosome 3D conformation structure remained the same after the chromosomal rearrangements. We hypothesized that this may explain the absence of post-zygotic reproductive isolation among *Sus* species. We further identified the different signatures of adaptive and domestication selection in Visayan warty pig and domestic pig, respectively. In particular, we investigated the evolution of olfactory and gustatory genes and reported the genetic basis of species-specific sensation.

In **Chapter 5** we went one step further and used a reference-guided assembly approach to generate genome sequences for three other *Sus* species (i.e., *Sus verrucosus*, *Sus celebensis* and *Sus barbatus*) and the outgroup species *Porcula salvania*. With the near complete phylogenomic framework of *Sus* species, we were able to perform admixture analyses directly from genome sequences. This unbiased admixture analysis reaffirmed the extensive inter-species gene flow between *Sus* is concomitant with past climatic fluctuations. In addition, we tested the shift of selection pressure along the *Sus* phylogeny and interpreted the results in the context of paleobiogeographic evidence. We provided candidate genes that might have contributed to adaptive radiation and domestication of pigs. This case study demonstrates that utilizing genome sequences is a powerful tool for evolutionary and functional genomic analyses.

Finally, in **Chapter 6**, I string all chapters together and provide additional discussion on the findings presented in **Chapter 2** to **5**. I discuss the molecular and genetic mechanism underlying the complex evolutionary history of *Suidae* species. I also discuss the evolution of genome architecture during speciation and hybridization from a general perspective. Lastly, I summarize methodological challenges of evolutionary analysis in the genomic era and explore the potential implication of the emerging methods.

概要

探索物种的起源和生物多样性的产生是进化生物学的核心课题之一。多年来生物学家一直坚持不懈地对新的物种进行基因组测序，并积累了大量数据。这些数据可以用于研究进化的生物学机制和协助生态系统保护工作。我的博士课题的研究目的是使用基因组数据来探究猪科动物的进化及其机制。在我的博士论文中，我详细描绘了猪科动物从其起源年代直至现今的进化历史。此外，我构建的高精度比较基因组框架，让我们能够从进化基因组学的角度理解复杂物种分化和物种杂交的生物学影响。

在第二章中，我们对高度濒危物种姬猪的基因组进行测序和分析。利用全基因组数据重建的系统发育树解决的此前存在的关于生物学分类的争议，证明了姬猪应该独立于猪属动物之外，被单独分类为姬猪属。分子钟分析判断姬猪最初出现在上新世和更新世交界时期，距今约 5 百万年前。此外，通过比较不同猪科动物的基因组，我们发现野猪在欧亚大陆上迅速扩张时曾发生过至少两次独立的物种间基因交流。尽管野猪和姬猪之间存在巨大的遗传距离，这两个物种曾经发生过杂交。同时野猪也和一个进化地位介于姬猪和非洲疣猪之间的已灭绝的物种发生杂交。从所得的结果我们推测，跨物种的基因渗入可能是成功的群体扩张和物种替代的关键条件。

在第三章中，我们对于现生姬猪群体的形成和历史群体变化的结果进行了深度地分析。利用六头姬猪的全基因组测序数据，我们发现姬猪从距今约 1 百万年起一直保持着极低的群体大小和遗传多样性。历史上曾经发生的母系瓶颈效应可能导致了现生群体线粒体基因多样性的缺失。基因组近交程度分析发现姬猪并没有在近期内发生过近交。另外，我们对有害遗传突变进行的基因组扫描和模拟分析，结果显示长期的极端小群体导致对有害突变的纯化选择的强度降低，遗传负荷得以累积。

在第四章中，我们运用 10X 和 Hi-C 技术对卷毛野猪的基因组进行从头拼装。最终我们的到了其染色体水平的参考基因组，包括 17 个染色体，基因组大小为 2.48Gb。卷毛野猪基因组和家猪基因组之间存在很高程度的共线性，同时也存在染色体融合和分裂的现象。尽管这两个物种已经分化了约 3 百万年，它们之间的染色质互作情况仍然相似。染色体的三维结构在染色体变异之后仍保持一致。我们推测这可能是猪属动物之间缺少生殖隔离的原因。我们还分别分析了这两个物种基因组中环境适应和驯化选择的信号。我们特别研究了嗅觉和味觉的进化，并报道了物种特异性感官进化的遗传基础。

