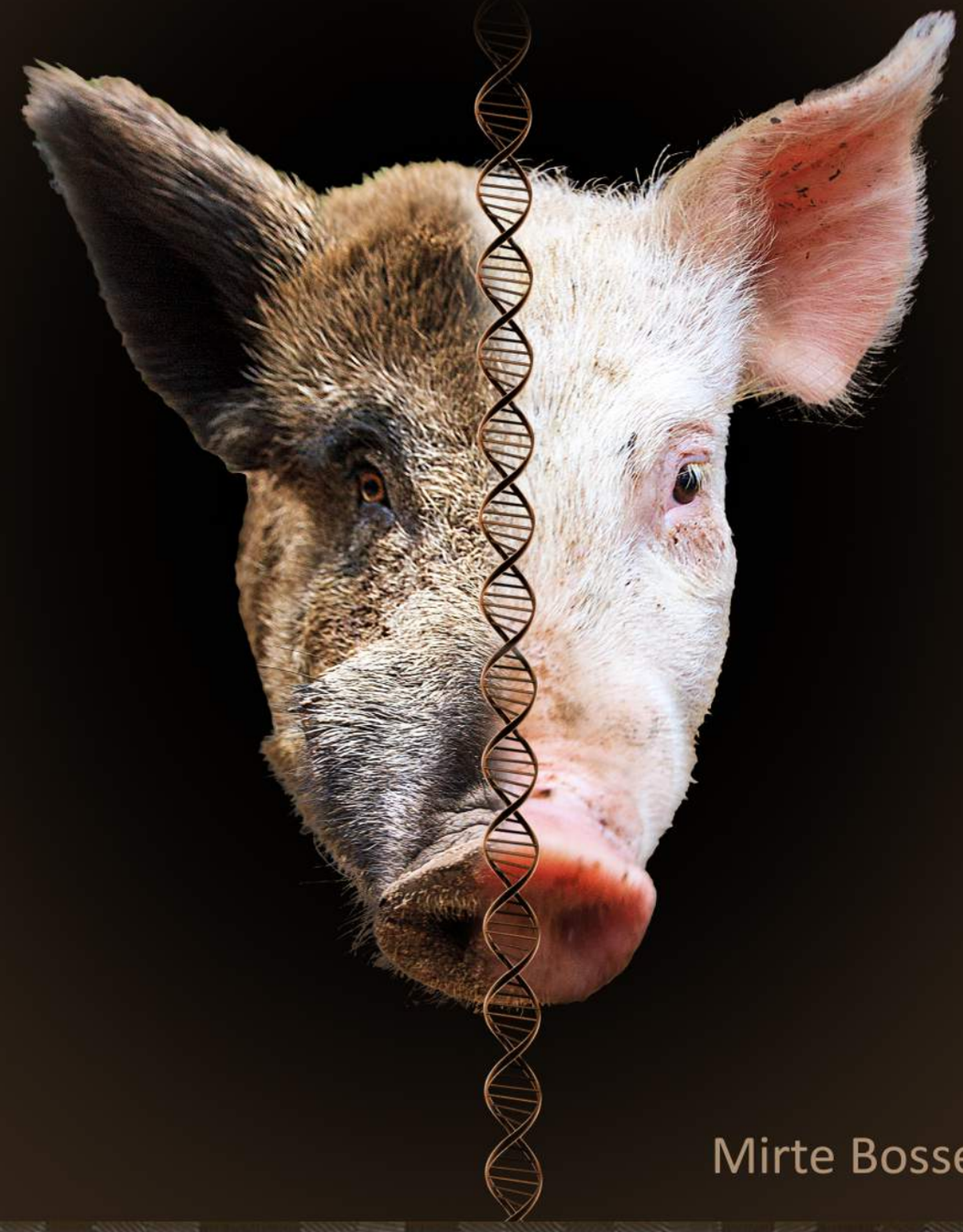
The hybrid nature of pig genomes

Mirte Bosse

The hybrid nature of pig genomes

Unraveling the mosaic haplotype structure in wild and domestic *Sus scrofa* populations



Mirte Bosse

The hybrid nature of pig genomes

Thesis committee

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

Co-promotors
Dr H.-J.W.C. Megens
Assistant professor, Animal Breeding and Genomics Centre
Wageningen University

Dr O. Madsen
Researcher, Animal Breeding and Genomics Centre
Wageningen University

Other members
Prof. Dr B.J. Zwaan, Wageningen University
Prof. Dr H.H.T. Prins, Wageningen University
Prof. Dr J. Ellers, VU University, Amsterdam, The Netherlands
Prof. Dr S.E.F. Guimarães, Universidade Federal de Viçosa, Brazil

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

The hybrid nature of pig genomes

Unraveling the mosaic haplotype structure in wild and commercial *Sus scrofa* populations

Mirte Bosse

Thesis

submitted in fulfillment of the requirements for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus
Prof.Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 22 May, 2015
at 4 p.m. in the Aula.

Bosse, M.
The hybrid nature of pig genomes,
254 pages.

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

ISBN 978-94-6257-300-0

Abstract

Bosse, M. (2015). The hybrid nature of pig genomes. Unraveling the mosaic haplotype structure in wild and commercial *Sus scrofa* populations. PhD thesis, Wageningen University, the Netherlands

A single genome contains information on the demographic history of a population, from ancient bottlenecks till recent inbreeding, hybridization and selection. In this thesis I provide an in-depth analysis of the genome-wide patterns of diversity in domestic pigs, wild boars and closely related species. Pigs originated around 4 million years ago in South-East Asia, and spread over the entire Eurasian continent from there. Because of its wide geographical range and because European and Asian wild boars diverged ~1.2 million years ago, the Eurasian wild boar is an excellent model species to study the effects of demography on genomic variation. Using the latest genomic tools, I show that past glaciations had a strong effect on the effective population size and diversification in many wild boar populations worldwide. I also detect signs of recent inbreeding. Pig domestication occurred independently in Europe and Asia about 10.000 years ago, leading to genetically and phenotypically highly distinct domesticated clades. It is well documented that Asian pigs have been imported into Europe in the early nineteenth century and were crossed with European pigs to improve the performance of local breeds. I demonstrate the genome-wide nature of these introgression and selection patterns in domesticated European pigs. The results presented in chapters 5 and 6 reveal selection on Asian haplotypes in the genome of European commercial pigs and significant effects of the Asian variants on multiple traits of commercial interest. The identified Asian introgressed haplotypes are associated with regions harboring genes involved in meat quality, development and fertility. On top of that, I show that nucleotide diversity in the genome of European domesticated pigs has increased as a result of the introduced Asian haplotypes, supporting the fundamental genetics theory behind outcrossing. Finally, I demonstrate how genomics-based measures of inbreeding can outperform classic pedigree-based breeding programs in maintaining variation and fitness in a population.

Table of Contents

1

General introduction

1.1 Introduction

1.1.1 Genetic variation

The genome is a mosaic of sequences, consisting of interconnected stretches of DNA that each represent a unique history. This history is enclosed in the specific combination of genetic variants in the sequence. Genetic variation provides the raw material for evolution. Some variants in the DNA will have an effect on the phenotype, and therefore also on fitness of an individual. Because variants occur in different combinations in all individuals in a population, fitness is unequally distributed over individuals. Therefore, genetic variation is the foundation for natural selection to act upon a population. While some variants will elevate the probability of survival and producing offspring, others will have a negative effect. To truly comprehend the mechanisms that underlie the evolutionary forces of survival and extinction, we need to understand the factors that shape the distribution of genetic variation in a population. The aim of my thesis is to provide insight into the mechanisms that shape the variation landscape in diploid organisms, and how this information can be used in selection and conservation efforts.

1.1.2 Sources of genetic variation

In diploid organisms, like most animals including mammals, a single copy (or haplotype) of the same chromosome is inherited from each parent. Although these two haplotypes originally represent the same chromosome, they are not completely identical. Two major phenomena in genetics are thought to cause most variation in the genetic code of diploid organisms: **mutation** and **recombination**. Single base substitution mutations result in single nucleotide polymorphisms (SNPs). SNPs are, in numbers, the most common type of genetic variant in the genome, and are relatively easy to study compared to other types of genetic variation like structural variations. The rate at which per-base pair mutations occur in cellular organisms is correlated to genome size (Lynch 2010). The prevailing assumption in the past four decades has been that the majority of mutations can be considered neutral (Kimura, 1968). Initially, only a fraction of the genome was thought to be functional. However, this view was challenged in 2012 by the publication of the ENCyclopedia Of DNA Elements (ENCODE) Project by the ENCODE consortium (Birney 2007, Dunham 2012), suggesting that a large proportion of the genome is actually functional to some extent. Apart from the conventional protein coding genes, the functional elements of the genome consist of non–protein-coding genes, transcriptional regulatory elements, and sequences

that mediate chromosome structure and dynamics in eukaryotes. Mutations can have effects on function, but it is important to realize that variation in functional regions does not necessarily indicate that the variant has any measurable effect. The different variants, or alleles, that can be observed at a SNP locus in the genome, have a specific genomic context. They occur at a particular physical region on one of the chromosomes, and since any stretch of DNA can have multiple such alleles, the alleles that co-occur on the same chromosome are physically (or genetically) linked. Genetic linkage was first suggested in 1905 by William Bateson, who discovered an excess of parental phenotypes and therefore a deviation from predicted Mendelian independent assortment in the F2 in pea plants crosses, hinting at a connection between the inherited traits. This 'connection' that Bateson described is the fundamental mechanism behind the concept of haplotypes: the combination of alleles on the same chromosome that are inherited together. The fact that the founder of the inheritance theory in the late 18th century, Mendel, did not find linkage is probably because the limited number of genes associated with the traits he studied were too far apart or on different chromosomes (Blixt 1975). A higher resolution and/or number of traits was needed in order to actually observe this phenomenon. Recombination is a process that occurs during meiosis between homologous chromosomes. Although recombination generally does not create *de novo* genetic variants that did not already exist in the population, it is the main mechanisms that shuffles genetic material present on the two parental copies of a chromosome, resulting in unique combinations of variation. The extent of genetic linkage, therefore, depends not only on the chromosomal location, but also on the local recombination frequency. With current techniques in genetics, basically all variants on a chromosome can be detected which allows us to study the occurrence and inheritance of combination of alleles in the highest possible detail. As we go back in time, more recombination events have shuffled the combination of alleles on a chromosome, and only the combination of alleles of markers very close together can be found back in the descendants. A central theme in this thesis is how long these fragments of co-inherited alleles are, and what they can tell us about the history of a population.

Each haplotype has a unique history of mutation and recombination that has resulted in the combination of alleles on that haplotype. Whether a particular variant will persist within a population, and at which frequency, therefore, is highly dependent not only on its own selective (dis)advantage, but also on its genomic context. Other factors, in particular demography, contribute to the distribution of variation in genomes as well. In the following sections of my introduction, I will

briefly discuss techniques for measuring genetic variation and the main factors that determine the genomic distribution of variation.

Glossary

SNP – Single nucleotide polymorphism or point mutation, resulting in another nucleotide at one single base in the genome.

Haplotype – The combination of alleles on the same (part of a) chromosome that are inherited together.

Recombination – The process that shuffles the alleles from parental haplotypes of the same chromosome into a new combination of alleles in the offspring.

Genetic variation – The number of loci that are variable between two haplotypes, proportional to the full number of loci that are measured.

Linkage Disequilibrium (LD) – A measure for the extend of genetic linkage, so how often alleles co-occur on the same haplotype in a population.

Coalescence – The convergence of alleles (or haplotypes) in a sample to a single ancestral allele, when tracing them back in time.

ROH – Region Of Homozygosity in the genome of an individual: a stretch of DNA for which both parental haplotypes are identical by descent.

Selective Sweep – A region in the genome that is or has been under selection and displays reduced genetic diversity based on the genotypes of multiple individuals in a population.

Hybridization – The process describing mixture of two genetically distinct populations, resulting in hybrid individuals that contain haplotypes from both gene pools.

Introgression – The introduction of a genetically distinct haplotype from another source into a genome (or population).

1.2 Measuring genetic variation

Although the concept 'genetic variation' is widely used in biology, it can refer to various levels in variation in a population or genome. Therefore, let me first describe what is actually meant by 'genetic variation' in this thesis. Nucleotide diversity, or π (Tajima 1983), can be referred to as the average number of per-site single basepair differences between two randomly sampled haplotypes in a

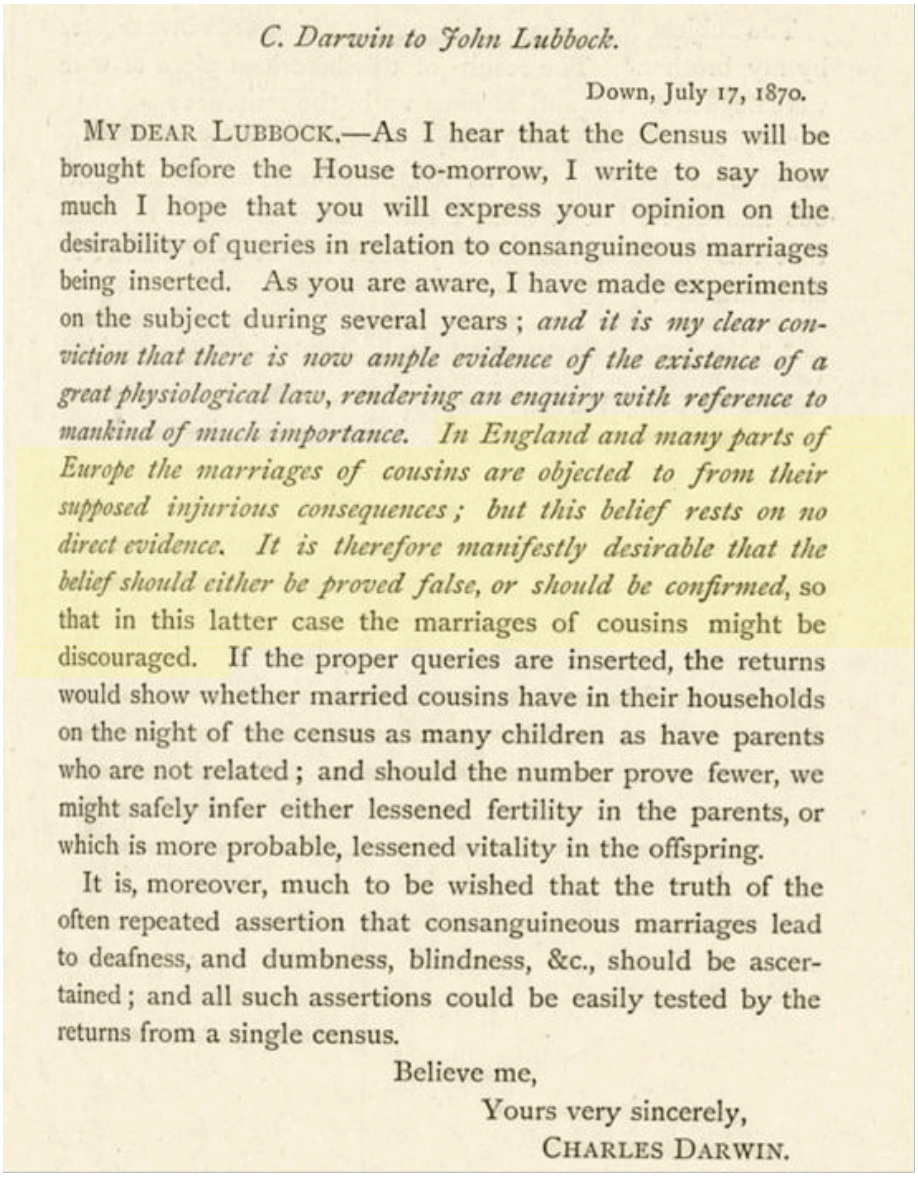
population. It is important to realize that this measure of variation can be used at multiple levels: one can measure variation between different species, but it can also refer to differences between haplotypes within the same individual. In a panmictic population, the genetic variation within a single individual is similar to the variation between two randomly chosen haplotypes in that population. Therefore, the pattern of variation *within a genome* can be used as representative of the variation in the population.

The first DNA-based studies that measured genetic diversity looked at variation at a single locus in the genome. The mitochondrial D-loop has been widely studied, because this eliminates within-individual variation, does not recombine and has a N_e of ¼ of autosomal nuclear markers, making it particularly suitable for phylogeography. As technology in genetics rapidly developed, multiple measurements from the same (individual) genome became feasible. A very popular type of marker that is still widely used is microsatellites, which are short copies of identical repeated sequence that can have many variants (alleles) at the same locus. With the advent of high-density genotype arrays (SNPchips), the resolution of genomic variation became dense enough to look at the interaction of variation, distributed over the length of a chromosome. Linkage disequilibrium (LD) is a measure to describe the co-occurrence of alleles at different marker positions in a population. If no genetic linkage is present, alleles at different markers are inherited independent of each other, and the frequencies of inherited allele combinations are similar thus LD is low. It was in the beginning of this century that the patterns of genetic linkage in the human genome were described in detail (Reich 2001, Altshuler 2005). LD is an important concept for population genetic studies, since it is determined by the number of generations (and, therefore, contains a time factor) and the effective population size. The current era of next-generation sequencing has revolutionized genome research since it allows almost full characterization of genomic variation. In this thesis, I use this genome-wide information to better understand the evolutionary forces shaping variation in genomes. In my opinion, the biggest advantage of genome-wide information over independent loci such as microsatellites is that it provides knowledge on the genomic context of a variant. Most information on demographic history and selection on a population scale lies within the population haplotype structure, e.g. the combination of alleles that are observed together on a haplotype, and the length and abundance of those haplotypes. Moreover, the genomic context of any variant determines the likelihood of persistence of the variant in the population.

1.3 Haplotypes and demography

Demography is one of the major determinants of the level of genetic variation in a population. When population size is large, many individuals contribute to the pool of alleles in the population and many different combinations are created by recombination. The longer this process has been going on, the more time there is for unique haplotypes to be created. Two haplotypes are said to coalesce when they eventually converge into one parent haplotype when they are traced back in time. So coalescent time is the time to the most recent common ancestor for two sampled haplotypes in a population. Multiple methods have been developed to estimate past and current population size based on the distribution of alleles in the population. One of the most widely used methods is the linkage disequilibrium (LD)-based method by Hill (1981) in which the extent of LD in a population, in conjunction with known recombination rate, predicts the effective population size. Although this method is very useful when a limited number of markers are available, implementing full-genome information should lead to the most accurate results. Recently, Li and Durbin (2011) and MacLeod (2013) proposed demographic inference based on the distribution of variation in a single diploid individual (=two copies of a genome). The key assumption here is that the genome is a mosaic of different haplotypes that are derived from different individuals at different time points. The time to the most recent ancestor of both haplotypes is a function of recombination and mutation, and based on the length of the shared segment, and differences between segments, demographic history can be inferred. Recent consanguineous mating will result in long stretches of homozygosity in the genome, since the same haplotype is inherited through both parents. The occurrence of these Regions Of Homozygosity (ROHs) are forced either by demographic processes or by selection elevating the frequencies of haplotypes surrounding a favored allele (Szpiech 2013). Associating genomic ROH regions to particular phenotypes can be utilized in different types of studies on selection, but also disease. Homozygosity mapping in artificially selected lines can pinpoint genomic regions harboring genes of commercial interest, but screening for ROHs in humans that are suffering from a genetic defect can also aid in recessive disease mapping in medical research.

Excessive homozygosity is thought to result in inbreeding depression, which has negative consequences for the fitness of inbred individuals. The key concept here is that, due to loss of heterozygosity, the change that recessive deleterious mutations become homozygous increase significantly. Even during times that genetics was still in its infancy, Charles Darwin already mentioned his concern regarding consan-



C. Darwin to John Lubbock.

Down, July 17, 1870.

MY DEAR LUBBOCK,—As I hear that the Census will be brought before the House to-morrow, I write to say how much I hope that you will express your opinion on the desirability of queries in relation to consanguineous marriages being inserted. As you are aware, I have made experiments on the subject during several years; *and it is my clear conviction that there is now ample evidence of the existence of a great physiological law, rendering an enquiry with reference to mankind of much importance. In England and many parts of Europe the marriages of cousins are objected to from their supposed injurious consequences; but this belief rests on no direct evidence. It is therefore manifestly desirable that the belief should either be proved false, or should be confirmed*, so that in this latter case the marriages of cousins might be discouraged. If the proper queries are inserted, the returns would show whether married cousins have in their households on the night of the census as many children as have parents who are not related; and should the number prove fewer, we might safely infer either lessened fertility in the parents, or which is more probable, lessened vitality in the offspring.

It is, moreover, much to be wished that the truth of the often repeated assertion that consanguineous marriages lead to deafness, and dumbness, blindness, &c., should be ascertained; and all such assertions could be easily tested by the returns from a single census.

Believe me,

Yours very sincerely,

CHARLES DARWIN.

Figure 1.1. Letter by Darwin. Year 1870 letter from Charles Darwin to his neighbor and member of the Census, John Lubbock. Darwin expresses his concern about the level of inbreeding in cousin marriages. Reprint from "Wyhe, John van ed., 2002- The Complete Work of Charles Darwin Online (http://darwin-online.org.uk)".

guineous marriages and the survival chances of offspring resulting from inbred matings. In a letter to his neighbor John Lubbock (Figure1.1) he stresses that more research was needed on the consequences of cousin marriages, especially lessened vitality in offspring. Even though the genetic revolution came too late for Darwin himself, who was actually married to his cousin, a wide range of literature has risen since, discussing the putative damaging effect of inbreeding. At the individual level, inbreeding can cause a reduction in fertility and juvenile survival (Witzenberger and Hochkirch 2011). Darwin actually lost three of his children at a young age, possibly due the deleterious effects of inbreeding. At the population level, inbreeding and loss of genetic diversity reduce the potential for evolutionary change (Allendorf 2010). Together, these effects make small populations highly vulnerable to extinction, as they tend to amplify each other ("extinction vortex"; Frankham 2010). However, the response of a population in terms of negative effects on fitness related traits when inbreeding increases can vary widely between different populations (Hedrick and Kalinowski 2000, Lacy 2012). Even within-population, the reduction of genetic variation after a bottleneck can vary at different loci as a consequence of alternating selective pressure in the genome. Whether inbreeding will have an effect on the population in the long term thus depends on multiple factors, complicating a priori- predictions about how a population will respond to a bottleneck.

1.4 Haplotypes and selection

If one of the variants that are part of a haplotype at a particular locus has an effect on the phenotype of the organism, it is likely that some selective pressure will act upon that variant. Those haplotypes that contain variants that elevate fitness will rise in frequency due to natural selection. The length and frequency of these haplotypes depend on a combination of factors. Selection is more effective in large populations with random mating (i.e. Charlesworth 2009). Patterns of neutral genetic diversity, especially in smaller populations, are determined by a combination of genetic drift and the loss of variation due to selection at linked sites. How variation is influenced by selection depends on the extent of linkage disequilibrium at selected loci and the type of mutation that is selected for. Selection on novel mutations and with high LD will result in a stronger reduction of local genetic variation than on standing variation with low LD.

Overall, the frequency and strength of selection will determine genome-wide variation patterns. However, which type of selection predominates and how these

loci are distributed throughout the genome varies. Although this process is extensively studied, many questions remain about how these patterns can vary across species and populations. Coalescence is defined as the convergence of alleles (or haplotypes) in a sample to a single ancestral allele, when tracing them back in time. Selection, therefore, may change coalescence time compared to neutrality since it influences the haplotype frequencies in a population. In some regions of the genome, heterozygosity is preferred over the homozygous state. This so-called balancing selection therefore increases local diversity levels, and coalescence time is increased because more diverse haplotypes are preferred. Alternatively, positive selection for a certain haplotype locally reduces coalescence time, because the genetic diversity or number of mutations between haplotypes at that locus are diminished. The more recent the selection on a particular haplotype is, the longer the associated selective sweep in the genome will be due to limited time for recombination. This phenomenon can have deleterious consequences for fitness of a population because putative damaging alleles are increased in frequency through genetic hitchhiking. Therefore, LD between two loci in the genome should be considered. The method of hitchhiking-mapping, i.e. screening a population for changes in allele frequency due to selection, depends on the density of the marker system and extent of LD near the selected variant. Next-generation sequencing enabled much progress in building the connection between genetic and phenotypic variation, since it provides the highest possible density.

Selective pressure might be very different in strength and direction in natural populations compared to commercial breeds under strong artificial selection, but I want to emphasize that the effects on genetic variation are the same in both cases. Artificial selection is, therefore, 'just' another case of positive selection for a particular allele. If the selective pressure is strong, as can be the case in artificial selection, variation near the selected locus will be reduced. In those instances, the low genetic variation at those loci is actually advantageous from the perspective of the breeder since all individuals will contain the preferred phenotype.

1.5 Hybridization and introgression

The genome of each individual in a panmictic population contains two randomly sampled collections of haplotypes from that population. Therefore, each individual genome can be considered a representative mixture of the genetic variation that is present in this population. The genetic variation in a population accumulated over time and shaped by selection and demography, and is therefore characteristic for

the population. If, however, genetic material from elsewhere is introduced, the uniformity of the genetic background no longer applies. Such introgressed haplotypes result in hybrid individuals that represent different population histories within a single genome. First generation hybrids will still contain a single copy from one particular (source) population, and another copy of the same chromosomes from the donor population. But after this first generation, recombination will create recombinant haplotypes so that the introgression is fragmented and reduced in frequency.

Speciation with the presence of gene-flow was originally thought to be rare and limited to few species, but this view is changing (Hey 2006, Nosil 2008). The increase in molecular techniques to measure fine-scale genomic variation has changed our perspective on admixture drastically. Interspecific hybridization appears to be relatively common in eukaryotic organisms (Schwenk 2008). Hybridization is gene-flow after the split of two lineages, which implies that the separation of the two populations into phylogenetically separate clades needs to be characterized and verified first. Hybridization may, therefore, complicate inferences of speciation, divergence and selection. If hybridization is indeed as common as appears according to recent estimates, species borders are permeable and we might need to change our traditional view on species boundaries. Hybridization, admixture and gene-flow all refer to combining genetic material from distinct lineages into one organism or population, and therefore are used alternately throughout this thesis.

The genomic and biological consequences of hybridization and admixture are largely unknown. Admixture of diverged populations generally leads to an increase in genetic diversity but not necessarily to higher fitness - this strongly depends on the environment and selection pressure. Hybridization can occur naturally, along hybrid zones of overlapping population ranges. Under these circumstances a steep decline in hybridization is often observed from the region of admixture, due to lower fitness of the hybrids. However, an increasing number of examples from wild and captive populations showed that haplotypes from another source indeed can confer a selective advantage. Originally, adaptive variation was thought to be derived from either standing variation in the population or new mutations. The discovery of the relatively new concept of adaptive introgression introduced a third source of adaptive variation (Hedrick 2013). However, the magnitude of adaptive introgression contributing to adaptive variation is largely unknown. The chance of an introgressed haplotype to remain in a population is highly increased if it

contains some selective advantage (Hedrick 2013). In natural populations of house mouse, about 10% of the genome was found to have another origin than the studied population, and in some genomic regions the introgressed haplotypes had risen to a higher frequency than expected under a neutral scenario (Staubach 2012). A well-known example of adaptive introgression is the warfarin resistance in mice described by Song (2011). Even sexual selection can act upon introgressed haplotypes. Mate choice in *Heliconius* butterflies has been shown to have a hybrid origin in the genome (Salazar 2010). It is therefore not surprising that also in human genomes, admixture with differentiated populations has been identified (Hellethal 2014, Hammer 2011, Alves 2012). For example, introgressed haplotypes with Neanderthal ancestry are thought to have driven evolution for lipid catabolism in Europeans (Khrameeva 2014). Humans have also played a role in stimulating hybridization in other species, either unintentional or on purpose. Human-induced hybridization can be a by-product of globalization as some species became widely distributed due to human mobility, but it can also be intentional in for example domesticated species (Kijas 2012, White 2011). In this thesis I disentangle the genome of hybrids at the finest scale possible with state-of-the-art genomics techniques, and draw conclusions about the mechanisms that distributed the introgressed haplotypes in the genomes of the source population.

1.6 Haplotypes in population management

Small populations are more susceptible to drift and associated change in frequency and distribution of recessive deleterious mutations (Lynch 1995). Captive and/or ex-situ populations are particularly sensitive since they are usually small and therefore genetic drift is large. Effective population sizes of ex-situ populations are often critically small (Ne <50, Baker 2007). Ex-situ populations require careful demographic and genetic population management to minimize loss of genetic diversity (Ballou and Lacy 1995). Without proper genetic management, such small populations will rapidly lose genetic diversity due to random drift and inbreeding depression. In livestock populations, apart from the clear selection goals, maintenance of variation to keep the population resilient to inbreeding depression is desired, just as in natural populations. In terms of strategies on how to manage a population, very different fields of population management face similar problems: intense farming systems and conservation efforts both require maintenance of variation and fitness to meet production or conservation goals. ROHs have high potential for estimating genome-wide autozygosity (McQuillan 2008). Using ROH coverage as a measure of the inbreeding, as an alternative for the inbreeding

coefficient estimated from pedigrees, has only recently been recognized by the livestock industry. Currently, only a hand-full of studies has estimated inbreeding through ROH identified with high density SNP chip data. Ferenčaković (2013a, 2013b) correlated inbreeding estimated from ROH (Froh) with the conventional pedigree coefficients (Fped) and showed that the accuracy of the inbreeding coefficient estimated with pedigree data varies. Pryce (2012), Bjelland (2013) and Purfield (2012) concluded that ROHs are a promising tool for management of inbreeding. Recently, de Cara (2013) developed a method for population management using optimal contributions (OC) calculated from ROHs. Simulation studies demonstrated that avoiding inbreeding was more effective with OC calculated from ROHs than from pedigrees.

Such detailed identification of shared identical-by-descent (IBD) haplotypes can also aid other aspects of population management. Maintaining variation can be of importance, even for the breeding industry, because it creates a framework for selection. The fast development of genomic techniques and available resources is a key element in the current genomic selection based breeding technology. Coancestry mapping can also aid in the determination of hybrids in a population, which may be undesirable. Overall, high-resolution genomic tools have proven to be very useful in population management and will likely become omnipresent and indispensable in any form of population management in the near future.

1.7 Domestic animals as model for evolutionary processes

To study haplotype structure, a model organism is desired in which the different factors influencing the variation landscape in diploid organisms are well known. Domesticated species are generally good models to study genomic and phenotypic consequences of demography and selection (Megens and Groenen 2012). Domestication and breed formation result in population bottlenecks, leaving clear traces of drift and inbreeding in the genome. Artificial selection for particular phenotypes creates opportunities for characterization of Mendelian traits. Genotyping and sequencing technologies have opened up many opportunities to reveal the complex history of domestication, admixture and selection in livestock (Bruford and Bradley 2003, Larson 2014).

1.7.1 Suitability of the pig as model species

Apart from a suitable history and documentation, the availability of detailed genetic information is crucial to be able to study genomic alterations due to demography and selection. The pig (*Sus scrofa*, Linnaeus, 1758) was the first

livestock species for which a genome consortium was established with the intention to completely map the genome (Haley 2009, Schook 2005). The design of a 60K SNP chip for pigs in 2009 greatly contributed to the applicability of genomics techniques in pig breeding, and simultaneously increased possibilities for population genomics studies (Ramos 2009). The publication of the pig reference genome in 2012 (Groenen 2012) opened up an even greater window of opportunities to study various aspects of the genetics of the pig, since the highest resolution possible became reality. Together with the evolutionary history of the pig, these provide an unprecedented study system for the questions regarding genome evolution that I address in this thesis. Apart from population-specific occurrences, I recognize four major events in the evolutionary history of pigs that are of importance for the distribution of genetic variation in the pig genome (Figure 1.2).

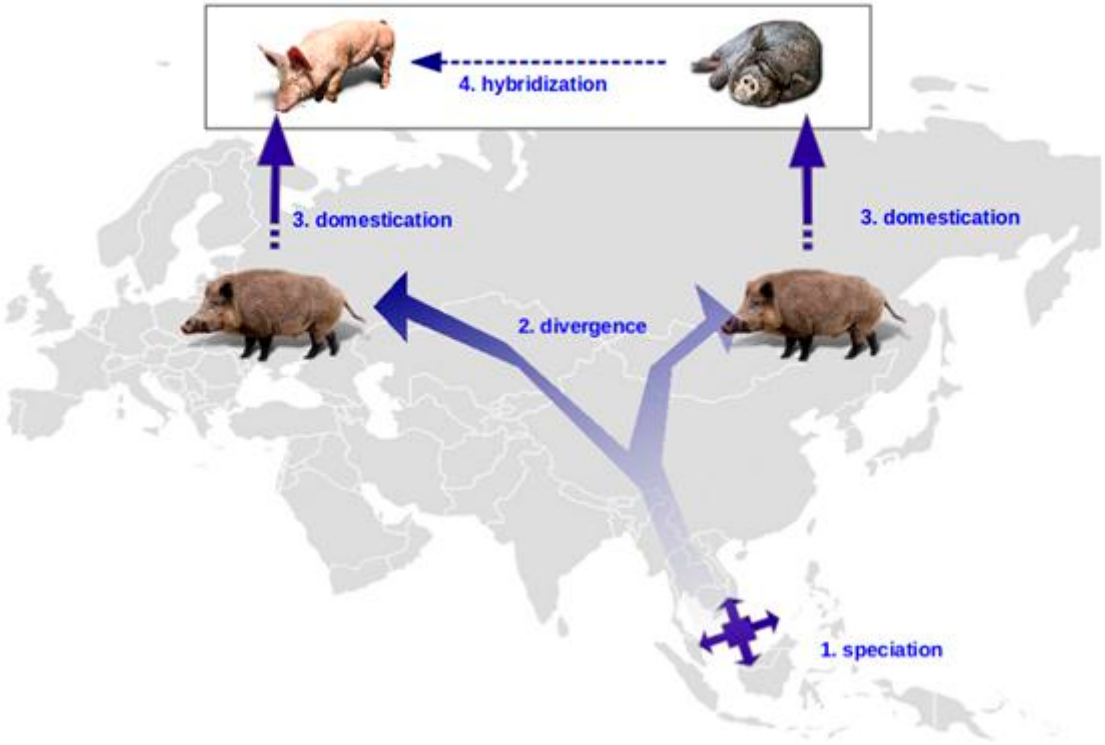


Figure 1.2. **Schematic overview of the history of the pig (*Sus scrofa*).** Four main events are indicated: 1) Speciation of Sus in ISEA. 2) Divergence between European and Asian *S. scrofa*. 3) Independent domestication leading to separate domesticated clades in Europe and Asia. 4) Hybridization between domesticated pigs from Asia and Europe.

1.7.2 Speciation of *Sus* in ISEA

Sus scrofa is widespread within Eurasia (figure 1.2) and consists of many isolated wild and domesticated populations. The Suidae family is particularly interesting for molecular genetic studies as it is one of the few mammalian lineages that have closely related species living today, and multiple species originated roughly ~4 million years ago on Island Southeast Asia (ISEA). The island structure in this region probably promoted speciation, since the bearded pig *Sus barbatus*, the warty pigs *Sus celebensis* and *Sus verrucosus* but also wild *Sus scrofa* occur on separate islands. The complex phylogenetic structure within the genus *Sus* has been studied intensively but until recently remained unclear (Larson 2005, Mona 2007). The past connection of landmasses at the Sunda shelf and isolation of Indonesian islands by the rapid sea level rise after the last glaciation period (Hanebuth 2000) created a dynamic process of (re)colonization, isolation and admixture of different *Sus* species and populations (Frantz 2013b, 2014). The African warthog *Phacochoerus africanus* is the closest living relative outside the genus *Sus*, together with other African genera such as *Potamochoerus* and *Hylochoerus*. The *Sus scrofa* reference genome has proven to be suitable for genomic analysis on all these closely related species.

1.7.3 Divergence between European and Asian *S. scrofa*

The species *Sus scrofa* has its origin in Southeast Asia some ~4 Mya and colonized almost the entire Eurasian mainland from there. The divergence between Western-European and Eastern-Asian populations has been estimated at about 1.2 Mya (Groenen 2012, Frantz 2013b), and has resulted in many fixed differences between both groups. This divergence not only resulted in a European and an Asian *S. scrofa* clade, but also in differences in demographic history and population size. The last glacial maximum probably reduced population size in both geographical regions, but was most severe in Europe (Groenen 2012). The geographic distribution of wild boar over Europe faced another severe decline starting in the middle ages and lasting until the late 18th century. In the mid-19th century, natural or human-mediated recolonization events resulted in isolated populations expanding their range. Some of these isolated populations were small in effective size for decades or longer, causing inbreeding and population differentiation. Local re-stocking of populations with geographically distinct wild boar resulted in complex genetic structures and signatures of population dynamics (Goedbloed 2013a,b). Such complex genetic architectures have been detected in Italian and Luxembourg wild

boar. However, these mixed genomes could have been shaped due to ancient glaciation events (Scandura 2008) or because of recent mixture (Frantz 2013a).

1.7.4 Independent domestication leading to separate clades

The demographic and geographic history of the domesticated pig may be just as complex as that of its wild counterpart. There is compelling evidence that pig domestication events occurred at multiple locations across Europe and Asia independently (Larson 2005, Megens 2008). Domestication has not been a single event, but rather a long period with recurrent admixture with wild populations (Frantz 2015). Even very recently, hybridization between wild and domesticated pigs has been reported (Goedbloed 2013a, 2013b, Frantz 2012b). The domesticated pig as it is used nowadays for agricultural purposes consists of many breeds that have been separated and kept isolated for decades, which has resulted in genetic differences between these breeds. Breeds have been subjected to different selection goals. Breed and population specific genetic studies have greatly enhanced the dissection of complex traits that are economically important. The influence and contribution of commercial pig breeds to local ecology and biodiversity is however debated (Hall & Bradley, 1995, Goedbloed 2013a,b). Knowing and understanding the origin and distribution of variation in (domesticated) species is important for conservation of genetic resources (Groeneveld 2010). In this thesis I will investigate the details of genetic variation in both wild and domesticated populations, and the proportions of shared and unique alleles.

1.7.5 Hybridization between domesticated pigs

It is well documented that during the Industrial Revolution in Europe, European pigs have been deliberately hybridized with Asian pigs. Urbanization in Europe increased the demand for meat such as pork, but during those times pig farmers would still have their pigs roaming in surrounding forests. Forest cover was decreasing and a different pig production system seemed inevitable (White 2011). Due to this changing environment, pig breeders sought a way to improve their stock in such way that pigs had to become adapted to living in small(er) enclosures, be more prolific and gain weight more rapidly. This led to selection for traits better adapted to the changed environment. Many of these traits were already present in Asian domestic pigs. Therefore, British farmers started crossbreeding their own pigs with these Asian pigs (White 2011). This introgression of Asian genetic material into European populations has been demonstrated by genetic markers (Guiffra 2000, Amaral 2011). We expect that many livestock species/breeds actually are a

mixture of highly divergent populations with a mixed demographic history, combined in one genome. Pig breeding and the formation of livestock breeds provide a good example of how Mankind can influence the genome of species.

1.8 This thesis

The main goal of my research is to investigate the distribution of genomic variation within and between different populations of pigs and closely related species. By analyzing re-sequence data from multiple individuals per breed and from wild populations, and some additional individuals from closely related suid species, I identify species- and population-specific regions with particular high or low variation, and relate these patterns to their demographic and selective history. This thesis gives a broad overview of genomic variation in pigs: how it is created, maintained, reduced and increased. This turned out to be a complex interplay of molecular processes, selection, demographic history, gene-flow and human interference. Also the importance of maintenance and possible applications are discussed. In **chapter 2** I evaluate the occurrence of ROH in the porcine genome and assign its non-randomly distributed nature over the genome of several pig species to a combination of demographic events and recombination rate. **Chapter 3** describes the genomic consequences of hybridization between two divergent pig populations from Europe and Asia. In **chapter 4** I demonstrate that the introgression landscape of Asian haplotypes is highly heterogeneous in the commercial Large White breed, and that the Asian introgression at the *AHR* locus increases litter size in this breed. In **chapter 5** I further explore these introgression signals and their association with commercially important traits, and conclude that the deliberate introgression and artificial selection broadly shaped the introgression landscape in commercial pigs. **Chapter 6** discusses the origin of the higher nucleotide diversity in commercial pigs compared to wild boars in Europe, disentangling the different contributions of European wild haplotypes and Asian introgressed haplotypes to the genomic variation in commercial pigs. In **chapter 7** I make use of the latest techniques to follow haplotypes in a breeding scheme and implement this information into selection and conservation management strategies. Finally, in **chapter 8** I discuss the relevance of my findings and place them in a broader context. I reflect on the advantages and shortcomings of using sequence data for these analyses and discuss future perspectives.

2

Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape

Mirte Bosse[1], Hendrik-Jan Megens[1], Ole Madsen[1], Yogesh Paudel[1], Laurent A.F. Frantz[1], Lawrence B. Schook[2], Richard P.M.A. Crooijmans[1], Martien A.M. Groenen[1]

[1] Animal Breeding and Genomics Group, Wageningen University, Wageningen, The Netherlands; [2] Department of Animal Sciences, University of Illinois, Illinois, USA

PLoS genetics (2012) 8(11): e1003100

Abstract

Inbreeding has long been recognized as a primary cause of fitness reduction in both wild and domesticated populations. Consanguineous matings cause inheritance of haplotypes that are identical by descent (IBD) and result in homozygous stretches along the genome of the offspring. Size and position of regions of homozygosity (ROHs) are expected to correlate with genomic features such as GC content, recombination rate but also direction of selection. Thus, ROHs should be non-randomly distributed across the genome. Therefore, demographic history may not fully predict the effects of inbreeding. The porcine genome is has a relatively heterogeneous distribution of recombination rate, making *Sus scrofa* an excellent model to study the influence of both recombination landscape and demography on genomic variation. This study utilizes next generation sequencing data for the analysis of genomic ROH patterns, using a comparative sliding window approach. We present an in-depth study of genomic variation based on three different parameters: nucleotide diversity outside ROHs, the number of ROHs in the genome and the average ROH size. We identified an abundance of ROHs in all genomes of multiple pigs from commercial breeds and wild populations from Eurasia. Size and number of ROHs are in agreement with known demography of the populations, with population bottlenecks highly increasing ROH occurrence. Nucleotide diversity outside ROHs is high in populations derived from a large ancient population, regardless current population size. In addition, we show an unequal genomic ROH distribution, with strong correlations of ROH size and abundance with recombination rate and GC content. Global gene content does not correlate with ROH frequency but some ROH hotspots do contain positive selected genes in commercial lines and wild populations. This study highlights the importance of the influence of demography and recombination on homozygosity in the genome to understand the effects of inbreeding.

Author summary

Small populations have an increased risk of inbreeding depression due to a higher expression of deleterious alleles. This can have major consequences for the viability of these populations. In domesticated species like the pig, that are artificially selected in breeding populations, but also wild populations that experience habitat decline, maintaining genetic diversity is essential. Recent advances in sequence technology enabled us to identify patterns of nucleotide variation in individual genomes. We screened the full genome of wild boars and commercial pigs from Eurasia for regions of homozygosity. We found these regions of homozygosity were caused by the demographic history and effective population size of the pigs. European wild boars

are least variable, but also European breeds contain large homozygous stretches in their genome. Moreover, the likelihood of a region becoming depleted depends on its position in the genome because variation has a high correlation with recombination rate. The telomeric regions are much more variable, and the central region of chromosomes has a higher chance of containing long regions of homozygosity. These findings increase knowledge on the fine-scaled architecture of genomic variation, and are particularly important for population genetic management.

Key words: Regions of Homozygosity, Demography, *Sus scrofa*, Inbreeding, Identity by descent, Recombination

2.1 Introduction

The effects of parental relatedness on the fitness of the offspring has long been recognized. Consanguineous matings cause the inheritance of haplotypes that are Identical By Descent (IBD), resulting in potentially long homozygous stretches across the genome of the offspring. These Regions Of Homozygosity (ROHs) increase the risk of recessive deleterious alleles to be co-expressed, reducing the viability of the organism. In human and canine populations, large homogeneous outbred populations have a lower proportion of genomic autozygosity than small isolated populations (Auton 2009, von Holdt 2011, Ku 2011). In addition, studies have shown a correlation between homozygous stretches in the genome and human diseases (Nalls 2009, Vine 2009, Lencz 2007). One of the long standing interests across various facets of biology is to understand the direct consequences of inbreeding. The inbreeding coefficient F is a commonly used statistic to estimate the degree of same alleles inherited as a consequence of parental relatedness (Wright 1921). However, inbreeding depression may greatly vary across the genome and studies using few molecular marker are unlikely to detect these differences. Thus, it is important to understand the genomic distribution of IBD alleles, to fully grasp the importance of inbreeding for the viability of a given population. The biological characteristics of a species, such as mating systems and reproductive rate, play an important role in maintaining genetic diversity in a population. Moreover, the interactions between standing genetic variation, and past and current demography effect the degree of inbreeding in a population. Homozygosity is used in artificial mate selection to minimize progeny inbreeding (Pryce 2012). Maintenance of the minimum viable population size (MVP) is essential for a population to ensure its persistence in time (Shaffer 1981). This is important for conservation efforts but also in commercial breeding. But, the intrinsic features of the genome that contribute to its architecture, such as recombination rate, are usually neglected in estimations of genetic variation and associated considerations for genetic conservation (Allendorf 2010, Laikre 2010).

In a randomly mating population, IBD tracts are expected to be broken down through time by recombination. In humans, ROH decay is thought to follow an inverse exponential distribution with each generation since the common ancestor halving the ROH size (Keller 2011, Nothnagel 2010). Thus, the size and position of ROHs in the genome are expected to correlate with recombination rate (Macleod 2009). Homozygous stretches should be non-randomly distributed if, as is expected, recombination rate varies throughout the genome and cannot be explained only by past demography. The occurrence of ROHs should rather be an interaction between

demography and the recombination landscape. Pemberton (2012) showed that ROHs may have swept through a population because of positive selection of a particular allele in the region. Moreover, ROHs derived from consanguineous matings may falsely appear to be a signature of positive selection, as these two effects are expected to display depletion of polymorphism in a given genomic region. Therefore, it is important to understand how ROHs segregate across the genome if we are to distinguish signal of selection from inbreeding.

Previous studies that investigated the pattern of ROHs in different mammalian species found that the occurrence of ROH correlates with recombination rate (Vonholdt 2010, Howrigan 2011). However, these studies were based on homozygosity scores from high-density single nucleotide polymorphism (SNP) chips. Recent advances in sequencing technology enable a thorough investigation of genome-wide SNP distributions, and can largely extend the use of high-density SNP arrays for ROH identification. Moreover, re-sequencing strategies should enable a less biased characterization of variation, whereas SNP chips usually suffer from ascertainment bias. In addition, subtle effects of recombination rate can be examined with a full genomic resolution. The porcine genome is known to have a relatively heterogeneous distribution of recombination rate and GC content (Tortereau 2012). Particularly the central parts of chromosomes have a much lower recombination rate than peripheral parts. Although this effect is present in other mammalian genomes, it seems much more pronounced in the porcine genome. In In addition, the species *S. scrofa* (domestic pigs and Eurasian wild boars) is known to have very diverse population structure across its natural and artificial habitat. These characteristics make *Sus scrofa* an excellent model to study the effect of recombination and demography on the distribution of ROHs in mammalian genomes.

The genus *Sus* originated in Southeast Asia during a speciation event in the late Miocene or near the Miocene/Pliocene boundary ~14-4 million years ago (Mya) (Larson 2007a, Mona 2007, Groenen 2012) and the wild boar expanded its range all throughout Eurasia in the Pleistocene ~ 1 to 0.5 Mya (Larson 2005). The European wild boar populations, which are geographically the most distal from the putative origin of the species, are thought to have separated from Southeast Asian *Sus scrofa* in the late Pleistocene between 0.5 and 0.9 Mya (Larson 2007a, Larson 2005, Scandura 2008). The latest glaciation events in Europe created population bottlenecks and subsequent post glacial demographic expansion from refugia in the Iberian Peninsula and the Balkans (Scandura 2008). The genetic diversity of Asian wild boars was probably less affected by the latest glaciation event because a larger area

of suitable habitat would have remained available, although it may have separated Northeastern and South-eastern wild boars (Zachos 2001). The pig has been domesticated at least twice, independently, from local wild boar populations in Asia Minor and China around 8,000 years ago, and there was probably recurrent introgression from the wild species and between breeds since the first domestication event. Because of possible introgression, or even de novo domestication, Near Eastern mitochondrial haplotypes have been completely replaced by European wild boar haplotypes in European commercial pigs (Larson 2007a). Known population histories of *Sus scrofa*, both wild and domesticated, provide a valuable framework for population genomic studies, as conclusions from sequence data can be supported by demography.

This study uses re-sequencing data for the analysis of ROHs and nucleotide diversity (π, (Nei 1979)), to explore how genomic distribution of ROH and π is shaped by additive effects of the recombination landscape, demography and selection. The polymorphism distribution in complete genomes from multiple individual pigs, from different breeds and wild populations from Europe and Asia, are studied in substantial detail. We expect the abundance of ROHs in the genome to be correlated to effective (past and current) population size. The size of ROHs in particular can be expected to correlate to recent and current population size, reflecting founder effects and population bottlenecks. Nucleotide diversity between non-IBD haplotypes, should rather reflect past or ancient population size. In addition, we investigate the influence of recombination rate on the genomic ROH patterns. This highly heterogeneous genomic recombination landscape make pigs and wild boar very well suited for studying the effects of recombination on shaping variation on a genome-wide scale. Furthermore, we investigate the integral effects of demography and recombination on the distribution of ROHs. Finally, we investigate ROH hotspots for traces of positive selection and gene content. Since these different factors are interconnected, the formation and degradation of ROHs is a dynamic genomic process. Overall, we found that a combination of past demographic events and the recombination landscape mostly shaped the pattern of ROHs in the genome.

2.2 Results

2.2.1 General statistics

Regions of homozygosity in the autosomes of individuals were determined by re-sequencing pigs and wild boars of Asian and European origin. We grouped our samples based on geographic origin and domestication status for further analysis

(Figure S2.1; Table S2.1). Pigs were separated into five groups, being Asian domesticated, Asian wild, European domesticated, European wild and other species. Grouping was based on geography and domestication rather than phylogeny. ROHs were identified in all 52 sequenced individuals (examples in Figure S2.2, details in Table S2.1). We found an average number of 778.8 ROHs/genome (+/- 349) with an average size of 1.11 Mbp. ROH size ranged from 10Kbp (minimum size considered) to 83.6 Mbp (29% of the chromosome). Genome-wide nucleotide diversity (π) was on average 1.733 SNPs/Kbp (+/- 0.57) and 2.49 SNPs/bp (+/- 0.57) in the genomic regions outside ROHs (π-out). The difference in π and π-out varied between 1.2 SNPs/Kbp in a European Large White pig, and 0.05 SNPs/Kbp in the *Sus barbatus* individual. The mean number and size of ROHs varied significantly between European and Asian domesticated pigs ($p<0.001$) as well as between wild boars and breeds within continents ($p<0.001$, Figure 2.1C and 2.1D). On average 23% of the genome was considered to be a region of homozygosity. Nucleotide diversity outside ROHs was not significantly different between domesticated pigs and wild boars within Asia, but did vary between the two continents and within the two European groups ($p<0.001$, Figure 2.1B). The most extreme ROH coverage was observed in the Japanese wild boar (78% of its genome).

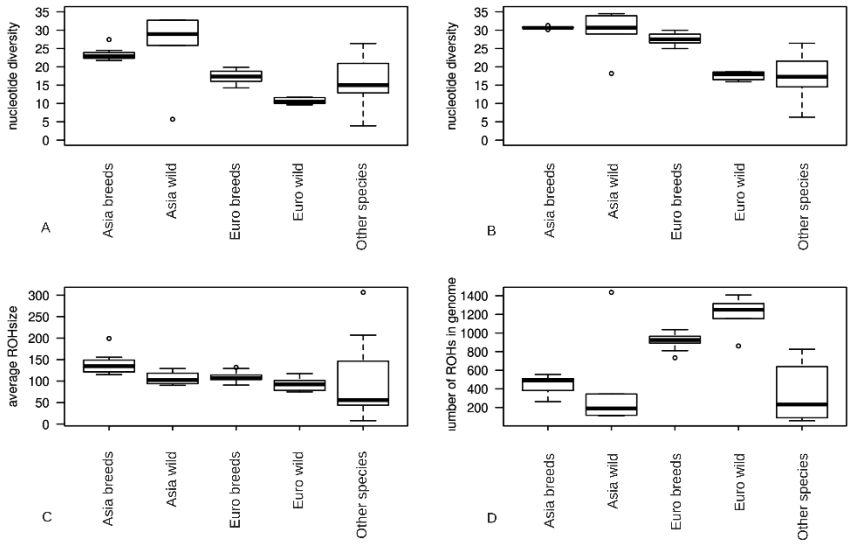


Figure 2.1 Summary statistics for genomic variation. The distributions of ROH statistics for Asian breeds (n = 7), Asian wild boars (n = 5), European breeds (n = 29), European wild boars (n = 6) and other species (n = 5). Groups are divided based on geography (Asians and

Europeans), domestication (pigs and wild boars) and speciation (Other-species include the African Warthog Phacochoerus africanus and other representatives of the *Sus* genus being *Sus barbatus, Sus celebensis, Sus verrucosus* and *Sus cebifrons*). Values are averaged within individuals resulting in a single data point per ROH characteristic for each individual. **A**. nucleotide diversity including ROHs (*10^{-4} bp) **B**. nucleotide diversity excluding ROHs (*10^{-4} bp) **C**. Average ROH size (*10^{-4} bp) **D**. number of ROHs in the genomes of individuals.

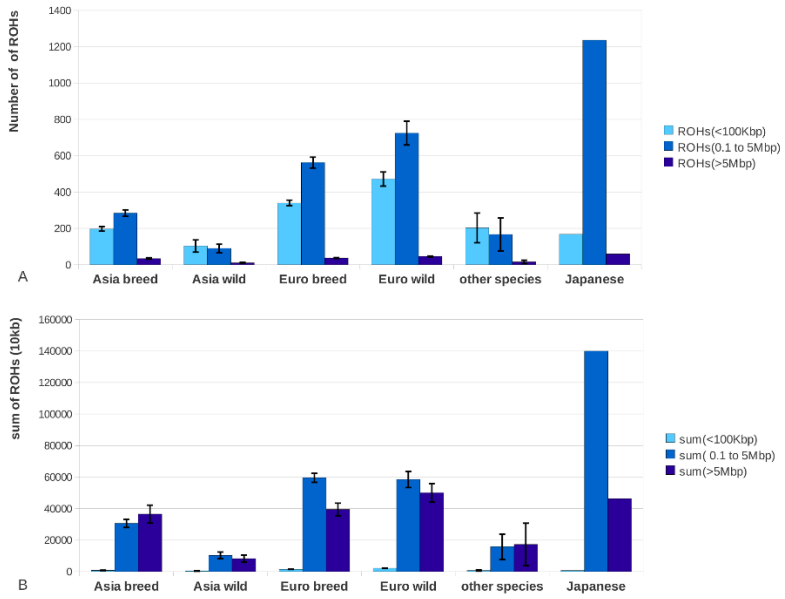


Figure 2.2 Total number of ROHs and proportion of the genome covered by ROHs. **A**. The average number of ROHs belonging to three size classes small (<100 Kbp) medium (0.1 to 5 Mbp) and large (>5 Mbp) for each of the different groups. **B**. The total size of the genome that is covered by the particular ROH size class in one individual, averaged per group. The Japanese wild boar is shown separately and is not included in the Asian group, as its demographic history from an island population and the associated ROH pattern is very distinct from all other sampled individuals. Asian wild boars (n = 4), Asian pigs (n = 7), European wild boars (n = 6), European pigs (n = 29), other species (n = 5).

2.2.2 Effects of Demography on ROH distribution

ROHs were separated into three size classes: 1) small (<100Kbp), 2) medium (0.1 to 5Mbp) and 3) large (>5Mbp). We computed the proportion of ROHs falling in each class in all our 52 samples. While small ROHs were abundant throughout the genome, their absolute contribution to the genome was relatively small (Figure 2.2). In contrast, medium sized ROHs were the less common but covered significantly more of the genome than small and large RoHs. The large ROHs were at least a tenfold less abundant than medium RoHs, but nevertheless covered a significant proportion of

the genome. Asian domesticated pig genomes were covered mostly by large ROHs. Asian wild individuals had much fewer genomic ROHs and also a smaller proportion of ROHs in their genome than all European pigs and the Asian domesticated pigs ($p<0.0001$). European wild boars had on average the highest number of ROHs and highest proportion of genomic autozygosity. The Japanese wild boar is an outlier in both number of ROHs and cumulative size likely due to its island bio-geographical background, so we treated it separately. The divergence between the wild boars in Europe and Asia was estimated to have occurred ~1.2 mya and a major drop in population size in both continental groups took place from ~50 kya and onwards, based on individual genome demographic inference as implemented in the *Pairwise* Sequentially Markovian Coalescent (PSMC) model (Figure S2.3). Population size in the Asian *Sus scrofa* is thought to not have been reduced as severely as for the European populations, which is supported by the nucleotide diversity outside ROHs and ROH analysis (Figure 2.1A-B and Figure 2.2). In addition, the Asian wild populations were estimated to have a larger effective population size than the Asian domesticated pigs, and the European wild populations had the smallest population size based on the ROH analysis.

We tested the utility of the Illumina porcine 60K beadchip to identify ROHs in the three size classes. Genotyping arrays are widely used and offer the possibility to cost-effectively study a much wider sample size and to test the usefulness of this technology for the detection of ROHs. Using this array we evaluated whether the results from whole genome re-sequencing analyses for a limited number of individuals could be extrapolated to an entire population. The chip-based methodology was capable of detecting the ROHs larger than 5 MB (Figure 2.3) but underestimated the cumulative size of ROHs in the genome, especially for the European samples. This phenomenon is likely to be due to the number of small sized ROHs in European populations that cannot be detected due to the limited resolution of the SNP chip. The Japanese wild boar had many ROHs, but the ROH size was not extremely large because the ROHs were interceded by short sections with variable sites (Figure S2.2 and Table S2.1). Therefore, the total sum of ROHs was probably overestimated in this individual (Figure 2.3) by the chip-based method and showed a weak correlation with the cumulative ROH size of >5Mbp homozygous blocks that were identified with the re-sequencing method. Since ROHs in the highest size class are fully detected (>5Mbp, Figure 2.3) comparing populations based on their 60K-defined ROH distribution is valid for analysis of large ROHs. Naturally, the limited capability of detecting shorter ROHs has implications for the inferred demography

and therefore we use the 60K defined ROHs only for comparison with our largest sequence based ROHs.

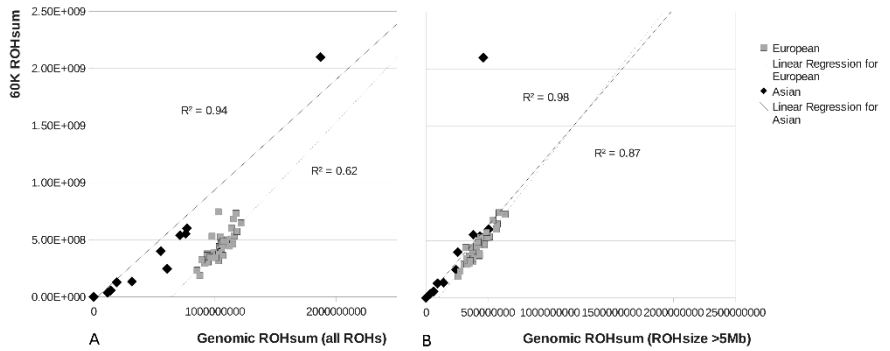


Figure 2.3 ROH size in pigs based on the 60K chip method and the re-sequencing method. **A** shows the correlation between both methods when the total sum of ROHs is taken from the re- sequencing method ('Genomic ROHsum') and the 60K chip method ('60K ROHsum'). In **B** the correlation is shown when only the ROHs over 5 Mbp are taken into account for the re-sequencing method. The outlier for the Asian group (the Japanese wild boar) is not included in the R^2 calculations.

241 individuals from the different *Sus scrofa* populations that had been re-sequenced, were genotyped using the 60K assay, and number and cumulative ROH size where scored. Details of the genotyped individuals can be found in table S2.2. Sequenced individuals were never extreme within their population in terms of ROH number or ROH size. In the Asian and European breeds, the number of ROHs ranged from 5 to 59 and cumulative ROH size was 10 Mb to 1 GB (Figure 2.4). European breeds had a narrower distribution of number of ROHs and cumulative ROH size. Both sum and number of ROHs in the Asian breeds Jianquhai and Xiang showed a modest bimodal distribution. The Chinese wild boars tended to have fewer ROHs and cumulative ROH size than their European relatives. Even though cumulative ROH size for the Japanese wild boars may have been overestimated because of the low resolution of the 60K chip, four individuals were extremely homozygous with more than 2/3 of their genome consisting of ROHs. Variances in ROH size and abundance in the Japanese wild individuals was much higher than in the other groups. Notable, two Dutch wild boars had significantly fewer ROHs than all other European wild boars from the same populations (indicated with an* in Figure 2.4B).

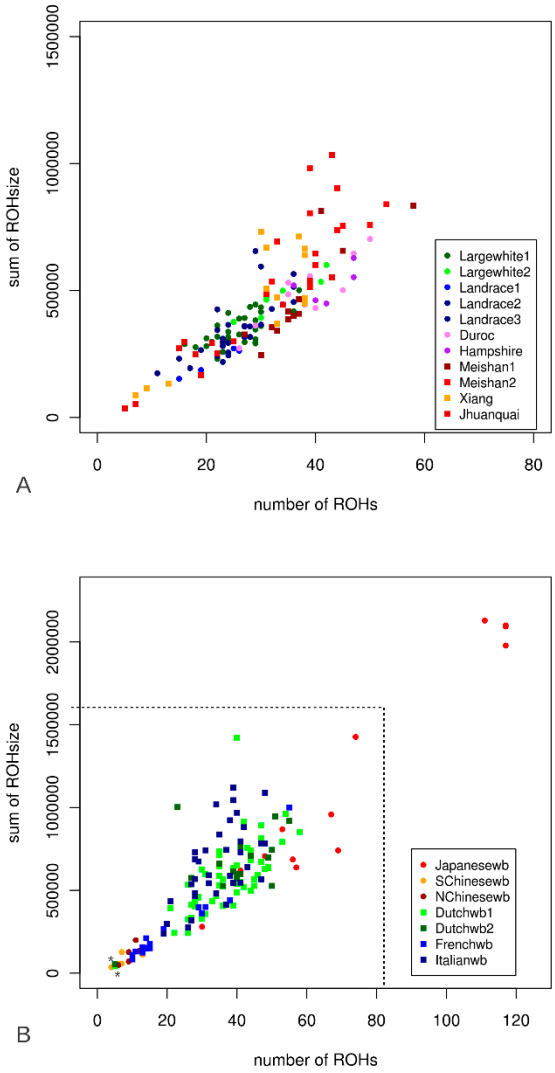


Figure 2.4 Number and cumulative ROH size (ROHs>5MB) for all genotyped individuals. Number of ROHs and sum of ROHs detected by PLINK for all 241 individuals genotyped by the Illumina porcine 60k beadchip. Sum of ROHsize is *1000 bp. **A**. ROHs in domesticated individuals. Asian pigs are shown in red, orange and purple and the European pigs are in blue and green. **B**. ROHs in wild individuals. Asian wild boars are shown in red and orange and the European wild populations are displayed in green and blue. The dashed line represents the range of ROH number and ROH size for the domesticated individuals. The individuals marked with * are putative hybrids.

The genomic variation pattern for the 52 re-sequenced individuals was analyzed in more depth. We found that statistics such as π-out (nucleotide diversity outside ROHs), average ROH size and total number of ROHs were good predictors to assign individual to their corresponding group (Figure 2.5). Interestingly, while all Asian wild boars formed a monophyletic clade on our phylogenetic tree (Figure S2.1), we found that the Japanese sample did not cluster with other Asian samples on our three dimensional plot (Figure 2.5). The Chinese wild boars represent the most variable cluster due to their high nucleotide diversity and few ROHs ($p<0.001$). We found that nucleotide diversity was higher in European breeds than European wild boars ($p<0.0001$). Moreover, total number of ROHs in the genome was also lower in European breeds. This resulted in two clusters in our three dimensional plot (Figure 2.5), contrasting with the monophyletic status of European populations on our phylogenetic tree (Figure S2.1). The Asian breeds were more inbred than their wild ancestors but displayed fewer ROHs and higher nucleotide diversity than the European animals ($p<0.0001$). The sequenced *Sus verrucosus* individual had the lowest genomic variation of all tested animals due to extremely low nucleotide diversity, intermediate ROH number and large ROH size. The sequenced *Sus barbatus* individual had the least ROHs and smallest ROH size of any of the sequenced individuals, suggesting the individual is highly outbred, with a high effective population size. The ROH pattern in *Sus cebifrons* was particularly interesting because the total number of ROH was very low, but it contained few very large ROHs and had relatively low nucleotide diversity.

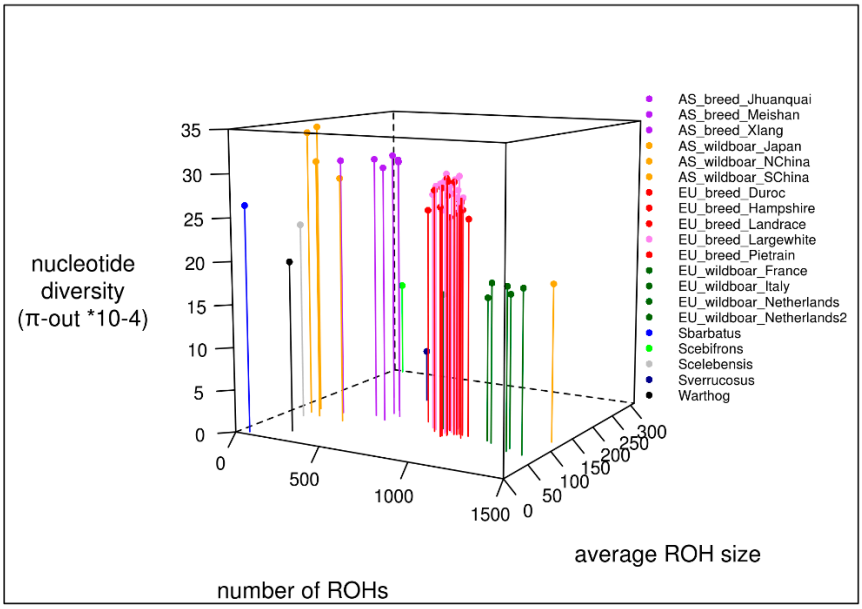


Figure 2.5 Three-point ROH statistics for all 52 sequenced individuals. On the x-axis, the number of ROHs in the genome per individual is plotted, the average ROH size ($*10^4$ bp) is displayed on the y-axis and the nucleotide diversity outside ROHs in a 10 kb window 'nucleotide diversity (π-out $*10^{-4}$)' on the z-axis. Coloration is based on relatedness and geography, with individuals from the same populations having the same color.

2.2.3 Effects of recombination rate on ROH distribution

To test for the effect of recombination rate and GC content on ROH formation and distribution, we computed GC content and recombination rate relative to the physical chromosomal position, for each chromosome separately, and averaged the results over all chromosomes (Figure 2.6C and 2.6D). The GC content was based on the porcine reference genome build 10.2 (Groenen 2012). GC content was generally higher when moving toward telomeric regions in metacentric chromosomes and toward chromosomal edge in acrocentric chromosomes (Figure 2.6A). Overall, GC content was inversely correlated with distance to the telomeres (Figure 2.6C). Recombination rate for pigs was calculated based on ~60.000 markers, obtained from Tortereau (2012) and averaged over all chromosomes. Variation in recombination fraction over the physical position of the chromosomes, with high recombination rates at telomeric regions and very low recombination rates at the central part of chromosomes, was most pronounced in pigs and virtually absent in mice (Figure

2.6D). A 'U-shaped' distribution of recombination rates was present in all chromosomes in pigs, while in humans this is only observed in metacentric chromosomes (data not shown). Nucleotide diversity correlated strongly with both recombination rate (cor=0.88, p<0.00001) and GC content (cor=0.61, p<0.005). Nucleotide diversity greatly increased in the European breeds and wild boars when ROH bins were excluded. However, this phenomenon was only observed in Asian breeds at the chromosome tips (Figure 2.6A, B). ROH distribution was negatively correlated with GC content, recombination rate and nucleotide diversity outside ROHs (cor=-0.71, -0.87 and -0.95 respectively, p<0.0001 for all). This is expected as these genomic features all appeared to be highly correlated.

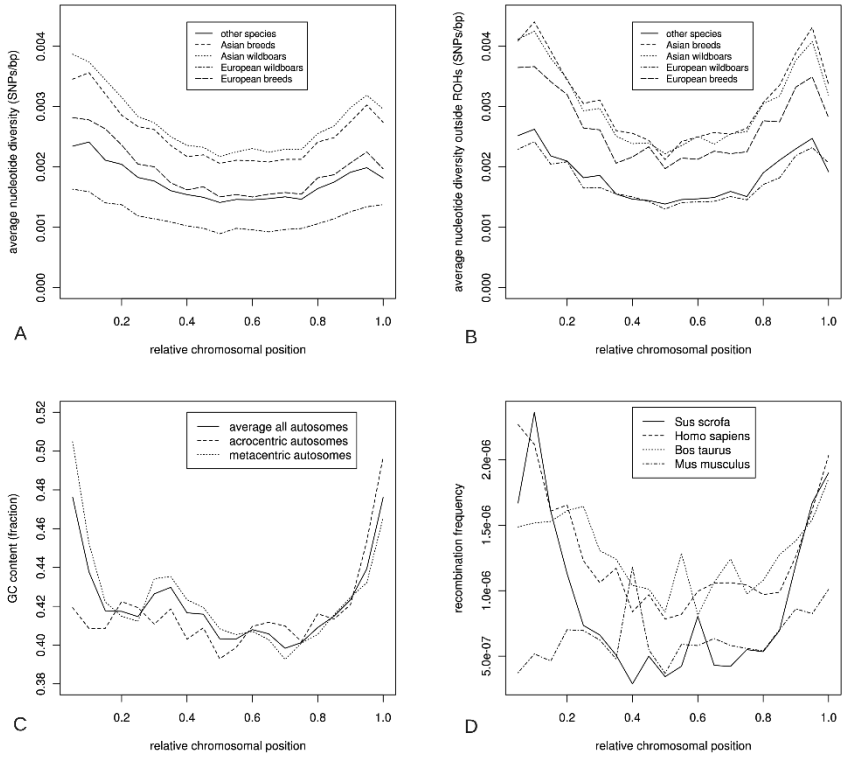


Figure 2.6 Characteristics of the porcine genome over relative chromosomal position. Physical distribution of total nucleotide diversity (**A**), nucleotide diversity outside ROHs (**B**), GC content (**C**), recombination rate (**D**) over chromosomes. Relative chromosomal position is averaged for all chromosomes so that 0.0 represents the left telomeric region and 1.0 the far right telomere.

The likelihood of ROH occurrence at a particular chromosomal position was dependent on the size of the ROH (Figure 2.7). The ROHs from the four *Sus scrofa* groups were separated into the three previously mentioned size classes (small, medium, big) and the relative distribution of ROHs over the genome was calculated for each size class (the Japanese wild boar is included in the Asian wild boar group, Figure 2.7D). The largest ROHs appeared more in the low recombination regions in the middle of the chromosome in European breeds and both Asian groups, and the smallest ROHs had a relatively higher distribution towards the telomeric regions ($p<0.001$). Medium sized ROHs seemed to be evenly distributed across the genome in all groups. The ROHs in European wild boars tend to be more evenly distributed than those in other groups (Figure 2.7B). The differences in ROH occurrence and nucleotide diversity between the extreme regions in recombination frequency were most profound in the European domesticated pigs.

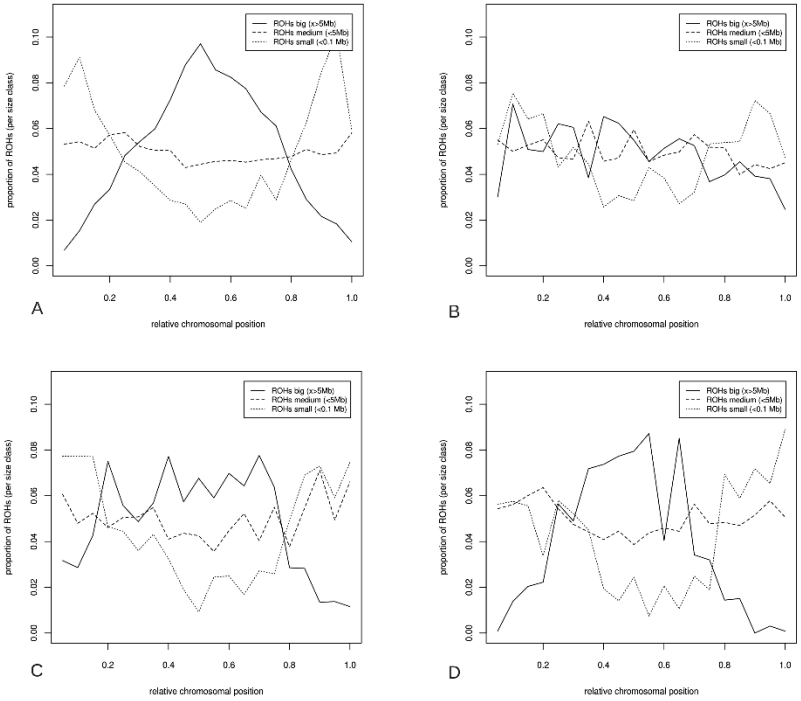


Figure 2.7 Distribution of ROHs over relative chromosomal position. ROHs were split into three size classes; big (x>5 Mbp), medium (0.1 to 5 Mbp) and small (x<0.1 Mbp). The distribution is relative only to the total number of ROH bins in that particular size class, and the distributions are averaged over all chromosomes. ROH distribution is given for four groups: European pigs (**A**), European wild boars (**B**), Asian pigs (**C**) and Asian wild boars (including the Japanese, **D**).

2.2.4 ROH, gene content, and shared homozygosity

We investigated the correlation of gene content and ROH occurrence on a genome-wide scale. We found no significant correlation between gene content and ROH count. This result was confirmed when repeating the same analysis in different sample groups (Figure S2.4). Moreover this was still non-significant when using different classes of ROHs (small, medium, large). No regions of homozygosity overlapped in all sequenced *Sus scrofa* individuals. We found two overlapping RoHs among our European breeds (cumulative size of 6.03Mb). The first region on chromosome 1 did not contain any genes, whereas the second region on chromosome 9 contained 11 genes, of which 7 were semaphorins. These genes are involved in cell differentiation (nervous system development) and have previously been identified as differentially expressed between the Large White breed and Iberian pigs (Ferraz 2008).

Among all homozygous regions in the pigs from Asian breeds, 4 regions were shared among all individuals with a cumulative size of 4.38 Mb. The largest region was on chromosome 1 and contains a total of 1136 genes. Two neighboring fractions of chromosome 3 contained a ROH in all Asian breeds. Interestingly, we found that one of these regions also partially overlapped with ROHs found in all Asian wild boars, but not in all European pigs. This gene dense region contained 91 genes total, and has previously been reported as a region possibly under positive selection in Asian pigs (Groenen 2012). The last homozygous region in the Asian breeds on chromosome 5 contained 3 genes including LEMD3, MSRB3, which is involved in oxidation reduction, and the WIF1 gene, involved in positive regulation of fat cell differentiation.

We found 28 ROHs that were overlapping among our 13 Large White samples (European breed; cumulative size of 54.38Mbp). All of these ROHs were carrying genes (336 genes in total). GO analysis revealed 37 significantly overrepresented GO-IDs, i.e. involved in the organization and assembly of nucleosome and involved in reproduction (Table S2.3). ROHs that were found exclusively in all individuals in the Large White – Landrace group of breeds contained 11 genes, including the PLAG1 gene on chromosome 4, which is related to growth.

The European wild boars displayed shared genomic homozygosity on a total of 47.52 Mb in 81 regions. Out of these mutually homozygous regions, only 26 were carrying genes. The 97 genes in these regions contributed to 24 significantly overrepresented GO-IDs. Some enriched terms found in the Large Whites were also overrepresented in the European wild boars, mostly histones. Histones are usually very syntenic, which

may explain the clustering of histone related genes and may not have any functional relevance. Macromolecular complex and cellular component assembly and organization were abundant in the list for wild boars, as well as defence responses (Table S2.4).

2.3 Discussion

Recombination maps are currently available for only a few vertebrate species. The pig shows very pronounced differences in recombination-rate throughout the genome, and highly diverse demographies in both wild and domesticated populations, making it an excellent model species to demonstrate the interaction between recombination landscape and demography.

2.3.1 Effects of demography

The size and abundance of ROHs in the porcine genome varied markedly between individuals from different (sub) populations. Animals from the same population tend to have similar ROH patterns in their genome (Figures 2.4 and 2.6), which indicates of the influence of shared demography on ROH distribution. The class of large ROHs is most sensitive to recent population changes. Thus, the bimodal distribution of large ROHs in Asian breeds may be explained by the sampling of two different populations. In European humans, the number and expanse of ROHs correlated strongly with latitude (Nothnagel 2010) but not longitude. Kirin et al. also showed a correlation between geography and genomic ROH occurrence in humans (Kirin 2010). In our study we also showed a significant difference in ROH abundance and size between European and Asian wild boars. However, the Japanese wild boar shared a long demographic history with the other Asian wild boars compared to all other individuals, and yet displayed a completely different ROH pattern. The ROH pattern in this individual is consistent with a small current population size. The high number of ROHs (rather than ROH size or nucleotide diversity) indicates that the population has been small for a longer period of time, but that the source population has been substantial in size. This ROH pattern is probably indicative of small isolated or island populations, as was also found to a smaller degree in relatively small human populations (Kirin 2010).

The fact that cumulative ROH size was dominated by large ROHs in Asian pigs indicates that the population size has only recently reduced and these pigs are derived from large source populations (Figure 2.2). The European wild boars, by contrast, showed a more uniform distribution of ROHs relative to chromosomal position (Figure 2.7), consistent with a long lasting low effective population size.

Findings by Scandura (2008) confirmed that genetic diversity in current European wild boars (apart from those on the Italian Peninsula) was mostly affected by glacial bottlenecks. Nothnagel (2010) stated that recent population growth in humans over the last ~200 years has not significantly contributed to genomic ROH distribution. For the current study on pigs, ROHs did appear to be affected, both by population growth, as previously larger ROHs are broken down by recombination in a non-uniform distribution (Figure 2.6 and 2.7), and population decline, as new ROHs are formed. These different results may be a consequence of the demographic histories of the two species, since humans have had a global population expansion, but wild boars and pigs experienced severe (local) population reductions. Thus it seems like ROHs reflect both the recent past and current status of a population as well as distant population history, and are very susceptible to population dynamics.

South-East Asia has been pinpointed as the centre of origin of *Sus scrofa* (Larson 2005, Groenen 2012, frantz 2013b). Thus, it is not surprising that the estimated haplotype diversity in the Asian source populations was found to be higher than in European populations. This phenomenon is also seen in the (out of Africa) human genetic patterns, where linear relationships exist between nucleotide diversity and distance to source of origin (Ramachandran 2005, Li 2008). Likewise, the westward migration across Eurasia by *Sus scrofa* likely involved numerous founder events. These events are expected to result in lower genetic diversity in Western Eurasian populations. Moreover, recent studies have found evidence for more intensive bottlenecks in Europe compared to Asia due to Pleistocene glaciation (Groenen 2012). Thus, we expect a degradation of the overall genetic diversity in European populations compared to Asian populations. This phenomenon was most apparent in the nucleotide diversity outside ROHs (Figure 2.5). The demographic decline of most European wild boar populations did not seem to cause a decline in genetic variation within these populations according to Scandura (2008). Here we show, however, that patterns of ROH distribution as well as nucleotide diversity outside of ROHs are consistent with a long and ongoing history of small local effective population sizes.

In the breeding industry, a possible consequence of artificial selection is a reduced effective population size and associated genetic diversity. Pig breeders, however, are generally concerned about retaining sufficient genetic variability to maintain a good selection response in the future (Habier 2009). Based on microsatellite markers and quantitative measures of genetic diversity such as the expected heterozygosity (H), Ramirez (2009) suggested that the genetic variation within domesticated pig breeds is not significantly lower than within the wild boar genome. This is confirmed by our

findings in that European wild boars contained more ROHs and lower nucleotide diversity outside ROHs compared to European domesticated pigs. Even though the Duroc and Hampshire breeds cluster within one clade with the European wild boars (Figure S2.1) their ROH pattern and nucleotide diversity are typical for the European breeds (Figure 2.5). The ROH pattern is a signature of similar treatment of the breeds. In closed populations, nucleotide diversity outside ROHs, named π-out, should reflect ancient haplotype variation that was present in the ancestral population and should not be strongly affected by (recent) occurrences of autozygosity. The difference in overall π and π-out was highest in European commercial pigs. This suggests high haplotype diversity in the ancestral population, despite the more recently formed populations with smaller effective population size that result in ROHs that reduce the overall nucleotide diversity π. This may be explained by a recent admixture between European and Asian breeds. Indeed, during the industrial revolution, Asian pigs were imported to Europe to improve local pigs. Such introgression of Asian pig genomes in Europe is expected to have increased overall genetic diversity in European pigs compared to their wild counterpart. The hybridization could have introduced distinct haplotypes resulting in fewer IBD tracts and probably higher variation. The higher nucleotide diversity may be a consequence of improvement of European breeds by hybridization with Asian pigs (Giuffra 2000, Megens 2008, SanCristobal 2006), showing the strong influence of outcrossing. Among the Dutch wild boars that were analysed using the Illumina porcine 60K beadchip data, two showed a highly distinct ROH pattern compared to other individuals from the same population (Figure 2.4B). A previous study identified these two individuals as being recently introgressed with domesticated pigs (Goedbloed 2012). These individual-specific ROH patterns and the relatively high nucleotide diversity in the European breeds underline the importance of parental ancestry for the levels and pattern of variation in the genome of the offspring.

We were able to use the same re-sequence based methodology to study the ROH pattern in a few closely related *Sus* species. The most outstanding individuals in terms of low nucleotide diversity and number of ROHs were found in this outgroup. The bearded pig *Sus barbatus,* most widespread in Borneo, had a minimal genomic ROH coverage. This indicates that the population has been large enough to avoid consanguineous matings for a substantial (recent) time period. Interestingly, the genome of *Sus cebifrons* displayed few ROHs, despite the fact that this species only occurs on a few small islands in the Philippines. Nevertheless, the average ROH size was the largest of all individuals, and it showed a low π-out indicative of small ancient population sizes. As this species is confined to a few small islands, founder effects

could explain the low observed π-out. The species *Sus verrocusus* had the lowest nucleotide diversity and ROHsize. This indicates a very small current and past effective population size, consistent with its endangered status on the IUCN red list. The inbreeding could have intensified due to this individual coming from a zoo. Other factors, such as different mating system may also influence ROHs distribution. For example, the mating pattern is expected to be different in artificially managed populations than in natural populations. In addition, closely related species or even separate wild populations can have different hierarchical systems that strongly influence the effective population size (Wright 1921). For instance, the Chinese domesticated pigs cluster closer to the European pigs in Figure 5 than the European wild boar, although the phylogenetic tree displays a different clustering (Figure S2.1). This indicates that the pattern of haplotype variation is similarly shaped in domestic populations, despite having a mostly independent domestication history.

2.3.2 Effects of recombination rate

From population genetic theory, the effects of linkage disequilibrium are important to understand variation in genomes (Megens 2009), particularly in the presence of selection. Large parts of genomes seem to be under selection, and all functional sites in a genome are potentially under purifying selection (Lohmueller 2011) or adaptive evolution, even in non- transcribed regions (Andolfatto 2005). The effects of selection and demography are expected to have an interaction with the recombination landscape in the genome, thereby shaping genome wide variation in individuals and populations. This interaction has so far been poorly studied even in species for which considerable genomic resources exist, and has been neglected in studies on genetic conservation (Allendorf 2010).

ROH distribution over the chromosomes was found to be non-random in other mammals, including humans (Nothnagel 2010, Vonholdt 2010, Kirin 2010, Curtis 2008). The proportion of ROHs in the genome was much higher in pig than in any other species studied so far, with individuals containing ROHs in over 75% of their genome. The U-shaped distribution of recombination rate was more profound in pig compared to other mammals. Despite a high degree of conserved synteny between human and pig (Sun 1999), it is surprising that this pattern is so pronounced in pigs compared to humans. Correlations between ROHs and LD exist for other species, and were also very strong in pigs, with a higher recombination rate outward of the central chromosomal regions and in short chromosomal arms (Rohrer 1996). We showed that heterozygosity is higher towards these peripheral regions, a phenomenon that was previously observed in pigs (Esteve-Codina 2011). In humans, chromosome size

seems to be an important determinant of ROH occurrence (Nothnagel 2010). In pigs the occurrence of ROHs was not proportionally higher on larger chromosomes, but seemed to be present throughout the genome and mostly influenced by the physical position on the chromosome, particularly in relation to local recombination rate. The higher abundance of small ROHs towards the telomeres probably stems from the central part of the chromosome being covered by the larger ROHs that have not been broken down due to lower recombination in this region. A bottleneck in the past with stable or on-going population growth ever since may lead to a more equal distribution of ROHs, as observed in the European wild boars.

Genomic features, GC content, nucleotide diversity and recombination rate, were all correlated and displayed similar U-shaped chromosomal distributions in the porcine genome (Figure 2.6A-D). This has important implications for the probability of autozygosity in different chromosomal areas. Large ROHs appeared significantly more often in regions with low recombination. The difference in pattern of ROHs between European domesticated pigs and European wild boars is probably related to more continuous inbreeding in European wild populations, which have only expanded their population and range in the past 60 years (Goedbloed 2012). Breed formation in European pigs has likely resulted from hybridization of different domestic and wild origins, including pigs originating from Asia. Pig populations, defined as breeds or commercial lines, are likely to have had an effective population size, in many cases measured in tens rather than hundreds or thousands, over the past decades. Many traditional breeds have been marginalized, with very small breeding stock (Herrero-Medrano 2013a). Even the commercial pure lines, particularly the boar lines that are usually applied in three- or fourway crosses to generate the finisher pigs that go to slaughter, are often kept closed with small effective population size. Larger ROHs were therefore mostly found in regions of low recombination rate in domesticated pigs, because time since formation has been short. Small ROHs are thought to be present in a population longer than large ROHs and are more often shared among individuals than large ROHs. The rationale behind this is that recombination will degenerate large ROHs with time, but in regions with little or no recombination, small ROHs will be retained. Therefore, despite the time factor, these non-recombining regions will preserve ROHs when created, while recently originated large ROHs may occur randomly in the genome before they are degraded. The number of ROHs and the size distribution of the ROHs are therefore important determinants of recent and more historical population bottlenecks and inbreeding events.

2.3.3 Effects of gene content and selection

Coding sequences are generally GC-rich regions in mammals, including pigs (Nie 2010, Wernersson 2005). We found a correlation between ROH occurrence and GC content in the genome, but not between global gene content and ROHs. The apparent lack of gene enrichment in ROHs suggests no direct correlation to the ROHs identified in our study and selection acting on genes. However, it is possible that some of the ROHs overlap with non-coding functional elements such as cis-regulatory modules. Although a few regions were identified where loss of genetic diversity may have been the result of selection, our study suggests that vast majority of the ROHs are likely to be neutral. The occurrence and distribution of ROHs, therefore, are mainly shaped by the interaction of past demographic events and recombination rate.

For the Large White breed, of which 13 individuals were sequenced, only 54Mb was found to be homozygous across all individuals combined, a fraction of the total of the genome embedded in ROHs across the same population. The total sum of homozygosity for each individual, therefore, is much larger than it is for the population. In the Large White breed, some genes were found in the homozygous regions that are possibly under positive selection associated with traits of commercial interest, such as fast reproduction. These genes are, however, found in regions that are large (many Mb in size). In other populations, such as the European wild boar, the cumulative shared homozygous regions are much shorter and not always carrying genes, which could indicate that, despite the high degree of homozygosity in individual genomes in wild populations, selective sweeps may not be very common. Some overlapping ROHs may contain selected genes that are associated with defense mechanisms and adaptations to novel environments, but the fact that no genes were found in many overlapping ROH regions between the wild boars elucidates the stochasticity of ROH occurrence. We conclude that only a small fraction of the ROH-containing regions in pigs are homozygous due to positive selection.

2.4 Conclusion

Our study shows that the formation of ROHs is mainly influenced by past demographic events and local recombination rate. This finding implies that inbreeding and recombination rate may act together in regions containing genes, mimicking selection. Genes in regions of low recombination, therefore, are at higher inbreeding risk, and could experience more rapid fixation. This phenomenon can have drastic influence on the fitness of individuals in small populations.

The genome-wide correlation of ROHs with the local GC content and recombination rate highlights the importance of genomic features such as recombination rate for autozygosity predictions. Many diploid species are likely to be heterogeneous in genome-wide recombination rate. This means that estimating inbreeding coefficients from effective population size, pedigrees, and even genetic data such as microsatellite genotype data (Keller 2011, Leutenegger 2003) does not completely measure the proportion and distribution of IBD homozygosity. Therefore of risk of inbreeding depression is underestimated. In addition, in a selective sweep analysis such local genomic regions of low recombination may wrongly be interpreted as being under selection.

Our re-sequencing based methodology to determine genomic variation implements genomic ROH distribution as a separate variable to nucleotide diversity. We show that the method is applicable even to closely related non-model species. Therefore, its utility exceeds species boundaries and combines different characteristics of diversity in diploid organisms. We show that both population demography and recombination landscape influences genomic ROH occurrence and these factors should both be taken into consideration when designing genetic conservation strategies in wild and domesticated species. We suggest more research on the genome-wide mechanisms that prevent the negative effects of inbreeding by influencing the localization of ROHs.

2.5 Materials and Methods

2.5.1 Experimental setup

A total of 52 animals were selected for re-sequencing and genotyped on the porcine 60K SNPchip. We re-sequenced one individual per species of *Sus barbatus, Sus celebensis, Sus cebifrons* and *Sus verrucosus,* and one warthog (*Phacochoerus africanus*) representing one of the closest relatives outside the genus *Sus*. Within *Sus scrofa*, the five European pig breeds Duroc, Hampshire, Large White, Pietrain, and Landrace were represented by 4, 2, 13, 5 and 5 individuals, respectively. A total of six animals from European wild boar populations from four distinct populations from the Netherlands, France and Italy were included as a separate group, as well as five Asian wild boars (two from North China, two from South China and one from a small population originated from a Japanese island). Finally, seven Chinese pigs, four from the Meishan breed, two from the Xiang breed and one from the Jianquhai breed were selected to represent the variation within Asian domesticated pigs. An additional 241 individuals from *Sus scrofa* populations, for which individuals were

sequenced, were genotyped for SNP assay based ROH analysis. Because of ascertainment bias and paucity of segregating SNPs on the 60K chip for other Suids than *Sus scrofa*, no other *Sus* species were genotyped (Figure S2.5).

2.5.2 DNA extraction, SNP genotyping and library preparation

DNA was extracted from whole blood by using the QIAamp DNA blood spin kit (Qiagen Sciences). Every DNA sample was checked for quantity and quality on the Qubit 2.0 fluorometer (Invitrogen) and run on a 1 % agarose gel. SNP genotyping was performed on the Illumina Porcine 60K iSelect Beadchip (Ramos 2009). DNA from all individuals was diluted to 100ng/ul and genotyped according the IlluminaHD iSelect protocol. Data was analyzed using Genome Studio software (Illumina Inc.). In case of re-sequencing, library construction and re-sequencing of the individual samples was performed with 1-3 ug of genomic DNA according to the Illumina library prepping protocols (Illunima Inc.). The library insert size was aimed for 300-500 bp and sequencing was performed with the 100 paired-end sequencing kit.

2.5.3 Sequencing and SNP discovery

All selected individual pigs from domesticated breeds and wild populations were completely sequenced to ~8X depth (details on coverage in Table S2.1). Reads were trimmed to a phred quality > 20 and minimum length of both pairs of 40 bp, and the quality trimmed reads were aligned to the *Sus scrofa* reference genome build 10.2 (Groenen 2012) using the unique alignment option of Mosaik Aligner (V. 1.1.0017) to avoid erroneous called SNPs due to copy number variations and repeats. The data has been deposited to the Sequence Read Archive (SRA) at EBI, under accession number ERP001813 (link: http://www.ebi.ac.uk/ena/data/view/ERP001813). SNPs were called and filtered with mpileup from the SAMtools (V.0.1.7 r510) software package (Li 2009) with default settings for diploid organisms. Additional filtering was applied to the called variants with VCFtools (minDP=7; minDP calling a SNP=2; maxDP=~ 2*average coverage; INDELs excluded). By setting the minimum depth to call a SNP to 7X and only consider a base sufficiently covered at 7X, we reduce the number of missed variants. Nucleotide diversity was calculated for bins of 10kbp over the entire genome within each individual. “SNPbin” is the SNP count per 10kbin, corrected for the number of bases within that bin that was not covered enough for the VCFtools filtering, so that the eventual SNP count per bin (SNPbin) is proportional to 10.000 covered bases.

SNPcount = total number of SNPs counted in a bin of 10kbp

DP = coverage in bp/bin (per base at least depth of 7X and maximum of ~2*average coverage)

binsize =10.000

Correction factor = DP/binsize

SNPbin = SNPscount /Correction factor

2.5.4 Phylogenetic tree construction

A phylogenetic tree was constructed for the 52 re-sequenced individuals. We genotyped these individuals on the Illumina Porcine 60K iSelect Beadchip. Based on these genotypes, an IBS similarity matrix was created using Plink 1.07 (Purcell 2007). Subsequently a neighbor joining hierarchical clustering was performed using the program Neighbor available from the Phylip package (Felsenstein 2005).

2.5.5 ROH definition

Regions of homozygosity were extracted for all autosomes of the 52 re-sequenced individuals. Sex chromosomes were excluded as their recombination landscape is known to be different from the autosomes and the genetic map resolution for the X-chromosome differed from the autosomes in pig. Moreover, males and females should have been treated differently when the X-chromosome would have been included, and such analysis falls outside the scope of this paper. Autozygosity (a genomic region that was inherited from a common ancestor by both parents, and therefore indicates a certain level of relatedness) can typically be traced back in the genome as a ROH. The autozygous stretch is eventually broken into shorter pieces by recombination. A region of homozygosity is a genomic stretch that contains less variation in an individual than is expected based on the genomic average. Autosomal homozygous stretches (ROHs) for the re-sequenced individuals were determined using a sliding window approach. SNPs were counted in bins of 10 kbp, and those bins that fall into a window of 10 consecutive bins with a total SNP average below the genomic average were extracted in both the forward and reverse orientation. All neighbor bins were concatenated to form homozygous stretches. Out of this selection, only those stretches that contained a SNP count below a set threshold were considered part of a true homozygous stretch. The threshold was set to a SNP count of maximum 0.25 times the genomic average, with a maximum absolute value per stretch of the false discovery rate plus the mutation rate ($\mu = 2.5*10^{-8}$) because in some cases that exceeded the value of 0.25 times the genomic average. The false discovery rate was calculated based on the homozygous loci genotyped on the Illumina Porcine 60K iSelect Beadchip (Ramos 2009) that were called as a heterozygous locus in our database by vcftools (average ~1.78 per bin). The rationale

behind a threshold for heterozygosity rather than no heterozygote allowance is based on the thought that mutations in originally autozygous stretches may mask autozygosity over time. The genome-wide heterozygosity of an individual expresses the present variation in the population, and the associated chance that a certain autozygous stretches will reunite. The sequenced individuals varied greatly in the genomic heterozygosity and in population history. In addition, not all populations were sampled equally. Therefore the height of the threshold was based on the genomic average of the tested individual only, rather than the total set of individuals or an allele frequency-based likelihood of ROH occurrence. The threshold of 0.25 times the genomic average is based on permutations where the individual SNP distribution is randomized over all chromosomes. At a value of <0.25 times the genomic average, the observed ROH distribution deviates from the randomized distribution (see Figure S2.6). Local assembly or alignment errors were avoided as much as possible by relaxing the threshold for individual bins within a homozygous stretch, allowing for maximum twice the average SNP count in a bin, if the local maximum of 10 bins did not exceed 2/3 times the genomic average, and if the average of the ROH surrounding the presumed erroneous bin(s) still matched the previously mentioned criteria. Insufficiently covered bins (DP=<10%) were excluded from the SNP average calculations but were included in the ROH size determinations, with accepted ROHs containing maximum 2/3 uncovered bins and containing covered bins at both ends (example in Figure S2.7). In an analysis where the coverage of all individuals was lowered, we used a range of 5 thresholds for bin coverage (DP=<5, <10, <20, <50, <80%) and proportion of uncovered bins within a ROH (<1/4, <1/3, <1/2, <2/3, <3/4). We compared the outcomes with the highly covered individuals, and the errors in ROH size and abundance due to low coverage were minimized when thresholds of DP=<10% for bin coverage and <2/3 of missing ROH bins were used.