在第五章中，我们利用已有的遗传资源，使用有参拼装的方法，对另外三个猪属动物和外群物种姬猪的基因组进行从头组装。由此我们构建了几乎完整的猪属动物系统发育框架。我们通过直接比较不同物种的基因组序列，得到了精确

的物种间基因交流信息。猪属动物间频繁的基因交流与历史气候波动是一致的。另外，我们检测了猪属动物中选择压力的变化，并联合古生物地理学信息进行解释。我们报道了与家猪适应性扩张和驯化相关的候选基因。本研究突出了参考基因组序列对于进化和功能基因组研究的重要推动作用。
最后，在第六章中，我结合前文所有章节中的结果进行了综合的讨论。我对猪科动物复杂进化历史的分子和遗传机制进行了讨论。我讨论了物种分化和物种杂交过程中基因组结构的进化。最后，我总结了基因组时代下进化分析中的方法学问题，并对前沿新兴方法的潜在应用进行探讨。

Acknowledgements

This thesis could not have been written without the support of all the people in the last four years. First, I would like to thank my promoter Martien. There are not enough words to explain how grateful I am to get an opportunity to work with you. I am very much indebted to your mentorship, expertise and kindness. Thank you for your trust in me, giving me the freedom to explore research topics, especially granting my cost on HPC. You have boost-started my career and truly developed my self-confidence as an evolutionary biologist. I look forward to future collaborations and social interactions to be had.

Ole, Ole, Ole, Ole, how lucky I am to have you as my supervisor. I have enjoyed every moment of working with you. Those discussions on our weekly meetings which we cannot convince each other will be the most memorable thing in these four years. Thank you for your open heart and selfless attitude to teach me everything, except your Danish cake recipe :P. Your critical scientific vision and your personal philosophy are inspirations for me not only as a scientist but also as a human being.

I would like to thank Mirte and Hendrik-jan. Although you two are not in my supervision team, I cannot have come this far without your support. Mirte, thank you for your contributions to my PhD project. Your extensive and broad knowledge have improved this work greatly. Thanks to Hendrik-jan for your directness and your constructive suggestion. I am grateful for the "crime" you committed, encouraging me to aim high and criticizing my unrealistic ideas. It has been an exciting and awesome learning experience.

Thanks to Laurent for not only the insightful suggestion and ideas from you, but also the codes and scripts you left on the HPC and Deepphylo, which led me to the systematic studies of programming and bioinformatics. I also take the opportunity to thank all other co-authors Evan, Qitong, and Manon for your contributions to my projects.

Thank you, Richard for your endless support, scientifically and mentally. You are such a helpful person all the time. Gwen, I am profoundly grateful for your impeccable administration of the HPC. Thank you for bearing with me as an annoying user liu194. You are the unsigned co-author for every work of mine.

I would like to thank my dear paranymphs. Lim, Younglim xi, thank your support and your friendship during these four years. Thanks for all these discussions about work, about life and about future. Wish you all the best in your future career, believe in yourself, you can make it. 老邢，非常感谢博士期间能交到你这个朋友。酒逢知己千杯少，感谢你的阅历，让你能理解我幼稚的喜怒哀乐。前程未卜，但看前面黑黑洞洞，定是那贼人巢穴，待俺赶上前去，杀它个干干净净。

Thanks to my colleagues and friends in the Genomics group. Martijn, thanks for your help and support during this work. Your expertise in genomics, bioinformatics and dropshots has been invaluable. Henri, wish our non-competition never reach the end of the game. Thanks to Vinicius, Maulik, Rayner, Jani, Fatma, Chrissy, Gibbs ... Also thanks to Lisette and Fadma for all your secretarial work. Thanks to many PhDs and staffs at ABG, sorry I cannot mention every name. Having all of you around made my PhD an enjoyable and smooth journey.

Thank you Harmen, although after four years ik spreek geen Nederlands, you made the Netherlands feels like home. Thank you for all these talks, walks, trips, dinners, boardgames... gosh, it is difficult for me to recall a joyful moment without you. Thank you for this heady friendshiskey :P. Chiara, our discussion about science, politics and culture really sharpened my mind and provided me a different perspective to understand the Western world. I admire your passion for science and good food (not only Italian). Wish you and Miguel all the best, and look forward to our future collaborations. I am grateful to Maria (and Ton), Robert, Raymond, Lisanne, and all other officemates for making E.0201 the best place to work.