We genotyped 241 individuals on the Illumina Porcine 60K iSelect Beadchip for ROH detection (details in table S2.2). ROHs were calculated with the Runs of Homozygosity tool in PLINK (v.1.07) with adjusted parameters (--homozyg-density 1000, --homozyg-window-het 1,--homozyg-kb 10, --homozyg-window-snp 20) (Purcell 2007). Markers were filtered for call rate >95%. The homozygosity tool in PLINK v.1.07 does not include removal of MAF <0.05 or LD pruning when assessing ROHs. We aimed at keeping the ROH detection methods for the 60K data and genomic data as similar as possible in order to make sound comparisons. Therefore, no additional filtering for low allele frequencies was done, because sampling was unequal across populations and removing rare alleles could result in an overestimation of ROH in individuals from undersampled populations. No adjustments according to recombination rate were

done because part of our goal was to analyze the influence of recombination rate on the occurrence of ROHs. For the re-sequenced animals, correlations with ROHs defined with PLINK were tested with the R (v.2.11.1) *cor* and *cor.test* functions.

2.5.6 Population size estimations

Estimates of effective population size and split between the European (n=2, from the Netherlands and Italy) and Asian (n=2, from North and South China) wild boars were inferred using a HMM as implemented in PSMC (Li 2011) on copy number neutral fragments with a cumulative size of 1Mbp, with a generation time of five years (g=5) and default mutation rate/generation ($\mu = 2.5*10^{-8}$).

2.5.7 Statistical analysis of the genome ROH distribution

All genomic features are based on the non-repeat masked *Sus scrofa* reference genome (build10.2). Values for GC content and nucleotide diversity were calculated for each relative chromosomal distance (0-1 with steps of 0.05) and averaged for all chromosomes. Based on the porcine genetic map (Tortereau 2012) we estimated recombination rates based on the ratio of genetic and physical distances of neighbouring markers within the relative bins, averaged over all markers in the bin. For comparisons with recombination rate in the human, mouse and cow genome we used the genetic distances and chromosomal sizes described by Myers (2005), Shifman (2006) and Arias (2009). Four groups (Asian wild, European wild, Asian breeds and European breeds) were analyzed separately and correlation coefficients for the relative ROHbin distribution within the groups and the genomic features were calculated and tested for significance by the R (v2.11.1) *cor* and *cor.test* functions with Pearson's product-moment correlation. The between-group differences in outside-ROH-nucleotide diversity, ROHnumber and ROHsize were tested with one-way analysis of variance in R(v2.11.1). Proportional differences of ROHs between groups and uniformity of ROHs over relative chromosomal position were tested with the $\chi 2$ test for proportions and goodness-of-fit in R. All plots were generated with the R (v.2.11.1) lattice package and Ubuntu OpenOffice 3.2.1.

2.5.8 ROH and gene content

Each chromosome was divided into 20 equal sized segments and the relative gene content per segment was calculated. ROHs were grouped according to the three size classes and per class their relative distribution over these chromosomal segments were calculated. Correlations of gene content and ROH content were tested with corr.test in R.

All the annotated porcine genes from *Sus scrofa* (build 10.2 Ensembl release 67), were extracted using Biomart (Haider 2009). The distribution of genes over chromosomes was calculated in a similar way as the ROH occurrence. Each chromosome was divided in 20 equal sized stretches (thus each stretch representing 5% of the chromosome), the total number of genes per stretch was counted and expressed as relative gene content per stretch, proportional to the total gene content on the chromosome. Since the human genome is better annotated, all the human Ensembl orthologues of porcine genes were considered for the gene ontology analysis. BinGO v2.44 (Maere 2005) a Cytoscape v2.8.3 (Shannon 2003) plugin was used to identify over-represented GO terms related to biological processes using the human annotation as background. A hypergeometric test was used to assess the significance of the enriched terms and the Benjamini and Hochberg correction was implemented for multiple comparisons.

2.6 Acknowledgements

We would like to thank the swine genome sequencing consortium for the prerelease of the reference genome build10.2. DNA samples were provided by Dr. Ning Li; China Agricultural University, China; Dr. Alain Duvro, UMR INRA-ENVT, France; Sem Genini, Parco technologico Padano, Italy; Dr. Gono Semiadi, Puslit Biologi, Indonesia; Dr. Naohiko Okumura, Staff Institute 446-1 Ippaizuka,Japan; Dr. Alan Archibald, Roslin Institute and the Royal (Dick) School of Veterinary Studies,University of Edinburgh ,Scotland; Institute of pig genetics BV, The Netherlands; Dr. Oliver Ryder, San Diego Zoo, USA; Cheryl L. Morri, Ph.D., Omaha's Henry Doorly Zoo, USA. We thank prof. dr. Bas J Zwaan, Laboratory of Genetics, and Gus Rose, Animal Breeding and Genomics group, Wageningen UR for editing and discussion.

3

Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations

Mirte Bosse[1], Hendrik-Jan Megens[1], Ole Madsen[1], Laurent A.F. Frantz[1], Yogesh Paudel[1], Richard P.M.A. Crooijmans[1], Martien A.M. Groenen[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, De Elst 1, 6700AH, Wageningen, The Netherlands

Molecular Ecology (2014) 23(16): 4089-4102

Abstract

The merging of populations after an extended period of isolation and divergence is a common phenomenon, in natural settings as well as due to human interference. Individuals with such hybrid origins contain genomes that essentially form a mosaic of different histories and demographies. Pigs are an excellent model species to study hybridization because European and Asian wild boars diverged ~1.2 Mya and pigs were domesticated independently in Europe and Asia. During the Industrial Revolution in England, pigs were imported from China to improve the local pigs. This study utilizes the latest genomics tools to identify the origin of haplotypes in European domesticated pigs that are descendant from Asian and European populations. Our results reveal fine-scale haplotype structure representing different ancient demographic events, as well as a mosaic composition of those distinct histories due to recently introgressed haplotypes in the pig genome. As a consequence, nucleotide diversity in the genome of European domesticated pigs is higher when at least one haplotype of Asian origin is present, and haplotype length correlates negatively with recombination frequency and nucleotide diversity. Another consequence is that the inference of past effective population size is influenced by the background of the haplotypes in an individual, but we demonstrate that by careful sorting based on the origin of haplotypes both distinct demographic histories can be reconstructed. Future detailed mapping of the genomic distribution of variation will enable a targeted approach to increase genetic diversity of captive and wild populations, thus facilitating conservation efforts in the near future.

Key words: conservation genetics, domestication, hybridization, identity by descent, introgression, *Sus scrofa*

3.1 Introduction

Separation and consecutive fusion of populations is common in both natural and managed populations. For instance, the waxing and waning continental ice sheets during the Pleistocene is known to have had a pronounced effect on shaping the population genetics of many species. While the glacial periods usually resulted in refugial populations and thereby promoted population differentiation, the interglacial periods that followed would result in renewed gene flow. Apart from natural causes, populations can also be reunited due to deliberate management. It is well known that the adaptive ability of a population or species to an ever changing environment is mainly determined by its standing variation, and susceptibility to a variety of diseases and environmental changes is assumed to increase if nucleotide diversity is low in the population (e.g. Jimenez 1994, Lacy 1996, Keller and Waller 2002). An increased probability of homozygosity for partially deleterious recessive mutations may lead to individuals with reduced fitness, i.e. inbreeding depression. Such inbreeding effects can be offset by directed population management aimed to facilitate outcrossing, which could result in higher haplotype diversity.

These patterns of reticulation can severely complicate the elucidation of population history. In the past decades, marker systems that have relative fast coalescence and do not or rarely undergo recombination (e.g. mtDNA, Y chromosome) have proven to be useful for phylo-geographic analysis. However, the ensuing pictures of population history that were thus constructed often turned out, or will turn out, to be literally only part of the demographic story. Because autosomes recombine, the genome of a single individual can contain haplotypes from distinct sources, each with another demographic history. Hybridization of populations therefore entails a great challenge to disentangle what has essentially become a mosaic of different demographies. In studies that focused on a relatively small number of nuclear DNA markers, results are usually concatenated to provide a ”genome average”, for instance by doing Structure analysis. Although such analyses may provide insight in the degree of mixing of populations, they do not contain details of the distribution of the introgressed haplotypes over the genome. For instance, the number of generations since the last common ancestor influences the probability of haplotypes in the genome of two individuals to be Identical By Descent (IBD), since the size of the IBD segment declines over time due to recombination. Therefore, the length of IBD haplotypes as a function of local recombination frequency is a measure for the time since the last common ancestor

(Palamara 2012, Ralph 2013, Henn 2012) and signatures of introgression are revealed by coalescence times of haplotypes that are shorter than expected (Staubach 2012). Current high-throughput genotyping and sequencing techniques enable such investigations on a whole-genome scale, providing information on how long ago the reticulation took place. Genomes have been studied in detail to elucidate population history for only a handful of species, e.g. Human (Harris and Nielsen 2013), polar bears (Miller 2012) and pigs (Groenen 2012). However, the effects of admixture in terms of nucleotide diversity on the genome but also on inferences of demographic parameters like past effective population size (N_e) are largely unknown.

Ever since Darwin, domesticated populations have served as important model organisms for evolutionary and population genetic questions (Megens and Groenen 2012). *Sus scrofa* – domesticated pigs and wild boars - is an excellent model species to examine the evolution of genome-wide patterns of haplotype sharing because of its complex but generally well documented demographic history, multiple domestication events and recent admixture between Asian and European breeds. The Eurasian wild boar has its origin in Southeast Asia where it diverged ~3-6 Mya from a clade that gave rise to several other species in the genus *Sus* that are mostly confined to Islands South East Asia (Frantz 2013b, Meijaard 2011). *Sus scrofa* spread throughout the entire Eurasian mainland ~1.2 Mya and an Eastern and Western clade diverged soon after colonization of the West during the cold Calabrian period, in which especially the European population experienced severe bottlenecks (Fang and Andersson 2006, Fang 2006, Alves 2010, Groenen 2012). Domestication of wild boars occurred independently in Europe and Asia, as early as 10,000 years ago and subsequent intensification of the pig breeding industry has led to a variety of breeds (Ottoni 2012, Larson 2005, Kijas 2001, Megens 2008, Groenen 2012). Hybridization between wild and domesticated *Sus scrofa* occurs sporadically nowadays (Giuffra 2000, Goedbloed 2013a), but is likely to have been common until pigs were kept in sties (e.g. Larson 2007b; Herrero-Medrano 2013a, b). Around the late 18th, early 19th century, pigs were imported from Asia to improve local European pigs for key traits such as fertility, growth and fatness.

As a consequence of this hybridization, two very divergent populations, that were separated around 1.2 million years ago, have artificially become merged again. Each of these populations from the Eastern and Western regions of the Eurasian landmass had their own demographic history, with the European wild boar in particular being very much less variable compared to the East Asian wild boars

(Groenen 2012, Bosse 2012), due to founder effects during migration throughout Eurasia and the sequential marginalization in refugia during glaciations. It is historically well documented that pigs from the UK in particular were improved by hybridization with Asian pigs in the 18th, 19th century, and subsequently, due to superior production traits, became founders of a number of the modern commercial pig breeds such as the Large White breed (LW, formally established as a breed in 1868). Therefore, the LW breed serves as an excellent model for studying divergence and subsequent hybridization between populations, since it originated from two highly distinct source populations, that have even been called subspecies (a.o. Groves 2008, Genov 1999), and the hybridization events have been well documented.

The aim of our study is to investigate the consequences of hybridization on genome-wide variation and on disentangling demographic parameters. On a genome-wide segment-by-segment basis we elucidate the origin of the haplotypes in LW pigs, investigating whether they have a Western Eurasian origin or an Eastern Eurasian origin. By this we aim to investigate patterns of introgression and to unravel genomic consequences of isolation and outbreeding. Because the time of divergence between Eastern and Western *Sus scrofa* has been estimated to be around ~1.2Mya (Frantz 2013b, Groenen 2012), Asian wild haplotypes in European commercial pigs are expected to be shorter and less abundant than European wild haplotypes. Since the European population suffered a severe bottleneck, genomic regions for which pigs have one haplotype of Western origin and one of Eastern origin, are likely to show a higher degree of nucleotide diversity than regions for which pigs have two haplotypes that both are of European origin. For comparison purposes, we also investigated the haplotype patterns in an Asian breed, Meishan (MS), as a representative of East Asian pigs. Not only do these pigs represent a domestication event independent from the Western Eurasian pigs, they also represent the demographic history of the East Asian wild boars (up until domestication). Because introgression of Asian haplotypes into European pigs has occurred fairly recently (White 2011, Merks 2012), it is expected that haplotypes shared by European and Asian pigs are longer compared to haplotypes shared by common ancestry in the Western pigs and Asian wild boar. Finally we investigate the effect of the composite nature of the LW genome on demographic inferences like N_e. This analysis of haplotype patterns in pigs provides a detailed insight into the genomic distribution of variation after recent hybridization.

3.2 Materials and Methods

The genomes of 70 domesticated pigs and wild boars were re-sequenced for this study (Table S3.1). These individuals originate from Asia and Europe and form four different functional and geographical groups; Asian wild boars (ASWB), Asian domesticated pigs (ASDom), European wild boars (EUWB) and European domesticated pigs (EUDom). We sequenced two wild boars from Sumatra as outgroup (Groenen 2012). The other Asian wild individuals come from North China (3), South China (4) and Japan (1). The 18 European wild boars originate from the Netherlands, France, Switzerland, Greece and Italy. We sequenced 13 Asian domesticated pigs from the MS, Jianquhai and Xiang breeds and 29 European domesticated pigs from the Duroc, Hampshire, Pietrain, Landrace and LW breeds.

3.2.1 Sampling and preparation

DNA was extracted from whole blood samples from all 70 individuals using the QIAamp DNA blood spin kit (Qiagen Sciences). Quality and quantity of DNA extraction was checked on the Qubit 2.0 fluorometer (Invitrogen). 1-3 ug of genomic DNA was used for the construction of the sequencing library (insert size range 300-500 bp), according to the Illumina library preparation protocol (Illumina Inc.). All samples were 100 bp paired-end sequenced on 1-3 ug of genomic DNA on Illumina HiSeq sequencing systems to a targeted ~10x depth of coverage. Details on all used samples can be found in Table S3.1.

3.2.2 Alignment and variant calling

Reads were quality trimmed to a phred quality >20 for both mates over 3 consecutive bases, and read length were >44 bp after trimming for each mate. Trimmed reads were aligned with the unique alignment option of Mosaik aligner (V. 1.1.0017) to the porcine reference genome build 10.2. SNPs were called for each sample individually with Samtools mpileup 0.1.12a (Li 2009), with the alternative base covered at least 2 times. We filtered the SNPs with VCFtools for a read-depth between 7x and twice the average depth, and discarded SNP sites with a genotype quality<20. We constructed a genotype matrix for all 70 individuals, for those sites that were heterozygous or non-reference in at least 1 individual. We included only sites that were covered >=4x in all the individuals to reduce biases, resulting in a total of 2,377,607 autosomal markers.

3.2.3 IBD detection

We phased all 70 individuals for each chromosome separately, based on the 2,377,607 markers, with Beagle fastPhase (V. 3.3.2). IBD detection between

individuals was executed with Beagle fastIBD for each chromosome, as described in Browning and Browning (2011). We ran 10 independent cycles of phasing and pairwise IBD detection, and merged the identified IBD tracts based on the Beagle probability scores, as suggested (Browning and Browning 2011). Since fastIBD was originally designed for human data, we tested different thresholds for IBD detection to examine which threshold fits our pig data best. We empirically determined that the relative IBD size and number of recorded IBD tracts remained stable with different thresholds, although absolute numbers varied. Our aim was to identify haplotypes that are IBS or IBD, and reflect demographic history over a relatively large time frame. Asian haplotypes are expected to be more diverse and fragmented, and therefore comparatively small in size. Because a higher threshold will enable us to identify Asian wild haplotypes within the genome of the LW pigs, we decided on a threshold of 5.0-6. This is higher than that used in the original paper (Browning and Browning 2011), but this threshold fits our data best.

3.2.4 Haplotype classification

The purpose of the haplotype classification is to be able to infer the geographic origin of the haplotypes that are present in the LW and MS pigs. Shared haplotype tracts were recorded for all pairwise comparisons between the individuals in our matrix. Then, only those haplotypes were extracted from this dataset that were shared between any individual and an individual belonging to either the MS or the LW breed. The haplotypes shared with a MS were grouped into one of three classes, i.e. haplotypes shared between MS and either a) European wild boars, b) European domestics and c) Asian wild boars. The haplotypes shared with a LW pig were also grouped into one of three classes, i.e. haplotypes shared between LW and either a) European wild boars, b) Asian domestics and c) Asian wild boars (figure 3.1). In total there are four reference groups of pigs, but the MS and LW pigs were only compared to three groups because they were not compared to the same group as they belong to themselves. With this setup, we have a total of 6 group comparisons and 351 unique pairwise comparisons between individuals. The rationale of the pairwise comparisons is further described in figure S3.1. The group of Asian breeds included 3 individuals of the MS group, and the group of European breeds included 3 individuals of the LW group. Only six individuals were used from the EUDom group, even though a larger number has been used for the phasing step, to keep number of animals in all four reference groups similar. Because the analysis is based on pairwise comparisons, any individual may share a haplotype with multiple individuals from different pig groups. The average length and number of shared haplotype tracts between the LW or MS pigs and the members of the

four pig groups were computed and significant levels were calculated with a two-sample Kolmogorov-Smirnov test in R version 2.13.1. Recombination frequency was obtained from Tortereau (2012) and correlation with IBD length was calculated with a Pearson's product-moment correlation test in R.

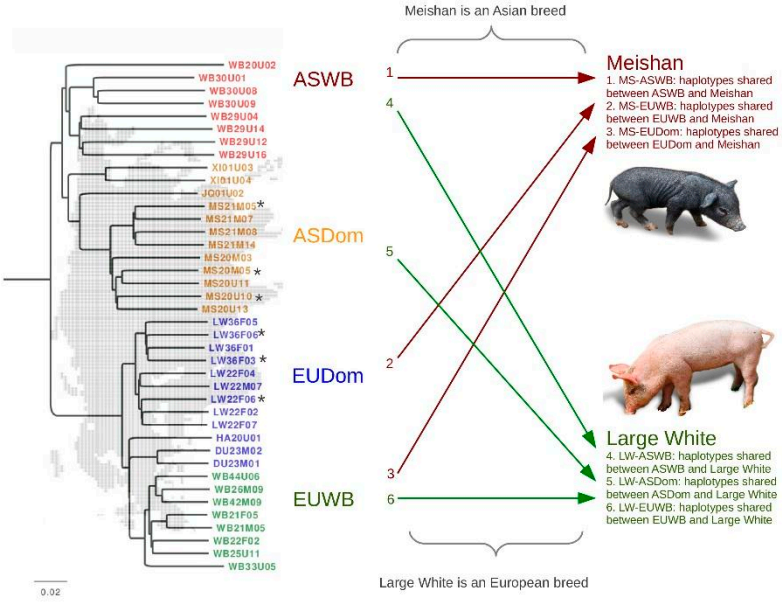


Figure 3.1 Experimental setup for the IBD detection. Arrows indicate all the pairwise comparisons between groups that are used for the IBD detection. Individuals from two breeds (LW: Large white and MS: Meishan) are used for all comparisons with four geographical and functional groups: European wild boars (EUWB), European domesticated pigs (EUDom), Asian wild boars (ASWB) and Asian domesticated pigs (ASDom). All individuals in the phylogenetic tree are used for the pairwise IBD detections. Three individuals from the LW and MS group (indicated with*) are also used for the representation of the EUDom and ASDom group, respectively, which means that they are used for IBD detection in both groups. Total numbers of IBD detection between two individuals for the six pairwise comparisons are: 1–72; 2–72; 3–54; 4–72; 5–54; 6–72 (see also Figure S3.1, Supporting information).

3.2.5 Nucleotide variation

To estimate nucleotide diversity within the individuals on a genome-wide scale, the genome was divided into bins of 10,000 bp and within each bin SNPs were called according to the criteria mentioned above. Nucleotide diversity was calculated as SNPs per called base in the bin (read-depth of 7x to 2 times the average coverage).

To compute the nucleotide diversity in the LW or MS pigs within regions that are IBD with the 4 different pig groups, the IBD tracts that were recorded during the IBD detection were likewise divided into bins of 10,000 bp. Nucleotide diversity within the LW and MS individuals was extracted for these bins as described above. Significance levels were calculated with a two-sample Kolmogorov-Smirnov test in R. We also computed the average nucleotide diversity for entire IBD tracts (without dividing the tracts into bins). The correlation of the nucleotide diversity and length of IBD tracts was calculated with Pearson's product-moment correlation test in R.

3.2.6 F_{st} analysis

We calculated pairwise F_{st} as defined by Weir and Cockerham (1984) in bins of 10,000 bp over the full genome with Genepop 4.2 (Rousset 2008), based on the 2,377,607 SNPs. The pairwise F_{st} between the LW and two wild boar groups (EUWB and ASWB) as well as pairwise F_{st} between the MS and the two wild boar groups (ASWB and EUWB) was computed.

3.2.7 Phylogenetic analysis

A phylogenetic tree was constructed for all the 42 re-sequenced individuals that were used in the pairwise comparisons (figure 3.1, figure S3.1) with the Sumatran *Sus scrofa* INDO22 as an outgroup. A distance matrix was constructed in PLINK (Purcell 2007) for all 2,377,607 genotypes spanning the full genome and a neighbor-joining tree was created in Phylip (Felsenstein 2005). The tree was depicted in FIGTREE (http://tree.bio.ed.ac.uk/software/figtree/).

3.2.8 Admixture analysis

For the admixture analysis, the outgroup individuals from Sumatra were removed, and all bi-allelic sites in the matrix were LD-pruned with the PLINK option –indep with a window size of 50, steps of 5 SNPs and a variance inflation factor of 1.5 and the remaining SNPs were filtered for MAF<0.05. Then an Admixture (Alexandre 2009) analysis, which uses the same statistical model as STRUCTURE (Pritchard, Stephens and Donnelly 2000), was computed for the remaining 68 individuals with K between 2 and 5.

3.2.9 PSMC analysis

The consensus sequence for one LW (LW22F07), one MS (MS20U10) and one European wild boar (WB25U11) was constructed using samtools mpileup and vcftools (Li 2009). To estimate past effective population sizes, we performed a Pairwise Sequential Marcovian Coalescent (PSMC) analysis (Li and Durbin 2011) on

these consensus sequences. Generation time was set at 5 years and mutation rate at $\mu=2.5*10^{-8}$ as used in previous analyses (Groenen 2012, Bosse 2012 and Frantz 2013b). The PSMC analysis was also performed for all three individuals on only those regions of the genome in which the LW contains at least one haplotype shared with ASDom. The same analysis was done for those regions where the LW did not contain an Asian haplotype, but did have a shared haplotype with an European wild boar. These genomic fragments were filtered for regions that contained only a EUWB signal for at least 100 kbp in length, because we expect these calls to be more reliable. Smaller IBD fragments are more difficult to detect and therefore these are more prone to false positives and negatives of Asian heritage, which in turn may influence the effective population size estimates.

3.3 Results

The Asian and European *Sus scrofa* in our dataset formed two distinct clades (figure 3.1). Our two focal populations, the European Large Whites (LW) and the Asian Meishans (MS), both represent the domesticated form on their continent. We show however that the LW contained a proportion of Asian haplotypes in their genome, indicative of the recent admixture probably stemming from the late 18th, early 19th century (figure 3.2). Although the Admixture analysis with K=4 had the highest likelihood (figure S3.2), the analysis with K=2 assigned the Asian or European heritage of the alleles (figure 3.2). The genetic differentiation between the LW and MS populations, measured as Wrights fixation index (F_{st}, Weir and Cockerham 1984), is 0.383(+/- 0.217). F_{st} between LW and ASWB is higher than the F_{st} between LW and the EUWB ($p<0.001$, figure 3.2). By contrast, the F_{st} between the MS and ASWB is lower than between MS and EUWB. The overall F_{st} between LW and the ASWB is lower than F_{st} between the MS and EUWB, which corroborates the Asian introgression. This phenomenon is the initial concept behind our further analyses.

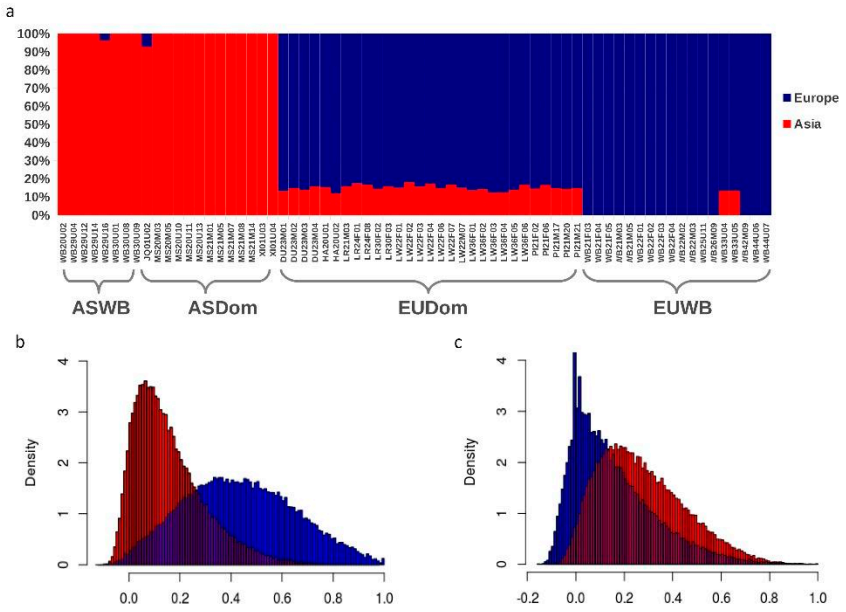


Figure 3.2 Genomic structure and F_{st} of the four main pig groups used for analysis. **A**. The percentage of Asian (red) and European (blue) genetic material is displayed on the y-axis, and each individual is displayed on the x-axis. We examined the genetic structure of all individuals when we force two populations (K = 2) for four groups: Asian wild boars (ASWB), an Asian domesticated breeds (including MS), European wild boars (EUWB) and the European domesticated breeds (including LW). Asian wild boars are the most diverse group in terms of both within-individual and between-individual variation. **B.** Distribution of F_{st} between MS and EUWB (blue) and MS and ASWB (red). F_{st} was calculated for each bin of 10.000 bp over the full genome. **C.** Distribution of F_{st} between LW and EUWB (blue) and LW and ASWB (red).

3.3.1 IBD haplotype occurrence

We extracted shared haplotype tracts between LW or MS pigs and pigs originating from the four wild and domesticated pig groups from Asia and Europe. An example of the distribution of IBD haplotypes in the genome of one LW pig is shown in figure S3.3. This example clearly shows a large proportion of Asian-derived haplotypes in the genome of the LW, sometimes in homozygous state, sometimes occurring together with a European haplotype. The length and number of shared haplotypes shows a distinct pattern for each of the pairwise comparisons as described in figure 3.1. Size and number of all shared haplotype groups differ significantly ($p<0.001$ for all; figure 3.3); the LW share more and longer haplotypes with the European wild boars than with both Asian *Sus scrofa* groups ($p<0.001$ for both).

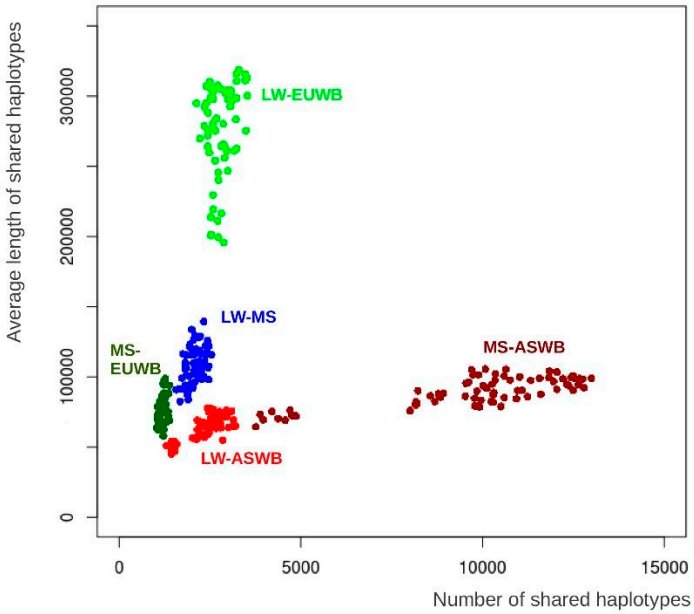


Figure 3.3 Size and number of pairwise IBD tracts. The average size of an IBD haplotype is plotted against the number of pairwise IBD haplotypes for each pairwise comparison between individuals. Coloration is based on five different groups of pairwise comparisons: the LW individuals compared with ASWB (red) and EUWB (light green); the MS individuals compared with ASWB (brown) and EUWB (dark green); and both domesticated groups compared with each other (LW-MS, in blue).

Likewise, the MS share more and longer haplotypes with the Asian pigs than with the European domestics and wild boars ($p<0.001$ for both), in agreement with their independent domestication history. The average size of LW haplotypes that were found to be IBD with the Asian domesticated pigs is significantly larger than the haplotypes shared with the Asian wild boars ($p<0.01$). In addition, the MS-LW haplotypes are, on average, longer than the MS-EUWB haplotypes. Haplotypes that are shared between LW and EUWB are longer than haplotypes shared between MS and ASWB, but the number of the relatively smaller MS-ASWB haplotypes is higher in the MS genome than the number of LW-EUWB haplotypes in the genome of the LW. The occurrence of all IBD haplotypes in the genome of LW pigs is not randomly distributed. For all three groups of IBD haplotypes in the LW, we found a negative correlation between length of the IBD haplotype and recombination frequency ($r=-0.3 +/- 0.12$, $p<0.001$, Pearson's product-moment correlation, example in figure 3.4).

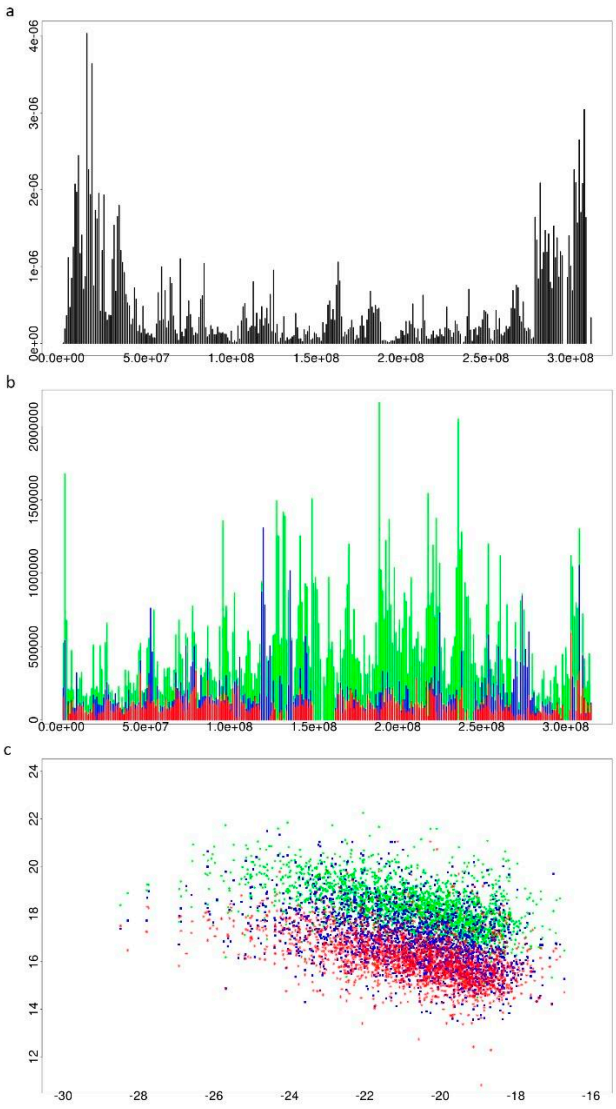


Figure 3.4 Recombination frequency and average length of IBD haplotypes over chromosome 1. Example of the distribution of recombination frequency and the average length of IBD haplotypes over the full length of chromosome 1 in bins of 1 Mb. **A.** Recombination frequency per bin of 1 Mb (Tortereau 2012), the x-axis displays location on chromosome 1. **B.** Length of IBD fragments in LW shared with EUWB (green), ASDom (blue) and ASWB (red) over chromosome1. **C.** Correlation between log-transformed recombination frequency (x-axis) and log-transformed haplotype length (y-axis) over all autosomes for haplotypes in LW shared with EUWB (green), ASDom (blue) and ASWB (red).

We compared the distribution of shared haplotypes with ASDom over the full genome for all 9 LW pigs. On a population wide scale most parts of the genome contain at least one Asian haplotype (Figure S3.4), but some parts of the genome contain no Asian haplotype and others are relatively high in Asian haplotype frequency. The regions in the genome without Asian haplotypes are longer in the middle of the chromosomes, which is in line with the observed correlation of haplotype length and recombination frequency.

3.3.2 Nucleotide diversity

We define nucleotide diversity in this paper as the proportion of SNPs between the two haplotypes of an individual in a particular region of the genome, relative to all the sites called in that region. The nucleotide diversity within the genome of an individual was computed for all LW and MS pigs. Average nucleotide diversity was higher within the genomes of the MS pigs than within the LW pigs ($p<0.001$). The geographic origin of the haplotypes influenced the local nucleotide diversity in the genome. Figure 3.5A shows an overview of the nucleotide diversity within one LW pig, over the full length of chromosome 1. Relatively recent consanguineous matings are reflected as Regions of Homozygosity (ROH) on this chromosome. The diversity between two haplotypes on chromosome 1 for this pig was significantly higher when at least one haplotype was shared with an Asian pig or Asian wild boar ($p<0.001$, figure 3.5A-E). The same pattern was observed when we extrapolated this to a genome-wide scale for all LW pigs ($p<0.001$, figure S3.5A). Those genomic regions in the LW that share at least one haplotype with an European wild boar are relatively less diverse than the regions that share at least one haplotype with the European domesticated pigs ($p<0.001$), but note that these regions are not mutually exclusive. All distributions of genome-wide nucleotide diversity contain multiple peaks at low nucleotide diversity, showing the presence of homozygosity in the genome, regardless the origin of the present haplotypes (figure S3.5A-B). A negative correlation can be observed between length of the IBD fragment and nucleotide diversity in the fragment ($r=-0.26$, $p<0.0001$, figure S3.5B). The strongest correlation was found for LW-EUWB haplotypes ($r=-0.35$).

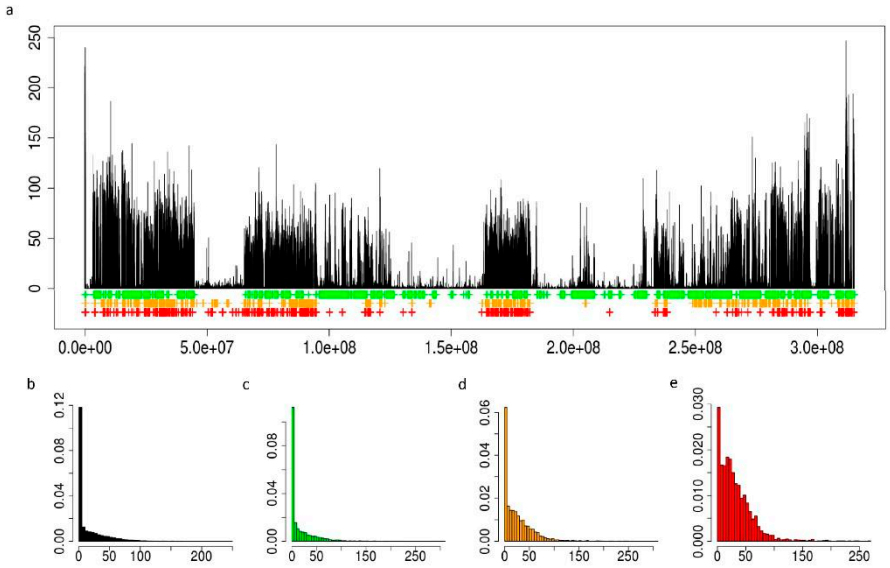


Figure 3.5 Example of nucleotide diversity and shared haplotypes for a LW pig over chromosome 1. **A.** The nucleotide diversity between two haplotypes per bin of 10Kbp is plotted against the physical position of the bin on chromosome 1. Regions in the genome in bins of 10Kbp where one of the two haplotypes is IBD with EUWB (green), ASDom (orange) or ASWB (red) are indicated in bars beneath the original plot. **B.** Histogram of nucleotide diversity between the two haplotypes in this individual in bins of 10Kbp. **C-E**. Histogram of nucleotide diversity between the two haplotypes in this individual in bins of 10Kbp, where at least one of the two haplotypes is IBD with EUWB (green, **c**), ASDom (orange, **d**) or ASWB (red, **e**).

The past effective population size was estimated for the full genome of one LW pig, one French wild boar and one MS (figure 3.6). Although the LW breed is known to be domesticated from the European wild population, its past effective population size is estimated to be larger than that of the French wild boar (figure 3.6A, B). When the same analysis is done for regions where the genome of this LW pig has a European haplotype (and no Asian), the population size for the LW is lower than for regions where the LW has an Asian haplotype (figure 3.6A). However, the French wild boar and the MS have the same estimated population size when estimated for these regions as compared to their full genome (figure 3.6B, C), which suggests there is no effect on the estimate of N_e due to the regions in the genome that these haplotypes were extracted from.

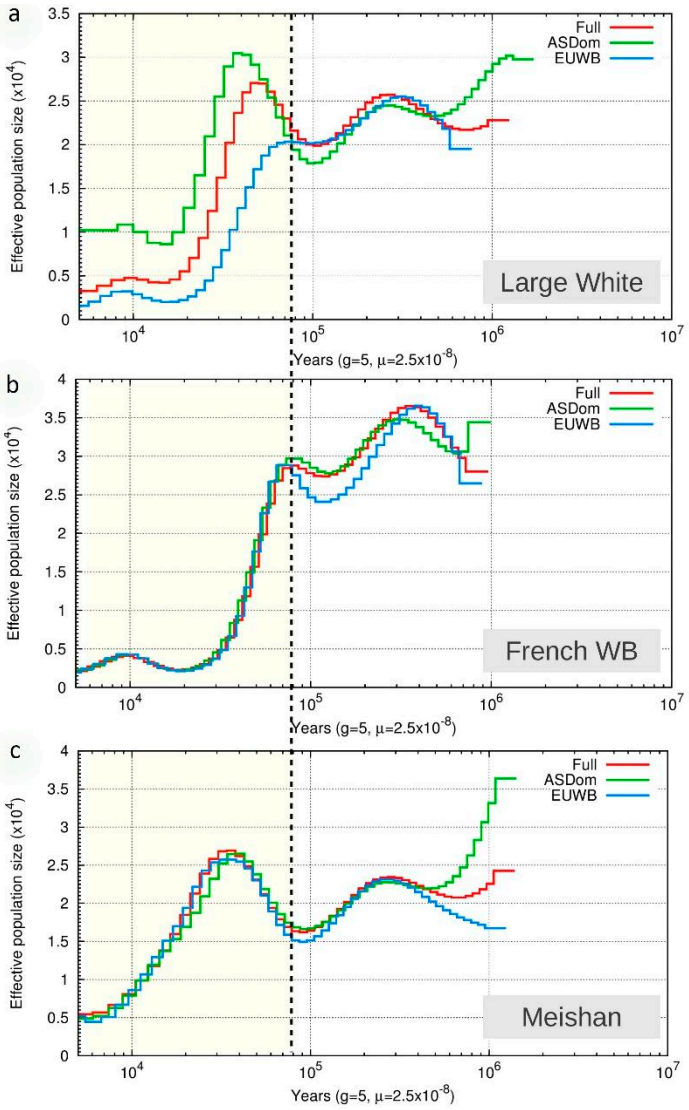


Figure 3.6 Estimates of past effective population size based on parts of the genome with different demographic history for three pigs. Red lines indicate the effective population size when a pairwise sequential markovian coalescent analysis is run on the consensus sequence for the full genome. Green lines are estimates for N_e based on only those segments in the genome where the LW pig contains an Asian haplotype, and blue lines indicate the N_e when the PSMC is run on only those segments where the LW contains an European wild haplotype and no Asian haplotype. These regions are filtered for a length of at least 100 kb, and therefore, we consider only the left side of the dotted black line to be accurate N_e estimates. **A.** N_e estimated for the LW individual LW22F07. **B.** N_e estimated for the French wild boar WB25U11. **C.** N_e estimated for the MS pig MS20U10.

3.4 Discussion

3.4.1 IBD haplotype occurrence

Multiple studies have shown that domestication of pigs took place at least twice, and independently, in Western and Eastern Eurasia (Larson 2005, Kijas 2001). In both cases, it was the local wild boar that was domesticated, and subsequent hybridizations with local wild boar populations as agricultural practices spread have been documented from ancient DNA studies (e.g. Larson 2005, 2007a). Pigs, therefore, represent a subset of the natural variation present in the east and west of the natural range of the wild boar that encompasses the Eurasian supercontinent. The LW breed used in this study serves as a model in which these diverged populations have been reunited. Incidentally, an ideal model for ancestral European pigs is not available, as it is presumed that even most of the traditional heritage breeds in Europe may have been influenced by Asian pigs over the past two centuries albeit probably indirectly through improving pigs using popular commercial stock. Intriguingly, the European wild boar, despite representing the non-domesticated form, may therefore be the best model for the ”original”, pre-18th century European pigs alive today.

The Asian breed for which we were able to obtain the largest number of sequenced individuals was the MS. MS pigs, as far as is known, have never been crossed with European pigs, and therefore serve as a good model for the imported Asian pigs. The F_{st} between MS and European *Sus scrofa* is higher than between MS and ASWB, confirming two independent domestication centers. Incidental exchange of genetic material between European and Chinese pigs has been suggested to happen as early as during Roman times (Porter 1993). However, during the intensification of Northern European agriculture in the eighteenth century, pig breeding expanded from forested areas to more urban environments, resulting in a changing selection pressure on multiple traits. Since in particular the European breeds that were crossed with Asian breeds seemed to perform best in this relatively new environment, there was an extensive period of genetic exchange between multiple European breeds in the early nineteenth century (White 2011). It was during this time of experimental crossing and breeding that the first modern pig breeds with mixed English and Asian origin emerged. Therefore it is not surprising that the proportion of Asian material that we identified in these breeds is roughly similar. The genetic signature of introgression from Asian into European pigs was first discovered using mtDNA sequence data (Giuffra 2000), and many pig

populations, particularly certain commercial breeds derived from British breeds such as LW, were found to contain large proportions (>>50%) of Asian-derived mitochondrial haplotypes. Interestingly, Asian Y-chromosome haplotypes appear to be very rare in European pigs, which suggests that the introgression was predominantly female-driven (Ramirez 2009). Our F_{st} analysis shows a greater divergence between LW and Asian individuals than between LW and EUWB, confirming that indeed there is an asymmetry in the hybridization event. The Asian component is less than half, which is consistent with earlier findings based on full genome data that suggested that the Asian fraction in European modern pigs can be up to 35% (Groenen 2012), and mitochondrial studies that estimate the average proportion of Asian mitochondrial (mt) haplotypes in European breeds at ~29% (Fang and Andersson 2006). However, the proportion of Asian mt haplotypes can vary considerably between breeds with Duroc and Hampshire containing less Asian mt material than Large White. In this study we show that the average Asian component in the autosomes is very similar in individuals belonging to different European breeds, suggesting that autosomal and mtDNA tell different stories regarding the introgression history as has been proposed previously (Ramirez 2009). Our Admixture analysis confirms the global introgression of Asian material into European breeds, although the estimated fraction is somewhat lower (~20%). The introgression of Asian domesticated breeds into European breeds may have reduced the genetic differentiation (F_{st}) between the LW and ASWB compared to the MS and EUWB.

If no hybridization had taken place since the original split of European and Asian *Sus scrofa*, one would expect that the shared haplotypes between all European and Asian *Sus scrofa* are similar in size and abundance, regardless of the domestication status of the individuals. The fact that haplotypes shared between LW and ASDom are, on average, larger than haplotypes shared between LW and ASWB, is an indication that indeed Asian domesticated haplotypes have been introgressed at a later stage. These differences in length in particular indicate a more recent common ancestor between these haplotypes, since haplotypes are broken up in time due to recombination. There is, however, a proportion of LW-ASDom haplotypes that have lengths overlapping the length distribution of the LW-ASWB haplotypes. These haplotypes may represent the original split between European and Asian wild boars, i.e. incomplete lineage sorting between Eastern and Western *Sus scrofa*. More IBD tracts are found between MS and ASWB than between LW and EUWB and haplotypes shared between EUWB and LW are, on average, longer. The difference in length between MS-ASWB haplotypes and LW-EUWB haplotypes

can be a signal of smaller effective population size in the EUWB and the LW (i.e all European *Sus scrofa*), resulting in a smaller haplotype diversity in European wild boars, compared to Asian *Sus scrofa*.

3.4.2 Nucleotide diversity

Because the effective population size in Asia was larger and the latest glacial bottlenecks was not as severe in Asia compared to Europe, on average, two Asian haplotypes drawn from a population are thought to be more divergent than two European haplotypes (Groenen 2012). Secondly, the relative old divergence (~1.2 Mya) between Asian and European *Sus scrofa*, is expected to result in more variation between haplotypes of Asian and European origin than any two European haplotypes. Therefore, those parts in the genome of LW pigs where an Asian haplotype has been detected were expected to be more diverged, as was corroborated in this study. It has been shown previously that European domesticated pigs contain more variation than European wild boars (Groenen 2012). The higher nucleotide diversity in the regions that contain an Asian haplotype in the LW pigs compared to regions that contain two European haplotypes, suggests that the higher diversity is due to hybridization with Asian individuals, rather than a post-domestication bottleneck in the European wild boar population. Both breeds show some regions of low variation, probably due to recent inbreeding resulting in ROH formation. ROHs tend to be longer and more abundant at the center of the chromosomes, chiefly following the distribution of recombination frequency (Bosse 2012). In our IBD analysis we show that length of shared haplotypes follows the same pattern and correlates negatively with recombination frequency. The negative correlation between length of the IBD fragment and nucleotide diversity in the fragment might also be influenced by the recombination frequency, since higher nucleotide diversity tends to be found in regions of high recombination (Bosse 2012). Less recombination results in longer haplotypes, and in many species there is a positive correlation between recombination rate and nucleotide diversity [i.a. Begun and Aquadro 1992, Fang 2008, Lercher and Hurst 2002], which may explain the negative correlation in pigs as well. Another explanation for this observation can be that long IBD fragments are an indication of (recent) selection for a particular haplotype, resulting more often in homozygosity, regardless the background of the haplotype. Figure S3.4 shows that some regions in the genome of the LW pigs are enriched in Asian haplotypes, while other parts do not contain any Asian material at all. Such variation of Asian haplotypes may also hint towards selection, which can be expected in a hybrid population that carries very divergent haplotypes. After

introgression, Asian haplotypes may either have had a neutral effect, could have been beneficial or have had a negative effect. This study focuses on the consequences for nucleotide diversity in the genome and inference of demographic history under a neutral introgression scenario, which is observed in the majority of the genome resulting in the general patterns that are described. However, introgression mapping could also be used to screen for regions with an excess of heterozygosity within individuals with introgressed and non-introgressed haplotypes, in order to detect regions under balancing selection, as has been shown for the major histocompatibility complex (Charbonnel and Pemberton 2005, Castric 2008, Abi-Rached 2011). Also regions with a lack of introgressed haplotypes or with more introgressed haplotypes than expected could answer interesting questions about selection in the focal population after introgression, as has been shown for i.e. mouse (Song 2011), and human (Jeong 2014). Our results clearly show that the genomes of LW and MS pigs are a mosaic of haplotypes, representing a variety of demographic and selection events.

We have shown that the genomes of LW pigs have a composite origin in which European and Asian haplotypes are combined. This phenomenon has important implications for demographic analyses on these genomes, since a single individual essentially represents multiple, distinct, demographies. By running a Pairwise Sequential Marcovian Coalescent analysis (PSMC, Li and Durbin 2011) on different fragments of the genome of one LW pig, we showed that it is possible to disentangle these separate demographies if the origin of the haplotypes can be properly assigned. If no introgressed haplotypes are included in this inference, the effective population size indeed resembles that of the source of domestication (the European wild boar). The effective population size is however greatly overestimated if one of the two haplotypes originated from an Asian pig. The effective population size of the LW resembles that of the MS in those genomic regions where European and Asian haplotypes are combined in the LW and used for the PSMC analysis. Since the European wild boars are descendants from Asian wild boar, the majority of genetic variation that is present in the European wild boar population has its origin in Asia. Therefore, one European and one Asian haplotype could indeed approximate the Ne estimates for two Asian haplotypes, as is found in the MS. Newly arisen mutations in the European and Asian clades after the original split will probably have resulted in a slightly higher Ne estimate when an Asian and European haplotype are combined, as we see in the LW, than when the N_e is based on two Asian haplotypes. These findings highlight the importance of

knowledge on the background of samples when these types of analyses are used to infer the demographic history of a population.

A combination of recombination, genetic drift, selection and introgression has resulted in a complex distribution of haplotypes in the two breeds. Knowing the genomic footprints of admixture can be used in commercial breeding and conservation management to increase the variation within populations. Introducing new haplotypes from one inbred population to another, highly divergent but also inbred population, may result in a strong increase in variation within the genomes of hybrid individuals. Detailed mapping of the genomic distribution of variation enables a targeted approach to increase genetic diversity of captive and wild populations, by selecting individuals that contain particular desired haplotypes in breeding programs. However, the identification of introgressed haplotypes may also be used in breeding efforts that intend to “purify” a particular breed or population. The integrity of a population can be very important for branding of particular regional products for example (e.g. Herrero-Medrano 2013a, b), but also for species conservation (e.g. Frantz 2013b). When the contribution of introgressed haplotypes to future generations can be actively managed, this approach may facilitate conservation and breeding efforts in the near future.

3.5 Acknowledgements

DNA samples were provided by Dr. Ning Li; China Agricultural University, China; Dr. Alain Ducos, UMR INRA-ENVT, France; Sem Genini, Parco technologico Padano, Italy; Dr. Gono Semiadi, Puslit Biologi, Indonesia; Dr. Naohiko Okumura, Staff Institute 446-1 Ippaizuka, Japan; Dr. Alan Archibald, Roslin Institute and the Royal (Dick) School of Veterinary Studies, University of Edinburgh, Scotland; Institute of pig genetics TOPIGS BV, The Netherlands; Dr. Oliver Ryder, San Diego Zoo, USA; Cheryl L. Morri, Ph.D., Omaha's Henry Doorly Zoo, USA. This project is financially supported by the European Research Council under the European Community's 256 Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n° 249894". We thank Barbara Harlizius and Naomi Duijvesteijn from TOPIGS for valuable discussion.

4

Hybrid origin of European commercial pigs examined by an in-depth haplotype analysis on chromosome 1

Mirte Bosse[1], Ole Madsen[1], Hendrik-Jan Megens[1], Laurent A.F. Frantz[1], Yogesh Paudel[1], Richard P.M.A. Crooijmans[1], Martien A.M. Groenen[1]

1 Animal Breeding and Genomics Centre, Wageningen University, The Netherlands

Frontiers in Genetics (2014) 1664-8021

Abstract

Although all farm animals have an original source of domestication, a large variety of modern breeds exists that are phenotypically highly distinct from the ancestral wild population. This phenomenon can be the result of artificial selection or gene flow from other sources into the domesticated population. The Eurasian wild boar (*Sus scrofa*) has been domesticated at least twice in two geographically distinct regions during the Neolithic revolution when hunting shifted to farming. Prior to the establishment of the commercial European pig breeds we know today, some 200 years ago Chinese pigs were imported into Europe to improve local European pigs. European domesticated pigs are genetically more diverse than European wild boars, although historically the latter represent the source population for domestication. In this study we examine the cause of the higher diversity within the genomes of European domestic pigs compared to their wild ancestors by testing two different hypotheses. In the first hypothesis we consider that European domestics are a mix of different European wild populations as a result of movement throughout Europe, hereby acquiring haplotypes from all over the European continent. As an alternative hypothesis, we examine whether the introgression of Asian haplotypes into European breeds during the Industrial Revolution caused the observed increase in diversity. By using re-sequence data for chromosome 1 of 135 pigs and wild boars, we show that an Asian introgression of about 20% into the genome of European domesticated pigs explain the majority of the increase in genetic diversity. These findings confirm that the Asian hybridization, that was used to improve production traits of local breeds, left its signature in the genome of the commercial pigs we know today.

Key words: *Sus scrofa*, hybridization, domestication, introgression, genetic variation, haplotype, homozygosity

4.1 Introduction

Domestication is a complex process that has major implications for both phenotypic and genetic variation. It is not an exception that the domesticated form appears to be very different from the wild species in terms of phenotype and genetic makeup. Examples include multiple crop species (Doebley 2006), dogs (VonHoldt 2010) and farm animals (Andersson 2001). The differences are caused mainly by two phenomena: 1) selection for particular desired traits in the domesticated population including domestication genes, which facilitate the maintenance of the species in question or have commercial interest; 2) hybridization with individuals from highly divergent populations to improve selected traits. The domesticated pig (*Sus scrofa*) is a good example of such a species, since the domesticated form as well as its wild relatives are widespread across the Eurasian continent but can be phenotypically highly distinct. Domestication of the pig is known to have its origin independently in the Near East and in Asia roughly 10.000 years ago (ya), which lead to at least two distinct domestication clades (Kijas 2006, Larson 2005).

Strong artificial selection after the initial domestication lead to a wide variety of breeds, each with distinct phenotypes, and selective signatures in the genome (Rubin 2012, Wilkinson 2013). Breed formation and artificial selection for particular traits can drastically reduce genetic diversity, which has been shown for multiple species (Taberlet 2008, Kristensen and Sorensen 2005). Surprisingly, in pigs the commercial breeds in Europe are generally more diverse than their wild counterpart (Groenen 2012, Bosse 2014a). In this research we examine which process contributed most to the difference in genetic diversity between European commercial breeds and European wild boar.

In Europe, pig domestication did not occur as a single, unique event, but rather was a continuous process of domestication, isolation and hybridization that lead to current European pigs (Larson 2007b). Furthermore, glaciations likely had a major impact on the genetic diversity in European wild boar (Scandura 2008). It has been suggested that there were multiple refugia in Europe during the last glaciation, resulting in many private haplotypes for the separate populations (Alves 2010). In the drawn-out process of domestication of the pig in Europe, the mixing of wild boar genetic variation from different regions in Europe, might explain the high diversity found in modern European pigs. Although variation has been lost locally in most European wild populations, the combined genetic diversity from

geographically isolated populations should display similar patterns of genetic diversification as is shown for European commercial haplotypes. The first hypothesis we test, therefore, is that the European breeds are a combination of separate European populations that have been amalgamated into a single population, resulting in higher levels of variation.

Introgression from Asian pigs into European breeds was first demonstrated with molecular data by Giuffra (2000), and indeed multiple international breeds have subsequently been found to contain Far Eastern mitochondrial haplotypes (Clop 2004, Fang and Andersson 2006). Ramirez (2009) suggested that this introgression was mostly female driven, because of the predominance of the European HY1 Y-chromosomal haplotype in Europe. An Asian origin for multiple commercially important phenotypes has been shown as the result of this hybridization (Ojeda 2008, Wilkinson 2013, Bosse 2014b and Hidalgo 2014). Alves (2003) showed that not all European local breeds, such as Iberian pigs, contain mtDNA of Asian origin, and based on genomic DNA varying levels of admixture in local breeds have been suggested (Herrero-Medrano 2013b). We recently found that parts in the genome of Large White pigs that contain DNA that is shared with Asian pigs are generally more diverse than regions that do not share DNA with Asian pigs (Bosse 2014a). Whether this is a direct result of the introgression (rather than, for example, incomplete lineage sorting) and the overall contribution by Asian haplotypes to variation, remain unanswered questions. Thus, the second hypothesis we test is that Asian introgression leads to higher diversity in the European commercial pigs.

For prioritizing farm animal genetic resources (FanGR) for conservation, it is important to know the distribution and the origin of variation in the (domesticated) species (Groeneveld 2010). With this work, we make a contribution by analyzing the details of genetic diversity on chromosome 1 within and between groups of pigs and wild boar in Asia and Europe.

4.2 Materials and Methods

4.2.1 Data

The data that we used for this paper consists of all variants on chromosome 1, 2 and 18 that were observed in 136 pigs. These variants were previously deposited into dbSNP (release 138). This data was obtained by aligning Illumina paired-end 100bp reads to the *Sus scrofa* reference genome (build 10.2) with Mosaik Aligner

(V.1.1.0017). Reads were trimmed to a minimum base PHRED quality of 20 averaged over 3 consecutive bases and only mate pairs with both reads at least 45 bp in length were included. Each individual was sequenced to ~10x depth of coverage. SNPs were called separately per individual with SAMtools (V. 0.1.13) pileup with a minimum coverage of 4x, with at least 2 reads supporting the alternative allele. Sites were filtered for a minimum genotype and mapping PHRED quality of 20. Most of our analyses were based on all 2,747,210 variants called on chromosome 1. From the original matrix containing all variable sites in all 136 pigs, indels were excluded and SNP loci were retained if called in >80% of all individuals. The minimum coverage of genotypes called within each group of pigs was set to >80%, resulting in 410,237 high-quality SNPs on chromosome 1. All individuals were imputed and phased for these 410,237 SNPs with Beagle v.3.3.2. Although it is unsure whether these two haplotypes represent the actual phases, we considered them as one full-length haplotype because the uncertainties in phase should be balanced out when we calculate homozygosity rates for all pairs of haplotypes in the dataset and look at the distributions. We pooled the haplotypes from pigs belonging to the 8 groups listed in table 4.1.

4.2.2 Phylogenetic analysis

To assess the relationship of haplotypes in our dataset, we constructed a phylogenetic tree based on the phased haplotypes. Each haplotype was considered as an independent sample, so that haplotypes belonging to the same individual do not necessarily need to cluster together. Because missing sites were imputed with Beagle, no missing alleles were present in the phased haplotypes. Sites with more than two alleles were removed from the data and a distance matrix was constructed in PLINK (Purcell 2007). NEIGHBOR (PHYLIP V. 3.695; Felsenstein 2005) was used to build a neighbour-joining tree for all haplotypes using two Sumatran *Sus scrofa* as outgroup, and the tree was depicted using FIGTREE (http://tree.bio.ed.ac.uk/software/figtree/).

Table 4.1 Number and haplotypes per group and background of sequenced individuals. The group name of the pigs under 'group' is how this group of individuals is referred to in the rest of the text. The codes of all pigs correspond to their labels in figure 4.1. The details of the populations or breeds that the pigs belong to are summarized in the column 'Population details'. Note that information for the European local and Asian local individuals can be limited, and therefore these are rather heterogeneous groups.

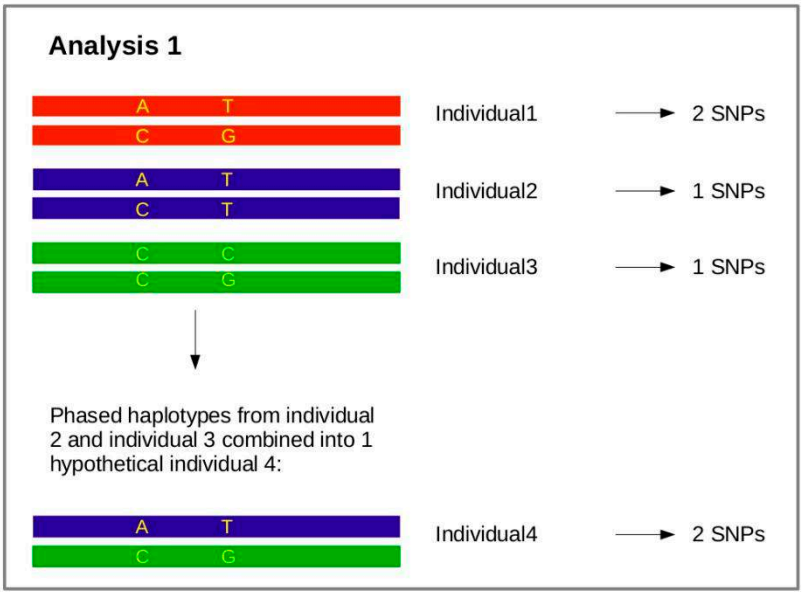
group	No. haplotypes	Codes	Population details
Outgroup	4	INDO	(wild) Sumatran *Sus scrofa*
European local	32	AS,BB,BK,BS,GO,LB,LE,LS,MW,TA,NS	Heritage breeds (Old Britisch breeds), Less global breeds (Linderodsvin, Bunte Bentheimer, Angler Sattelschwein, Leicoma, Nera Siciliana)
European Iberian	22	CA,CM,CS,CT, MA,NI	Pigs from the Iberian peninsula (Mangalica, Negro Iberico, Casertana, Chato Murciano, Calabrese, Cinta Senese)
European commercial	94	DU,HA,LR,LW, PI	Widespread commercial breeds (Duroc, Hampshire, Landrace, Large White, Pietrain)
European wild	52	WB21,22,25,26,28,31,32,33, 42, 44,72	Wild boar from Western, South-Eastern and Southern Europe (Netherlands, France, Spain, Italy, Switserland, Greece, Samos, Armenia)
Asian commercial	30	JQ,MS,XI	Asian breeds known to be commercially important (Meishan, Xiang, Jianquahai)
Asian local	18	JI,LSP,TH,WS,ZA	Local breeds and wild pigs (Jinhua, Leping spotted, Wannan spotted, Zhang, Thai)
Asian wild	20	WB20,29,30	North China, South China, Japan

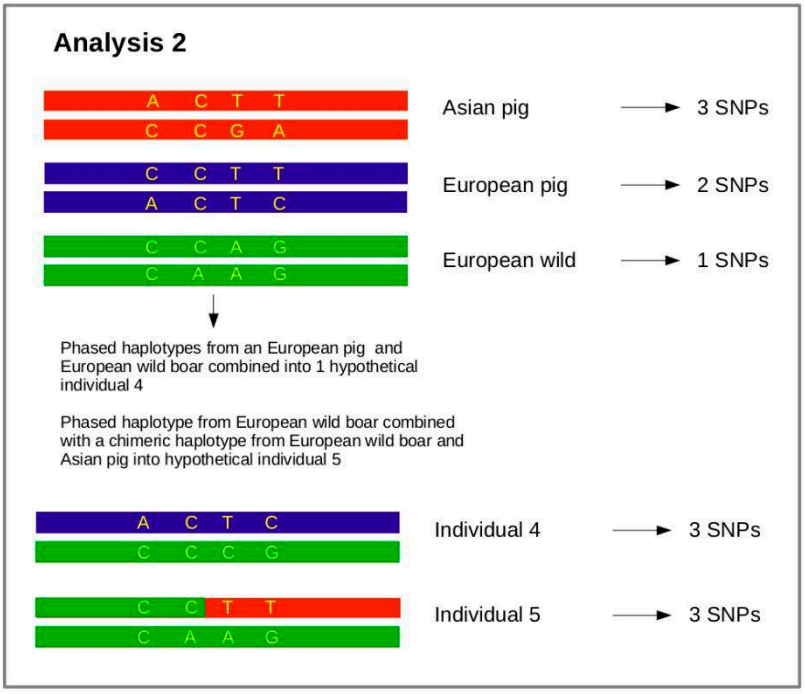
4.2.3 Haplotype homozygosity analysis

Analysis 1 After individuals were phased for the full length of chromosome 1, we analysed the homozygosity between two haplotypes spanning the full chromosome for all possible combinations of two haplotypes in the dataset. Haplotype homozygosity is defined as the proportion of homozygous sites between two paired haplotypes, and ranged between 0 and 1. We calculated haplotype homozygosity as the proportion of all sites (410,237) that occurred in homozygous state, so that 0 represents only heterozygous loci and 1 represents complete homozygosity between both haplotypes. We then paired all possible combinations of two haplotypes in the dataset and determined the homozygosity of these hypothetical diploid individuals in R (see box 4.1). We pooled the haplotype homozygosity for pairs of haplotypes belonging to the same group (table 4.1), so that we ended up with a distribution of homozygosity within a group that represents the full range of variation between haplotypes in a group. Within-group haplotype homozygosity was then compared between the different groups. In the second part of this analysis we paired haplotypes from different groups and computed the haplotype homozygosity for these mixed pairs, so that we have a distribution of homozygosity between haplotypes from two different groups which is compared with the distribution of homozygosity between haplotypes from two other groups

Analysis 2 Previous estimates on the fraction of Asian DNA ranged from 20 to a maximum of 35% (Bosse 2014b, Groenen 2012). In the second analysis we want to assess the influence of Asian introgression into a European haplotype. In order to do this, we simulated introgression by transferring 15, 20 and 25% of a haplotype belonging to the Asian commercial group into a haplotype that belongs to the European wild group (see box 4.1). We used a custom perl script to construct these chimeric haplotypes in which 15, 20 or 25% of the alleles coming from an Asian commercial haplotype replace the alleles in a European wild haplotype. All possible pairs between European wild and Asian commercial haplotypes to construct a chimeric haplotype were included. Then, these chimeric haplotypes were again paired with all possible European wild haplotypes (except for the one that the chimeric haplotype is constructed of) and the homozygosity between the two haplotypes was calculated as described for analysis 1. These haplotype homozygosities were pooled so that we obtained a distribution of haplotype homozygosity in the artificially created Asian-European hybrids.

Box 4.1 Principles of the analyses





4.2.4 Consistency over chromosomes

All analyses presented in this paper are based on haplotypes spanning the full length of chromosome 1. We selected this chromosome because it is the longest pig chromosome and therefore the introgression signals are probably most representative for the full genome and less prone to occasional aberrations due to a limited recombination/drift. However, to check whether chromosome 1 is representative for the complete genome, we compared the haplotype homozygosities for the same pairs of individuals between chromosome 1 and two other chromosomes: chromosome 2 (the second longest chromosome), and the shortest and acrocentric chromosome 18. We tested the correlation coefficient between the haplotype homozygosities of the different chromosomes with Pearson's product-moment correlation in R.

4.2.5 Runs of homozygosity

We extracted runs of homozygosity (ROH) from all combinations of paired haplotypes coming from the European pigs and wild boars. ROHs were called with the –homozyg option using PLINK v1.07, allowing for one heterozygous site in the ROH and a minimum ROH size of 10Kb.

4.3 Results and discussion

4.3.1 Variation within groups

We analyzed the phylogenetic relationship of all haplotypes spanning the full chromosome 1 by constructing a neighbor-joining tree (figure 4.1). The Asian and European haplotypes form two distinct clusters, which is consistent with the hypothesis of independent domestication (Kijas 2006, Larson 2005, Groenen 2012 and Ramirez 2014). European wild boars constitute a monophyletic clade within the European domesticated pigs. The pig reference genome sequence (Groenen 2012) clusters within a group of Duroc pigs, which is expected because the reference genome is based on a female Duroc. The Chinese commercial and local haplotypes cluster with the Northern and Southern Chinese wild haplotypes. The only exception is the Zhang pig, which is closer to European pigs (labelled “ZA” in figure 4.1). This individual is possibly introgressed with European breeds and therefore we mention explicitly when this individual is included in the analysis. Haplotypes from the same individual generally cluster together, but within the European commercial group this is not always the case, showing the close

relationship of these individuals. The Japanese wild boar ("WB20") and the Mangalica pigs ("MA") are the most inbred individuals, with homozygosity between the two haplotypes within each individual above 0.99. Branches within the Asian cluster are longer than those for European haplotypes.

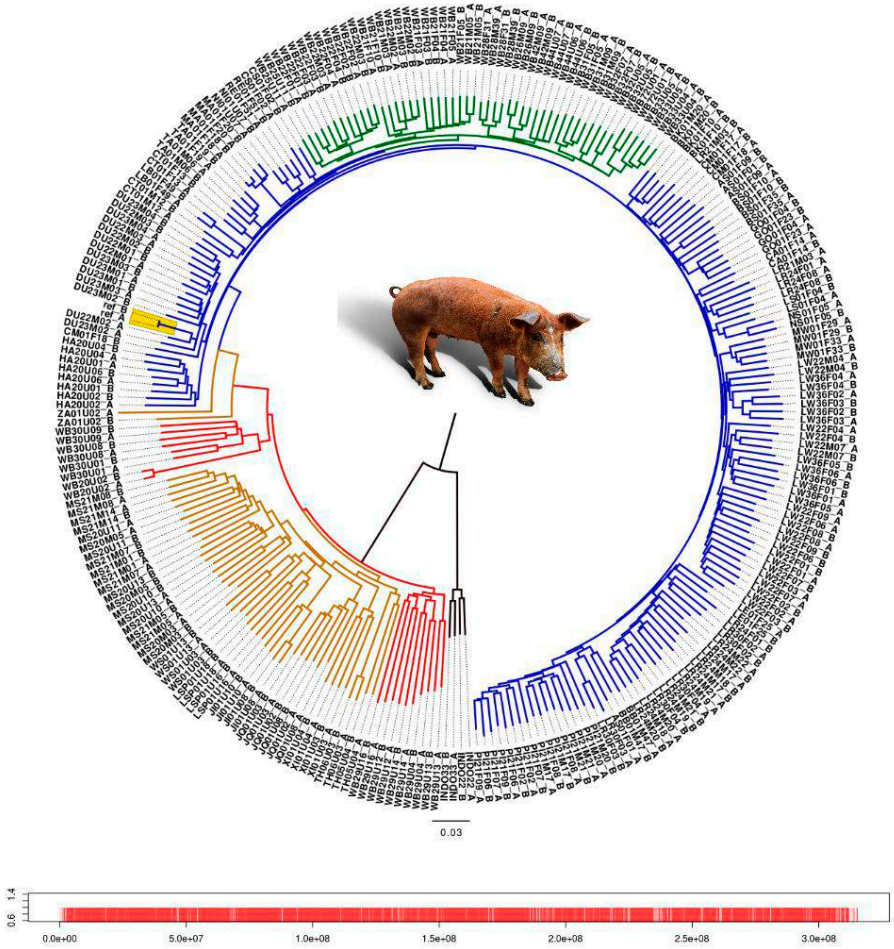


Figure 4.1 Neighbor-joining tree of all haplotypes of chromosome 1. Each individual has 2 haplotypes, one labeled after the name of the individual with the suffix "A" and the second haplotype contains the suffix "B". Red line=Asian wild haplotype; orange line =Asian commercial or local haplotype; blue line=European commercial or local haplotype; green line=European wild haplotype. Locations of the markers on chromosome 1 are indicated by red bars. Alleles from the pig reference genome are included as two separate haplotypes without variation between them, and are highlighted in yellow.

When the homozygosity between two haplotypes from individuals with the same background is measured, the variation (within groups) between two Asian haplotypes is indeed higher than between two European haplotypes from the same group, except for the Japanese wild boar (figure 4.2). This is congruent with previous findings that *Sus scrofa* has its origin in Asia (Groenen 2012, Frantz 2013b) and that European pigs experienced a stronger bottleneck during the last glaciation, resulting in reduced variation (Bosse 2012). Independent domestication should lead to Asian domesticated pigs being more variable than European pigs, which has been shown previously based on microsatellite data (Megens 2008) and sequence data (Bosse 2012) and is also supported by our analysis (figure 4.2).

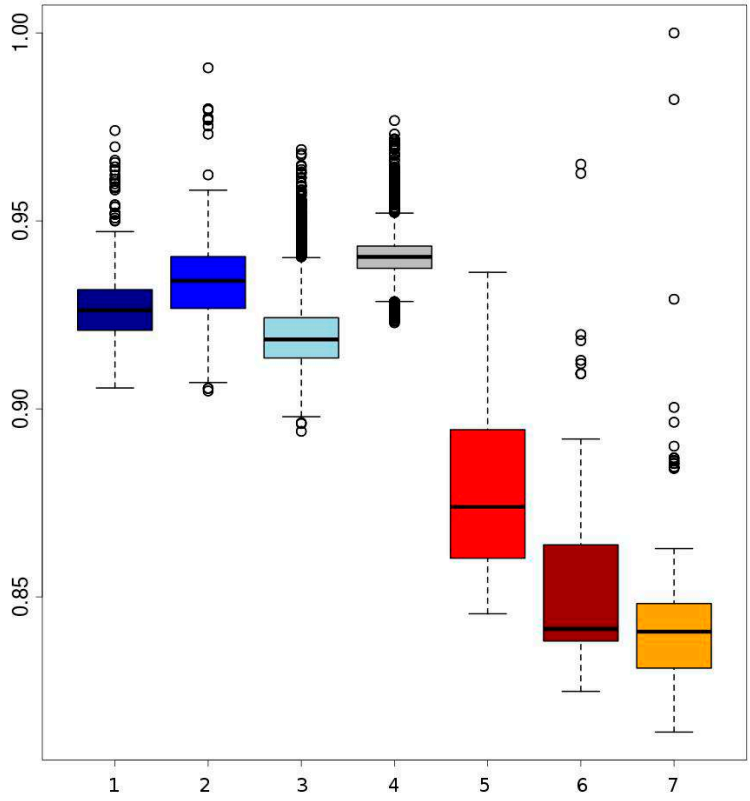


Figure 4.2 Boxplots of homozygosity between two randomly paired haplotypes within groups. 1) Darkblue = two European local haplotypes; 2) blue = two European Iberian haplotypes; 3) lightblue = two European commercial haplotypes; 4) grey = two European wild haplotypes; 5) red = Asian commercial haplotypes; 6) brown = Asian local haplotypes; 7) orange = Asian wild haplotpyes (the highest dot indicates haplotype homozygosity within the Japanese wild boar).

Before and even after the establishments of modern breeds, hybridization between different European populations was common practice. Therefore European commercial pigs are all thought to contain Asian haplotypes. However, this is not necessarily the case for all local breeds in Europe. Our results show that variation between haplotypes from European local breeds is lower than between European commercial haplotypes, which could be due to less Asian introgression or because they have a less mixed European origin (Herrero-Medrano 2013b). Some breeds from the Iberian peninsula and old British heritage breeds cluster with the European wild boar (figure 4.1) which suggest that the source population for domestication more closely resembles these breeds and wild boar, and that genetic differentiation between those pigs is low as recently described by Ramirez (2014). In line with our expectations, we find that variation between two European wild haplotypes is generally lower than between two European commercial haplotypes, especially when variation within individuals is not considered. These findings serve as initial concept of our further analyses.

4.3.2 Consistency over chromosomes

We did an in-depth analysis of haplotypes on chromosome 1, but first we wanted to know whether chromosome 1 is actually a representative model for the rest of the (autosomal) genome. The correlation between haplotype homozygosity for pairs of haplotypes of chromosome 1 and haplotype homozygosity for chromosome 18 is 0.9848, and between chromosome 1 and chromosome 2 is 0.9874. Looking at the homozygosities for pairs of haplotypes on chromosome 1 and pairs of haplotypes on chromosome 18 (figure S4.1), two small clouds of dots stand out: one having a higher homozygosity on chromosome 18 (red) and the other having a lower homozygosity on chromosome 18 compared to chromosome 1 (orange). These clouds actually represent the haplotypes from only two Asian pigs WS01U03 (red) and ZA01U02 (orange) in combination with all European haplotypes, suggesting a different level of European introgression into the different chromosomes for these two pigs. Since the overall correlation coefficients are so high for the rest of the paired haplotypes in the dataset, we conducted the rest of the analyses only on chromosome 1 and exclude these two individuals from further analyses.

4.3.3 Variation in wild boars

Sus scrofa probably originated in South-East Asia. To assess the full width of variation that is present within the species in the wild, we measured variation for all possible pairs of haplotypes in the dataset. The lowest homozygosity between

haplotypes is observed when a haplotype is paired with an outgroup haplotype (the peak at ~0.72 in figure S4.1). The geographic region closest to the center of origin is often the richest in genetic diversity, as shown for other species like dogs and humans (VonHoldt 2010, Long and Kittles 2003). Indeed, our analysis corroborate that the divergence between haplotypes is larger when at least one haplotype is Asian than when no Asian haplotypes are present (figures 4.2 and 4.3). Eastern and Western *Sus scrofa* diverged around 1.2 Mya and this divergence resulted in a multitude of fixed differences between both wild populations (Groenen 2012). Naturally, this divergence also contributes to genetic variation within the species, and to quantify the unique contributions of both continents to variation within the species we looked at the difference in homozygosity between paired haplotypes from the same continent and paired haplotypes from Europe and Asia. For mainland *Sus scrofa*, most divergence between haplotypes is found when a European wild haplotype is pooled with an Asian haplotype, regardless its domestication status. The fact that we do not find a significant difference in homozygosity between a Asian wild or Asian domesticated haplotype paired with a European wild haplotype suggests that the time since the most recent common ancestor is similar and that generally no or very little introgression from Europe into our sampled Asian domesticated breeds occurred. The homozygosity of European wild haplotypes paired with Asian wild is lower than that of two Asian wild haplotypes (averages of 0.825 and 0.84, figure 4.3), but the difference is far less profound than the difference in homozygosity between two European wild haplotypes and the mixture between European and Asian (0.94 vs 0.825, figure 4.3). This indicates that the largest source of variation comes from the Asian wild boars, and that despite the ~1.2My divergence between European and Asian populations, the European clade contributes marginally to the genetic diversity of the species as a whole. The finding that populations further away from the source population capture less genetic diversity is consistent with other species.

4.3.4 Variation between European haplotypes

We had a closer look at the cause of the difference in variation within Europe. One of our hypothesis was that if the higher variation in the commercial lines is mainly caused by a mixture of different European populations, the distribution of variation between two European haplotypes should overlap with the distribution of variation between European commercial haplotypes. The European wild boar used in the current study are derived from different glaciation refugial origins and should therefore represent well extant wild boar variation throughout Europe. All possible pairs of haplotypes from European wild origin should therefore result in a distribu-

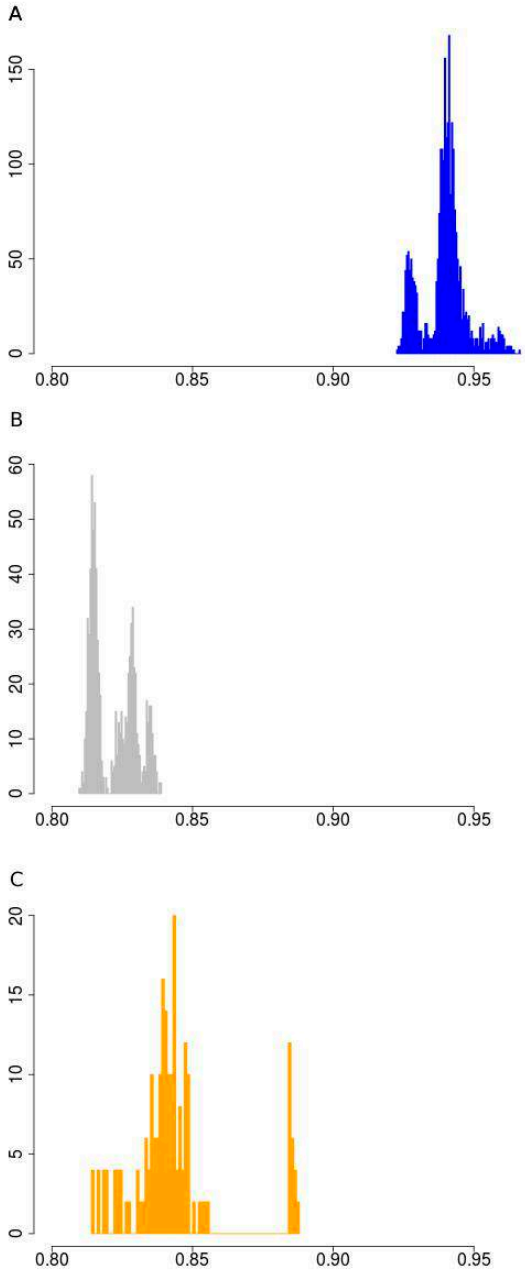


Figure 4.3 Homozygosity between paired wild haplotypes. **A**. Haplotype homozygosity between all possible pairs of European wild haplotypes. **B**. Haplotype homozygosity between all possible pairs of 1 European wild and 1 Asian wild haplotype. **C**. Haplotype homozygosity between all possible pairs of two Asian wild haplotypes.

tion that exceeds the lowest haplotype homozygosity of all pairs of European commercial haplotypes, because the most divergent haplotypes from Europe are included in the European wild distribution. The far tail of the distribution of European wild haplotypes with most variation does not even overlap the mean of variation between two commercial European haplotypes (figure 4.4A), indicating that two wild European haplotypes show more homozygosity than two random European commercial haplotypes, even if these wild haplotypes are sampled from very divergent populations. This suggests that the variation within the European commercial group cannot be completely explained by a mixture of European wild haplotypes. Therefore, it is highly unlikely that the relatively high degree of variation (compared to European wild boar) that is generally found within the European commercial breeds, is due to a mixture of European wild haplotypes, as assumed in hypothesis 1. The distributions for paired haplotypes within the European local and European Iberian group have lower means than the European wild group as well, and their extremes also exceed the European wild distribution. These findings suggest that even some local breeds may contain introgressed haplotypes.

4.3.4.1 Runs of homozygosity (ROH)

Another possibility of the higher variation in European commercial breeds is that European wild boar populations experienced strong recent bottlenecks and associated loss of diversity after the split with European domestic pigs (domestication). We compared the correlation between total ROH coverage in the genome on chromosome 1 (as inferred from PLINK) and homozygosity between haplotypes for the European commercial breeds and European wild boar. ROHs between two commercial European haplotypes are slightly more abundant and longer than ROHs between one European commercial and one European wild haplotype (figure S4.2 A, B). By contrast, more ROHs are found between two European wild haplotypes than between a European wild and a European commercial haplotype (figure S4.2 C, D). Average length of ROH between two European wild haplotypes is generally the same as between a European wild and a European commercial haplotype, unless haplotypes belong to the same European wild population (i.e. within the Netherlands). If the higher level of homozygosity between European wild haplotypes would have been caused by recent inbreeding, the coverage of ROH on chromosome 1 should be higher between two European wild haplotypes than between two European commercial haplotypes. As can be seen in figure 4.4B, the haplotype homozygosity between two European wild

haplotypes is higher than between two European commercial haplotypes with the same level of ROH coverage. These findings suggest that recent inbreeding (i.e. the

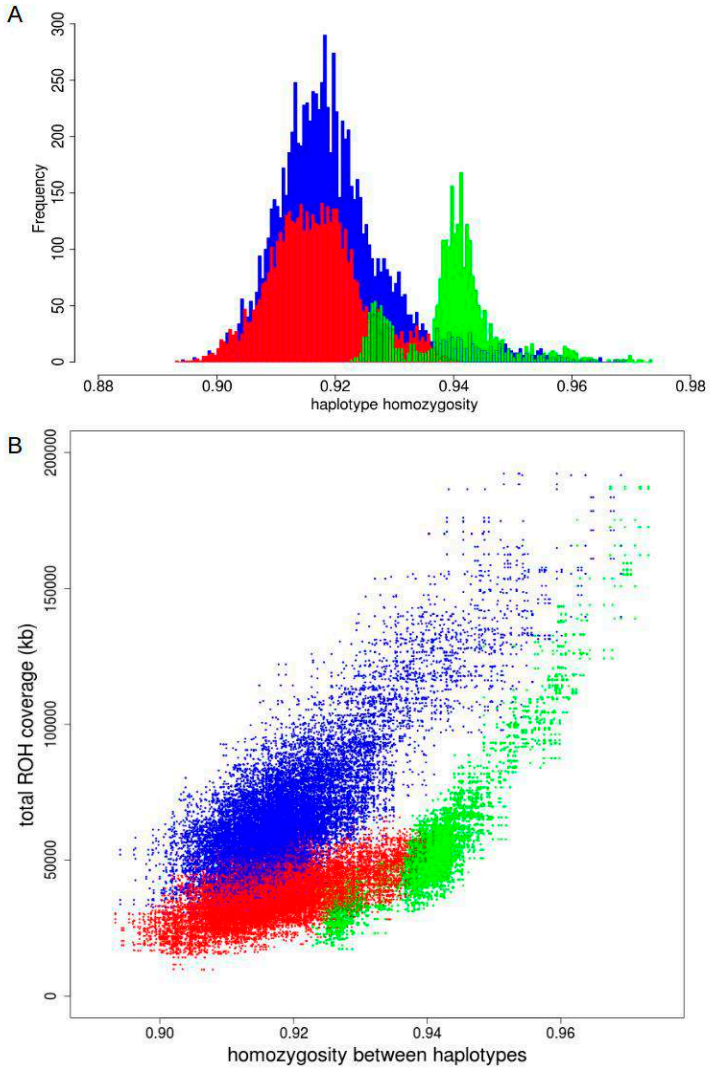


Figure 4.4 Homozygosity between paired haplotypes in Europe. **A**. Homozygosity between two European wild haplotypes is displayed in green. Homozygosity between two European commercial haplotypes is in red and the blue bars indicate homozygosity between one European wild and one European domesticated haplotype. **B**. Homozygosity between haplotypes over the full chromosome on the x-axis is plotted against total ROH coverage between haplotypes on the y-axis for three combinations: two European commercial haplotypes (blue); two European wild haplotypes (green); one European wild and one European commercial haplotype (red).

occurrence of ROH) does not explain the higher homozygosity between wild haplotypes compared to domestic haplotypes.

4.3.5 The effect of introgression

4.3.5.1 Pairing with Asian haplotypes

Although the hypothesis that different source populations in Europe caused the higher diversity in domestic pigs can be rejected based on these previous analyses, our second hypothesis, that Asian introgression caused the higher diversity, is not immediately confirmed. In a previous study (Bosse 2014b) we showed that within the genome of a commercial European pig, the variation is higher when at least one Asian haplotype is present. This observation however does not confirm the role of Asian introgression either, since the presence of an Asian haplotype can be due to incomplete lineage sorting or recent introgression. Another potential cause of the increased variation is hybridization with an unknown population, so called 'ghost admixture'. Introduced haplotypes from an unknown source are likely to increase variation in the European commercial population. Since this source should be unrelated to any of the pig groups here studied, pairing of a commercial European haplotype and an Asian haplotype should not result in less variation than an European wild haplotype paired with an Asian haplotype. If, however, the higher variation in European commercial genomes is due to Asian introgression, pairing with an Asian haplotype should results in higher homozygosity when a commercial European haplotype is used than when a wild European haplotype is used. We do find a small but significant difference between the European wild and European commercial haplotypes when they are paired with a commercial Asian haplotype (figure 4.5A). As expected, the pairing with a European commercial haplotype results in less variation than the European wild haplotypes. Together with the lower haplotype homozygosity in the European commercial group, these findings indeed suggest that the introgression is Asian derived, or at least that the introgressed haplotypes are genetically more similar to Asian haplotypes.

4.3.5.2 Variation with chimeric haplotypes

In order to test whether the influx of Asian haplotypes caused the increase in homozygosity between the Asian wild and European domestic group, and to quantify this amount, we created composite haplotypes that contained 15, 20 and 25% of an Asian breed haplotype and 85, 80 and 75% of a European wild haplotype

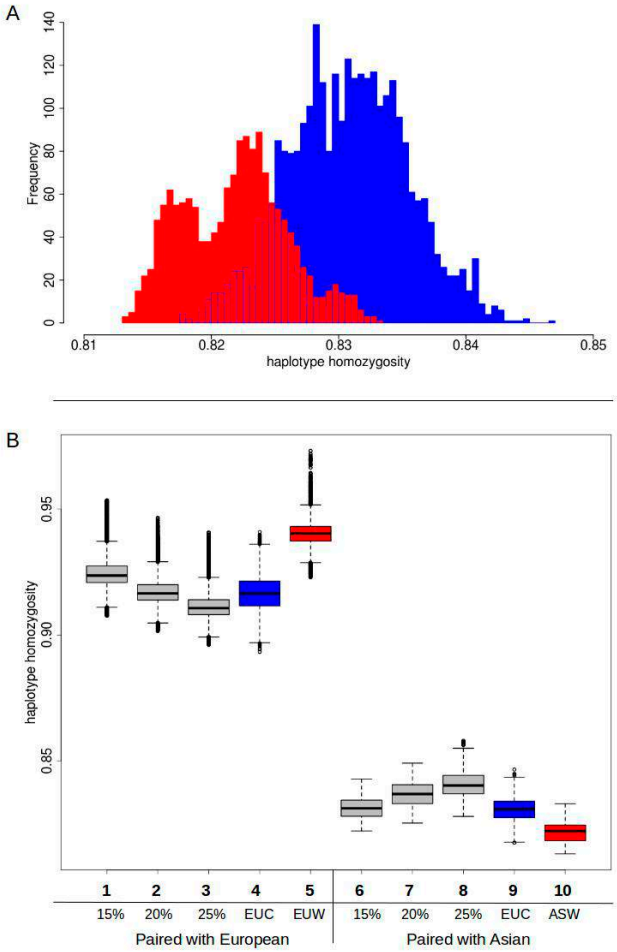


Figure 4.5 Haplotype homozygosity with Asian introgression **A**. Homozygosity between haplotypes when Asian commercial haplotypes are paired with European commercial (red) or European wild (blue) **B**. Boxplots of haplotype homozygosity. Haplotypes are paired with European wild haplotypes (left) or Asian commercial haplotypes (right). Red boxes indicate haplotypes paired with European wild haplotypes. Blue boxes represent haplotypes that are paired with European commercial haplotypes. Grey boxes represent the distribution of homozygosity when the haplotype is paired with a chimeric haplotype that is a combination of a European wild haplotype and an Asian commercial haplotype (see also box 4.1). 1) European wild paired with 15% Asian chimeric haplotype 2) European wild paired with 20% Asian chimeric haplotype 3) European wild paired with 25% Asian chimeric haplotype 4) European wild paired with European commercial 5) European wild paired with European wild 6) Asian commercial paired with 15% Asian chimeric haplotype 7) Asian commercial paired with 20% Asian chimeric haplotype 8) Asian commercial paired with 25% Asian chimeric haplotype 9) Asian commercial paired with European commercial 10) Asian commercial paired with European wild.

as described in box 4.1.2. These percentages were chosen because the introgression fraction from Asia into the European commercial pigs has previously been estimated to be between 15-35% (Fang and Andersson 2006, Groenen 2012, Bosse 2014a). If the percentage of introgression is indeed around 20%, the distribution of haplotype homozygosity when a European domestic haplotype is paired with a European wild haplotype is expected to strongly overlap the distribution when a chimeric haplotype in which 80% of the markers contain European wild alleles and 20% of the markers contain Asian domestic alleles is paired with a European wild haplotype. On top of that, the distribution of the chimeric haplotype paired with an Asian haplotype should overlap that of a European commercial haplotype paired with an Asian haplotype. The results show (figure 4.5B) that pairing of a chimeric haplotype of European wild and Asian domestic with a European wild haplotype indeed results in a similar distribution of homozygosity as a pair between a European wild and a European domestic haplotype. Mean haplotype homozygosity shifts from 0.941 to 0.917, suggesting 20% introgression of Asian haplotypes. Our results confirm the previous estimates of around 20% admixture and demonstrate that the Asian introgression decreased haplotype homozygosity within Europe. In addition, we show that the haplotype homozygosity when a chimeric haplotype is paired with an Asian domestic haplotype increases compared to a European wild haplotype paired with an Asian domestic haplotype. The mean of the 15% Asian chimeric haplotypes is closest to the mean of a European commercial haplotype paired with an Asian domestic haplotype (figure 4.5B), supporting the hypothesis that the introgression comes from Asia indeed.

4.4 Conclusions

We confirmed Asia as the biggest source of genetic variation in *Sus scrofa*, in line with its geographical origin. The higher variation in the European domesticated pigs compared to the European wild boar is largely explained by introgression of Asian haplotypes, rather than a mixture of European backgrounds.

4.5 Acknowledgements

This work was financially supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 249894. We thank Bert Dibbits and Kimberley Laport for labwork, B. van de Water for graphics production and Kyle Schachtschneider for editing the manuscript.

5

Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression

Mirte Bosse[1], Hendrik-Jan Megens[1], Laurent A.F. Frantz[1], Ole Madsen[1], Greger Larson[2], Yogesh Paudel[1], Naomi Duijvesteijn[1,3], Barbara Harlizius[3], Yanick Hagemeijer[1], Richard P.M.A. Crooijmans[1], Martien A.M. Groenen[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, The Netherlands; [2] Durham Evolution and Ancient DNA, Department of Archaeology, Durham University, Durham, UK; [3] TOPIGS research Center IPG, Beuningen, The Netherlands

Nature Communications (2014) 5:4392

Abstract

The independent domestication of local wild boar populations in Asia and Europe about 10,000 years ago led to distinct European and Asian pig breeds, each with very different phenotypic characteristics. During the Industrial Revolution, Chinese breeds were imported to Europe to improve commercial traits in European breeds. Using whole genome sequence data, we demonstrate the presence of introgressed Asian haplotypes in European domestic pigs and selection signatures on some loci in those regions. The signatures of introgression are widespread and the Asian haplotypes are rarely fixed. The identified Asian introgressed haplotypes are associated with regions harboring genes involved in meat quality, development and fertility. More specifically, we identified Asian-derived non-synonymous mutations in the *AHR* gene that are associated with increased litter size in multiple European commercial lines. These findings demonstrate that increased fertility was an important breeding goal for early nineteenth century pig farmers, and that Asian variants of genes related to this trait were preferentially selected during the development of modern European pig breeds.

Key words: Domestication, hybridization, selection, introgression, *Sus scrofa*

5.1 Introduction

Thousands of years of selection on numerous traits in domesticated species like dogs (Boyko 2010, vonHoldt 2010), cows (Diamond 2002) and chickens (Miao 2012) has led to a wide range of distinct phenotypes that are not (or only rarely) observed in the wild. Improvement of breeding practices, which involves making crosses between different breeds and even (sub)species, has greatly improved productivity in intensive farming systems. Examples include the white seed colour in rice, shown to originate from a single mutation that swept through different subspecies following hybridization (Sweeney 2007), and the yellow skin allele that is fixed in the majority of modern western chicken breeds originating from admixture with the Grey junglefowl (Eriksson 2008). Human-mediated introgression of alleles is likely to have played a major role in the genomic architecture of many modern domestic species including pigs.

The wild boar (*Sus scrofa*) originated ~4 million years ago in Southeast Asia and has since expanded its range over Eurasia leading to the emergence of numerous geographically and genetically divergent populations (frantz 2013b, Meijaard 2011). The independent domestication of two of these populations in East Asia and western Eurasia led to distinct domesticated populations (Larson 2011, Megens 2008 and Groenen 2012). Hybridization and introgression between domestic pigs that originated from highly divergent wild populations has resulted in modern genomes that possess a mosaic of different haplotypes (Giuffra 2000, Goedbloed 2013a). While some gene-flow may have taken place before the 19th century, it was certainly extremely rare given the geographic distance between Asia and Europe and the lack of any historical records describing the importation of Asian breeds into Europe before the 19th century (or vice-versa). Mitochondrial studies have suggested that the introgression was mostly female driven (Giuffra 2000) and the introduction of Asian pigs into Europe at the onset of the Industrial Revolution in the late 18th and early 19th centuries has been particularly well documented (White 2011, Jones 1998). In parallel with increasing intensification of farming at that time, British pig breeders sought to improve productivity of the local breeds, and did so, in part, by importing Chinese pigs. Chinese pigs were renowned for having great mothering characteristics, superior meat quality, strong resistance to diseases, better adaptation to living in sties, and producing larger litters (>15 live born young). The selection for specific traits in European pig breeds has resulted in multiple selective sweeps in the genome of domesticates (Rubin 2012). Since European pig breeders deliberately introgressed Asian haplotypes into European

local breeds, it is expected that the origin of haplotypes for which evidence of selection exists, often stems from Asian introgression. Known examples include the *EDNRB, IGF2* and *KITLG* regions (Wilkinson 2013, Okumura 2008 and Ojeda 2008), all of which the identified variants have considerable effect on the phenotypes. Interestingly, the selection criteria shifted through time. When Asian lines were first introgressed into European pigs, fatness was selected for, while now leanness is preferentially selected. The Large White (LW) breed, one of most widely used breeds in commercial pig production, originated in the United Kingdom and is renowned for high growth rate, desirable carcass lean meat percentage and a desirable feed to body weight conversion ratio, traits potentially the result of selected Chinese haplotypes. Because of deliberate introduction of Asian germplasm into European pigs and subsequent intensive artificial selection, Asian haplotypes in Large White genomes are expected to be non-randomly distributed, but rather to be overrepresented in regions that contain genes or regulatory elements that are linked to traits relevant for production. Here, we test this hypothesis and identify specific gene variants bred into European pigs involved in key production traits. More specifically, we interrogated the genomes of Large White pigs to reveal patterns introgressed and selected haplotypes to unravel the genomic consequences of human-mediated hybridization and artificial selection.