当然还要感谢瓦村各位中国同胞，社会包姐、喻运、小飞、老司机、海博、曹露、袁桃林、Mandy、王卓识、彭业博。人在异乡，是你们让“四年寒窗苦”变得充满生机，充满欢乐。

感谢国家留学基金基金委（CSC），资助我完成博士学业。感谢我的祖国，国家的发展和个人的进步让爱国不再只是一句空谈。

感谢四姑妈、王叔叔和安安妹妹，谢谢你们这些年来给我的关爱和照顾。感谢王大凡哥为我设计论文封面。虽然我离家千里，但却没有远离爱我的家人们。

吴舟，你见证了我写下这本论文的每一个字，见证了这些年我迈出的每一步。感谢你陪伴我度过了这些年所有的高潮低谷快乐忧愁。你不只是伴侣，还是同事，更是最好的朋友。谢谢你陪伴我一同成长，希望在接下来的人生阶段里也能携手同行。谢谢你 color my life with you。

最后，我要特别感谢我的爸爸妈妈，感谢你们在我没有申请到奖学金的时候毫不犹豫资助我出国攻读博士。没有你们的理解，我不可能开始博士学业。没有你们的鼓励，我不可能完成博士学业。爸爸妈妈，谢谢你们在我的成长道路中倾注了所有的爱，谢谢你们无条件地支持我追求自己的理想。儿子无所为报，在此，将我完成这本论文的喜悦与自豪，与你们分享。

刘琅青

2021 年 1 月 11 日于 Wageningen

Curriculum Vitae

Langqing Liu was born on February 25th 1992 in Chengdu, China. He has developed passion for biology since junior high. In 2010 Langqing studied Animal science in China Agricultural University. He obtained his bachelor degree in 2014 July with thesis entitled "Using Drosophila Melanogaster conplastic strain model to verify the exnuclear gene effects on growth traits" under the supervision of Prof. Xingbo Zhao. Langqing continue his master study in China Agricultural University under the supervision of Prof. Xingbo Zhao. During the course of his master degree, he investigated the domestication of pig using ancient DNA. He also learned evolutionary genetics and molecular biology. In 2016 July he obtained his master degree with thesis entitled "Exploring the origin and domestication of pigs in China by ancient DNA information". In 2016 October he joined Prof. Martien Groenen's group at Animal Breeding and Genomics in Wageningen to continue PhD study and result in this thesis entitled "Genome evolution of Suidae: a looking glass of speciation and hybridization".

Training and Supervision Plan (TSP)

Graduate School WIAS

A. The Basic Package (3 ECTS)	
WIAS Introduction Day **(mandatory)**	2016
Course on philosophy of science and/or ethics **(mandatory)**	2017
Course on essential skills (Frank Little) *(recommended)*	2017

B. Disciplinary Competences (15 ECTS)	
Writing research proposals	2016
Advanced Bioinformatics	2017
Emerging technologies	2017
Characterization,management and exploitation of genomic diversity in animals	2018

C. Professional Competences (2 ECTS)	
Scientific publishing	2018
The Final Touch: Writing the General Introduction and Discussion	2019
Poster and pitching	2020

D. Presentation Skills *(maximum 4 credits)*	
Zoology 2017, oral	2017
WIAS science day,oral	2018
Livestock Genomics IV Meeting 2018, oral	2018
WIAS science day,oral	2019
NLSEB 2019, poster	2019
ISAG 2019,Spain,oral	2019

E. Teaching competences *(max 6 credits)*	
Supervising practicals and excursions-Genomics WUR	2018 P5
Supervising practicals and excursions-Genomics WUR	2018 P2
Supervising practicals and excursions-Genomics WUR	2019 P5
Supervising theses-MSc thesis	2018
Supervising theses-MSc thesis	2019
Supervising theses-MSc thesis	2019

Education and Training Total (minimum 30 credits)*	**30**

This research was financed by Wageningen University. Sequence data used within this research was obtained from the SelSweep project financially supported by a European Research Council grant (ERC-2009-AdG: 249894).

Langqing Liu was sponsored by Lichuan Liu & Xu Zhang and a Chinese Scholarship Council (CSC) fellowship.

Artwork on cover by Yujie Wang. Animal illustrations on cover and Chapter 5 credit to Sheila McCabe, https://faunalfrontier.com .

Printed by: DigiForce | Proefschriftmaken.nl