5.2 Results & Discussion

5.2.1 Evidence of introgression

We identified haplotypes in the Large White pigs that were identical by decent (IBD) with individuals from both the original source of domestication, the European wild boars (EUWB), and the source of introgression, the Asian domestic group (ASDom, figure 5.1). Individuals from different locations in Europe were used to represent the source of domestication, whiles individuals from three different Asian breeds were used to represent the pool of putative introgressed haplotypes (see table S5.1 and Methods for details). Average genetic differentiation (F_{st}, as defined by Weir and Cockerham (1984)) between the LW and ASDom was 0.33 (sd 0.23, se 0.0008), while the average F_{st} value between LW and EUWB was 0.16 (sd 0.17, se 0.0006). These results show that the genomes of the LW pigs still share greater similarity with their EUWB ancestors than with ASDom. We used another independent method to further verify the existence of gene-flow between ASDom and LW after lineage divergence between Asian and European *Sus scrofa* (D-statistics (Green 2010), see Methods). We computed this statistic for each possible

trio between LW, ASDom and EUWB and (LW, EUWB, ASDom) so that significantly negative D (Z<-4) imply admixture between ASDom and LW. Our results demonstrate that all Large White individuals possess roughly an equal degree of admixture with Asian pigs over their entire genomes, reflecting the human-mediated hybridization with Asian domestics in the late 18th and early 19th century (D=-0.083 +/- 0.015, Z=-20).

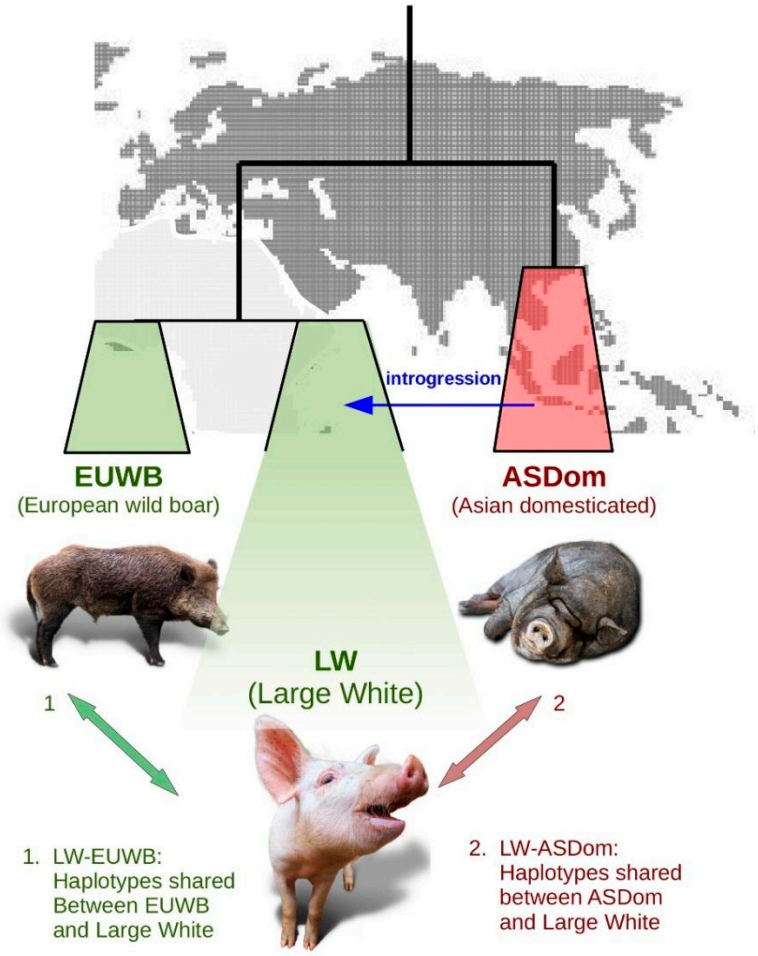


Figure 5.1 Experimental setup for the introgression detection. Arrows indicate the comparisons between groups that are used for the IBD detection. Individuals from the LW breed are used for all pairwise comparisons with individuals from two geographical and functional groups: EUWB and ASDom. The blue arrow indicates the human-mediated introgression from ASDom in to LW.

5.4.2 Introgression mapping

To infer whether a region was introgressed in multiple individuals, we calculated the frequency of all LW haplotypes that were of Asian or European origin for each bin of 10.000 bases in the genome. The relative fraction of Asian vs European haplotypes in the LW group is expressed as rIBD. Asian haplotype frequency in the LW population, for a given locus, ranged from 0.7 (where 1 indicates all haplotypes are ASDom and none are EUWB) to -1 (all haplotypes are IBD with EUWB, figure 5.2a). The majority of the genome displays more similarity with the European wild boars that with the Asian domesticated pigs. Despite this, every chromosome contained regions in which the signal for Asian ancestry was stronger than the European signature. A cutoff of two standard deviations from the mean in the Z-transformed rIBD distribution allowed us to identify regions, which spans ~1.3% of the genome of LW pigs that were likely to be of Asian origin (figure 5.2c-d; table S5.2).

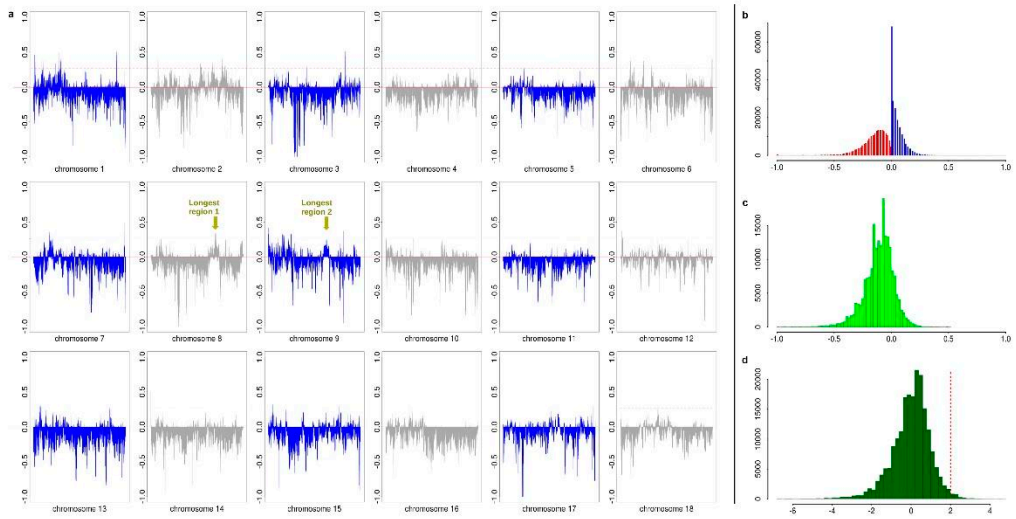


Figure 5.2 Distribution of regions in the genome where the LW contain introgressed haplotypes from ASDom. **A.** The x axis shows the full length of all chromosomes, and the y axis represents the relative frequency of LW haplotypes IBD with ASDom or EUWB, ranging from 1 (all haplotypes are IBD with ASDom, and none with EUWB) to -1. The two longest regions of consecutive introgression are indicated with arrows. **B.** Distribution of the relative proportion of IBD haplotypes in LW and the EUWB (green, IBDEUWB, 0 to -1) or ASDom (blue, IBDASDom, 0 to 1) in bins of 10 Kbp. **C.** Distribution of the rIBD scores for the LW haplotypes (rIBD¼IBDEUWB–IBDASDom). **D.** Z-transformed distribution of rIBD.

This introgression pattern is likely the result of a combination of drift and selection. In contrast to dogs, where selection seems to have acted on a relatively small number of loci with large effects (Boyko 2010, vonHoldt 2010), the introgression signal in the pigs is found at many loci, and the putatively selected Asian haplotype is rarely fixed. This pattern suggests that selection on Asian haplotypes, if present, mostly involved complex multi-genic traits or genes influencing traits selected in opposing directions. A high rIBD signal in our analysis refers to a region that contains predominantly introgressed haplotypes, but this does not imply that the introgressed haplotypes are identical or similar. Regions that contain more Asian haplotypes just by chance, but have not been under selection, could result in a high rIBD signal. We used an extended haplotype homozygosity test (Sabeti 2002) to check for extended haplotypes in the Large White population and compared the iHS signal with our rIBD values. This way, we can distinguish between regions that contain multiple Asian haplotypes and regions that contain one or few particular elongated Asian haplotypes. We inferred a significant correlation between rIBD and iHS in bins of 500kb over the full genome (figure S5.2A). To check whether the extended haplotype homozygosity in the LW pool was specific for the breed or observed in more European breeds, we contrasted the LW signal with a reference pool of other European commercial pigs (figure S5.2B-C).

5.4.3 Genome-wide patterns of introgression

On a genome scale, many of the genes located within the regions where the LW pigs share more haplotypes with ASDom than with EUWB (σZrIBD>=2), are associated with commercial traits such as meat quality (*DNMT3A, SAL1, ME1, IGF2BP1*), fertility (*PGRMC2, KIF18A CDK20*, *AHR*), and development (*NRG1*, *AHR*, table S5.2), although no significant enriched GO-term was found. Gene-dense regions on chromosome 1 and 2 display a high rate of alternating between ASDom and EUWB haplotypes. For instance, the regions containing the *CDK20* and *SAL1* genes, which both have been associated with reproduction traits (Liu 2011, Seo 2011), are only 10-20kb long. These short tracts of shared haplotypes either indicate a high recombination frequency (corroborated by the recombination map for pig (Tortereau 2012), a more temporally distant hybridization episode, and/or favorable European haplotypes surrounding these genes that could lead to positive selection on recombinant haplotypes. The recombination landscape in *Sus scrofa* is known to be highly heterogeneous, and this probably results in an unequal distribution of haplotype length (Bosse 2012). Longer Asian haplotypes will be

found in regions of low recombination and therefore the introgression signal is easier to identify in regions with a low recombination rate.

5.4.4 Longest regions of introgression

Chromosomes 8 and 9 contain the largest consecutive regions of inferred introgression in the LW genomes (defined as regions where rIBD>0). To check whether the extended haplotype homozygosity in the LW pool was specific for the breed or observed in more European breeds, we contrasted the LW signal with a reference pool of other European commercial pigs with the Rsb statistic (Tang 2007) (figure S5.2B-C). This analysis demonstrates that the region of introgression on chromosome 8 contains a stronger EHH signal in the reference panel, and that the region on chromosome 9 contains a particularly strong signal in the Large White population. We used two independent methods, D-statistics and F_{st}, to support the detected introgression in these regions in the LW (figure 5.3a-e). To show that divergence between LW and ASDom was reduced in the introgressed regions, we calculated F_{st} for these regions separately. The F_{st} between ASDom and the LW was lower in both introgressed regions than between EUWB and LW (figure 5.3c-e), thereby supporting the signal of Asian introgression (high rIBD). The D-statisctics for the regions on chromosome 9 was lower than the genome-wide average, which corroborated our rIBD analysis (figure 5.3b). The region on chromosome 8 shows a wide distribution, indicating that some LW-haplotypes contain the Asian signature, while others do not. Inconsistent clustering of European haplotypes within an Asian clade at this locus supports this hypothesis (figure S5.1). Curiously, the ~4 Mb sequence shows a clear signal of introgression, although a large part of the region is devoid of annotated genes. Because this part of the genome has a relatively low recombination frequency (Tortereau 2012), the region may extend considerably beyond the position of the actual favorable allele that has been selected for, due to genetic hitch-hiking and the short time since introgression. Alternatively, drift could have resulted in the presence of Asian haplotypes in this region. The *PGRMC2* gene, coding for the progesterone receptor, lies within the highest peak of Asian haplotypes in that region. Progesterone is an important hormone involved in female reproduction and maternal behavior (Chen 2010), traits that Asian pigs have been selected for extensively. The Asian haplotype containing the *PGRMC2* gene, therefore, could be associated with higher reproductive success in LW pigs and may have been subjected to selection pressure as a result. The rSB signal suggest that in other European breeds the proportion of Asian haplotypes is even higher for this locus (figure S5.2B-C). We used genotype data from the Illumina 60K iSelect porcine beadchip (Ramos 2009) for an additional

5143 pigs from three European commercial purebred lines to screen allele frequencies in this region. Two genetic lines have been selected for reproductive traits since establishment of the lines (A and B), and one line for finishing traits (C). The 60K markers in this 4Mb region show a clear difference between the two reproduction-associated lines and the growth-associated line (figure S5.3). These findings could indicate that the Asian haplotypes in this region are associated with fertility, but further analyses are needed to support this hypothesis.

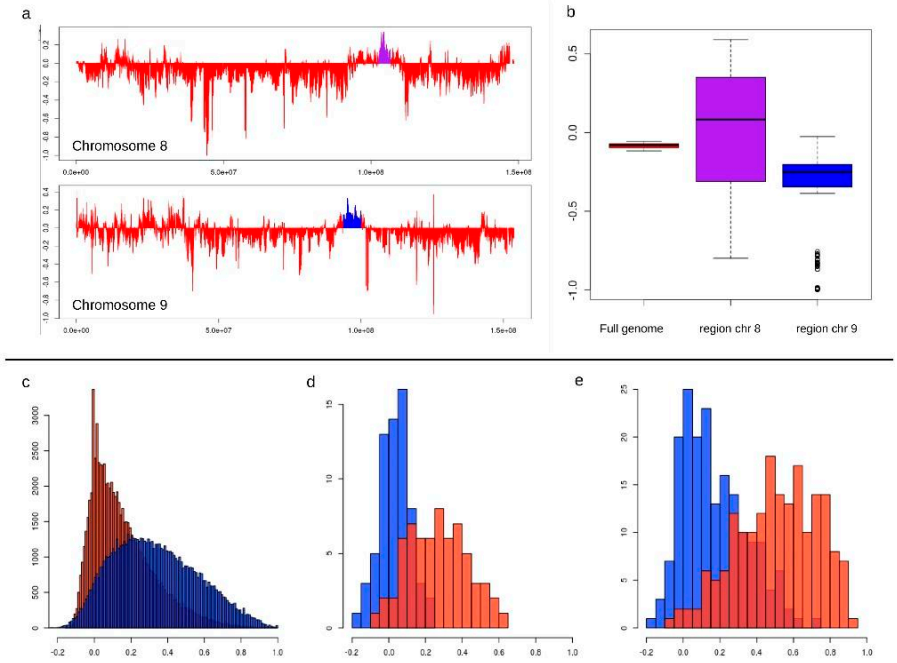


Figure 5.3 Levels of differentiation between LW and ASDom or EUWB in regions of introgression. **A.** Relative introgression fraction (rIBD) over the full length of chromosome 8 and 9. The longest regions of introgression are indicated with purple and blue. **B.** Boxplot of D-statistics for the full genome (red) and the two longest regions of introgression as indicated in a on chr8 (purple) and chr9 (blue). The minimum, first quartile, median, third quartile and maximum are indicated with the box and whiskers with outliers >1.5*IQR. D-statistics are computed for each possible trio with LW¼P1, ASDom¼P2 and EUWB¼P3, with the Sumatran *S. scrofa* as outgroup (O) resulting in 378 trios. **C-E**. Distribution of F_{st} between LW-ASDom (blue) and LW-EUWB (red) in bins of 10 Kbp. The left histogram shows the F_{st} distributions based on the full genome (**c**), and the other two show the F_{st} distribution for the regions of introgression on chr8 (**d**) and chr9 (**e**).

5.4.5 Introgression at the *AHR* locus

The 6.8 Mb region on chromosome 9 has a large proportion of Large White haplotypes that are nearly identical to two haplotypes found in the Asian Jianquhai breed (figure S5.1). Among the genes in this region of putative Asian introgression are multiple members of the TWIST gene family, transcription factors known to be involved in a variety of processes including embryonic development (Gitelman 2007). This highest introgression peak in this region contains the *AHR* gene that has previously been associated with female reproductivity (Hernández-Ochoa 2009). Originally, the *AHR* gene has been identified as mediator of xenobiotic induced toxicity (Fujii-Kuriyama 2010, Denison 2011). *AHR* has been shown not only to be involved in the response to toxicity, but also to be associated with fertility in mammals (Baba 2005, Pocar 2005). The *AHR* gene seems to play an important role in the female reproductive system at multiple levels (Hernández-Ochoa 2009). In pigs, expression of *AHR* during the estrous cycle and putative involvement of *AHR* in the regulation of reproduction has been observed (Jablonska 2011, Onteru 2012). Furthermore, polymorphisms in human are known to occur predominantly in exon 10, which contains an important transactivation domain (Harper 2002). We screened this gene in pigs for non-synonymous mutations and identified four non-synonymous mutations in exon 11 of the *AHR* gene, which corresponds to exon 10 of the human *AHR* gene. The variants of Asian origin were all in strong LD. Since the *AHR* is a strong candidate gene for which Asian variants have been selected since introgression during the Industrial Revolution, we examined this gene further (see methods).

The iHS signal within the Large White population is strongest for the *AHR* locus. The fact that the Rsb signal is also strong in the LW compared to other European breeds, indicates that the Asian haplotypes at the *AHR* locus were never fixed in the ancestral population that led to current commercial breeds, because the frequency and iHS signal differs between breeds (figure S5.2B-C). We used an additional method to screen for selection, nSL that has recently become available and should be robust to variation in recombination rate (Ferrer-Admetlla 2014). We averaged the rIBD, the nSL statistic and the corrected iHS p value over bins of 500kb for chromosome 9 (figure S5.4A-C). All three statistics contain a high signal at the previously identified region on chromosome 9 that also contains the *AHR* gene. We show that the haplotype containing the ancestral allele at the *AHR* locus is more persistent than haplotypes containing the derived allele at this locus (figure S5.4D-F). A survey of wild boar populations and four other *Sus* species for these loci revealed that the ancestral haplotype is homozygous in all closely related *Sus*

species, and at high frequency in Asian domesticated pigs and European breeds. However, only derived haplotypes were found in European wild boar and the ancestral type was present only at a low frequency in Asian wild boars. This suggests a history of selection for the ancestral state in domestics, after the derived state reached high frequency in the wild populations.

5.4.6 Effect of Asian haplotypes at the *AHR* locus

To examine the effect of the amino acid changes in the *AHR* protein on reproductive success, we used genotype data from the Illumina 60K iSelect porcine beadchip (Ramos 2009) for the same 5143 pigs from three European commercial purebred lines for which estimated breeding values (EBV) for the total number of piglets born (TNB) was available. We extracted genotypes for those markers that fell within the region of high introgression (rIBD>2). The *AHR* gene was the only annotated gene in this selected part of the genome. Either Asian or European heritage was assigned to the 60K haplotypes for these 5143 commercial pigs at the *AHR* locus, based on our re-sequenced individuals and confirmation in the lab (see methods). Haplotypes containing the Asian *AHR* variants had a significant effect on the EBV for total number born over all three lines (EBV-TNB 0.162, std.error 0.04, $p<0.0001$, Table1). Although the EBV-TNB can be rather different between different lines, also within-line effects of the Asian allele were estimated. A difference of 0.1 in EBV equates to a difference of 0.1 piglets born. Since total number of piglets born is a complex multi-locus trait, an increase of 0.16 piglets born (across all three breeds) is substantial in the current breeding industry. If the costs of maintaining a sow on a farm are spread over a larger number of piglets being weaned from that sow, the marginal cost reduction of producing a finishing pig is just over 3 euros per extra piglet (Hanenberg 2010), that is, 2% of that total. Even though we cannot rule out that the effect could be due to some extended haplotypes covering other genes in the region, these details in combination with known literature contribute to *AHR* as the strongest candidate gene.

Interestingly, the *AHR* locus is an example of selection on the ancestral state that was either lost or never present in the European wild population. Re-introduction of the variants by introgression of Asian haplotypes and the positive effect of these alleles on litter size contributed to the high frequency of Asian haplotypes at the *AHR* locus in the current population. Since the *AHR* gene seems to be involved in multiple life history traits, it could very well be that some long-term balancing selection acted upon the alleles. The *AHR* gene is involved in multiple traits and

Table 5.1 Estimated breeding values for total number born per line for haplotypes containing Asian or European alleles at the AHR locus. Two models were used to estimate the effect of the Asian allele: a linear model and an animal model. As-EBV is the estimated breeding value for haplotypes containing the Asian alleles and Eu-EBV is the EBV for European haplotypes under the linear model. "Estimate" is the estimated effect of the origin of the haplotypes on TNB for the linear model, and "Effect" refers to the effect of the origin of haplotypes under the animal model. The effect was calculated for two reproduction-associated lines (line A, N=1053 and Line B, N=568) and one growth-associated line (Line C, N=965). N is number of haplotypes in the line. Significant codes: '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1; ' ' 1.

		Linear model						Animal model			
Line	**N**	**As-EBV**	**Eu-EBV**	**Estimate**	**Std. Error**	**t value**	**Significance**	**Effect**	**Se effect**	**F value**	**P value**
ALL	2586	0.038	-0.115	-0.163	0.039	-4.221	2.54e-05***	-0.07	0.03	5.65	0.018*
Line C	965	-0.091	-0.178	-0.089	0.040	-2.191	0.0287*	-0.04	0.09	0.19	0.66
Line B	568	0.242	0.113	-0.128	0.088	-1.461	0.144	-0.02	0.02	0.91	0.342
Line A	1053	0.039	-0.167	-0.227	0.068	-3.329	0.000901***	-0.15	0.06	6.5	0.012*

during ever changing adaptation to e.g different environments, some alleles might be more desirable that others under different circumstances (Connallon 2013). The significant association between the Chinese haplotypes and an increased estimated breeding value for total number born, which in our opinion is a strong independent indication for selection, in combination with the selection sweep results provide convincing evidence for the *AHR* locus to have been under selection after introgression. Similar examples of selection on Asian haplotypes in European pigs exist in literature, like the signals of selection associated with coat color (Kijas 1998), ear morphology (Wilkinson 2013) and increased lean content (Ojeda 2008).

The evidence presented here demonstrates how crossing of divergent populations may shape the variation on a genome wide scale in populations. The introduction of Asian haplotypes into European breeds in the late 18th and early 19th centuries (Meijaard 2011, Larson 2005) and consecutive selection for desired traits in these breeds thereby provides a robust, historically documented model system for these instances. We identified numerous genomic regions where Asian haplotypes were introgressed into a larger European background, including the *AHR* locus. The *AHR* gene has been known to be involved in reproduction (Baba 2005, Pocar 2005, Jablonska 2011, Onteru 2012), and our study corroborates that earlier report by

demonstrating a significantly increase in litter size in European commercial pigs that possess the Asian haplotype. Our findings provide a unique insight into the genomic haplotype patterns resulting from breeding practices from first domestication until the intensive breeding industry we know today. The observed introgression pattern is a combination of drift and selection, and detailed analyses like those demonstrated for the *AHR* locus will shed more light onto the importance of other introgressed Asian haplotypes on signatures of selection in modern pig breeds.

5.5 Materials and Methods

5.5.1 Sample collection and DNA preparation

Blood samples were collected from a total of 70 individual wild and domesticated *Sus scrofa*. Among these individuals were two wild boars from Sumatra, eight Asian wild boars from China and Japan, eighteen European wild boars from the Netherlands, France, Switzerland, Greece and Italy, thirteen Asian domesticated pigs from the Meishan, Jianquhai and Xiang breeds and 29 European domesticated pigs from the Duroc, Hampshire, Pietrain, Landrace and Large White breeds. DNA was extracted from the blood samples with the use of the QIAamp DNA blood spin kit (Qiagen Sciences) and checked for quality and quantity on the Qubit 2.0 fluorometer (Invitrogen). Library construction for the re-sequencing was performed with 1-3 ug of genomic DNA according to the Illumina library prepping protocols (Illumina Inc.) and the insert size varied from 300-500 bp. Sequencing was performed on 1-3 ug of genomic DNA with the 100 paired-end sequencing kit for all samples. SNP genotyping was performed on the Illumina Porcine 60K iSelect Beadchip (Bosse 2012). DNA from all individuals was diluted to 100 ng/ul and genotyped according the IlluminaHD iSelect protocol. Data was analyzed using Genome Studio software (Illumina Inc.).

5.5.2 Alignment and variant calling

All individuals were re-sequenced with the Illumina paired-end sequencing technology (Illumina Inc.) to ~10x depth of coverage (details in table S5.1). The read pairs were trimmed to have a minimum phred quality > 20 over three consecutive bases, while each mate should have a minimal size of 45 bp after trimming. Alignment was performed with Mosaik Aligner (V. 1.1.0017) with the unique alignment option to the Porcine reference genome build 10.2. Variants were called using Samtools mpileup 0.1.12a (r862). The alternative allele should be covered at least two times to call a SNP and INDELS were excluded. VCFtools was used for

filtering for a genotype quality of >20, a minimum read depth of 7x and a maximum read depth of twice the average read depth. For the list of all those sites that were heterozygous or non-reference within at least one individual, genotyping was performed with Samtools mpileup for all 70 individuals to create a matrix containing an unbiased representation of the variation present in the samples. Homozygous reference alleles were also included in the matrix at this stage, and only those sites that were covered >=4x in all individuals were included, resulting in 2.377.607 markers.

5.5.3 Pairwise IBD detection

The matrix of 70 individuals genotyped for 2.377.607 positions in the genome served as input for the IBD detection pipeline. All individuals were phased with the fastPhase function in Beagle version 3.3.2. Pairwise shared haplotypes were extracted with the Beagle fastIBD function as described by Browning and Browning (2011). Phasing and IBD detection was executed 10 times independently and identified IBD tracts were merged from all 10 runs, as suggested by the authors. Partially overlapping runs were extracted and the IBD runs with the highest probability scores were added to the pool of IBD tracts. The 10 cycles of IBD detection were run with different thresholds for assigning IBD to the haplotypes of two individuals. The numbers varied from zero detected pairwise IBD tracts to complete IBD genomes. We empirically determined that the middle range of thresholds resulted in pairwise IBD tracts that remained stable in terms of relative number and length of detected IBD tracts between members of different pig groups. To this end , the final threshold that was used for IBD detection (5.0−6) was elevated compared to that of the original paper, to allow extracting similar, but not necessarily identical, haplotypes between individuals. Because the focus of this analysis was to identify regions containing haplotypes that are more similar between distantly related individuals than expected based on their inheritance, and the frequencies of similar haplotypes are leveled out, this threshold was thought to fit the data best.

To estimate the frequency of shared haplotypes in different regions of the genome, the genome was divided into bins of 10.000 bp and the number of recorded IBD tracts between the LW pigs and the two different pig groups (ASDom and EUWB) was computed per bin. Because the total number of pairwise comparisons differed between the groups, these numbers were normalized, ranging from 0 (no IBD tract detected) to 1 (all individuals IBD with all individuals within the group). Relative IBD between the LW and the two competing pig groups ASDom and EUWB was then

calculated by extracting per bin the normalized IBD with ASDom from the normalized IBD with EUWB.

Normalized IBD for one pig group: nIBD = cIBD / tIBD (cIBD = count of all haplotypes IBD between LW and one pig group, tIBD = total pairwise comparisons between LW and one pig group)
Relative IBD between two pig groups: rIBD = nIBDASDom – nIBDEUWB

The distribution of rIBD for the comparison between LW-ASBr and LW-EUWB IBD haplotypes resembled and normal distribution and therefore was Z-transformed as follows: ZrIBD = (rIBD – μ)/ σrIBD. ZrIBD therefore represents the number of standard devations that rIBD deviates from the mean rIBD. The threshold for extreme IBD with the breeds from the other continent compared to the wild boars from the same continent was set to 2 standard deviations from the mean in the far right tail of the distribution.

5.5.4 GO-enrichment analysis

All annotated genes in build 10.2 (Ensembl release 67) from the *Sus scrofa* reference genome were extracted. GO-enrichment analysis was performed for genes in the top 1.3% (σ>2 for ZrIBD) of regions with an over-representation of ASDom haplotypes in LW pigs. The Cytoscape v.2.8.3 plugin BinGO v2.4444 was used to identify over-represented biological process related GO terms. The human one-to-one orthologues (Ensembl db) for all pig genes were used for the analysis, since human genes are annotated more comprehensively. Significance levels were adjusted based on the Benjamini and Hochberg correction for multiple comparisons.

5.5.5 F_{st} analysis

To measure genetic differentiation the individuals from the Matrix were assigned to one of the following pig groups (if applicable): Large White breed (LW), Asian domesticated pigs (ASDom) and European wild boars (EUWB). Pairwise F_{st} as described by Weir and Cockerham (1984) between the LW breed and the other two groups were computed with Genepop 4.2 in bins of 10kb over the full length of the genome45. Relative F_{st} (rF_{st}) and Z transformation of rF_{st} (ZrF_{st}) were computed similarly as rIBD and ZrIBD. Correlations between ZrF_{st} and ZrIBD values were calculated with Pearson's product moment correlation in R.

5.5.6 Admixture fraction

In order to prove the existence of admixture between LW and MS in our potentially introgressed regions, but also genome-wide, we computed D-statistics (Green 2010) as implemented in qpDstats from the ADMIXTOOL software package (Durand 2011). In short, the D-statistics provide a robust test for admixture by challenging the strictly bifurcating nature of a phylogenetic tree. For a triplet of taxa P1, P2 and P3, and an outgroup O, that follows the phylogeny (((P1,P2),P3).O), one can compute the number of sites where P1 and P3 (BABA) or P2 and P3 (ABBA) share the derived state (B; assuming ancestral state, A, in the outgroup). Under a null hypothesis of no gene-flow or no sub-structure (strict bifurcation), the count of ABBA and BABA should not be significantly different. Alternatively, a significant excess of either ABBA or BABA site pattern provide a conclusive proof of gene-flow or sub-structure. However, because sub-structure is very unlikely to affect our analysis of domestic pigs from Asia and Europe (because of independent domestication), we can conclude that significant D implies gene-flow. We computed D-statistics between LW (P1), European wild-boars (P2) and Asian domestics (P3), for every possible combination of samples, using the sequence of wild boar from Sumatra as an outgroup (O). For each possible combination we first computed a genome-wide value. Significance level was computed using a standard block jackknife, with blocks of 1cM (assuming 1Mb=1cM). We also computed D-statistics in potentially admixed regions separately.

5.5.7 Haplotype association test

To examine the putative effect of the amino acid change we used genotype data from an additional 5143 individuals from 3 different commercial purebred lines, genotyped on the Illumina Porcine 60K iSelect Beadchip (Ramos 2009) for 21 markers surrounding the mutation. Line A and B are dam lines selected for reproductive traits and line C is a sire line selected for finishing traits. In all three lines, total number born (TNB) was routinely recorded on sows. The estimated breeding values (EBV) of the genotyped animals were obtained via routine genetic evaluation using MIXBLUP in a multitrait model (Mulder 2012). The model for obtaining the EBV of TNB included fixed effects (herd-year-season, insemination number, parity, cross-fostering (Y/N), interval weaning (class)) and random effects for service sire, permanent sow and animal. Reliabilities per animal were extracted from the genetic evaluation and were based on the methodology of Tier and Meyer (2004) and animals with a reliability <0.15 were excluded from the analyses. We genotyped 64 individuals for the mutation by Sanger sequencing with the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) on a ABI3730

sequencer. The following primers were used: *AHR*gene_Forward AGCAGCAGCAACAACTGTGT and *AHR*gene_Reverse GACACAGCTCCACCATAGCA. The haplotypes, based on the 21 markers that were associated with the alleles at the site of mutation were reconstructed. Then all 5143 individuals were phased with Beagle and the G or T allele at the site of the mutation was assigned to these phased haplotypes when possible. We identified 19 unique haplotypes in the dataset for which the allele at the site of mutation could be verified. The following linear model was fit to the data using R to test for a significant effect of the allele on EBV for TNB:

lm(formula = EBV_TNB ~ LINE + Origin, weights = REL_TNB)

EBV_TNB is the estimated breeding value for total number born and these values are weighted for the reliability for this trait (REL_TNB). Genetic line was included to account for differences between selection strategies between the three genetic lines A, B and C.

Finally, an animal model was fit (to account for family relatedness) to the data using ASReml (Gilmour 2009) to test for a significant effect of the allele on EBV for TNB:

$y = Xb + Za + e$

where b is a vector of the fixed effect for line and origin, a is the vector of the random animal genetic effect. The term e is a vector of the random residual effects assumed to be normally distributed, but weighted by the reliability of the EBV.

5.5.8 Mutation characteristics

The functional annotation of the genomic variants in the high IBD regions (σrIBD>2) was determined using Annovar (Wang 2010). The nature of the non-synonymous mutation in the gene was obtained from the webtool polyPhen 2 (Adzhubei 2010) by using the human ortholog *AHR* for the pig gene *ENSSSCG00000030484*.

5.5.9 Phylogenetics

We performed our primary phylogenetic analysis using MrBayes 3.2 and our matrix of variable sites (Ronquist 2003). To estimate correct branch length solely from variable sites, we used the Mkv model implemented in MrBayes, which provides a likelihood framework for data-sets that contains only variable characters. We

recorded SNPs in 4 potential states (0-3). Rate of evolutions, for each SNPs, were drawn from a gamma distribution. We ran 2 independent runs of 4 MCMC chains with 2 million samples. We repeated this analysis solely based on SNPs found in the 2 introgressed regions on chr 9 and 8 separately.

5.5.10 Extended haplotype homozygosity tests

The identification of extended haplotype homozygosity was tested on 56 LW individuals that were genotyped on the Illumina Porcine 60K iSelect Beadchip (Ramos 2009). First, a genome-wide scan for iHS within line was performed with the R package rehh (Sabeti 2002, Gautier 2012). Significance levels within line were averaged for 500kb and correlation with the rIBD values for the same bins of 500kb was tested with the cor.test R. Secondly, a reference panel of 100 individuals belonging to the Landrace, Pietrain and another Large White breed was used to polarize the iHS signal in the original 56 LW individuals with the ies2rsb function in rehh (Tang 2007, Gautier 2012). To check the signal on chromosome 9, we also performed the recently developed nSL test that uses a slightly different approach to screen for extended haplotype homozygosity than the original iHS test and is robust to variation in recombination frequency (Ferrer-Admetlla 2014). The bifurcation diagram option in rehh was used to visualize the LD from a focal SNP on the Beadchip that was closest to the *AHR* gene.

5.6 Acknowledgements

DNA samples were provided by Dr. Ning Li; China Agricultural University, China; Dr. Alain Ducos, UMR INRA-ENVT, France; Sem Genini, Parco technologico Padano, Italy; Dr. Gono Semiadi, Puslit Biologi, Indonesia; Dr. Naohiko Okumura, Staff Institute 446-1 Ippaizuka,Japan; Dr. Alan Archibald, Roslin Institute and the Royal (Dick) School of Veterinary Studies,University of Edinburgh ,Scotland; TOPIGS Research Center IPG BV, The Netherlands; Dr. Oliver Ryder, San Diego Zoo, USA; Cheryl L. Morri, Ph.D., Omaha's Henry Doorly Zoo, USA. We thank Bert Dibbits for the genotype validation assay in the lab, Miguel Perez Enciso for valuable comments on the manuscript and B. van de Water for graphics production. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n° 249894".

6

Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs

Mirte Bosse[1], Marcos S. Lopes[1,2], Ole Madsen[1], Hendrik-Jan Megens[1], Richard P.M.A. Crooijmans[1], Laurent A.F. Frantz[1], Barbara Harlizius[2], John W.M. Bastiaansen[1] and Martien A.M. Groenen[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen 6708WD, the Netherlands; [2] Topigs Norsvin, Beuningen 6640AA, the Netherlands

Proceedings B (Under review)

Abstract

Early pig farmers in Europe imported Asian pigs to cross with their local breeds in order to improve traits of commercial interest. Current genomics techniques enabled genome-wide identification of these Asian introgressed haplotypes in modern European pig breeds. We propose that the Asian variants are still present because they affect phenotypes that were important for ancient traditional, as well as recent commercial pig breeding. Genome-wide introgression levels were only weakly correlated with gene content and recombination frequency. However, regions with an excess or absence of Asian haplotypes contained genes that were previously identified as phenotypically important such as *FASN*, *ME1* and *KIT*. Therefore, the Asian alleles are thought to have an effect on phenotypes that were historically under selection. We aimed to estimate the effect of Asian haplotypes in introgressed regions in Large White pigs on two commercial traits: fatness and reproduction. The majority of regions in the genome that retained Asian DNA showed significantly increased backfat from the Asian alleles. We demonstrate here that the introgression in Large White pigs has strongly been determined by the selective pressure acting upon the introgressed Asian haplotypes. We conclude that human-driven hybridization and selection broadly shaped the genomic architecture of commercial pigs.

Key words: Hybridization, Adaptive introgression, Selection, Domestication, *Sus scrofa*, Commercial breeding

6.1 Background

Introgression and subsequent selection on introgressed variants is thought to be a widespread phenomenon among many species (Currat 2008). The introgression can occur naturally, due to mixture of populations in hybrid zones or occasional invasions. During those instances, selection for introgressed haplotypes can occur, a process known as adaptive introgression (Hedrick 2013). However, introgression can also be human-driven through hybridization, either accidental or on purpose (Crispo 2011, Harrison 2014). Domestic animals are a clear example of species that have experienced population admixture due to human interference (Larson 2013). Introduction of and selection on novel alleles into a population has been observed in e.g. chicken (Eriksson 2008), cagebirds (Rheindt 2011) and cattle (Flori 2014). Also in pigs, human mediated hybridization has introduced haplotypes that cause desired effects on phenotypes (Bosse 2014a).

Pig farming has undergone a true metamorphosis from first domestication till the intensified industry we know today. Pigs were domesticated independently leading to separate European and Asian domestic pigs some 10,000 years ago (Larson 2005, 2007a). Subsequent selection and breeding resulted in highly distinct breeds on these continents (Kijas 2001, Megens 2008). Especially in Europe, farmers were herding their swine in surrounding forests, and it was not before the Industrial Revolution during the eighteenth century that pigs were kept in sties and became an important farm animal (White 2011). Ever since pig farming as a profitable profession became widespread, breeders have been aimed at optimizing their production process. Pigs were selected based on their phenotypes regarding traits of commercial interest.

In the last 100 years, with the improvements in performance recording and making use of genetic evaluation methods based on pedigree information, breeding programs have achieved a remarkable genetic progress in reducing backfat for carcass quality and improving growth rate for production efficiency. Since the 1990s, using the same traditional breeding strategies (pedigree-based), genetic progress has also been observed in reproduction traits, especially litter size at birth (Merks 2000, 2012). Part of the success of these breeding programs can also be attributed to the introduction of genes from Chinese breeds in commercial European breeds. During the time that global trade increased, farmers in Europe realized that Chinese pigs possessed particular characteristics that would be beneficial to introduce to their breeding stock. Therefore, pigs from Chinese breeds were imported to Europe and multiple crosses between European and Asian breeds were made with the purpose of combining beneficial traits, such as backfat

thickness (BF) and litter size (LS) from Asian pigs, and body length from European pigs (White 2011, Jones 1998).

With the advent of genomic selection (Meuwissen 2001) the genetic progress is expected to speed up even more. The fast development of genomic techniques and available resources is a key element in the current genetics-based breeding technology. The design of a 60K SNPchip for pigs in 2009 (Ramos 2009) and the publication of the pig reference genome in 2012 (Groenen 2012) greatly contributed to the applicability of these techniques in pig breeding. In genome-wide association studies (GWAS), high-density genotypic information of populations is used to estimate the effect of genotypes on a particular phenotype, and identify the associated region in the genome. This genomic information can be used also to pinpoint regions in the genome that have been under selection pressure. The resulting changes at the DNA level have been detected as selective sweeps in a multitude of breeds (Wilkinson 2013, Rubin 2012). Interestingly, some of these loci have been identified as not only being under selection, but also introgressed. Asian alleles at the *EDNRB* (Wilkinson 2013), *IGF2* (Ojeda 2008) and *KITLG* (Okumura 2008) locus have proven effects on phenotypes (e.g. meat content and coat color) of European commercial breeds.

With the current genomics techniques it has become possible to trace back the haplotypes that were introduced during the Industrial Revolution. In Bosse 2014a, we examined the occurrence of Asian haplotypes in a population of European pigs that belong to the commercial Large White breed. Our results showed that Asian haplotypes are widely present in the genomes of these commercial pigs and highlighted the effect of the introgressed Asian variant of the *AHR* gene on litter size. Since the Asian haplotypes were introduced for a specific purpose, the effect on the phenotype should co-occur with presence of Asian variants in regions of the genome that are associated with the traits known to have been under selection. However, how much influence the Asian introgression had on the selective history of commercial traits remains unknown.

We hypothesize that the introgression landscape in commercial breeds is shaped mostly by artificial selection and therefore most introgressed regions should have an effect on commercial traits (BF and LS). Also, an absence of Asian haplotypes in some parts of the Large White genomes could be the result of purifying selection. In this study, we examine the introgression signatures that we identified previously in Bosse 2014a in more detail, showing that the majority of these regions have a significant effect on BF in a commercial Large White population. These findings have important implications for the knowledge on natural and human-driven evolutionary forces shaping genomes after hybridization.

6.2 Methods

The analyses in this paper build upon the dataset and results that were obtained in Bosse 2014a and 2014b. Introgression mapping was performed on a group of 9 Large White pigs and the background of their haplotypes was assigned to be European or Asian (Bosse 2014a). In bins of 10kb over the genome, the relative Asian introgression signal (rIBD) in the Large White population was obtained (figure 6.1A). We used these genome-wide rIBD signals to understand the details of the Asian introgression.

6.2.1 Genome characteristics

To assess the correlation between gene density, recombination frequency and introgression signal, we averaged the rIBD in 1MB bins and counted the number of genes within each bin. The recombination map from Tortereau 2012 was used to obtain the recombination frequency per Mb. To test whether the probability of introgression decreases with an increase in number of genes in a region, we used the Pearson's product-moment correlation in R.

6.2.2 Selection of regions

We used the ~400 regions of introgression previously identified by Bosse (2014a) with a Z-transformed rIBD (ZrIBD) >2. The regions were extended with one 10kb bin at the time to the left and right flank of each identified region of introgression, until the threshold of >2 ZrIBD was no longer met and/or the rIBD value for one particular 10kb bin was <0. We found 33 regions of introgression that were longer than 150kb, and checked whether they physically overlapped with markers on the Illumina Porcine 60K iSelect Beadchip (60k chip). Three regions were discarded because they contained less than 3 segregating markers on the chip. Table S6.1 contains the list of 30 regions that were included in the further analyses.

6.2.3 Genotyping and phasing

We genotyped a total of 9970 pigs and wild boar for 488 markers on the 60K chip that fell within the 30 identified regions of introgression. Phasing of haplotypes was done independently for each region with Beagle V. 3.3.2 (Browning 2007), using the genotype information for all individuals. After the phasing step, we used haplotype data from three groups: Asian (N= 448), European wild (N= 920) and European commercial (N= 18,572). Because the introgression analysis was done on Large White pigs, we extracted only those individuals from the European commercial group that were known to be purebred Large Whites, leaving us with a total of 4,764 Large White haplotypes.

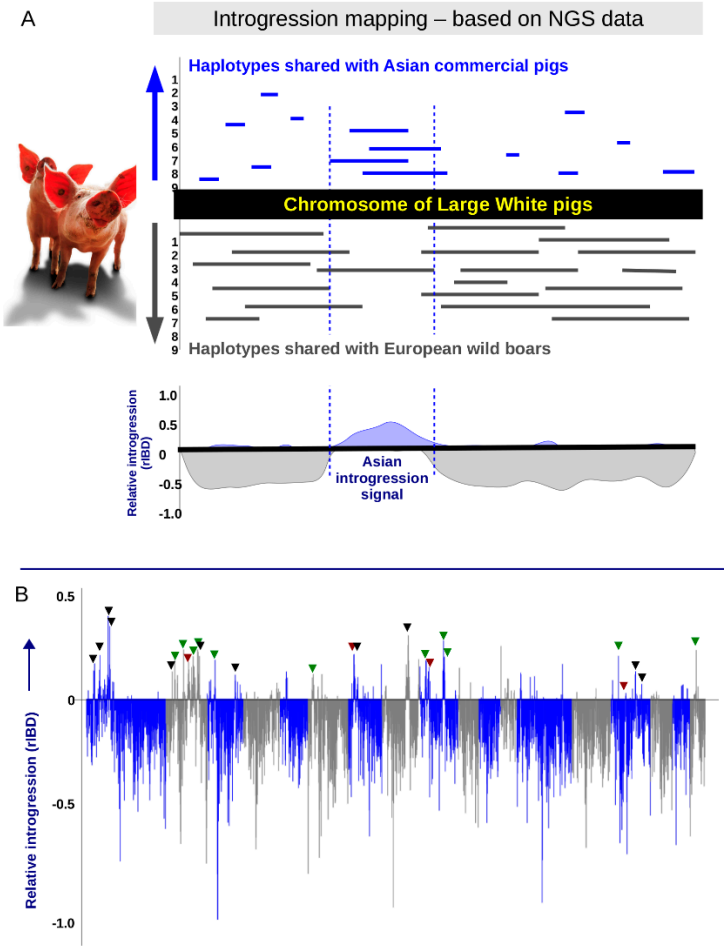


Figure 6.1 The principle of introgression mapping. The purpose of introgression mapping is to determine what the background is of haplotypes in a particular part of the genome. **A**. Haplotypes that were shared with either Asian or European pigs were mapped to the genome for all 9 Large White pigs. The total numbers of overlapping haplotypes were counted and the relative introgression signal (rIBD) was obtained by taking the difference of haplotype frequencies as described in Bosse (2014a). **B**. The y-axis displays the rIBD signal averaged in bins of 1Mb, and the x-axis contains the physical position for all 18 autosomes. Regions that were selected for the commercial trait analysis are indicated, with green triangles indicating included regions, black triangles indicating selected regions with strong HW disequilibrium and red triangles indicating regions discarded because of allele frequencies.

6.2.4 Determination of haplotype origin

The total number of observed reference haplotypes in the group of European wild boar was 920, and the total number of observed Asian reference haplotypes was 448. For each of the 4,764 haplotypes observed in Large White animals, we determined the Asian (AS) or European (EU) origin based on the frequencies of this haplotype in the European and/or Asian group of animals. For each region, we counted the number of unique haplotypes among the 448 haplotypes in the Asian group. Because we have unequal sampling, the number of unique haplotypes in a random sample of 448 haplotypes from the European group was also counted. Then, for each unique haplotype that was observed within the group of Large White haplotypes, we counted the number of times the haplotype was observed in the European and in the Asian group. To avoid a bias due to the (generally) higher diversity in Asia, we adjusted the counts for the amount of diversity in the Asian and European groups. This adjustment was done by calculating the ratio of unique haplotypes in both AS and EU groups. The number of times that particular haplotype was observed in the European group was multiplied by the proportion of unique EU compared to AS haplotypes, and the number of times the haplotype was observed in the Asian reference group was multiplied by the proportion of total observed Asian haplotypes. We then checked whether the corrected number of observed haplotypes in the European and Asian reference groups differed at least by a factor 4. If so, the haplotype was assigned to the group in which it was observed most. If not, it was assigned to the group for which both backgrounds were considered ("Both").

6.2.5 Cleaning of the data

Introgression regions in which the genotypes showed strong deviation from Hardy-Weinberg Equilibrium (P<0.00001) and the frequency of Asian haplotypes was <0.60 or >0.99 were excluded from further analysis. In order to test the independence of the evaluated regions, we also estimated the pairwise Pearson's correlations (r) for all regions using the recoded diplotypes. When two regions showed r >0.80, the shortest regions was excluded. After cleaning procedures, a total 1,384 animals with haplotype information on 11 regions were available for further steps (see supplementary information for more details).

6.2.6 Breeding values and association analyses

In this study we evaluated the traits backfat (BF) and litter size (LS). Deregressed estimated breeding values (DEBV) were used as response variable for each trait under study. The EBV was deregressed for each trait separately using the

methodology described by Garrick (2009). The EBV of each animal was obtained from the routine genetic evaluation by Topigs Norsvin using an animal model (pedigree-based). The model for BF included genetic line, sex, herd-year-month and weight as fixed effects and an additive genetic effect (animal) and a common litter effect as random effects. For LS, the model included genetic line, parity number, interval weaning-pregnancy (days), whether more than one insemination procedures were performed (yes or no) and herd-year-season, while the random effects consisted of service sire, a permanent effect to account for the repeated observations of a single sow and an additive genetic effect (animal). The reliabilities per animal for the purpose of deregression were extracted from the genetic evaluation based on the methodology of Tier and Meyer 2004. The heritabilities used for the deregression were also extracted from the routine genetic evaluation.
The association analyses were performed using the software ASReml (Gilmour 2009) applying the following model:

$$\boldsymbol{DEBV_{ij}\ w} = \boldsymbol{\mu} + \boldsymbol{R_i} + \boldsymbol{a_j} + \boldsymbol{e_{ij}}$$

where $\boldsymbol{DEBV_{ij}}$ is the observed DEBV for the animal $\boldsymbol{j}$, $\boldsymbol{\mu}$ is the overall DEBV mean of the population, $\boldsymbol{R_i}$ is the count of Asian haplotypes (AS) of the region $\boldsymbol{i}$, $\boldsymbol{a_j}$ is the additive genetic effect estimated using a pedigree-based relationship matrix and $\boldsymbol{e_{ij}}$ the residual error. The weighting factor $\boldsymbol{w}$ was used in the association analyses to account for the differences in the amount of offspring information available for the estimation of the DEBV (Garrick 2009). To ensure the quality of the DEBV, only animals with a $\boldsymbol{w}$ higher than zero and a reliability of the DEBV >0.20 were used. The association analyses were performed per region. In addition, a combined analysis was done where R represented the count of AS summed over all regions.

6.3 Results and Discussion

The purpose of this study was to determine whether Asian introgressed haplotypes in commercial Large White pigs mainly persist because they have an effect on selected traits. This was accomplished by performing an in-depth analysis of the introgression signals in the Large White population that were originally described by Bosse (2014a). Introgression mapping was performed by Bosse (2014a) to obtain the relative IBD signal (rIBD) in the Large White population for Asian and European background in 10kb-bins in a particular region in the genome. The principle of introgression mapping is described in figure 6.1A. The genome-wide rIBD values were used for the further analysis in this paper.

6.3.1 Effect of introgression on commercial traits

6.3.1.1 Selection of regions

We hypothesized that the pattern of introgression in the Large White population is mainly determined by artificial selection acting upon the Asian haplotypes. Following this rationale, the Asian introgression should persist mainly in those regions of the genome where the Asian variant has a favorable effect on a phenotype of interest. To test this, we extracted haplotypes in the introgressed regions and estimated the effect of their origin (European or Asian) on two commercially important traits: backfat thickness (BF) and litter size (LS). A total of 2,382 individuals from the commercial Large White line were genotyped for markers on the 60K SNPchip (Ramos 2009) that cover those regions. More specifically, we extracted 11 regions that had an introgression signal that persisted over more than 150 Kbp from the data presented by Bosse (2014a), and that passed our thresholds for data cleaning (see Methods and supplementary information for more details). The further analyses for these regions were based on this selection of 60K markers.

6.3.1.2 Effects per region

We evaluated whether these 11 regions were significantly associated with the traits BF and LS. None of these regions were found to have a significant effect on LS (table 6.1). However, six of these 11 regions showed a significant association with BF (table 6.1). For all these significant regions, we observed an increase in BF when an European haplotype was replaced with an Asian haplotype.

Most introgressed regions were identified on chromosome 2 (figure 6.1B, 6.2A) with the strongest effect in the gene-dense region 2_2 (0.22 mm). This region contains multiple genes coding for intercellular adhesion molecules (*ICAMs*) that have been shown to have an effect on obesity (Dong 1997). Whether the effect of BF is caused by these genes is however unclear, since the regions contains a total of 39 annotated genes in the current Ensembl release 76.

Region 2_7 is the region with the strongest introgression signal on chromosome 2 and therefore it was used as an example of the applied method in figure 6.2. The substitution of an European haplotype for an Asian haplotypes in this region on average increased backfat by 0.17 mm. As can be seen in the Ensembl annotation for this regions (figure 6.2B), one candidate gene, *COMMD10*, lies within the peak of region 2_7. COMMD proteins contain a conserved and unique 'COMM' domain involved in cellular homeostasis including copper and the NFkβ pathway, and at

least 10 COMM family members exist that are conserved in all vertebrates (Maine 2007). Murgiano (2010) found another *COMMD* gene differentially expressed in *longissimus lumborum* muscle samples between Large White and Casertana pigs, and suggest that *COMMD* negatively regulates NFkβ signaling which in turn can result in triggering the adipogenetic cascades.

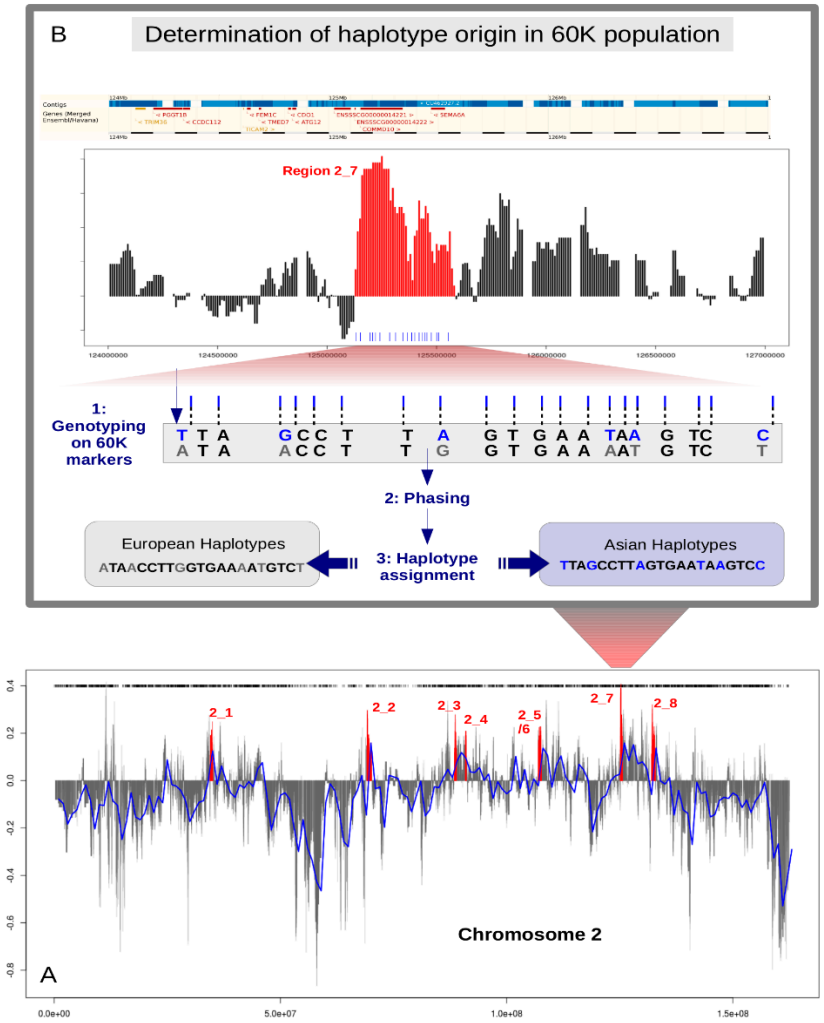


Figure 6.2 Determination of haplotype origin in introgression regions. Chromsome 2 contains most of the introgression regions, which are indicated in red in **A**. The x-axis represents the full length of chromosome 2 and the y-axis contains the rIBD. The blue line displays the smoothed rIBD signal based on 1Mb bins, and the grey bars indicate the raw

rIBD values for each 10kb bin. The selected regions of introgression are highlighted in red, and the location of markers on the Illumina porcine 60K beadchip are indicated as vertical bars in black above the chromosome. **B**. Determination of the origin of haplotypes in introgressed regions consists of 3 steps: 1) Genotyping of all 9970 individuals for the markers that cover the introgression region; 2) Phasing of haplotypes in the introgression region with the full dataset; 3) Comparison of the haplotypes in the commercial line to the haplotypes in the European population and the haplotypes in the Asian population, and assignment of the haplotypes to one of these two groups. In this example we focus on region 19 (Ssc2: 125.12-125.59 Mb) that covers the full coding sequence of the candidate gene COMMD10, as can be seen in the Sscrofa10.2 Ensembl annotation (V.76.102).

The other significant region (2_6) on this chromosome contains 7 candidate genes for increased BF (*CAST, ERAP1, ERAP2, LNPEP, LIX1 RIOK2* and *RGMB*). *LNPEP* is an insulin-regulated amino peptidase which acts as membrane protein associated with glucose transporter vesicles in cultured mouse adipocytes according to Larance (2005). *CAST*, calpastatin, has no direct known function in fat synthesis, but interestingly it is a well-known locus involved in meat tenderness in multiple species and studies, see: Ropka-Molik (2014) and meat quality traits of pigs in general (Ciobanu 2004).
Regions 9_3 and 9_5 are both located on chromosome 9, in the vicinity of the *AHR* gene that was identified in Bosse (2014a), with an effect of ~0.15 mm of BF per Asian haplotype (table 6.1). Region 9_3 overlaps with the AHR gene, suggesting that this gene is involved in multiple biological processes, as discussed by Denison (2011) and Hernandez-Ochoa (2009). Region 9_5, in addition to its proximity to *AHR*, contains two *TWIST* neighbor genes (*TWISTNB*) that are involved in transcription and the *TMEM196* gene coding for transmembrane protein 196. The last significant region 18_1 on chromosome 18 contains only one gene, protection of telomeres1 (*POT1),* that has previously been shown to have a higher expression level in multiple tissues from the fat-type Wujin pigs, compared to Large White pigs, including *longissimus dorsi* muscle (Yong 2012).
In summary, for the 11 regions with a strong introgression signal, the Asian haplotypes displayed a significant positive effect on BF in the majority of regions. Zooming into the genes in the BF-associated regions, a multitude of candidate genes could be identified that possibly caused the effect on BF. We suggest further experiment that focus on these specific genes to confirm their role in the accumulation of BF in pigs.

Table 6.1 Effect of Asian haplotypes on backfat (BF) and litter size (LS) in introgression regions. An animal model was used to estimate the effect of Asian haplotypes on the deregressed estimated breeding values for the two traits. P-values of significant regions (P<0.10) are given in bold. Effect for BF is in mm BF per Asian haplotype, and effect for LS is in number of piglets per Asian haplotype. 'All combined' indicates the regression analysis over all 11 regions.

Region	BF		LS	
	P-value	Effect	P-value	Effect
15_1	0.560	-0.07	0.110	0.21
18_1	0.024	0.15	0.617	-0.03
2_2	0.002	0.22	1.000	0.00
2_4	0.420	0.07	0.471	0.07
2_6	0.027	0.15	0.203	0.09
2_7	0.011	0.17	0.234	-0.09
3_1	0.527	-0.05	0.729	0.03
6_1	0.299	-0.42	0.348	-0.41
9_1	0.823	-0.05	1.000	0.00
9_3	0.081	0.14	0.590	0.05
9_5	0.054	0.13	1.000	0.00
All combined	<0.001	0.09	0.610	0.01

6.3.1.3 Additive effect over regions

The effect of the Asian haplotypes is an increase in BF in all significant regions, suggesting that the Asian haplotypes could have an additive effect on BF over all regions. We examined the association of Asian haplotypes with BF and LS combining all regions (summing the count of Asian haplotypes of all regions). We performed the regression of the counts of Asian haplotypes (ranging from 9 to 22, figure S6.1) on both BF and LS. For LS, the association test was not significant, but for BF the association was even stronger than when individual regions were analyzed (table 6.1). For BF, an additive effect of 0.09 mm of BF was observed per Asian haplotype that replaced an European haplotype (table 6.1). The overall phenotypic standard deviation of BF in the Large White population was 1.61 mm. The contrast between Asian and European homozygotes was around 0.20 mm of BF when averaged over haplotypes, equivalent to about 0.12 phenotypic standard deviations of BF. Based on individual regions, an animal with only Asian haplotypes (n=2) will have 0.40 mm of BF more than an animal that carries only European haplotypes. Analyzing all 11 regions together, an individual that presents only Asian haplotypes (n=22) will show 1.98 mm of BF more than an animal that presents only European haplotypes, which means 1.23 phenotypic standard deviations of BF.

Chinese breeds were thought to be superior for the traits fatness and litter size according to the early European pig farmers, and these traits were artificially selected after introgression (White 2011). Our initial hypothesis was, therefore, that the regions with a strong introgression signal would have a significant effect on both BF and LS. However, our results showed that regions of introgression only have a significant effect on BF. Selection signatures for complex traits do not necessarily leave a sweeplike signature in the genome (Kemper 2014, Heidaritabar 2014b). This could explain why none of the introgressed regions display a significant association with LS. Indeed, if we look at the previous finding for the *AHR* gene (Bosse 2014a) it is one particular Asian allele rather than all Asian haplotypes at that locus that have the effect on LS. Another explanation why the introgressed Asian haplotypes have no effect on LS could be that the specific loci that are involved in a complex trait like litter size contain genes that are also involved in other (life history) traits. The pleiotropic nature of these genes may restrict the selection on Asian haplotypes, resulting in less obvious signatures in the genome than for BF related genes, although this is speculative.

Our results demonstrate that by screening a population for signals of introgression, regions can be pinpointed where introduced haplotypes have an effect on selected traits. Popular methods that are developed to detect selective sweeps in a population, like F_{st} (Weir 1984) or homozygosity tests, use increase or reduction of genetic variation as a signal. Ongoing selection for introduced haplotypes that are genetically more diverse or distant than haplotypes from the source population, will not be picked up by these methods. We therefore suggest consideration of alternative methods when the studied population has a known history of admixture, or when the goal is to screen specifically for adaptive introgression.

6.3.2 Introgression and genome characteristics

6.3.2.1 General patterns

Our hypothesis was that artificial selection is the main factor in shaping the introgression pattern in Large White pigs. The results of the association analysis support the hypothesis that introgression signals are enriched for associations with a commercially interesting trait. In line with this hypothesis, general genome characteristics like gene density and recombination frequency should contribute little to the introgression pattern. To assess whether the introgression signal was correlation with gene density or recombination frequency, the rIBD was averaged over 1MB bins over the genome. We found a very modest significant negative correlation of -0.05, as well as a significant correlation between rIBD and (log

transformed) recombination frequency of 0.10 (figure 6.3). In a recent study on Neanderthal introgression in modern humans, gene deserts were enriched for Neanderthal ancestry (Sankararaman 2014). This finding suggests that purifying selection removed the majority of Neanderthal haplotypes from the population, and that introgressed haplotypes mainly occur in regions with relaxed selection pressure. In pigs, these general patterns in the genome explain only a fraction of the variation in introgression signatures. The circumstances of introgression in these two species are however very different, since the admixture in commercial pigs has been deliberate, and selection for some of the introgressed haplotypes is expected to be positive because of the known differences between Asian and European pigs. Because we have shown that the majority of identified introgression regions have a significant effect on BF, our expectation is that it is the characteristic of a single gene or gene cluster in a particular region, rather than the gene density or recombination frequency that causes the level of introgression in commercial pigs. According to Hedrick (2013), the probability of an introgressed haplotype to be maintained in a population is strongly increased when it has some selective advantage.

Although the purpose of the introgression of Asian haplotypes was to maintain beneficial variants in the population, probably not all introduced Asian haplotypes had the desired effect on the European stock. Selection pressure on Asian haplotypes could have been either positive or purifying, and should have resulted in either an excess or an absence of the Asian variants, depending on the location in the genome and the associated genes in those regions. Commercial pig breeds are known to be under strong artificial selection, and our results have shown that indeed a majority of the strongly introgressed regions had a significant effect on BF. Therefore, we expect that regions with an excess or absence of Asian haplotypes contain genes of commercial interest. The 1% extreme tails of the introgression distribution (table S6.2) were scanned for genes that are known to be related to commercially important traits. Introgressed regions should contain genes that have an effect on traits that are present in the Asian breeds and that had an selective advantage in the European breeds. By contrast, within those regions that do not contain Asian haplotypes, we expect to identify genes that have an effect on traits that are typical for European breeds.

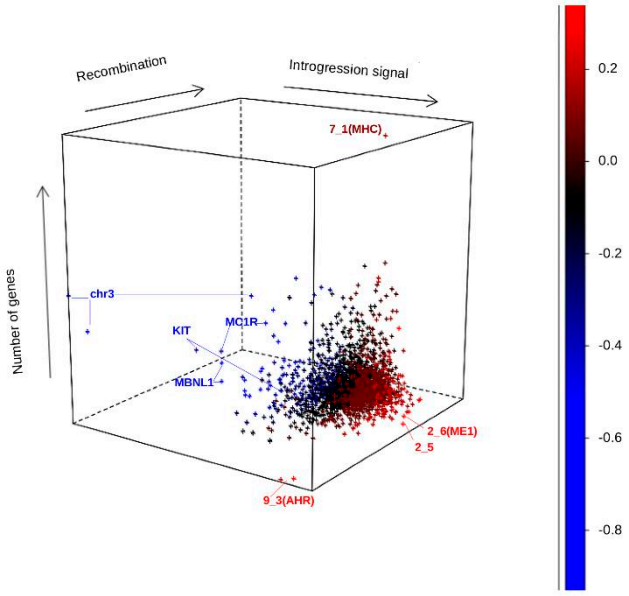


Figure 6.3 rIBD and genome characteristics. For each 1Mb bin in the genome, the rIBD signal was plotted against the number of genes and the recombination frequency in the bin. Coloration is based on the rIBD signal so that bright red indicates bins with the strongest Asian introgression signal, and blue bins indicate the strongest underrepresentation of Asian haplotypes. Number of genes per bin ranged from 0 to 128 and recombination frequency ranged from -16.6 to 3.3 (log2-transformed) cM/Mb. Bins containing interesting genes that are discussed in the main text are indicated.

6.3.2.2 Fat related genes

The two 1Mb bins containing the highest rIBD scores are both on chromosome 1. When we look for candidate genes in the first bin, the malic enzyme 1 (*ME1*) gene has previously been described as an important QTL region for backfat thickness and meat quality (Vidal 2006, Bartz 2013, Ramirez 2014). This bin overlaps with region 1_6 from the association analysis, but that region was discarded because of an excess of heterozygotes. The second bin contains *AMD1* (cell proliferation, polyamine synthesis pathway) and zeta catalytic subunit DNA polymerase (rev3L in human) as candidate genes. This bin overlaps with region 1_5 from the association analysis, but it was also discarded due to an excess of heterozygotes.

On the beginning of chromosome 12 we identified another bin with a high rIBD signal, hinting at a locus that contains Asian introgressed haplotypes. This region

overlaps the FASN gene that was previously described as fatty acid synthase and an important gene involved in fat deposition (Braglia 2014). In addition to the genes described in this paragraph, our 11 regions that are used in the association analysis can be found in the 25 bins that span the top 1% of introgressed regions in the genome. These findings hint at a prominent role for selection on fatness in shaping the introgression landscape in Large White pigs.

The reason for the observation that more heterozygous individuals are observed than one would expect based on Hardy-Weinberg equilibrium remains unclear. Apart from the technical issues, these regions are potentially very interesting if the signal is genuine. Balancing selection or a heterozygote advantage can result in more heterozygous individuals than expected. Also, if selection for a previously low-frequency haplotype is ongoing, an excess of heterozygous individuals could be observed. As shown by Merks (2000 and 2012), BF was included in the selection index of breeding companies over 100 years ago, and nowadays leanness is preferred in commercial pigs. This switch in preference and direction of selection could result in more heterozygosity in some regions. Merks (2000 and 2012) also describe that selection for LS and other commercially interesting traits has only heavily started in the 1990's. If some regions have an effect on both BF and another ore multiple other traits, this selection in different directions could result in more heterozygous animals at these loci as well. Further experiments for the regions with an excess of heterozygotes should indicate what causes this peculiar pattern.

6.3.2.3 Pigmentation genes

Skin pigmentation is an important trait for modern pig breeders, and therefore we expect a distinct introgression signal at pigmentation loci (figure S6.2). In the top 1% of bins with low Asian introgression are two regions that have previously been identified as important candidate regions for coat coloration in pigs, containing the *KIT* gene and the *MC1R* gene. The *KIT* gene is a very well-known gene that is involved in coat color (Okumura 2008). The European pigs contain a copy number variable dominant white allele, mostly in homozygous form (Rubin 2012, Paudel 2013). We clearly see European haplotypes surrounding this gene rather than Asian, suggesting selection against introgression at this locus. Also present in this peak region is *MAP9*, a gene involved in mitotic spindle formation (figure S6.2). *MC1R* is known to be involved in pigmentation in multiple species including pigs. No introgression is expected in Large Whites for this region based on previous results (Fang 2009), and indeed we see a clear lack of introgression at this locus (figure S6.2). Interestingly, among those bins that contain the highest introgression

signal in the Large Whites, were two other genes found to be involved in pigmentation (*TYR* and *RAB38*). *RAB38* and *TYR* are both involved in pigmentation patterns of skin, eyes and hair, according to a multitude of different studies (Loftus 2002, del Marmol 1996). Tyrosinase seems to have some temperature-dependent coloration patterns, but different forms exist. *RAB38* lies within region 9_1, and has been identified in Wilkinson et al. 2013 as well as a region in Large White pigs that is introgressed and selected.

6.3.2.4 Morphology

The 1Mb bin covering *MBNL1* (muscleblind-like splicing regulator 1 gene) has the lowest introgression signal. In human and mouse, this gene has been shown to be associated with muscle dystrophy (Kanadia 2003). Since European commercial pigs and Asian pigs are known to be very different in terms of muscle content, selection for muscle related traits in European pigs might select against Asian variants in this region. In the introgression peak at the very end of chromosome 8 we identified another interesting gene, *BMP3* (bone morphogenesis protein 3), that has previously been identified as being involved in growth restriction in human (Bonnet 2010) and a mutation in this gene has an effect on skull shape in zebrafish and dogs (Schoenebeck 2012). This region was also found to be introgressed and selected in LW pigs in Wilkinson (2013).

6.3.3 Concluding remarks

With this work we demonstrate that the introgression landscape in Large White pigs is strongly determined by the selective pressure acting upon the introgressed Asian haplotypes. The majority of the regions with a high frequency in Asian haplotypes turn out to have an effect on backfat thickness, and many of these regions overlap with previously identified fat-related genes. The fact that the proportion of Asian material is relatively similar for most European commercial breeds (Groenen 2012, Bosse 2014b) suggest that the introgression occurred before the establishment of modern breeds. After many generations since the introgression, Asian haplotypes could have been purged if they had a selective disadvantage. On a genome-wide scale, however, we observed a general pattern of low frequency of Asian haplotypes, suggesting a more neutral scenario for the introgressed genetic material in this Large White pig population. Regions with an excess or absence of Asian haplotypes indeed contain genes where the Asian variants are thought to have an effect on phenotypes of interest, and therefore we illustrate that human-driven introgression and selection broadly shaped the genomic architecture of commercial pigs. Our findings provide a unique insight

about how the selection history of pig breeding influenced the genomic haplotype patterns of the commercial breeds that we know today. How general this introgression pattern is, should be pointed out by future studies on other organisms that likely experienced introgression and consecutive selection.

6.4 Acknowledgements

The DNA samples that we used for for genotyping were provided by Dr Ning Li, China Agricultural University, China; Dr Alain Ducos, UMR INRA-ENVT, France; Sem Genini, Parco technologico Padano, Italy; Dr Gono Semiadi, Puslit Biologi, Indonesia; Dr Naohiko Okumura, Staff Institute 446-1 Ippaizuka, Japan; Dr Alan Archibald, Roslin Institute and the Royal (Dick) School of Veterinary Studies, University of Edinburgh , Scotland; TOPIGS Research Center IPG BV, The Netherlands; Dr Oliver Ryder, San Diego Zoo, USA; Cheryl L. Morri, Ph.D., Omaha's Henry Doorly Zoo, USA. We thank Bert Dibbits for lab work, B. van de Water for graphics production and Yogesh Paudel and Egbert Knol for valuable discussion.

7

Genomic Data in Population Management: Implications for Conservation and Selection Programmes

M Bosse[1], H.J. Megens[1], O Madsen[1], R.P.M.A. Crooijmans[1], O.A. Ryder[3], F. Austerlitz[2], M.A.M. Groenen[1] and M.A.R. de Cara[2]

[1] Animal Breeding and Genomics Group, Wageningen University, Wageningen, The Netherlands; [2] Muséum national d'histoire naturelle, Paris, France; [3] San Diego Zoo Institute for Conservation Research, San Diego, USA

Genome Research (under review)

Abstract

Conservation and breeding programmes aim at maintaining most diversity, therefore avoiding deleterious effects of inbreeding while maintaining enough variation from which traits of interest can be selected. Theoretically, most diversity is maintained using optimal contributions based on many markers to calculate coancestries, but this can decrease fitness by maintaining linked deleterious variants. The heterogeneous patterns of coancestry displayed in pigs make them an excellent model to test these predictions. We propose a method to measure coancestry and fitness from resequence data, and implement this into population management. We analysed resequencing data of *Sus cebifrons*, a highly endangered species from the Philippines, and genotype data from the Pietrain domestic breed. We demonstrate that the maintenance of both diversity and fitness depends on the genomic distribution of deleterious variants, which is shaped by demographic and selection histories. By analysing the demographic history of *Sus cebifrons* we inferred two past bottlenecks that resulted in some inbreeding load. In Pietrain, we analysed signatures of selection possibly associated with commercial traits. We *in silico* managed each population to assess the performance of different optimal contribution methods to maintain diversity, fitness and selection signatures. Maximum genetic diversity was maintained using marker-by-marker coancestry, and least using genealogical coancestry. Using a measure of coancestry based on shared segments of the genome achieved best results in terms of diversity and fitness. However, this segment-based management eliminated signatures of selection. Our findings show the importance of genomic and next-generation sequencing information in the optimal design of breeding or conservation programmes.

Key words: coancestry, genomic data, management, pigs, conservation, selection

7.1 Introduction

The main goal in population management is to maintain genetic diversity, which in turn can maximize survival potential of the population, and provides the opportunity to select variants that have fitness consequences (Frankham 2002). Conservation programmes usually use small numbers of breeding individuals, which means genetic variation within the population is likely to decrease rapidly. In commercial breeding programmes, it is known that artificial selection can lead to a reduction in overall diversity and an increase in inbreeding. This can have highly detrimental consequences if breeds lose their ability to adapt to different environmental conditions, or if disease alleles are linked, pleiotropically or through genetic linkage, to selected trait alleles. Conservation and commercial breeding programmes, therefore, are not so different in their approach of managing their populations, even though their ultimate goals differ. Inbreeding depression (Charlesworth and Charlesworth 1987) is a common phenomenon in captive populations for many wild species like wolves (Laikre and Ryman 1991) and can be severe as well in domesticated species (e.g. dogs, Leroy 2011, O'Neill 2014). Thus, breeders are aware of the need to maintain diversity, while also preserving genetic variants that confer desired, distinct, phenotypes.

For this purpose, controlling the inbreeding rate and, therefore, optimising the effective population size, is required. Currently the best known method to achieve these goals is optimal contributions (OC). This method relies on determining the number of offspring that each individual of the current population should contribute to the next generation, so as to minimize global coancestry (Ballou and Lacy 1995, Meuwissen 1997, Grundy 1998). Relatedness, i.e. coancestries between individuals, is needed to apply OC in any management programme. Traditionally, genealogical coancestries were used when OC were first proposed, as marker data were scarce (Ballou and Lacy 1995). Currently, OC based on molecular coancestry (identity by state, IBS) is the best way to maintain the most diversity in terms of heterozygosity, provided that genotypes with a high enough marker density are available (de Cara 2011, Gomez-Romano 2013). However, management of populations using OC with molecular coancestry may lead to a fitness decrease, since deleterious alleles linked to the markers used to measure coancestry will be maintained (de Cara 2013a). Recently, a measure of coancestry based on shared genomic regions has been proposed, as a compromise to maintain both fitness and genetic diversity when the population in the programme has a medium to high inbreeding load (de Cara 2013b). One of the aims of this approach is to avoid the

occurrence of long runs of homozygosity (ROH) in the offspring. Long ROHs are characteristic of reduced diversity, due either to selection or bottlenecks, and, therefore, may confer inbreeding depression (Keller 2011, Szpiech 2013). Determining the occurrence of segments of identity by descent (IBD) in potential parents, thereby measuring their relatedness and coancestry, can be used to minimize the occurrence of ROHs in the offspring.

Predictions for management based on genealogical, molecular or IBD segments have been made with simulated data (de Cara 2013b), but have so far not been tested with actual genotype data. Pigs are an excellent model to test the use of genetic data in population management. Various molecular data sources are available, like a high-quality genome reference (Groenen 2012) and genotyping arrays (Ramos 2009). In addition, pedigree information is available for a variety of breeds and other captive populations. Pigs display a high degree of heterogeneity in the occurrence of ROHs (Bosse 2012), which to a large extent reflects differences in demographic histories of populations. The domesticated pig *Sus scrofa* consists of many commercial breeds that are under strong artificial selection for commercial traits, but that should simultaneously maintain high levels of fitness and diversity. While this particular species is widespread in captivity as well as in the wild, other pig species within the genus *Sus* only occur on a few islands in South-east Asia and are critically endangered, like the Visayan warty pig *Sus cebifrons*. Breeding programmes for *S. cebifrons* in zoos are now part of the conservation programme in order to maintain the species at least in captivity.

In this study, we combined pedigree information, genotype data and next-generation sequencing data to perform *in silico* management of two pig populations with distinctly different management histories: a commercially maintained population of the Pietrain breed of *S. scrofa*, and a captive zoo population of the critically endangered *S. cebifrons* species. By comparing the decay of heterozygosity over 10 generations, we assessed which of three management strategies maintained the most diversity. These three management strategies were based either (i) on genealogical coancestry, or (ii) on molecular marker-by-marker coancestry, or (iii) on shared regions of the genome. As *S. cebifrons* is known to have undergone recent bottlenecks which have led to establishing the conservation programmes for this species in captivity, we analysed the demography of *Sus cebifrons* to determine the effect of population-specific demography on the management outcome. On the other hand, to understand the effect of the management strategy on ongoing selection in the population, we

identified signatures of selection in the Pietrain breed before and after management. In this way, we investigated whether the best strategy is sensitive to demographic history or initial patterns of variation in the population and how this information could be relevant to conservation practitioners, while by analysing signatures of selection we addressed whether any of the management strategies may erase these signatures, which could interfere with selection goals. Finally, by predicting fitness based on deleterious variants in the genome in both populations, we analysed the performance of each of these management strategies not only on diversity but also on fitness, as ignoring the latter could lead to the accumulation of deleterious variants, loss of viability and potentially extinction of the conservation population.

With the fast development of genomic tools and the reduction in sequencing costs, it is not unrealistic that, within the near future, full genomes of many individuals and species can be sequenced cost efficiently. The identification of deleterious variants and assessment of shared variation in populations is becoming a valuable tool for conservation and management purposes. This study is thus a significant contribution towards the implementation of genomic data into breeding programmes.

7.2 Results

Data from two pig populations were used for the *in silico* management: re-sequencing data from five *S. cebifrons* individuals from San Diego zoo (5) and genotypes from 46 and sequence data from 11 individuals of the Pietrain breed of *S. scrofa*. Pedigree data were available for both populations, covering several generations for the Pietrain breed, while for *S. cebifrons* the pedigree available only covers the individuals since the foundation of the conservation programme. At the starting point for the management, the *S. cebifrons* population consisted of 5 individuals (1 male and four females) that we expanded to 10 individuals (2 males and eight females) before the first generation of management, with genotypes for 104.035 sites, and the Pietrain population contained 47 individuals genotyped for 51165 sites.

7.2.1 Status before management

7.2.1.1 Genetic diversity

We assessed the genetic characteristics of both populations. The average nucleotide diversity over the full genome was higher within Pietrain pigs (mean

π=0.00175), although the distribution of variation was much more homogeneous in the genome of *S. cebifrons* pigs (mean π=0.00105 figure 7.1). Both populations contained signatures of recent coancestry between haplotypes in their genome, either as ROH within single genomes or as shared IBD segments between individuals. Observed heterozygosity, based on the selected markers that were used for the *in silico* management, was 0.308 for Pietrain pigs and 0.302 for *S. cebifrons*, and genetic diversity was 0.296 in the Pietrain population and 0.258 in the *S. cebifrons* population. These numbers were used solely for comparison purposes between the different management strategies after 10 generations. The relatedness within the *S. cebifrons* population, based on their genealogy, ranged from unrelated to half-sibs, as can be seen in their pedigree and phylogenetic tree (figure S7.1). Relatedness within the Pietrain population was more evenly distributed as can be seen in the phylogenetic tree in figure S7.2.

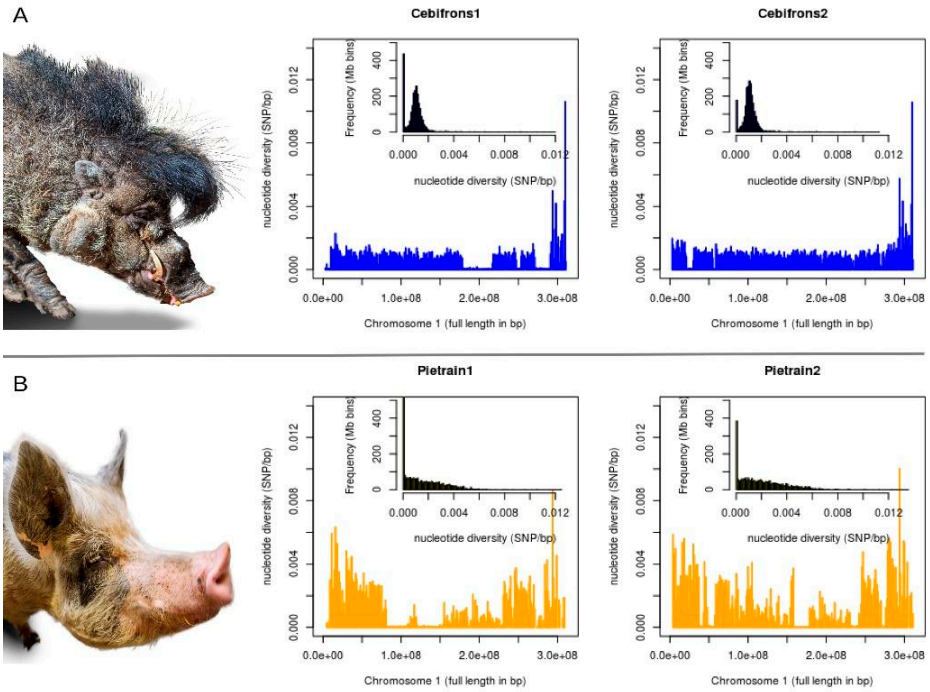


Figure 7.1 Genome-wide variation in individual pigs. The nucleotide variation within the genome of two individuals for each population is displayed. The large box shows on the x-axis the full length of chromosome 1 in bp, and on the y-axis the nucleotide diversity within an individual genome in number of heterozygous sites (SNPs) per bp. Values are averaged over bins of 1Mbp. The small histogram within each box represents the distribution of nucleotide diversity per bin of 1Mbp for all autosomes. **A**. The zoo population of *Sus cebifrons*. **B**. The commercial Pietrain population.

7.2.1.2 Demography of Sus cebifrons

The zoo population of *S. cebifrons* consisted of a few individuals that have only recently been transferred from the wild. To understand their low diversity compared to other closely related species (see also Bosse 2012), we analyzed their demographic history, for which we used two independent methods to reconstruct the past effective population size over time: pairwise sequential Markov chain (PSMC) developed by Li and Durbin (2011) and a method based on the distribution of lengths of ROHs in the current population developed by MacLeod (2013). Both

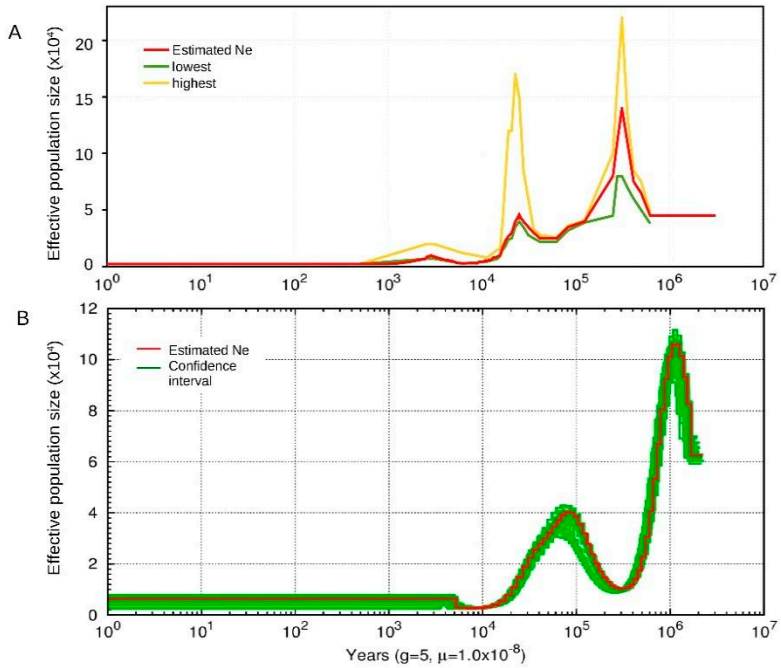


Figure 7.2 Demographic history of *Sus cebifrons*. Two methods were used to infer the demographic history of the wild *Sus cebifrons* population, based on the genome sequence of one male: **A**. The method by MacLeod (2013). The original estimated Ne is visualized in red, and the upper and lower limits of the confidence interval are indicated in yellow and green, respectively. **B**. The method by Li and Durbin (PSMC). The original estimated Ne values for the same individual are indicated in red and the green lines represent 10 bootstrap estimates for this particular individual and the other 4 individuals from the same population. For both methods, the generation time was set at 5 years, and the mutation rate at 1.0×10^{-8} and scaled for the false negative rate. The x-axis displays the time before present in years, and the y-axis displays the estimated effective population size.

methods indicated two population expansions, followed by a bottleneck (figure 7.2). All five individuals displayed similar demographic patterns (figure 7.2B). The effective population size has been relatively low since the last bottleneck ~10.000 years ago.

7.2.1.3 Deleterious variants

Because demographic history largely determines the inbreeding load in a population, we looked at putative damaging mutations in the five re-sequenced *S. cebifrons* individuals and 11 Pietrain pigs. The total number of predicted deleterious variants within the *S. cebifrons* population was 3129, and a gene ontology (GO) enrichment analysis showed generally an under-representation of genes coding for nucleotide binding proteins and an over-representation of genes coding for cell adhesion molecules (table S7.1). The number of deleterious sites within the 11 re-sequenced Pietrain pigs was 3468. Interestingly, an obvious overlap in GO-terms was found in the enrichment analysis in the Pietrain and *S. cebifrons* populations (table S7.1). Genes involved in transcription, RNA metabolic processes and nucleic acid binding had significantly less deleterious mutations than expected in both groups. Most predicted deleterious sites were found in heterozygous state in only one individual. These deleterious sites were not used for the calculation of coancestries during the *in silico* management, but their presence in individuals during management was used for the fitness calculations (see Methods for details).

7.2.1.4 Signatures of selection in Pietrain

The Pietrain population consists of commercially bred pigs that have been under strong artificial selection pressure for at least 30 generations. This selection is thought to result in long haplotypes in some regions in the genome with reduced variation. Therefore, we screened the genomes of Pietrain pigs for presence of ROHs and identified numerous regions containing ROHs (figure S7.1). Length distribution of ROHs ranged from 3.0 Mb to 132.1 Mb and the average ROH length was 12.8 Mb. Chromosome 10 contained the least ROH coverage and chromosome 8 the most. Some ROHs are clustered within a particular region of the genome, while others appear more randomly distributed. Chromosomes 8 and 15 contain many ROHs at the same locus, indicating selection for a specific haplotype that is close to fixation. We then used the extended haplotype homozygosity (EHH) test to detect regions under ongoing selection resulting in partial sweeps. The strongest signatures of ongoing selection were found on chromosome 13 and 15 (figure 7.3).

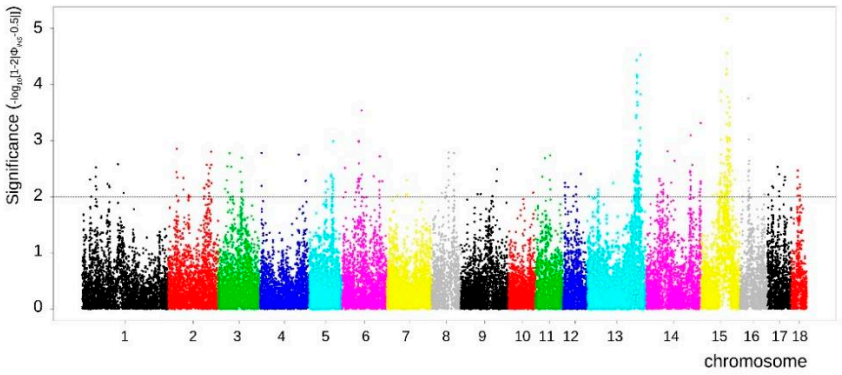


Figure 7.3 Signatures of selection in the Pietrain population. Extended haplotype homozygosity per chromosome for the Pietrain population. Values are based on the raw iHS signal before management over all chromosomes in the Pietrain population. The x-axis displays the location on all 18 autosomes and the y-axis shows the p-value of the iHS signal before management for each marker. Values >2 are considered to be significant according to Voight (2006).

7.2.2 In silico management

7.2.2.1 Genetic diversity

We recorded the observed heterozygosity and genetic diversity in both populations during the ten generations of simulated management of the two species. Management based on measures of molecular coancestry maintained the highest level of genetic diversity in both populations (figure 7.4). Management based on IBD segments performed better in terms of diversity than management based on genealogical coancestry, with management based on longer segments maintaining less variation than when shorter segments were used to measure coancestry. The loss of genetic variation over time was stronger in the *S. cebifrons* population than in the Pietrain population, and the difference in decay of variation between management strategies was larger in the *S. cebifrons* population as well (figure 7.4A, B). After ten generations of optimal contributions based on genealogical coancestry, the observed heterozygosity in the *S. cebifrons* population dropped from ~0.3 to below 0.18, while the molecular-based management strategies maintained the level of observed heterozygosity above 0.21 (figure 7.4A). After the first generation of molecular-based management, the Pietrain population contained slightly more genetic variation than the base population in terms of

heterozygosity, but the variation dropped from the second generation onwards (figure 7.4B).

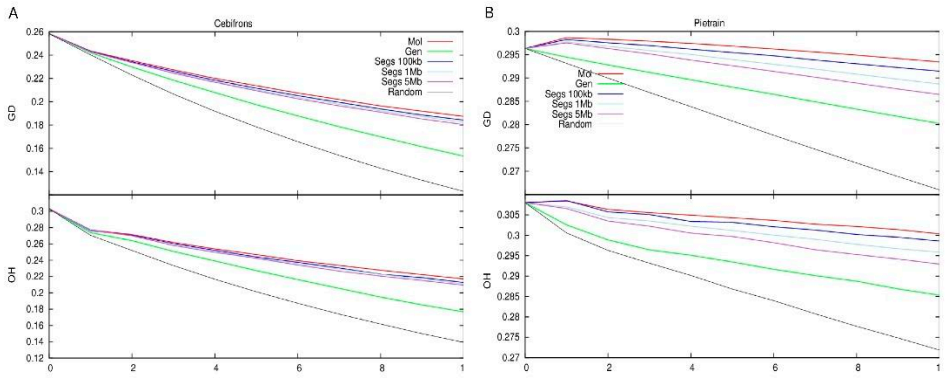


Figure 7.4 Variation in both pig populations during management. Evolution of genetic diversity (GD) and observed heterozygosity (OH) during 10 generations under five different management strategies. **A**. Variation in the *S. cebifrons* population. The population size is kept constant at 10 individuals with a male to female ratio of 1:4. **B.** Variation in the Pietrain population. The population size is kept constant at 50 individuals with a male to female ratio of 1:1.

7.2.2.2 Effect of management on fitness

Fitness of individuals in the *S. cebifrons* population was measured based on the deleterious sites in their genome. We recorded the total number of deleterious sites in the population, and also the proportion of homozygous and heterozygous deleterious variants. Then, based on different selection and dominance coefficients we calculated the average fitness for each management strategy through time. The proportion of deleterious variants in individuals remained relatively stable over time, regardless the management strategy (figure 7.5A). However, the proportion of homozygous carriers of the deleterious variants gradually increased, while the proportion of heterozygous carriers conversely decreased. The slopes were steeper under the genealogy-based management strategy than under marker-assisted management, and steepest with random management. In figure 7.5B and 7.5C it can be seen that the effect on fitness of the population depended on the mean selection (*s*) and dominance (*h*) coefficients of the deleterious variants. The overall fitness decline after ten generations was stronger with *h*=0.35 than with *h*=0.5. Based on predictions made on simulated data (de Cara 2013b), we expected that the strategy that maintained the most diversity would be the one that performed the worst in terms of fitness. However, and maybe partly due to the small

population size here managed, the genealogical strategy resulted both in maintaining less fitness and diversity than any marker-based strategy (figure 7.5B, C, D). Management based on shared segments and marker-by-marker performed relatively similarly in terms of fitness for the *S. cebifrons* population, but the exact size of shared segments for which the most fitness was maintained depended on a balance between segment size to measure coancestry and the distribution of mutational effects (i.e., both on the mean values of s and h and their distribution),

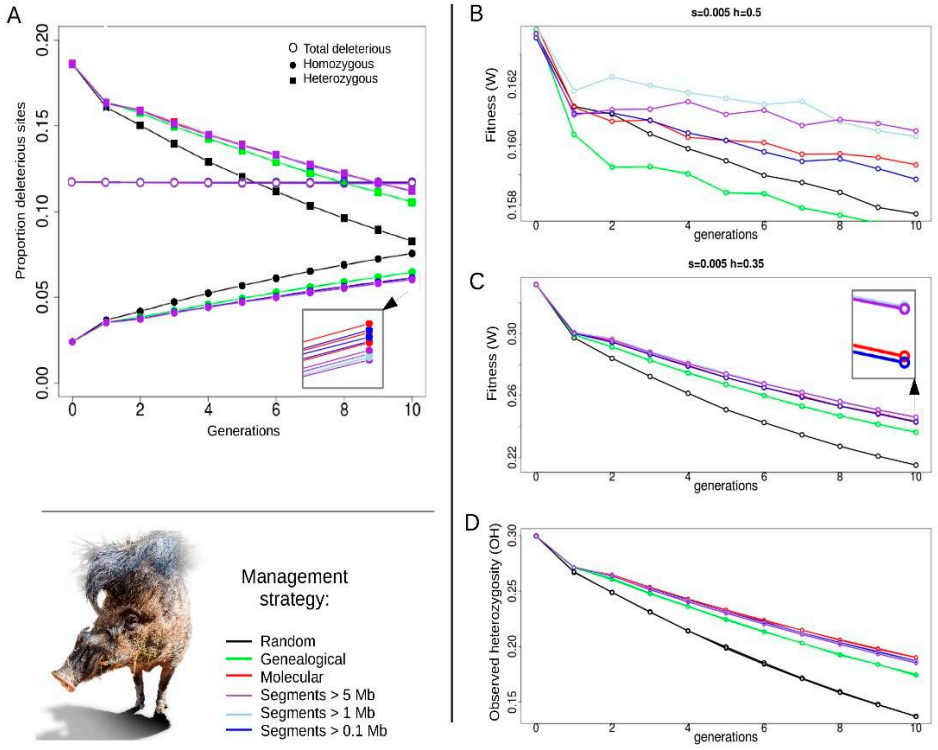


Figure 7.5 Fitness and diversity during management of the *S. cebifrons* population. The change in fitness and observed heterozygosity (OH) during 10 generations of management is displayed for 5 different management strategies. **A**. The proportion of deleterious sites is displayed on the y-axis and the generations are shown on the x-axis, for the scenario in 5B (s=0.005 and h=0.5). Deleterious sites are split into homozygous sites (filled circles), heterozygous sites (filled squares) and total number of deleterious sites (open circles). **B**. Fitness change over 10 generations of management when a dominance coefficient of 0.5 is applied. **C**. Fitness change over 10 generations of management when a dominance coefficient of 0.35 is applied. **D**. Observed heterozygosity during 10 generations of management.

as the length of a particular segment is related both to the distribution of mutational effects and to the demographic history. Concerning fitness, the loss of heterozygous deleterious variants was slowest in the marker-by-marker coancestry and the increase in homozygous deleterious variants was slowest in the segment-based management for *S. cebifrons* (figure 7.5A). These factors result in a slightly higher fitness in the segment-based managed population, especially for the small *h* value (figure 7.5B, C).

After 10 generations of management, the fitness in the Pietrain population was lowest in the population that was managed based on the pedigree under both scenarios (figure S7.4). This strategy maintained the least fitness under both scenario 1 (*s*=0.005 and *h*=0.5 figure S7.4A) and scenario 2 (*h*=0.35, figure S7.4B), and it performed even worse than managing at random under scenario 1. Although the marker-by-marker-based management outperformed the segment-based management in terms of observed heterozygosity (figure S7.4C), more fitness was maintained with the intermediate segment-based management under scenario 1 (figure S7.4A), and roughly the same with the intermediate and long segments in scenario 2.

7.2.2.3 Effect of management on selective sweeps in Pietrain

We used the iHS statistic (Voight 2006) to screen for selective sweeps in the Pietrain population before and after 10 generations of management. This statistic is designed to identify partial sweeps, and depends heavily on the occurrence of a long haplotype at high frequency in the population. The ongoing selective sweeps signal can thus be reduced if the management strategy decreases the frequency of the long haplotypes or selects for recombinant haplotypes. Some haplotypes were fixed in the population, resulting in regions without a signal, similar to the pericentral gaps observed on chromosome 8 (figure S7.2). As shown in figure 7.6, regions that before management were identified as having the highest p-value for a selective sweep have after management the larger difference in iHS signal before and after management compared to regions without selective sweeps. For example, the sweep on chromosome 13 was no longer present in the population after 10 generations of management (figure 7.6A). This indicates that the signature left by a selective sweep is counteracted when the management strategy aims at optimizing variation in the genome. Indeed, we observed a positive correlation of 0.68 between the p-value for each marker before management and the difference in iHS signal before and after management (figure 7.6B), meaning that incomplete sweeps signals were reduced because of the management. The tails of the

distribution of this difference in iHS signal before and after management were fatter when using the segment-based management strategy with a threshold segment of 5Mb compared to those that result when managing with the genealogy-based strategy (figure S7.5). This means that the segment-based strategy was more efficient in reducing the presence of long, similar haplotypes in the next generation. Therefore, in previously identified regions under selection, the reduction of the selection signature was stronger for the segment-based management strategies than for the genealogy-based method.

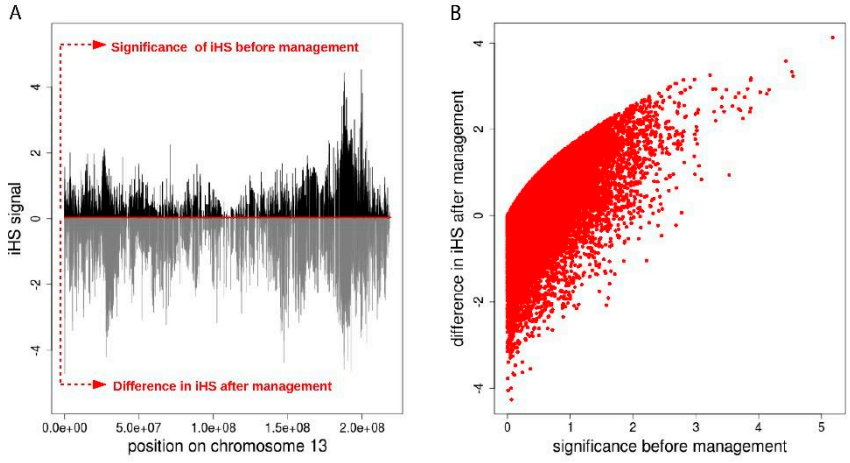


Figure 7.6 Change in selective sweeps in the Pietrain population. Signatures of selection are measured as extended haplotype homozygosity (iHS signal) in the Pietrain population before and after management. **A**. Example of the effect of management on the selective sweep on chromosome 13. The significance level of the iHS signal before management are indicated in black (-logp, y-axis) and range between 0 and 4, so that markers with a signal >2 are considered to be significant. Differences in iHS signal per marker before- and after management are indicated in grey and range from 0 to -5, with a strong negative number indicating a large difference. **B**. Genome-wide correlation between the significance level of the iHS signal before management (x-axis) and the strength in iHS signal per marker before- and after management (y-axis). Negative values on the y-axis indicate a stronger signal in the population after management for the associated marker.

7.3 Discussion

Theoretical and simulation studies have shown that implementing genetic marker information in the management of populations can lead to maintaining a higher degree of variation than would be established without this information (de Cara

2011). Here we have performed an empirical study to address the impact of whole-genome marker-assisted management strategies on 1) maintaining diversity, 2) whether the demographic history has an effect on which strategy to choose, 3) the impact on linked deleterious variants in the viability of the population, and 4) the maintenance of signatures of selection.

7.3.1 Management strategies: maintaining diversity

Our first objective was to determine whether the use of marker-by-marker based coancestry in management strategies outperforms genealogical coancestry and the recently postulated IBD-based coancestry (de Cara 2013b) in terms of maintaining variation in a population. By managing the two pig populations through *in silico* management, but using their actual genotypes for the initial population at the start of this *in silico* management, we have shown that the best management strategy in terms of maintaining diversity was indeed to use molecular coancestry (marker-by-marker similarity). Using genealogical coancestry maintained the least diversity, while using segment-based coancestry, measured as the average shared segments between individuals, yielded intermediate results, the difference with molecular coancestry being small in some cases. These two measures are similar in theory, because genealogical coancestry is highly correlated with segment-based coancestry for long segment sizes (Keller 2011). However, genealogical coancestry is the expected relationship between individuals, while segment-based coancestry will vary between individuals with the same genealogical relatedness depending on the actual gene variants that they carry and where meiosis has occurred.

The high performance of the molecular coancestry method stems from the fact that OC optimizes expected heterozygosity based on the differentiation between individuals, and this differentiation can reach the largest values with the highest number of markers, that is, managing a population with the marker-by-marker coancestry. These results are thus consistent with a previous simulation study (de Cara 2011). It must be noted that, in both cases, we had a high density of markers available, as it is known that a low density of markers using molecular coancestry will not maintain as much diversity as genealogical coancestry (Fernandez 2005, Gomez-Romano 2013). Thus, management of population demography and effective population size should integrate with pedigree management in efforts to ensure population sustainability. Together with optimal contributions as described here with the different management strategies, and controlled matings, the loss of genetic variation in small populations can be mitigated by advances in reproductive technologies, including gamete banking and cloning of appropriately identified individuals.

Nevertheless, when high density genotypes are available, optimizing genetic diversity through management based on molecular coancestry does not always yield the desired result, even though most variation is maintained. The goal in the management of an endangered population like *Sus cebifrons* is to maintain as much variability as possible without accumulating deleterious mutations in single individuals that could lead to a loss of viability and thus extinction. Thus, it is important to have a priori knowledge of the viability of the population or fitness-related traits, as populations which show some viability issues should be managed possibly based on genomic information that can provide details on deleterious mutations and segment-based coancestries. On the other hand, in commercial breeds like Pietrain, maintaining diversity is a goal, but breeders also have specific selection goals, which may be at odds with maintaining diversity and minimizing inbreeding. Therefore, knowledge of the demographic history, population fitness and selection signatures are important in the implementation of the management programme. We discuss each of these issues below.

7.3.2 Demography and genetic diversity

Our second objective was to understand the effects of the demographic history of the population on the effectiveness of each management strategy. Inbreeding is not necessarily a problem in a population. Negative effects on fitness related traits have not always been observed when inbreeding increases, and populations can respond differently to the same amount of inbreeding (Hedrick and Kalinowski 2000, Lacy 2012). Whether this will have an effect on the population in the long term depends on multiple factors, one important factor being the deleterious variants that are present in the population. Demographic history plays an important role in the number and distribution of fitness related variants in a population. Deleterious variants can be most problematic if the population suffers a sudden bottleneck and if the mutations are abundant and recessive, especially if their selection coefficient is high (Lynch 1995).

The two pig populations here studied both have high levels of inbreeding, but for very different reasons. As we have shown in our analyses of demographic history, *Sus cebifrons* has experienced two large reductions in its effective population size. In comparison to other *Sus* species, the effective population size dropped drastically during the first bottleneck about ten thousand years ago (Frantz 2013b). The second bottleneck we identified is recent, possibly due to reduction in its habitat and over-hunting, resulting in its current status as critically endangered

(Oliver 2008). These events probably caused that the base level of variation in *S. cebifrons* is relatively low compared to other *Sus* species (Bosse 2012), and that these low levels of variation are unlikely to increase in the (near) future. The variation in *Sus cebifrons* is already very low and without proper management it will quickly decay further. Keeping the diversity as such in the captive population is absolutely essential since it is a crucial species characteristic. Thus, we need to maintain as much diversity as we can while avoiding potential inbreeding depression, that is, we need to maintain the adaptive potential, especially because the possibility exists that the species will become extinct in nature within a few decades.

The Pietrain breed of *Sus scrofa* is widely used commercially, and has been strongly selected. Although this breed is widely distributed, the number of individuals that contribute to a specific selection line is limited, which is a general pattern for livestock populations. Brotherstone and Goddard (2005) estimated the current effective population size of some cattle breeds to be around 50, which is thought to be the minimum viable population size in the short term. They raised their concern regarding the diminishing effective population size in cattle and the associated consequences of inbreeding, and suggest active management to maintain genetic diversity in breeding programmes. As we have shown, the different histories of the two pig populations studied here, result in different patterns of genetic diversity, *Sus cebifrons* having a lower overall genetic diversity, while Pietrain shows larger runs of homozygosity but an overall larger diversity than *Sus cebifrons*. Many ROHs in the Pietrain population could be due to selection rather than inbreeding, as has been observed for other commercial pigs (Bosse 2012), while the ROHs in the *S. cebifrons* population should mainly be due to consanguineous matings as can be seen in their pedigree. However, both populations have a limited contemporary effective population size, because they are kept in enclosures, and therefore avoiding inbreeding to maintain as much as diversity as possible is desired since they belong to the most vulnerable category in terms of inbreeding depression. This will apply to most captive (zoo) populations and commercially kept farm animals.

The demographic history showing recent bottlenecks can lead to the accumulation of deleterious mutations by drift. Thus it is important to know both the current viability and the history. A population that has undergone a bottleneck and that is perfectly viable, is likely to either have undergone purging or just never had viability issues. On the other hand, a population with viability issues may have

accumulated recessive variants of small effect that are difficult to eliminate by natural selection. Which variants are maintained can be essential for the success of the breeding programme. Actively changing the genetic composition in a population through management may, as a side effect, reduce the frequency of advantageous variants or increase the number of deleterious alleles. The management thus may not have the desired result, and the impact on fitness and selection should therefore be considered.

7.3.3 Management strategies: Impact on fitness

Thirdly, by identifying putative deleterious variants in the population and plugging them into the simulations, we have investigated whether the strategy that maintains the most diversity leads to maintaining the highest frequency of deleterious variants, which in turn could lead to inbreeding depression. Previous simulation results had shown that when the population exhibited some inbreeding load, managing the population with segment-based coancestry was the best strategy in terms of maintaining a compromise between maintaining diversity without accumulating too many deleterious mutations (de Cara 2013b). We have explored the robustness of these results here with real data and verified that this seems to be the case, i.e. that the segment-based approach can provide a compromise between maintaining fitness and diversity.

In this work, we use predicted deleterious variants in the genome as a proxy for fitness. The distribution and number of deleterious sites, and also their selective effect and dominance, depend on the demographic history of a population. Fitness, in turn, depends on the shape of the distribution of mutational effects and the mean selection and dominance coefficients. Changes in the frequencies of homozygous and heterozygous deleterious alleles can have big consequences on the overall fitness of the population, as we have shown. Thus, the management strategy, by acting on linked deleterious variation, can have a massive impact on fitness and fitness-related traits like viability. Long runs of homozygosity are thought to be enriched in deleterious mutations (Szpiech 2013). Our results show that avoiding these ROHs in the offspring, which is the essential concept in the segment-based management, seems an effective method to maintain a reasonable fitness in the managed population.

The putative deleterious variants in both populations occur significantly less often in particular gene families like nucleotide binding proteins, suggesting that these gene families either are less prone to mutation or, more likely, are more effective

in purging the deleterious variants from the population. On the other hand, this may indicate that if a mutation occurs in any of these genes, the effect is more severe. Multiple methods have already been developed in order to predict the severity of particular mutations (e.g. Ng and Henikoff 2001, Wang 2010, Choi 2012). Therefore, the use of sequence data will have an increasingly important value in conservation and selection programmes because these variants and their effect can be identified and implemented into the breeding scheme, vastly improving our predictions and *in silico* analyses.

In our management strategies, little difference was observed in the evolution of fitness when applying molecular or segment-based OC to the *S. cebifrons* population. This is probably due to the very limited sample size, which restricts the number of possible choices that can be made under both scenarios. However, segment-based coancestry results in somewhat higher fitness because the proportion of deleterious homozygous sites increases at a slower pace than with marker-by-marker coancestry. This can have an effect on fitness, particularly in the case of low dominance (h). We do show that most fitness is lost when using genealogical coancestry methods. This is in contradiction with our previous theoretical simulations results, which have shown that this strategy maintains the highest level of fitness in other scenarios with considerably larger population size (de Cara 2013b). This difference between the predictions and our results in the performance of the genealogical strategy were most likely due to the very small population size here managed, and to a different balance between the number of homozygous and heterozygous deleterious mutations, that will have a different contribution to overall fitness dependent on their effect and dominance. The small population size here managed and the sex ratio make natural selection less efficient at eliminating these deleterious variants, so it is likely that when managing such small population sizes with such load, the population enters an extinction vortex. These results emphasize the use of resequence data not only to estimate coancestries as we have done here, but also to better understand the distribution of detrimental or beneficial variation, which will be crucial to decide which management strategy to use.

Most breeding programmes in zoos are still based on pedigrees (Fienieg and Galbusera 2013). Sequencing costs are decreasing rapidly and therefore NGS provides an increasingly affordable means to measure genetic variation and potentially fitness at the individual and population levels (Ouberg 2010). This is especially the case if a (closely related) reference genome already exists, where it

might be worthwhile to sequence all available founders in management strategies of small zoo populations, in order to get insight into the genetic variation and the inbreeding load. It is not unrealistic to assume that it will become possible in the near future to predict the effect of particular (deleterious) mutations more accurately than we can now. This could then be used to get detailed predictions of the consequences for the population when these deleterious mutations occur in homozygous state. When haplotypes containing these mutations are identified, genotyping the offspring on a high-density SNP chip will suffice to follow these mutations in the population, and the management strategy can be adjusted accordingly. Naturally, other aspects of small population management like population structure but also health examinations should be considered simultaneously.

7.3.4 Impact of management on selection goals

The genome is always under selection, in both wild and domestic populations, and active, marker-assisted management, may interfere. Domestic animals provide a case where selection is explicit and strong, and therefore we used the commercial Pietrain breed to study the effect of management on selection signals. It has been recognized that genomic tools are increasingly important in livestock conservation and selection (Lenstra 2012). For breeding companies, many individuals are genotyped already and information from a high-density SNPchip is sufficient to outperform the pedigree-based optimal contributions (Solberg 2008, de Cara 2011, Gomez-Romano 2013). The Pietrain population is under strong management regarding performance of the pigs, and therefore clear selection signals are present in the population, as has been shown by e.g. Stratz (2014). Management based on molecular markers aims at reducing long homozygous segments in the genome of an individual, but this will have as side-effect that any selection signal on particular haplotypes will be reduced if the selected haplotype is not yet fixed. Indeed we show here that selection signatures are weakened when the focus is solely on optimizing diversity. The measures of coancestry used in this study have consequences not only in management of populations to calculate which individuals should contribute the most offspring to the next generation to maintain the most variation, but also in genomic prediction (GP). In GP, the proportion of variance explained by the markers, and how well the distribution of mutational effects is known, has been shown crucial for GP to work well (Goddard 2009). The need for implementation of genomic-based inbreeding control in genomic selection has been recognized (Sonesson 2012, Toro 2014). Genomic selection can lead to acceleration erosion of variation at specific loci in the genome (Heidaritabar

2014b), and there might be substantial effect on the efficiency of genomic selection when taking the maintenance of diversity into account (Clark 2013). Recent simulation results (MacLeod 2014) indicate that resequence data could recover the missing genetic variance, and a better accuracy in the prediction is achieved with resequence data for populations with large Ne. We believe that incorporating sequence-based measures of diversity for OC together with the prediction of deleterious variants as proxy for fitness in genomic selection programmes provides a way forward to combine the goals of maximising gain without leading the population to extinction.

7.3.5 Concluding remarks

Is has been shown previously that each selection strategy comes with different consequences for a population in terms of maintained diversity and fitness. Different initial scenarios will lead to different results after management because this is dependent on the inbreeding load, population demographic history, contemporary effective population size and other factors shaping population genetic composition. This work incorporates the important additional information that can be gained from NGS data into management strategies. Our simulation study on two realistic cases shows that variation in managed populations can be selectively maintained by whole-genome, high-density, marker-assisted methods. Specifically, methods that apply molecular coancestry seem to be the most efficient for this purpose. However, the effectiveness of marker-assisted methods depends on the demography and effective population size of the target population.

Our analysis shows that if the population has an inbreeding load, as is probably the case for *Sus cebifrons* due to the bottlenecks the population has suffered, then managing the population using molecular coancestry maintains the most diversity. Theoretical predictions had shown that such management could lead to the accumulation of deleterious variants and massive loss of fitness (de Cara 2013a). The results here obtained are more positive towards the use of molecular or marker-based coancestries in the management of populations, as we show that the loss in fitness is marginal compared to other methods when the population is managed to maximise genetic diversity. Therefore it is clear that to decide which measure of coancestry to apply in a management programme, one ideally has dense genotypes for each individual as well as information on the distribution of genomic variants and their mutational effects. Should the distribution of mutational effects and deleterious variants be available, as might be feasible in the future, a targeted approach to remove these variants from the population can be

implemented in the management programme. The initial presence and re-distribution of detrimental variants, particularly the accumulation of homozygous deleterious genotypes, seems also to be most favorably addressed by the application of marker-assisted coancestry methods, especially segment-based coancestry. Therefore, if details on the distribution of deleterious variants are not known, segment-based management strategies may in fact be most efficient in avoiding fitness reduction due to homozygous deleterious variants. This information should be combined with the estimation of the demographic and selective history of the population. When all this information is not available, then our initial preferred choice is using molecular coancestry, as it maintains the most diversity, while paying special attention to possible losses in viability or fitness-related traits.

Lastly, genomic marker-based methods may offset (past) selection events by effectively removing signatures of selection, and thereby possibly diminishing the phenotypic value of the population. This is certainly of concern for domesticated populations, but may in fact also be of concern for wild populations, if, for instance, local adaptation or local phenotypic variation is affected. We want to emphasize on the importance of these advances in technology for a wide range of disciplines. We strongly encourage the development of management programmes that include both fitness and diversity in their calculations of OC. Also, genomic selection strategies may benefit from integrating fitness and diversity in the choice of the contributing individuals, especially in the long term.

7.4 Material and Methods

7.4.1 Sample background

Two pig populations were used for the in silico management: re-sequencing data from five S. cebifrons individuals from San Diego zoo (5) and genotypes from 46 and sequence data from 11 individuals of the Pietrain breed of S. scrofa. Pedigree data were available for both populations, covering several generations for the Pietrain breed, while for S. cebifrons the pedigree available only covers the individuals since the foundation of the conservation programme. At the starting point for the management, the S. cebifrons population consisted of 5 individuals (1 male and four females) that we expanded to 10 individuals (2 males and eight females) before the first generation of management, with genotypes for 104.035

sites, and the Pietrain population contained 47 individuals genotyped for 51165 sites.

7.4.1.1 Sus cebifrons

Sus cebifrons is a critically endangered pig species that is endemic to the Philippine islands and currently occurs still on two islands. Its estimated population size in the wild is low: 200-500 (Negros island) and 500-1000 (Panay island). It is unknown how genetically distinct the two island populations are. We used pedigree data from two studbook-keeping zoos: Rotterdam zoo, which was founded with 6 individuals from Negros, and San Diego zoo, which was founded with 7 individuals from Panay. We used data from7 re-sequenced individuals, all sequenced to ~10x depth of coverage with the Illumina paired-end sequencing technology (Illumina Inc.): 2 from Rotterdam zoo, and 5 from San Diego zoo. The information on the 5 individuals from the San Diego Zoo was used for the in silico population management.

7.4.1.2 Sus scrofa

Pietrain is a commercial European breed that is extensively used for pork meat production worldwide. They are kept in sties with an average effective population size of around 50 individuals. The breeding company contains a dam line that was established in the early seventies, and consists of ~100 live animals. We had access to the pedigrees for the *Sus scrofa* Pietrain line, where we had data from 11 re-sequenced individuals, all sequenced to ~10x depth of coverage with the Illumina paired-end sequencing technology (Illumina Inc.), and 47 individuals genotyped on the Illumina porcine 60K iSelect Beadchip (the 11 re-sequenced individuals are a subset of this 47).

7.4.2 Methodology

7.4.2.1 Pedigree reconstruction

The pedigree for the *S. cebifrons* San Diego zoo population was obtained from the studbook keepers, and we used the drawing software Pedigraph (Garble and Da 2008) to infer and draw the pedigree. Although the pedigree for the Pietrain line was also available from the breeding company since its establishment in 1970, we did not create a graphical overview because of its complexity. Relatedness between individuals was extracted from the studbooks.

7.4.2.2 Sample collection and DNA extraction

Blood samples were collected from a total of 7 *Sus cebifrons* from the Rotterdam and San Diego zoo, and 11 *Sus scrofa* belonging to the Pietrain breed. The QIAamp DNA blood spin kit (Qiagen Sciences) was used to extract DNA from the blood samples the Qubit 2.0 fluorometer (Invitrogen) was used to check the isolated DNA for quality and quantity. Library construction for the re-sequencing was performed with 1-3 ug of genomic DNA according to the Illumina library prepping protocols (Illumina Inc.) and the Illumina 100 paired-end sequencing kit was used for sequencing.

7.4.2.3 Genotyping on 60K

DNA for a total of 156 animals was diluted to 100 ng/ul and samples were genotyped on the Illumina porcine 60K iSelect Beadchip (Ramos 2009) according to the IlluminaHD iSelect protocol. Data was analyzed using Genome Studio software (Illumina Inc.).

7.4.2.4 Alignment and SNP calling

All Illumina 100bp paired-end read libraries were quality trimmed with sickle -l 50. Trimmed reads were aligned to the *S. scrofa* reference genome build 10.2 with bwa 0.7.5a with -t 4 and bamfiles from multiple libraries were merged and deduplicated with samtools 0.1.19. Local realignment was executed with GATK v2.6 and finally the bamfiles were filtered with samtools using the samtools view options -F 12 and -q 30. Variants were called with samtools mpileup for each individual separately with a minimum read-depth of 5, a maximum of twice the average coverage and genotype quality PHRED score of at least 20.

7.4.2.5 Genomic variation

The variation between two haplotypes within individual genomes was assessed in bins of 1Mb according to the method described in Bosse (2012). Briefly, the number of filtered variants per bin was corrected for the proportion of sites with accurate coverage for each bin, and the number of heterozygous sites per bp was calculated. Bins with less than 20% of accurate coverage were removed from the analysis.

7.4.2.6 Matrix construction

For the *S. cebifrons* data, all individuals were genotyped for all sites where at least one individual contained a non-reference allele with the same settings as for the SNP calling, and only the sites where all individuals contained a reliable genotype

call with at least a PHRED score of 20 were retained. All singletons and sites with a clear Hardy-Weinberg deviation were removed with PLINK v1.07. Finally, we extracted the list of putative copy number variable regions from Paudel (2013) and removed variants within these regions, leaving us with a genotype matrix containing approximately ~100.000 high quality sites.

The genotypes for all 47 Pietrain individuals that were generated on the Illumina porcine 60K iSelect Beadchip (Ramos 2009) and were obtained from Topigs Research Center IPG BV, The Netherlands. Genotypes were filtered for MAF>0.01 and genotype calls of >0.9 with PLINK v1.07 to a total of 51165 sites.

7.4.2.7 Phylogeny

A pairwise distance matrix (1-IBS) was constructed using the –cluster option in PLINK v1.07 (Purcell 2007) for both matrices and a neighbor-joining tree with random input order was created in PHYLIP (Felsenstein 2005). The tree was represented using FIGTREE http://tree.bio.ed.ac.uk/software/figtree/.

7.4.2.8 Phasing

Genotype data for each chromosome was extracted from the matrices using PLINK v.1.07, and chromosomes were phased independently by Shapeit v2.r727 (Delaneau 2012, 2013) with 7 burn-in iterations, 8 iterations of the pruning stage, 20 main iterations, 100 states and a window size of 5Mb. For *Sus cebifrons*, phasing was performed on the filtered matrix. Although only the five individuals from the San Diego zoo were used in the *in silico* management, we used all 14 available haplotypes for the phasing step to increase accuracy. For the Pietrain breed, we used the 60K genotype data for the haplotype reconstruction.

7.4.2.9 Deleterious variants

The number of homozygous deleterious variants in the population before and after management was used as a proxy for fitness. We used the Ensembl Variant Effect Predictor tool v.74 on the filtered VCF files for each individual to assess the nature of the variants. For the *S. cebifrons* data, sites with a deleterious SIFT prediction score were extracted and included in the matrix; note that this step introduced a small amount of singletons to the matrix.

For the Pietrain population, the list of deleterious sites was based on the deleterious sites that were actually called within 11 individuals; on average each individual contained 656 deleterious sites. Then the average number of shared deleterious sites between 2, 3, 4, … individuals was obtained, to be able to infer

how many unique deleterious sites each individual would contribute to the total. The actual number of observed deleterious sites among the 11 individuals (=3468) was extrapolated by fitting a power curve to the number of unique contributions per extra individual, so that we expected with 47 individuals a total of ~10.000 deleterious sites. Then, 6532 more sites were randomly extracted from the genome and we assigned a deleterious effect to these sites. Then, for each individual 656 sites were randomly picked from this list, so that to some degree it fitted the observed distribution in the 11 re-sequenced individuals. These 10.000 markers and genotypes for 47 Pietrain pigs were added to the matrix containing the 60K markers for the management.

7.4.2.10 Ne estimation

We estimated past and recent effective population size in the *S. cebifrons* data using two methods independently. First, we used the method described by MacLeod (2014) to filter the heterozygous sites in our VCFfiles for false positives. Briefly, this method does not remove actual false positives but based on the false positive rate it randomly removes a number of heterozygous sites in a particular window along the genome. False positive rate was estimated based on 60K genotype data and the rate of heterozygotes on the non-pseudo-autosomal regions on the X chromosome for males. Then we recoded the distribution of heterozygous sites to psmc-fasta files and conducted the demographic analysis with a hidden Markov model approach as implemented in PSMC, using Tmax=20 and n=64 (4+50*1+4+6). The generation time was set at 5, mutation rate $1.0*10^{-8}$ which is identical to that used for human and cattle re-sequence data (MacLeod 2014), and false negative rate was estimated from the distribution of depth of coverage per site, ranging from 15% to 45%. Confidence intervals for each sample were estimated based on bootstrapping, as suggested in Li and Durbin (2011).

To be able to compare the inferred demography from the PSMC method, we used the iterative approach described by MacLeod (2014) to infer past effective population size for the male *S. cebifrons* from Panay. False positive rate was estimated based on the non-pseudo-autosomal regions on the X-chromosome and false negative rate was based on the proportion of sites in the genome where the average coverage was between 0.5 and 2 times the average coverage. Based on these inferences, scaled mutation rate was 4.0×10-9. The number of phases was based on the segment sizes that were used to infer demography in the original paper, so that segment size n was 1–1,000 bp, and then 1, 2,…, 1,000 kb etc. The

threshold delta for goodness of fit was set to 0.001. Upper and lower limits were estimated by changing local Ne until the delta threshold criteria were violated.

7.4.2.11 Regions of Homozygosity

The presence of regions of homozygosity (ROH) on the individuals' level was tested for the 47 Pietrain individuals that were genotyped on the Illumina porcine 60K iSelect Beadchip. SNPs were filtered for MAF 0.05 and a maximum proportion of missing genotypes of 0.1. ROHs were extracted with the –homozyg option in PLINK v.1.07, allowing for one heterozygote within a ROH.

7.4.2.12 Signatures of selection

To assess whether signatures of selection were present in the Pietrain population before management, we performed an extended haplotype homozygosity (EHH) test as implemented in the R package rehh (Gautier and Vitalis 2012). All 46 Pietrain pigs that were genotyped on the Illumina porcine 60K iSelect Beadchip were included in the EHH analysis. After filtering for missing data and MAF a total number of 51165 markers were included. Ancestral states of alleles were determined based on the genotypes from four outgroup Suids from Island South-East Asia (ISEA) obtained from Groenen 2012 and Frantz 2013b. Significance of iHS was calculated by the ihh2ihs option, where high significance levels indicate that the persistence of haplotype phase for either the ancestral or the derived allele is longer than expected.

7.4.3 In silico management

We simulated the management of both populations using optimal contributions (OC) during ten generations. OC minimises global coancestry by minimising the expression $\sum c_i c_j f_{ij} T^2$, where f_{ij} is the coancestry between individuals *i* and *j*, c_i is the number of offspring that individual *i* leaves to the next generation and *T* is the sum of contributions $\sum c_i$ which is set at 2 *N* to maintain population size constant. Every generation, we calculate the *c*'s that minimise that expression, by using a simulated annealing algorithm. We used three measures of coancestry:

1 - Molecular coancestry, also called throughout the text marker-by-marker coancestry, which is the probability that two alleles at a locus drawn at random are identical by state, averaged over all markers.

2 - Genealogical coancestry: as derived from the pedigree.

3 - Segment-based coancestry: as in eq. (2) de Cara (2013b), which is a measure of shared segments of identity by descent across individuals:

$$f_{R_{il}} = \frac{\sum \sum \sum l_{SEG_k}(a_i, b_l)}{4\mathrm{L}}$$

Where $l_{SEG_k}(a_i, b_l)$ here is the length of the k-th shared IBD segment measured over homologue *a* of individual *i* and homologue *b* of individual *j*, and *L* is the length of the genome. We used three thresholds for what we considered a shared segment: 100kb, 1Mb and 5Mb.

For the *in silico* management of the *S. cebifrons* population, we used the five individuals of the San Diego zoo, which consisted of one male and five females. Prior to management we performed random mating to expand the population to 10 individuals as otherwise the management is heavily constrained by having one male only, keeping the sex ratio of 1 male for every 4 females. We have also performed *in silico* management with even sex ratios, but the consequences remain the same for such small population size (data not shown). For the management of the Pietrain breed, we assumed half of the individuals were female and half male, as it is approximately the case in the real population. As the population consisted of 46 individuals, we did not expand its size.

We performed management in which all individuals were considered perfectly viable and we also analysed the cases in which deleterious mutations had been included. For these deleterious mutations we assumed that their effects followed a Gamma distribution with shape parameter 1 and mean effect 0.005. Dominance coefficients were drawn from a uniform distribution with mean exp(-k s), where s is the effect at each locus, and k is a constant that gives the desired mean dominance coefficient. We show results here for mean dominance (*h*) of 0.35 and 0.5. These values are taken from experiments on Drosophila (Mukai 1972), but with mean selection coefficient one order of magnitude smaller as otherwise the population would not be viable. This distribution is in line with deleterious mutations being common and of small effects. The fitness of each individual is multiplicative across loci. Every generation, we calculated the *c*'s that minimized global coancestry and then performed matings at random between contributing individuals. Their offspring was kept if its fitness was smaller than a random number drawn from (0,1) and discarded otherwise. The sex ratio was kept constant throughout the simulations. We performed 1000 replicates for each management strategy, and

every generation we recorded the genetic diversity (as observed heterozygosity and gene diversity), the mean fitness of the population and the mean numbers of heterozygous and homozygous deleterious variants. We stored the resulting genotypes of one of the Pietrain replicates to analyse the effect on signatures of selection. The replicate was evaluated based on its summary statistics and fell right within the confidence intervals of the 1000 replicates.

7.5 Acknowledgements

This work was financially supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC Grant agreement number 249894. M.A.R. de Cara and F. Austerlitz were funded by LabEx grant ANR-10-LABX-0003-BCDiv from Agence Nationale de la Recherche Investissements programme ANR-11-IDEX-0004-02, and additional funding for M. Bosse to realize this collaboration was provided by the ESF exchange grant Advances in Farm Animal Genomic Resources (GENOMIC-RESOURCES) no. 4579 to visit M.A.R. de Cara and F. Austerlitz group. We thank Simon Boitard, Flora Jay and Iona MacLeod for valuable discussion on the methodology. DNA samples were provided by Topigs Research Center IPG BV, The Netherlands and Rotterdam Zoo, The Netherlands.

8

General discussion

8.1 Introduction

The genome consists of a mosaic of haplotypes, representing a variety of demographic histories. Disentangling the different haplotypes in the DNA can provide a valuable source of information about the evolutionary history of a population. With this thesis I aim to provide better insight into the mechanisms that shape the variation landscape in diploid organisms, using the pig *Sus scrofa* as a model species. I focus on which genomic features influence the pattern of variation within individual genomes, and how the genomic distribution of haplotypes is influenced by factors like demography and selection at the population level. I emphasize the influence that humans have had on the genomic patterns in pigs through habitat reduction and fragmentation, deliberate and unintentional hybridization and selection. In the final part, I discuss how the information on haplotype distribution can be used in selection and conservation efforts. In this chapter, I explore my findings from chapter 2-7 by discussing the main conclusions, and by elaborating on how this thesis contributes to existing literature. Finally, I identify next steps to further extend and implement the obtained knowledge.

The work described in this thesis contributes to our understanding with regard to factors influencing genomic variation in pigs and wild boars. Genomic variation refers to the average number of variable sites per base pair between two haplotypes. Therefore, genomic variation and the patterns in which different haplotypes occur in a population are interconnected. The factors that shape the genomic variation in individuals can be grouped into two classes: 1) intra-genomic processes that contribute to the local distribution of haplotypes and generate or eliminate variation at the molecular level; 2) external factors that influence the persistence of haplotypes in a population, and that act upon the population as a whole. There will usually be interplay between the two classes. In the following paragraphs I mainly discuss the second class, in particular the effect of demography, selection and hybridization on the haplotype patterns in a genome. Furthermore, I highlight the benefits and drawbacks of methods used to identify these processes. The role of humans in shaping patterns of diversity in the genome is discussed, specifically with regard to future perspectives on the use of whole-genome sequence information for conservation of genetic diversity.

8.2 Demography and selection inferred from genomes

8.2.1 Levels of genetic variation

A central theme in population genetics is which factors shape the genetic diversity levels in species and populations (Leffler 2012). What the relative influence is of intrinsic features of a species, like genome characteristics, species biology and ecology, as opposed to external forces such as population history and selection on genetic diversity levels, is still under debate. Simple models predict that under a neutral scenario, the rate of genetic drift is inversely proportional to the population size. Therefore, population genetic variation can be seen as a balance between novel mutations and loss of variation due to finite population size (for a comprehensive review, see Charlesworth 2009). Genome size positively correlates with per-base and per-generation mutation rate (Lynch 2010). Recently, Romiguier (2014) argued that ecological strategy is the main predictor of genetic variation in a population or species, under the assumption that the population itself is at mutation/drift equilibrium. According to this predictor, short-lived and highly fecund species generally display higher genetic diversity. The fact that no consensus exists in literature about the relative contribution of these predictors can be attributed to the abundance of external confounding factors. This leads to the question of how these confounding factors, mostly ecological disturbance and selection, influence genetic diversity (Banks 2013).

Both the species characteristics and external factors influence the effective population size, which is in turn reflected in the diversity levels in the genome. How much each factor influences the actual levels of variation is difficult to estimate. Genomics facilitates studies on every aspect of genetic variation, including species specific features, since characterization of variation in the genome can effectively be accomplished with minor bias. My thesis contributes to the understanding how variation in the genome is influenced by internal and external factors, focusing on the distribution of haplotypes as a basic principle. In chapter 2 I relate within-genome nucleotide diversity to past and recent effective population size and bottlenecks. I find a difference in average nucleotide diversity of an order of magnitude between different pig species of Island Southeast Asia (ISEA). This order of magnitude difference is apparent even if recent inbreeding, estimated from the occurrence of ROHs, is excluded. If parental investment (propagule size, longevity, fecundity) is indeed a good predictor, how do these observed diversity patterns then relate to their key traits related to parental investments: are the different species on ISEA variable in terms of these key traits? Or are the high differences in

genetic variation between *Sus* species better explained by environmental factors, influencing their N_e? Sexual dimorphism is somewhat stronger in *S. verrucosus,* the species with the lowest nucleotide diversity, compared to most other pigs. Litter size has been reported between 3-9 young for *S. verrucosus*, which is not very different from other pig species. Other signs of parental investment show different patterns between pig species, providing little evidence that this is indeed a good predictor of the observed patterns of genetic variation. The observed diversity levels in chapter 2 correspond well to the estimated demographic trends in Frantz (2013b). The levels of nucleotide diversity are congruent with past N_e in other species al well, such as in great apes (Prado-Martinez 2013). The buildup of genetic variation is a slow process. However, variation levels can drop very rapidly when the effective population size is reduced. Colonization and dispersal, habitat loss and fragmentation, but also domestication, can leave strong signatures in the genome. A population bottleneck will initially result in a highly heterogeneous distribution of variation in individual genomes. The basic concept here is that the probability of homozygosity resulting from two haplotypes that are Identical By Descent (IBD) increases. Specifically, some parts of the genome will represent basically no variation, because the same haplotype is present on both copies. Other parts, however, where the homologous chromosomes contain two different haplotypes, will display similar variation levels just as before the bottleneck, suggesting a much larger effective population size (figure 8.1). External factors can change rapidly, and I have shown in chapter 2 that rapid changes in population size have a tremendous effect on genetic diversity reduction. I therefore argue that genome size and life history strategies are probably good predictors of genetic diversity levels under a stable and simplified scenario, but in reality, external factors are more realistic predictors of population genetic variation. With current techniques detailed screening of the variation landscape has now become possible. Relating the genomic variation landscape to both intrinsic and external determinants of N_e will undoubtedly aid in our understanding of the evolutionary forces shaping variation in the genome.

Similarity in measures of genetic variation within genomes between populations is not a good predictor for the relatedness between those populations. In chapter 2, I demonstrate that the Japanese wild boar is similar to the European wild boars, based on their distribution of the level of variation in terms of ROH occurrence and variation outside ROHs (figure 2.5). However, phylogenetically, the Japanese wild boar falls, as expected, within the group of Asian wild boar. Likewise, I find similarity in the variation landscapes in the African warthog and the bearded pig (*S.*

barbatus). Because the shape of levels of variation throughout the genome is similar, it is likely that the demographies of these two species are similar even though the actual variants found in each of the species are different. Levels of variation and sequence identity, therefore, indicate two different things: demography after divergence and time since divergence. Even though genetic diversity levels are similar, they probably do not predict well how the species will behave in another environment because that depends on which variants are actually present.

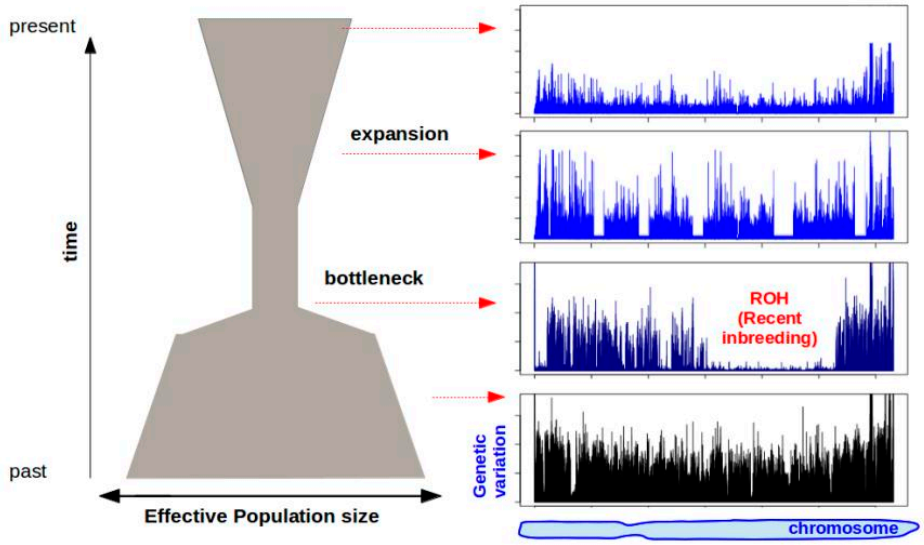


Figure 8.1 Distribution of genetic variation across a chromosome over time. When effective population time changes over time, the distribution of nucleotide diversity changes accordingly. A large panmictic population results in high base levels of variation. Consanguineous matings during a bottleneck forms ROHs in the genome, that are gradually broken down and genetic variation is equaled out through time.

8.2.2 Multiple evolutionary histories in a single genome

The mosaic structure of the genome is created through the interplay of genetic recombination, drift and selection (Paabo 2003). A haplotype can be defined as a continuous stretch of DNA containing inherited alleles that have shared evolutionary background, and the end of the haplotype is where the sequence reaches another demographic history. Discrete stretches of DNA (haplotypes) on a chromosome can differ in their relatedness to another copy of the homologous chromosome. It is important to realize that the scale at which we compare

relatedness between individuals is basically comparing two collections of haplotypes. Even within a single diploid genome, the two copies of the same chromosome pair inherited from two parents can be seen as collections (N=2) of haplotypes that, in case of random mating, are two randomly drawn samples from the population. Interestingly, in siblings, two copies of the same chromosome are probably more similar between individuals than within individuals, since siblings carry haplotypes that are Identical By Descent (IBD). In chapter 3 I separate European wild and Asian haplotypes within the genome of commercial European pigs, based on their IBD status. This assignment of IBD haplotypes requires model populations that represent the different origins of the haplotypes, to allow inference of geographic origin of haplotypes in the putative hybrid animals. But shared ancestry in this sense does not require that all of the haplotypes that share the same geographic origin (Asia or Europe) have similar coalescence times. Even those continuous stretches of DNA that represent European ancestry only, are still a collection of different ancestries that have been combined through recombination into one haplotype. For example, a region under strong selection for one particular variant that lies on a European haplotype will coalesce much faster than a neutral region with high mutation and recombination rate. These genomic regions can be in close proximity on the same chromosome, but still represent very different coalescent times. This chapter demonstrates that studies about shared ancestry in the genome should carefully consider the scale at which the shared ancestry is defined, since genomes are a collection of different ancestries.

8.2.3 Selection and demography

The effective population size in a genome segment that is subject of selection is altered compared to genome segments that do not experience selection (Charlesworth 2009). The difference between drift and selection is, that selection acts locally, while drift acts globally on the genome. Estimates on past and recent demography, therefore, can be best inferred from (near-) neutral positions in the genome. Conversely, inferring selection on haplotypes can be done because selection changes the frequency of haplotypes in a population. Various methods exist that, explicitly or implicitly, make use of locally reduced coalescence times to screen for positive selection in a population (e.g. Weir 1984, Gautier 2012, Rubin 2010). Apart from reducing local genetic variation and associated coalescence time, selection can also increase diversity levels. Climate change is thought to select for heterozygosity in fur seals (Forcada 2014). Genetic diversity is promoted by increased fertility in this case. Another well-known example of locally increased genetic variation in the genome is the major histocompatibility complex (MHC) in

vertebrates. The role of the MHC in the immune response promotes heterozygote advantage (Aguilar 2004). Heterozygote advantage results in balancing selection, and associated higher levels of genetic variation. This heterozygosity advantage probably results in longer coalescence times than expected under a neutral scenario. Therefore, inferences of effective population size based on loci under balancing selection such as the MHC will result in much higher estimates than from the other parts of the genome.

The number of studies on ROHs is increasing, and indeed I identify ROHs as a powerful measure for estimating inbreeding and selection. The occurrence of homozygosity, in the form of ROHs in individual genomes, can have multiple causes. In chapter 2 I conclude that ROH distribution in pig populations is mainly determined by local recombination frequency and demography. This does not imply that the effect of selection is negligible. Selection for particular haplotypes, in particular if pertaining a selective sweep, can result in ROHs as well (see glossary in chapter 1). In commercial pig populations the occurrence of ROHs in the same part of the genome in multiple individuals, indicative of haplotypes under selection, was more apparent than in wild boar, but still minor. In cattle, however, we see a contrasting pattern: many ROHs co-occur in similar genomic regions in different individuals (Ferencakovic 2013a). Populations under strong selection, therefore, may display ROHs, but this does not imply small N_e outside ROHs. Conversely, overlapping ROHs in a population with high inbreeding do not necessarily suggest selection: this can result from inbreeding and therefore the haplotypes that occur in homozygous state may not be the same. Genomic homozygosity seems to be population-specific because it results from the interplay between of demography, selection and genome features. Future research will likely reveal that this applies to other diploid species as well.

8.2.4 Demography and domestication

Domestication results in a deliberate separation of the domesticated population and its parent population. Domestication is, therefore, initially indistinguishable from any other event that results in cessation of gene flow between populations. This simple definition of domestication defines a domestic population as a subset of the wild population with cessation of gene flow. Therefore one can expect that domestication results in a reduction of the genetic variation in the domesticated population. The use of higherDNA marker densities has enabled researchers to reveal the complexity of livestock domestication, which was shown to be far more complex than a single sampling from the wild (Bruford and Bradley 2003). In pigs,

domestication does not seem to have left a clear population bottleneck (Frantz 2015). This suggests that the majority of the genetic variation that is present in European wild boar is also present in domestic breeds, even though modern pigs are phenotypically clearly different from the wild boars. This is probably because domestication, and subsequent interbreeding with wild boar, happened at different times and locations. Within this thesis I investigated what the impact of the domestication process and breed formation was on the levels and patterns of genetic variation within pigs. I demonstrate in chapter 4 that recent inbreeding (ROHs) and low genetic variation exists in some breeds, but nucleotide diversity between breeds in European breeds is comparable to, and even exceeds, variation in European wild boar.

8.3 The effects of hybridization

8.3.1 The hybrid nature of genomes

In chapter 3 I show that the genomes of modern commercial pigs are a mosaic of different haplotypes, each with a distinct demography linked to its sequence. Goedbloed (2013) and Frantz (2012b, 2013a) have shown that not only commercial pigs, but also wild boar populations in Europe can be a mixture of different domesticated and wild populations. Frantz (2013b, 2015) showed that this hybridization pattern even exceeds species boundaries, resulting in hybrids with admixed genomes containing haplotypes that diverged roughly 4 Mya. Speciation with gene-flow is not restricted to pigs; compelling evidence exist for example for *Drosophila* (Garrigan 2012), mice (Teeter 2010), and even human (Patterson 2006). Incongruence in hybridization frequency estimates between different parts of the genome is frequently observed. Specifically, phylogenetically incongruent results are often seen between mtDNA, the Y chromosome and autosomal DNA, suggesting some sex-bias in hybridization. In pigs, estimates of introgression based on mtDNA (i.e. Fang and Andersson 2008) are indeed higher, compared to my estimates in chapter 3 and 4, and estimates based on autosomes from other studies (Groenen 2012). That incongruence suggests that mitochondrial introgression of mtDNA is facilitated due to lower effective population size and associated larger drift effects (Currat 2008). Haldane (1922) proposed a biased sex ratio with an excess of the homozygous sex after hybridization, which has been observed in bison-cattle for example (Hedrick 2009). These findings highlight the importance of having multiple independent markers to infer the magnitude of hybridization and introgression, and to elucidate the mechanisms of introgression.

In chapter 4 I use the information of within-population genetic diversity compared to between-population variation in order to estimate the presence and number of introgressed haplotypes. Even without a direct source of introgression, it might become possible to detect putative introgressed haplotypes on an outlier-basis. Ideally, one would divide a putative hybrid genome into segments, as displayed in figure 8.2. Under a simple scenario with a single admixture event between divergent species, four distributions of genetic variation can be recognized: 1) Recent inbreeding or selection results in homozygous stretches, so no genetic variation; 2) Different haplotypes from the same population will result in a particular distribution of nucleotide diversity such as the blue distribution in figure 8.2; 3) Variation between two introgressed, but not identical, haplotypes result in

another distribution of either higher or lower average diversity levels such as the orange distribution; 4) The level of variation between hybrid parts of the genome that contain haplotypes representing both genetic backgrounds. The elevated nucleotide diversity in regions that represent mixed origins can aid to determine the origin of the haplotypes. Other classes can be added, for instance when hybridization occurred with different species. Wu (2014) demonstrates that such different distributions of heterozygosity are present within the genome of citrus species. There is an optimal divergence between populations that will lead to the best estimates: as soon as distributions start to overlap, inference of ancestry becomes much more difficult. In Chapter 3 I disentangle the Asian and European origin of European pigs. I emphasize the possible bias that can occur when analyzing the demographic history of a population based on hybrid genomes. I show that population size estimates are inflated when Asian haplotypes are pre-

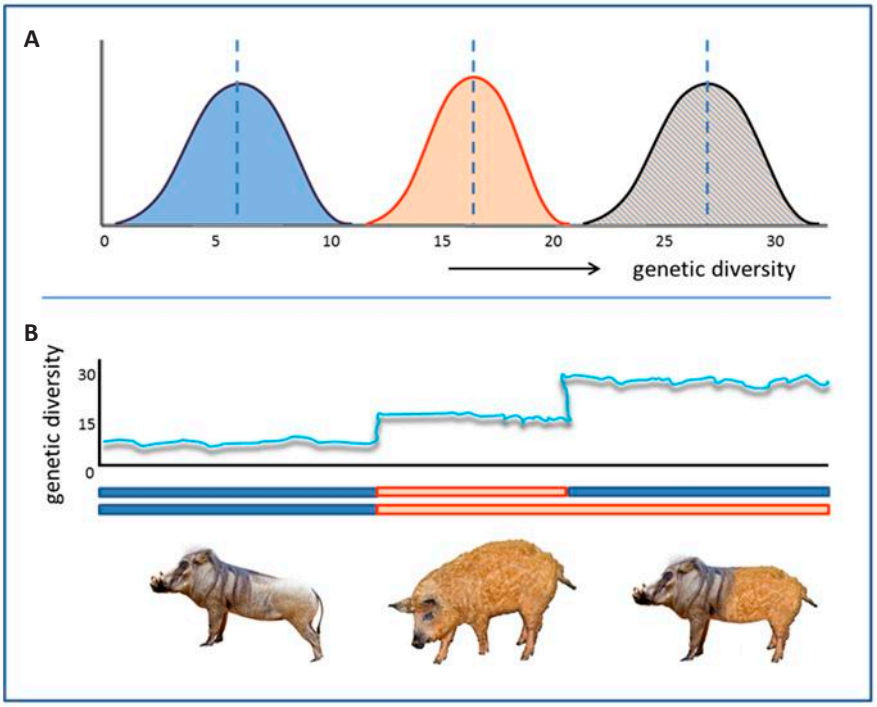


Figure 8.2 Schematic overview of hybridization in the genome. 8.2A. The hypothetical distribution of nucleotide diversity between two haplotypes from the same species (blue); another, more variable species (orange); and between two haplotypes from both species (hybrid, blue-orange). **8.2B.** Local nucleotide diversity within the genome when two haplotypes are from species 1 (blue), both local haplotypes are from species 2 (orange) or when the haplotypes come from two different species (hybrid blue-orange).

sent in European pigs. The solution is to sort the haplotypes in the genome by origin, and subsequently perform demographic analysis separately for, in the case of the domestic pigs, the European and the Asian haplotypes.

The existence of a shared polymorphism between members of different populations can have multiple causes. Convergent or parallel evolution may have promoted selection for this particular variant in both lineages, so the variant originated independently in both groups. Alternatively, the shared polymorphism in both populations is introgression through hybridization. A third possibility is that this polymorphism already existed in the ancestral lineage, also called incomplete lineage sorting (ILS). Incomplete lineage sorting can be recognized when a phylogeny, based on the haplotypes at a particular locus, is incongruent with the known phylogeny. The incongruence stemming from ILS is a well-known problem in phylogenetic studies. With the advent of genome-wide analyses, however, the incongruences have become a source of information on the complex nature of species formation and population differentiation. It is becoming increasingly recognized that ILS is widespread, contributing to the mosaic haplotype structure of genomes. The segments of shared ancestry between commercial pigs in Europe and Asia that I identify in Chapter 3 are probably a combination of ILS and introgression. Distinguishing hybridization, which signifies the introgression of haplotypes *after* separation of the two lineages, and ILS, which indicates sharing of haplotypes that stems from *before* the lineage separation, is a difficult task. Although incomplete lineage sorting is not easily distinguished from gene flow after separation of the two lineages, some methods have been developed (Joly 2009, Choleva 2014). The key of these methods is that the length of shared haplotype fragments reflects the number of recombination events since introgression, and therefore represents time. In combination with local recombination frequency and overall coalescent time estimated from the rest of the genome, this length can aid in the distinction between introgression and ILS. More context with regard to the background of the putative admixed populations can also help in resolving this issue. In chapter 3, for example, I find that the length of Asian haplotypes in commercial European pigs exceeds the length distribution between Asian wild boars and European commercial pigs, suggesting hybridization instead of ILS.

8.3.2 The consequences of hybridization

Hybridization combines (genetic) characteristics of divergent populations into one. Hybrids may contain phenotypes that display intermediate characteristics of two hybridizing populations. Under natural circumstances, hybrids may respond

differently to environmental pressure than parent populations. If reproductive isolation between two species is diminished, hybridization eventually may lead to the merging of two species into one. Classic examples of this reverse speciation include finches on the Galapagos Islands that show signs of convergence (Kleindorfer 2014), and the lineage fusion of two separate Galapagos tortoise populations (Garrick 2014). A rare phenomenon in reticulate evolution is enhanced speciation, in which a third, hybrid, lineage actually develops as a separate lineage. This is occasionally observed, such as a hybrid lineage that developed into a morphologically distinct third species in bats (Larsen 2010). Sometimes the increased heterozygosity in hybrids has positive effects on the phenotype, a phenomenon called heterosis. In livestock production, the advantages of hybridization are optimally used by combining characteristics of different parent populations into so-called finishers. The effect of heterosis is attributed to the fact that co-adapted gene complexes are still intact in the F_1, but recessive homozygotes are minimized (Chen 2013). Because of that, the heterosis effect is strongest in the first generation. The molecular mechanisms of heterosis however remain elusive, and will likely receive more attention in the near future now better genomics techniques are available. There is probably an optimal degree of relatedness between parents resulting in highest fitness of the offspring (Price and Waser 1979, Escobar 2008). Outbreeding depression can be seen as the introduction of maladapted genes into a population that decrease the overall fitness, or disrupting pleiotropic gene complexes so that local adaptation is distorted. Overall, introgressed haplotypes are, just as any other haplotype, subject to the same mechanisms of selection and drift that will determine whether they will remain in the population.

8.3.3 Adaptive introgression and selection

There are multiple examples of adaptive introgression, which suggest that the introgression of a novel haplotype can create increased adaptive potential. Hedrick 2013) states that the chance of an introgressed haplotype to persist in a population is highly increased when the haplotype has some adaptive potential. In chapter 5 I show that the introgression landscape is highly heterogeneous in the commercial Large White population and some regions in the genome are generally more Asian than European, suggesting selection for the Asian haplotypes. We show a significant effect of the Asian haplotypes on litter size for the introgressed region at the *AHR* locus, supporting this hypothesis. In chapter 6 I explore how much the introgression landscape has been influenced by selection for commercial traits, and demonstrate that the majority of regions containing an excess of Asian haplotypes

has an effect on backfat thickness in the Large White population. Together with lack of clear correlation between introgression signal and recombination frequency and gene content, this finding suggests that the frequencies of introgressed haplotypes are mainly determined by selection. In commercial pigs, therefore, introgression signatures seem to be clustered in regions of the genome where the introgressed haplotypes have a phenotypic effect that is selected favorably, but the genome-wide introgression signature is low. Other 'Success stories' of adaptive introgression include the yellow skin allele in chicken (Ericsson 2008) but also high-altitude adaptation in Tibetans that is thought to be adaptively introgressed from Denisovan-like humans (Huerta-Sanchez 2014). Resistance against the pesticide warfarin in house mice (*Mus musculus domesticus*) has been acquired from the Algerian mouse (*M. spretus*) through introgressive hybridization. However, no other evidence of hybridization between these two species exists (Song 2011). In guppies, locally adapted phenotypes can be maintained despite extensive gene flow from divergent populations (Fitzpatrick 2015).These examples of introgression generally pertain 'qualitative' traits, so traits with probably few genetic loci involved. The work described in chapter 6 demonstrates that introgression can also involve quantitative traits. The records of adaptive introgression follow the same trend as I describe in chapter 5 and 6: adaptive regions contain high introgression signatures, but the rest of the genome has a low frequency of introgressed haplotypes, and therefore selection is one of the major determining forces.

Interestingly, admixture between divergent species offers the possibility to screen for introgression as an alternative selective sweep analysis. If introgressed haplotypes indeed persist mainly because of selection, identifying loci with introgressed haplotypes can reveal selected sites. As ongoing selection for introgressed haplotypes will likely result in increased diversity levels instead of reduced diversity, classic sweep analyses are unlikely to pick up selection signals at these loci. This signal is somewhat comparable to selection on standing variation, because reduction in variation under that scenario is limited as well. In chapter 5 and 6 we use local frequencies of Asian haplotypes in a commercial population to search for selective sweeps for Asian variants. I conclude that the introgressed haplotypes indeed had a selective advantage because of their effect on litter size and fatness. Whether this selection acted upon a single haplotype or multiple introgressed haplotypes does not matter for the detection, but these regions would have remained unidentified without this introgression mapping. Our study therefore highlights the potential of introgression mapping for the identification of regions under selection.

8.4 Genomics in Management

It has become obvious that many species and populations face extinction due to human activities, to the extent that we are thought to approach the sixth mass extinction in the history of our planet (Barnosky 2011). Recent estimates of extinction of currently living species in 2050 are as high as 15 to 40% (Thomas 2013). Not only is much habitat destroyed by the ongoing expansion of the human population, local infrastructure has fragmented ecosystems resulting in small isolated populations. In addition, unintentional transportation of individuals led to invasive species that outcompete, or hybridize with, local species. Disregarding the moral question whether we are obliged to counteract what we have initiated, time is running out for many endangered species and populations. Proper management is crucial if we are to keep them from going extinct.

8.4.1 Populations under genetic management

It is important to obtain a good characterization of the population before applying management. Levels of inbreeding, but also associated inbreeding effects and future perspectives with regard to habitat availability are important factors that should determine conservation priorities. Prediction programs like Vortex have proven particularly useful in such prioritizations (Lacy 1993). However, the implementation of genomics as a tool to characterize populations is still limited. Which measurements and actions can be taken highly depends on the population. I distinguish three different types of populations under management:

1. **Wild:** Managing wild populations mostly encompasses the monitoring of the current status. Fragmentation, even by small disturbances such as motorways, can significantly reduce gene flow between populations. However, the effects on genetic diversity may be species-specific (Frantz 2012a). Species-specific knowledge may therefore be relevant for directing conservation efforts. Estimating numbers can be done through surveys, but genomics can aid in assessing inbreeding levels, genetic diversity, genetic load and admixture. Based on these outcomes, further decisions can be made for promoting gene-flow by increasing mobility of individuals through corridors, translocation or reintroduction.
2. **Captive:** In captive populations, generally, individuals can be distinguished and a well-founded estimate of the population size can be generated. Captive populations usually are not subjected to selection, apart from (unintentional) selection because of differential captive survival potential of individuals. In this aspect lies the distinction with domestic populations. If reproduction can be

managed in captive populations, optimal contributions (OC) can be applied. If not, the population should be tightly monitored, possibly with pedigrees, to keep insight in the population composition (i.e. determine paternity). If inbreeding and/or diseases are problematic, detection and removal of the causative haplotypes can be an option. Hybridization or allocations should only be done if another captive or wild population of the same species is available AND has been genetically characterized for suitability.

3. Domesticated: Domesticated species are usually under strong artificial selection directed to a specific breeding goal. The question to answer by genetic screening is whether the utility of the domestic population can be maintained or optimized. Maintaining variation is not a goal *per se*, but is considered important because it may be needed to keep inbreeding at an acceptable level and provides the raw material to select on. Very extensive genetic screening and genomic selection can be applied to some of these populations. Reference genome or high-density SNPchips are often available. Actually, genomic selection is a particular type of genetic management and a technique that is indispensable in the current breeding industry.

8.4.2 Advances of genomics in management

Next-generation sequencing (NGS) has provided unprecedented opportunities in a wide range of different fields in biology, like conservation biology, selective breeding and medicine. Sequencing costs will likely decrease rapidly, and accuracy of genotyping species will be much improved (Ekblom 2011). Genomics provides several advances to improve population management and conservation. Taxonomic status and phylogenetic relationships can be better assessed, and putative hybridization can be revealed and pinpointed in the genome. Genotype and sequence information can be used to maximize variation with breeding schemes. Also, particular variants under selection or deleterious mutations can be identified.

All forms of population management aim to maintain genetic diversity levels as high as possible. Maximizing variation, however, may lead to an increase of deleterious variants in the population, and according to the rule of Hardy-Weinberg (Hardy 1908), these will also occur in homozygous state if their selective disadvantage is ignored. Therefore, aiming for the highest possible variation is not always the best strategy to manage variation in population (deCara 2013b). In chapter 7 I use the distribution of haplotypes in a commercial population and a zoo population to optimally manage the population *in silico*. I demonstrate that

coancestry measures based on molecular markers outperform pedigree-based measurements. In addition, ROH-based method maintained a higher fitness level in the population than OC based on all markers, because more homozygous recessives are removed. In our simulations we used predicted deleterious sites as a proxy for fitness, and randomly assigned their effect based on a realistic distribution (Mukai 1972). Identifying deleterious variants in a genome will become easier and more reliable in the near future. Implementing estimates of effects of deleterious variants into breeding programs is therefore not unrealistic. In the future, it might become possible to estimate the selection coefficient *s* and dominance *h* from sequence data, but for now we still need to make assumptions about the putative effects of the variants. I show in chapter 7 that management strategies may influence ongoing selection by reducing the frequency of selected haplotypes. Therefore, haplotypes under selection should be considered in the breeding scheme as well, so that their high frequency is maintained despite the effort to optimize diversity. These types of marker-assisted management are feasible for populations in which the genomic information can be implemented in a breeding scheme. For wild populations, other management strategies should be applied.

A common reason why native populations facing rapid environmental changes become endangered is phenotype-environment mismatch (Carroll 2014). Reduction of this mismatch is desirable, also for failing (re-) introduced populations or captive populations. Genomics can aid in this matter by, for example, identifying variants that result in maladaptation. Particular genomic segments containing disadvantageous alleles can be selectively removed from the population by minimizing their contribution to future generations. However, controlling contributions to future generations is not always possible. In addition, standing genetic variation may very well not contain the desired variants that will lead to better adaptation. Therefore, other populations can be screened for suitable variants that could reduce the maladaptation. Improvement of diversity levels and/or fitness through the introgression of genetic material from another source is referred to as ‘genetic rescue’ (Whitley 2015). Some well-known success stories of genetic rescue are the Florida panther (Johnson 2010, Hostetler 2013) and South Island robin populations (Heber 2013). Such genetic rescue efforts do come with potential risks (Thomas 2013). Some species are thought to be resilient against inbreeding depression (Fountain 2015). Especially if small population size is a characteristic of the species, ‘genetic rescue’ might not be desirable because of the potentially more harmful effects of outbreeding depression (Lynch 1991). General

predictions about the outcome of hybridization are difficult, since specific circumstances seem to have a high impact on the persistence and effects of introgressed haplotypes in a population (Abbott 2013). Patel (2015) uses the characteristics of two hybridizing, cryptic, snail species to predict the direction of allele frequency change in the hybrid zone because of climate change. These types of predictions on the suitability of genetic material belonging to closely related species could also be used to determine the most suitable source for genetic rescue of endangered populations in a changing environment. This strategy will probably be most useful when visible phenotypic differences are small, so that genomics techniques add to the determination of suitability.

A general concern regarding genetic rescue is whether admixture threatens the integrity of a species. With regard to genetic diversity, admixture with another population or even species might eventually maintain more of the original genetic variation from the source population than when no genetic rescue is applied. An interesting example is the North American bison population that nowadays consists of roughly half a million individuals, but descended from only 100 individuals in the late 19^{th} century. Admixture with domestic cattle to genetically rescue the bison population has occurred more than once, and nearly all bison populations contain some cattle ancestry. But without this introgression, probably more bison-specific haplotypes would have been lost (Hedrick 2009). The pattern of the introgression landscape in pigs (chapter 5 and 6) shows that undesired haplotypes, or haplotypes without effect on the phenotype, may disappear from the population, leaving mostly those that confer a selective advantage. This is in agreement with the predictions by Hedrick (2013), and demonstrates that indeed introgressed haplotypes can be maintained only at those loci where they are desired. However, a reintroduction event of beavers in France unintentionally caused hybridization with North American beavers. The non-native and hybrid beavers are thought to outcompete the native lineage, threatening their survival (Dewas 2012). The level of asymmetry between mixing populations may influence how much introgressed haplotypes will persist, but also the selective advantage of the introgressed haplotypes. A potential solution here is to reduce the proportion of introduced haplotypes as much as possible. Could fine-scale haplotype identification also lead to the second step of removal of most of the introgressed haplotypes so that the population remains as ‘pure as possible’? I do believe that by identifying introgressed haplotypes as described in this thesis and optimizing contributions of individuals containing less introgression, the proportion of introgressed haplotypes can potentially be reduced. Such controlled breeding is often impossible in the

wild, and therefore the suitability of a donor population should be carefully considered.

8.4.3 Future perspectives in genomic management

Characterizing the complete genomic sequence of individuals will become feasible for all species. Promising new tools are under development and it is a matter of time before genome scans can be cost effectively executed. This may create the possibility to obtain full genome information for all individuals in a population. Unfortunately, identification of variation is not sufficient to maintain all variants in a small population. The genomic management techniques that I described in chapter 7 will be able to slow down the degradation of genetic variation in an isolated small population, but eventually this is just a way to buy time. Even if the best conservation method cannot guarantee viability of the population in the long run, buying as much time as possible can nevertheless be of vital importance. E.g., circumstances may change in time so that the population can recover in the wild by itself. Alternatively, genome-editing technology may, in the not too distant future, rectify accumulation of deleterious variants. This obviously would require that the population or species at risk survives long enough to make these techniques economically feasible. My expectation is that habitat recovery, required for population growth, will probably come too late, if ever, for most endangered species. Therefore, tweaking genetic variation for the benefit of species or population survival is a more realistic scenario, and becomes necessary for the future conservation of species.

The fate of a population can be determined by a single mutation only. In Californian condors, for example, one variant is thought to cause lethal dwarfism, but this variant is present in high frequency due to founder effects (Ralls 2000). Excluding individuals with this variant from breeding will reduce the effective population size even further. If the effect of a certain mutation, like in this example, is very well predicted and severe, the modification of a gene might be the best solution (Thomas 2013). Modification could either be direct by changing one or a few nucleotides in the sequence, or transplanting any genetic material from another source into the genome. Crop breeding has a relatively long history in genetic modification, and a successful example is the insertion of genes that induce drought tolerance in plants (Ashraf 2010). Also in the medical world, the benefits from genetic modification and gene therapy are already explored in medical trials (Tachibana 2013). These practices can be adapted to fit current needs in conservation biology. How to genetically modify organisms so that they are better

adapted to their current/changing environment? The most conventional manner is to transport a known “good gene” from a closely related population or species into the genome of individuals contributing to the endangered population. Such facilitated adaptation can have great benefits over crossbreeding, because the possible complications of outbreeding depression are reduced. Therefore, repairing genetic accidents by artificially inserting some genetic material from another source at one particular locus in the genome might be a better option than crossbreeding, because more authentic material can be maintained. Genome editing does not even require external material and simply using information of which variants used to be (or should be) present may be enough to (re)create the desired genotype. Overall, threatened populations should first be well characterized, since the best management strategy seems highly specific. More research on the genomic consequences of inbreeding and hybridization, such as the work in this thesis, will help in predicting the effects of different management strategies and making the best decisions to conserve threatened populations.

8.5 Concluding remarks

In this thesis, I disentangled the mosaic haplotype structure of multiple pig and wild boar populations. This thesis provides a detailed example of how genomes are affected by demography, hybridization and selection. These results are valuable for characterization and management of endangered species as well as livestock populations. However, detailed genomic information in combination with potential phenotypic consequences will not be enough to be able to conserve relevant evolutionary units under threat of extinction. We have entered an era in which fine-scale predictions about genomic effects of mixing populations and the putative effect of even single nucleotide polymorphisms are crucial in conservation biology. The insights gained in this thesis stress the need for genetic diversity measures at multiple levels, in order to best manage populations and potential genetic resources.

References

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman JW, Brelsford A, Buerkle CA, Buggs R et al. 2013. Hybridization and speciation. *J Evolution Biol* **26**(2): 229-246.

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA et al. 2011. The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science* **334**(6052): 89-94.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7**(4): 248-249.

Aguilar A, Roemer G, Debenham S, Binns M, Garcelon D, Wayne RK. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *P Natl Acad Sci USA* **101**(10): 3490-3494.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**(9): 1655-1664.

Allendorf FW, Hohenlohe PA, Luikart G. 2010. Genomics and the future of conservation genetics. *Nat Rev Genet* **11**(10): 697-709.

Altshuler D Brooks LD Chakravarti A Collins FS Daly MJ Donnelly P Gibbs RA Belmont JW Boudreau A Leal SM et al. 2005. A haplotype map of the human genome. *Nature* **437**(7063): 1299-1320.

Alves E, Ovilo C, Rodriguez MC, Silio L. 2003. Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Animal Genetics* **34**(5): 319-324.

Alves PC, Pinheiro I, Godinho R, Vicente J, Gortazar C, Scandura M. 2010. Genetic diversity of wild boar populations and domestic pig breeds (Sus scrofa) in South-western Europe. *Biol J Linn Soc* **101**(4): 797-822.

Amaral AJ, Ferretti L, Megens HJ, Crooijmans RPMA, Nie HS, Ramos-Onsins SE, Perez-Enciso M, Schook LB, Groenen MAM. 2011. Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA. *Plos One* **6**(4).

Andersson L. 2001. Genetic dissection of phenotypic diversity in farm animals. *Nat Rev Genet* **2**(2): 130-138.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437**(7062): 1149-1152.

Arias JA, Keehan M, Fisher P, Coppieters W, Spelman R. 2009. A high density linkage map of the bovine genome. *Bmc Genetics* **10**.

Ashraf M. 2010. Inducing drought tolerance in plants: Recent advances. *Biotechnol Adv* **28**(1): 169-183.

Austerlitz F, Gleiser G, Teixeira S, Bernasconi G. 2012. The effects of inbreeding, genetic dissimilarity and phenotype on male reproductive success in a dioecious plant. *P Roy Soc B-Biol Sci* **279**(1726): 91-100.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN et al. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* **19**(5): 795-803.

Baba T, Mimura J, Nakamura N, Harada N, Yamamoto M, Morohashi K, Fujii-Kuriyama Y. 2005. Intrinsic function of the aryl hydrocarbon (Dioxin) receptor as a key factor in female reproduction. *Molecular and Cellular Biology* **25**(22): 10040-10051.

Baker A. 2007. Animal ambassadors: an analysis of the effectiveness and conservation impact of ex situ breeding efforts. In *Zoos in the 21st century—catalysts for conservation?, vol 15 Conservation biology*, (ed. HM Zimmermann A, Dickie LA, West C), pp. 139-154. Cambridge University Press, Cambridge.

Ballou JD LR. 1995. Identifying genetically important individuals for management of genetic diversity in pedigreed populations. In *Population Management for Survival & Recovery Analytical Methods and Strategies in Small Population Conservation*, (ed. MG J.D. Ballou, and T.J. Foose). Columbia University Press, New York.

Banks SC, Cary GJ, Smith AL, Davies ID, Driscoll DA, Gill AM, Lindenmayer DB, Peakall R. 2013. How does ecological disturbance influence genetic diversity? *Trends Ecol Evol* **28**(11): 670-679.

References

Barnosky AD, Matzke N, Tomiya S, Wogan GOU, Swartz B, Quental TB, Marshall C, McGuire JL, Lindsey EL, Maguire KC et al. 2011. Has the Earth's sixth mass extinction already arrived? *Nature* **471**(7336): 51-57.

Bartz M, Kociucka B, Mankowska M, Switonski M, Szydlowski M. 2013. Transcript level of the porcine ME1 gene is affected by SNP in its 3' UTR, which is also associated with subcutaneous fat thickness. *Journal of Animal Breeding and Genetics*.

Bateson W. 1905. Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* **2**: 1–55, 80–99.

Begun DJ, Aquadro CF. 1992. Levels of Naturally-Occurring DNA Polymorphism Correlate with Recombination Rates in Drosophila-Melanogaster. *Nature* **356**(6369): 519-520.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng ZP Snyder M Dermitzakis ET Stamatoyannopoulos JA et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Bjelland DW, Weigel KA, Vukasinovic N, Nkrumah JD. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *Journal of Dairy Science* **96**(7): 4697-4706.

Blixt S. 1975. Why Didnt Mendel,G Find Linkage. *Nature* **256**(5514): 206-206.

Bonnet C, Andrieux J, Béri-Dexheimer M, Leheup B, Boute O, Manouvrier S, Delobel B, Copin H, Receveur A, Mathieu Ma. 2010. Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked growth restriction, mental retardation and absent or severely delayed speech. *Journal of medical genetics* **47**(6): 377-384.

Bosse M, Megens HJ, Frantz LAF, Madsen O, Larson G, Paudel Y, Duijvesteijn N, Harlizius B, Hagemeijer Y, Crooijmans RPMA et al. 2014a. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature Communications* **5**.

Bosse M, Megens HJ, Madsen O, Frantz LAF, Paudel Y, Crooijmans RPMA, Groenen MAM. 2014b. Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. *Mol Ecol* **23**(16): 4089-4102.

Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LA, Schook LB, Crooijmans RP, Groenen MA. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet* **8**(11): e1003100.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao KY, Brisbin A, Parker HG, vonHoldt BM et al. 2010. A Simple Genetic Architecture Underlies Morphological Variation in Dogs. *Plos Biol* **8**(8).

Braglia S, Zappaterra M, Zambonelli P, Comella M, Dall'Olio S, Davoli R. 2014. Analysis of g. 265T> C SNP of fatty acid synthase gene and expression study in skeletal muscle and backfat tissues of Italian Large White and Italian Duroc pigs. *Livestock Science* **162**: 15-22.

Brotherstone S, Goddard M. 2005. Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philos Trans R Soc Lond B Biol Sci* **360**(1459): 1479-1488.

Browning BL, Browning SR. 2011. A Fast, Powerful Method for Detecting Identity by Descent. *Am J Hum Genet* **88**(2): 173-182.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**(5): 1084-1097.

Bruford MW, Bradley DG, Luikart G. 2003. DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet* **4**(11): 900-910.

Carroll SP, Jorgensen PS, Kinnison MT, Bergstrom CT, Denison RF, Gluckman P, Smith TB, Strauss SY, Tabashnik BE. 2014. Applying evolutionary biology to address global challenges. *Science* **346**(6207): 313-+.

Castric V, Bechsgaard J, Schierup MH, Vekemans X. 2008. Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection. *Plos Genetics* **4**(8).

Charbonnel N, Pemberton J. 2005. A long-term genetic survey of an ungulate population reveals balancing selection acting on MHC through spatial and temporal fluctuations in selection. *Heredity* **95**(5): 377-388.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**(3): 195-205.

Charlesworth D. 2003. Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **358**(1434): 1051-1070.

Charlesworth D, Charlesworth B. 1987. Inbreeding Depression and Its Evolutionary Consequences. *Annual Review of Ecology and Systematics* **18**: 237-268.

Chen CY, Sargent C, Quilter C, Yang ZQ, Ren J, Affara N, Brenig B, Huang LS. 2010. Cloning, mapping and molecular characterization of porcine progesterone receptor membrane component 2 (PGRMC2) gene. *Genetics and Molecular Biology* **33**(3): 471-474.

Chen ZJ. 2013. Genomic and epigenetic insights into the molecular bases of heterosis. *Nat Rev Genet* **14**(7): 471-482.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. *Plos One* **7**(10): e46688.

Choleva L, Musilova Z, Kohoutova-Sediva A, Paces J, Rab P, Janko K. 2014. Distinguishing between Incomplete Lineage Sorting and Genomic Introgressions: Complete Fixation of Allospecific Mitochondrial DNA in a Sexually Reproducing Fish (Cobitis; Teleostei), despite Clonal Reproduction of Hybrids. *Plos One* **9**(6).

Ciobanu D, Bastiaansen J, Lonergan SM, Thomsen H, Dekkers JC, Plastow GS, Rothschild MF. 2004. New alleles in calpastatin gene are associated with meat quality traits in pigs. *Journal of animal science* **82**(10): 2829-2839.

Clark SA, Kinghorn BP, Hickey JM, van der Werf JH. 2013. The effect of genomic information on optimal contribution selection in livestock breeding programs. *Genet Sel Evol* **45**: 44.

Clop A, Amills M, Noguera JL, Fernandez A, Capote J, Ramon MM, Kelly L, Kijas JMH, Andersson L, Sanchez A. 2004. Estimating the frequency of Asian cytochrome B haplotypes in standard European and local Spanish pig breeds. *Genetics Selection Evolution* **36**(1): 97-104.

Connallon T, Clark AG. 2013. Antagonistic Versus Nonantagonistic Models of Balancing Selection: Characterizing the Relative Timescales and Hitchhiking Effects of Partial Selective Sweeps. *Evolution* **67**(3): 908-917.

Crispo E, Moore JS, Lee-Yaw JA, Gray SM, Haller BC. 2011. Broken barriers: Human-induced changes to gene flow and introgression in animals. *BioEssays* **33**(7): 508-518.

Curik I, Ferencakovic M, Solkner J. 2014. Inbreeding and runs of homozygosity: A possible solution to an old problem. *Livestock Science* **166**: 26-34.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution* **62**(8): 1908-1920.

Curtis D, Vine AE, Knight J. 2008. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Annals of Human Genetics* **72**: 261-278.

de Cara MA, Fernandez J, Toro MA, Villanueva B. 2011. Using genome-wide information to minimize the loss of diversity in conservation programmes. *J Anim Breed Genet* **128**(6): 456-464.

de Cara MA, Villanueva B, Toro MA, Fernandez J. 2013a. Purging deleterious mutations in conservation programmes: combining optimal contributions with inbred matings. *Heredity (Edinb)* **110**(6): 530-537.

de Cara MA -. 2013b. Using genomic tools to maintain diversity and fitness in conservation programmes. *Mol Ecol* **22**(24): 6091-6099.

del Marmol V, Beermann F. 1996. Tyrosinase and related proteins in mammalian pigmentation. *FEBS letters* **381**(3): 165-168.

Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**(2): 179-181.

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**(1): 5-6.

Denison MS, Soshilov AA, He GC, DeGroot DE, Zhao B. 2011. Exactly the Same but Different: Promiscuity and Diversity in the Molecular Mechanisms of Action of the Aryl Hydrocarbon (Dioxin) Receptor. *Toxicological Sciences* **124**(1): 1-22.

Dewas M, Herr J, Schley L, Angst C, Manet B, Landry P, Catusse M. 2012. Recovery and status of native and introduced beavers Castor fiber and Castor canadensis in France and neighbouring countries. *Mammal Rev* **42**(2): 144-165.

Diamond J. 2002. Evolution, consequences and future of plant and animal domestication. *Nature* **418**(6898): 700-707.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol Ecol Resour* **14**(1): 209-214.

Dobney K, Larson G. 2006. Genetics and animal domestication: new windows on an elusive process. *Journal of Zoology* **269**(2): 261-271.

Doebley JF, Gaut BS, Smith BD. 2006. The molecular genetics of crop domestication. *Cell* **127**(7): 1309-1321.

Dong ZM, Gutierrez-Ramos J-C, Coxon A, Mayadas TN, Wagner DD. 1997. A new class of obesity genes encodes leukocyte adhesion receptors. *Proceedings of the National Academy of Sciences* **94**(14): 7526-7530.

Dunham I Kundaje A Aldred SF Collins PJ Davis C Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. *Mol Biol Evol* **28**(8): 2239-2252.

Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **107**(1): 1-15.

Eriksson J, Larson G, Gunnarsson U, Bed'hom B, Tixier-Boichard M, Stromstedt L, Wright D, Jungerius A, Vereijken A, Randi E et al. 2008. Identification of the Yellow skin gene reveals a hybrid origin of the domestic chicken. *Plos Genetics* **4**(2).

Escobar JS, Nicot A, David P. 2008. The Different Sources of Variation in Inbreeding Depression, Heterosis and Outbreeding Depression in a Metapopulation of Physa acuta. *Genetics* **180**(3): 1593-1608.

Esteve-Codina A, Kofler R, Himmelbauer H, Ferretti L, Vivancos AP, Groenen MAM, Folch JM, Rodriguez MC, Perez-Enciso M. 2011. Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. *Heredity* **107**(3): 256-264.

Fang L, Ye J, Li N, Zhang Y, Li S, Gane KSW, Wang J. 2008. Positive correlation between recombination rate and nucleotide diversity is shown under domestication selection in the chicken genome. *Chinese Science Bulletin* **53**(5): 746-750.

Fang M, Berg F, Ducos A, Andersson L. 2006. Mitochondrial haplotypes of European wild boars with 2n=36 are closely related to those of European domestic pigs with 2n=38. *Animal Genetics* **37**(5): 459-464.

Fang M, Larson G, Ribeiro HS, Li N, Andersson L. 2009. Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genetics* **5**(1): e1000341.

Fang MY, Andersson L. 2006. Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. *P Roy Soc B-Biol Sci* **273**(1595): 1803-1810.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. . . Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ferencakovic M, Hamzic E, Gredler B, Solberg TR, Klemetsdal G, Curik I, Solkner J. 2013a. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. *Journal of Animal Breeding and Genetics* **130**(4): 286-293.

Ferencakovic M, Solkner J, Curik I. 2013b. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genetics Selection Evolution* **45**.

Fernandez J, Villanueva B, Pong-Wong R, Toro MA. 2005. Efficiency of the use of pedigree and molecular marker information in conservation programs. *Genetics* **170**(3): 1313-1321.

Ferraz ALJ, Ojeda A, Lopez-Bejar M, Fernandes LT, Castello A, Folch JM, Perez-Enciso M. 2008. Transcriptome architecture across tissues in the pig. *Bmc Genomics* **9**.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Mol Biol Evol* **31**(5): 1275-1291.

Fienieg ES GP. 2013. The use and integration of molecular DNA information in conservation breeding programmes: a review. *JZAR* **1**(2).

Fitzpatrick SW, Gerberich JC, Kronenberger JA, Angeloni LM and Funk WC. 2015. Locally adapted traits maintained in the face of high gene flow. Ecology Letters (2015) 18: 37–47

Flori L, Thevenon S, Dayo GK, Senou M, Sylla S, Berthier D, Moazami-Goudarzi K, Gautier M. 2014. Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol Ecol* **23**(13): 3241-3257.

Forcada J, Hoffman JI. 2014. Climate change selects for heterozygosity in a declining fur seal population. *Nature* **511**(7510): 462-465.

Frankham R BJ, Briscoe DA. 2002. *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge, UK.

Frankham R BJ -. 2010. *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge, UK.

Frantz AC, Bertouille S, Eloy MC, Licoppe A, Chaumont F, Flamand MC. 2012a. Comparative landscape genetic analyses show a Belgian motorway to be a gene flow barrier for red deer (Cervus elaphus), but not wild boars (Sus scrofa). *Mol Ecol* **21**(14): 3445-3457.

Frantz AC, Massei G, Burke T. 2012b. Genetic evidence for past hybridisation between domestic pigs and English wild boars. *Conserv Genet* **13**(5): 1355-1364.

Frantz AC, Zachos FE, Kirschning J, Cellina S, Bertouille S, Mamuris Z, Koutsogiannouli EA, Burke T. 2013a. Genetic evidence for introgression between domestic pigs and wild boars (Sus scrofa) in Belgium and Luxembourg: a comparative approach with multiple marker systems. *Biol J Linn Soc* **110**(1): 104-115.

Frantz LA, Schraiber JG, Madsen O, Megens HJ, Bosse M, Paudel Y, Semiadi G, Meijaard E, Li N, Crooijmans RP et al. 2013b. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. *Genome Biol* **14**(9): R107.

Frantz LAF, 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.

Frantz LAF. 2015. Speciation and domestication in Suiformes: a genomic perspective. PhD thesis, Wageningen University, Wageningen, The Netherlands. ISBN 978-94-6257-254-6

Fuji-Kuriyama Y, Kawajiri K. 2010. Molecular mechanisms of the physiological functions of the aryl hydrocarbon (dioxin) receptor, a multifunctional regulator that senses and responds to environmental stimuli. *Proceedings of the Japan Academy Series a-Mathematical Sciences* **86**(1): 40-53.

Garbe JR DY. 2008. *Pedigraph: A Software Tool for the Graphing and Analysis of Large Complex Pedigree. User manual Version 2.4.* Department of Animal Science, University of Minnesota, Minnesota.

Garrick DJ, Taylor JF, Fernando RL. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol* **41**(55): 44.

Garrick RC, Benavides E, Russello MA, Hyseni C, Edwards DL, Gibbs JP, Tapia W, Ciofi C, Caccone A. 2014. Lineage fusion in Galapagos giant tortoises. *Mol Ecol* **23**(21): 5276-5290.

Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, Presgraves DC. 2012. Genome sequencing reveals complex speciation in the Drosophila simulans clade. *Genome Res* **22**(8): 1499-1511.

Gautier M, Vitalis R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**(8): 1176-1177.

Genov PV. 1999. A review of the cranial characteristics of the Wild Boar (Sus scrofa Linnaeus 1758), with systematic conclusions. *Mammal Rev* **29**(4): 205-238.

Gilmour AR. 2009. *ASReml user guide release 3.0.* VSN International Ltd, Hemel Hempstead, UK.

Gilmour AR, Gogel B, Cullis B, Thompson R. 2009. ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Gitelman I. 2007. Evolution of the vertebrate twist family and Synfunctionalization: A mechanism for differential gene loss through merging of expression domains. *Mol Biol Evol* **24**(9): 1912-1925.

Giuffra E, Kijas JMH, Amarger V, Carlborg O, Jeon JT, Andersson L. 2000. The origin of the domestic pig: Independent domestication and subsequent introgression. *Genetics* **154**(4): 1785-1791.

Goddard M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**(2): 245-257.

Goedbloed DJ, Megens HJ, van Hooft P, Herrero-Medrano JM, Lutz W, Alexandri P, Crooijmans RPMA, Groenen M, van Wieren SE, Ydenberg RC et al. 2013a. Genome-wide single nucleotide polymorphism analysis reveals recent genetic introgression from domestic pigs into Northwest European wild boar populations. *Mol Ecol* **22**(3): 856-866.

Goedbloed DJ, van Hooft P, Megens HJ, Langenbeck K, Lutz W, Crooijmans RPMA, van Wieren SE, Ydenberg RC, Prins HHT. 2013b. Reintroductions and genetic introgression from domestic pigs have shaped the genetic population structure of Northwest European wild boar. *Bmc Genetics* **14**.

Gomez-Romano F, Villanueva B, de Cara MA, Fernandez J. 2013. Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size. *Genet Sel Evol* **45**: 38.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai WW, Fritz MHY et al. 2010. A Draft Sequence of the Neandertal Genome. *Science* **328**(5979): 710-722.

Groenen MA Archibald AL Uenishi H Tuggle CK Takeuchi Y Rothschild MF Rogel-Gaillard C Park C Milan D Megens HJ et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**(7424): 393-398.

Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, Negrini R, Finlay EK, Jianlin H, Groeneveld E et al. 2010. Genetic diversity in farm animals - a review. *Animal Genetics* **41**: 6-31.

Groves C.. 2008. Current views on the taxonomy and zoogeography of the genus Sus. In in Pigs and Humans: 10,000 Years of Interaction, (ed. KD U. Albarella, A. Ervynck, and P. Rowley-Conwy), pp. 15-29. Oxford University Press, Oxford.

Grundy B, Villanueva B, Woolliams JA. 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. *Genetical Research* **72**(2): 159-168.

Habier D, Gotz KU, Dempfle L. 2009. Breeding programme for Pietrain pigs in Bavaria with an estimation of genetic trends and effective population size. *Livestock Science* **123**(2-3): 187-192.

Haider S, Ballester B, Smedley D, Zhang JJ, Rice P, Kasprzyk A. 2009. BioMart Central Portal-unified access to biological data. *Nucleic Acids Research* **37**: W23-W27.

Haldane JBS. Sex ratio and unisexual sterility in hybrid animals. J Genet 1922;12:101-109

Hall SJG, Bradley DG. 1995. Conserving Livestock Breed Biodiversity. *Trends Ecol Evol* **10**(7): 267-270.

Hanebuth T, Stattegger K, Grootes PM. 2000. Rapid flooding of the Sunda Shelf: A late-glacial sea-level record. *Science* **288**(5468): 1033-1035.

Hanenberg. 2010. 2008 Pig cost of prod. in selected countries. In *Milton keynes, BPEX*

Harper PA, Wong JMY, Lam MSM, Okey AB. 2002. Polymorphisms in the human AH receptor. *Chemico-Biological Interactions* **141**(1-2): 161-187.

Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet* **9**(6): e1003521.

Harrison RG, Larson EL. 2014. Hybridization, Introgression, and the Nature of Species Boundaries. *J Hered* **105**: 795-809.

Heber S, Varsani A, Kuhn S, Girg A, Kempenaers B, Briskie J. 2013. The genetic rescue of two bottlenecked South Island robin populations using translocations of inbred donors. *P Roy Soc B-Biol Sci* **280**(1752).

Hedrick PW. 2009. Conservation Genetics and North American Bison (Bison bison). *J Hered* **100**(4): 411-420.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol* **22**(18): 4606-4618.

Hedrick PW, Kalinowski ST. 2000. Inbreeding depression in conservation biology. *Annual Review of Ecology and Systematics* **31**: 139-162.

Heidaritabar M VA, Muir WM, Meuwissen T, Cheng H, Megens HJ, Groenen MA, Bastiaansen JW. 2014. Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. *Heredity (Edinb)* **113**: 503–513.

Helbig AJ. 1991. Se-Migrating and Sw-Migrating Blackcap (Sylvia-Atricapilla) Populations in Central-Europe - Orientation of Birds in the Contact Zone. *J Evolution Biol* **4**(4): 657-670.

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**(6): 1527-1535.

Henn BM, Botigue LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J et al. 2012. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. *Plos Genetics* **8**(1).

Hernandez-Ochoa I, Karman BN, Flaws JA. 2009. The role of the aryl hydrocarbon receptor in the female reproductive system. *Biochemical Pharmacology* **77**(4): 547-559.

Herrero-Medrano JM, Megens HJ, Crooijmans RP, Abellaneda JM, Ramis G. 2013a. Farm-by-farm analysis of microsatellite, mtDNA and SNP genotype data reveals inbreeding and crossbreeding as threats to the survival of a native Spanish pig breed. *Animal Genetics* **44**(3): 259-266.

Herrero-Medrano JM, Megens HJ, Groenen MAM, Bosse M, Perez-Enciso M, Crooijmans RPMA. 2014. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. *Bmc Genomics* **15**.

Herrero-Medrano JM, Megens HJ, Groenen MAM, Ramis G, Bosse M, Perez-Enciso M, Crooijmans RPMA. 2013b. Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *Bmc Genetics* **14**.

Hey J. 2006. Recent advances in assessing gene flow between diverging populations and species. *Curr Opin Genet Dev* **16**(6): 592-596.

Hidalgo AM, Bastiaansen JWM, Harlizius B, Megens HJ, Madsen O, Crooijmans RPMA, Groenen MAM. 2014. On the relationship between an Asian haplotype on chromosome 6 that reduces androstenone levels in boars and the differential expression of SULT2A1 in the testis. *Bmc Genetics* **15**.

Hill WG. 1981. Estimation of Effective Population-Size from Data on Linkage Disequilibrium. *Genetical Research* **38**(3): 209-216.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol Ecol* **22**(11): 3002-3013.

Hostetler JA, Onorato DP, Jansen D, Oli MK. 2013. A cat's tale: the impact of genetic restoration on Florida panther population dynamics and persistence. *J Anim Ecol* **82**(3): 608-620.

Howrigan DP, Simonson MA, Keller MC. 2011. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *Bmc Genomics* **12**.

Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He MZ, Somel M et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**(7513): 194-+.

Jablonska O, Piasecka J, Ostrowska M, Sobocinska N, Wasowska B, Ciereszko RE. 2011. The expression of the aryl hydrocarbon receptor in reproductive and neuroendocrine tissues during the estrous cycle in the pig (vol 126, pg 221, 2011). *Animal Reproduction Science* **129**(1-2): 104-104.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu YT, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* **14**(4): 528-538.

Jeong CW, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, Beall CM, Di Rienzo A. 2014. Admixture facilitates genetic adaptations to high altitude in Tibet. *Nature Communications* **5**.

Jimenez JA, Hughes KA, Alaks G, Graham L, Lacy RC. 1994. An Experimental-Study of Inbreeding Depression in a Natural Habitat. *Science* **266**(5183): 271-273.

Johnson WE, Onorato DP, Roelke ME, Land ED, Cunningham M, Belden RC, McBride R, Jansen D, Lotz M, Shindle D et al. 2010. Genetic Restoration of the Florida Panther. *Science* **329**(5999): 1641-1645.

Joly S, McLenachan PA, Lockhart PJ. 2009. A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting. *Am Nat* **174**(2): E54-E70.

Jones G, Rothschild M, Ruvinsky A. 1998. Genetic aspects of domestication, common breeds and their origin. *The genetics of the pig*: 17-50.

Kanadia RN, Johnstone KA, Mankodi A, Lungu C, Thornton CA, Esson D, Timmers AM, Hauswirth WW, Swanson MS. 2003. A muscleblind knockout model for myotonic dystrophy. *Science* **302**(5652): 1978-1980.

Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. *Trends Ecol Evol* **17**(5): 230-241.

Keller MC, Visscher PM, Goddard ME. 2011. Quantification of Inbreeding Due to Distant Ancestors and Its Detection Using Dense Single Nucleotide Polymorphism Data. *Genetics* **189**(1): 237-U920.

Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. 2014. Selection for complex traits leaves little or no classic signatures of selection. *Bmc Genomics* **15**.

Kijas JMH, Andersson L. 2001. A phylogenetic study of the origin of the domestic pig estimated from the near-complete mtDNA genome. *Journal of Molecular Evolution* **52**(3): 302-308.

Kijas JMH, Wales R, Tornsten A, Chardon P, Moller M, Andersson L. 1998. Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics* **150**(3): 1177-1185.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic Runs of Homozygosity Record Population History and Consanguinity. *Plos One* **5**(11).

Kirkpatrick M, Jarne P. 2000. The effects of a bottleneck on inbreeding depression and the genetic load. *Am Nat* **155**(2): 154-167.

Kleindorfer S, O'Connor JA, Dudaniec RY, Myers SA, Robertson J, Sulloway FJ. 2014. Species Collapse via Hybridization in Darwin's Tree Finches. *Am Nat* **183**(3): 325-341.

Kristensen TN, Sorensen AC. 2005. Inbreeding - lessons from animal breeding, evolutionary biology and conservation genetics. *Animal Science* **80**: 121-133.

Ku CS, Naidoo N, Teo SM, Pawitan Y. 2011. Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* **129**(1): 1-15.

Lacy RC. 1993. Vortex - a Computer-Simulation Model for Population Viability Analysis. *Wildlife Res* **20**(1): 45-65.

Lacy RC -. 2013. Achieving true sustainability of zoo populations. *Zoo Biol* **32**(1): 19-26.

Lacy RC, Alaks G, Walsh A. 1997. Hierarchical analysis of inbreeding depression in Peromyscus polionotus (vol 50, pg 2187, 1996). *Evolution* **51**(3): 1025-1025.

Lacy RC, Ballou JD, Pollak JP. 2012. PMx: software package for demographic and genetic analysis and management of pedigreed populations. *Methods in Ecology and Evolution* **3**(2): 433-437.

Laikre L, Allendorf FW, Aroner LC, Baker CS, Gregovich DP, Hansen MM, Jackson JA, Kendall KC, McKelvey K, Neel MC et al. 2010. Neglect of Genetic Diversity in Implementation of the Convention on Biological Diversity. *Conservation Biology* **24**(1): 86-88.

Larance M, Ramm G, Stöckli J, van Dam EM, Winata S, Wasinger V, Simpson F, Graham M, Junutula JR, Guilhaus M. 2005. Characterization of the role of the Rab GTPase-activating protein AS160 in insulin-regulated GLUT4 trafficking. *Journal of Biological Chemistry* **280**(45): 37803-37813.

Larsen PA, Marchan-Rivadeneira MR, Baker RJ. 2010. Natural hybridization generates mammalian lineage with species characteristics. *P Natl Acad Sci USA* **107**(25): 11447-11452.

Larson G, Albarella U, Dobney K, Rowley-Conwy P, Schibler J, Tresset A, Vigne JD, Edwards CJ, Schlumbaum A, Dinu A et al. 2007a. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc Natl Acad Sci U S A* **104**(39): 15276-15281.

Larson G, Burger J. 2013. A population genetics view of animal domestication. *Trends Genet* **29**(4): 197-205.

Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, Rolett B, Spriggs M, Dolman G, Kim TH et al. 2007b. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in island southeast Asia and Oceania. *P Natl Acad Sci USA* **104**(12): 4834-4839.

Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E et al. 2005. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**(5715): 1618-1621.

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Vigueira CC, Denham T, Dobney K et al. 2014. Current perspectives and the future of domestication studies. *P Natl Acad Sci USA* **111**(17): 6139-6146.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *Plos Biol* **10**(9).

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK. 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci U S A* **104**(50): 19942-19947.

Lenstra JA, Groeneveld LF, Eding H, Kantanen J, Williams JL, Taberlet P, Nicolazzi EL, Solkner J, Simianer H, Ciani E et al. 2012. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. *Anim Genet* **43**(5): 483-502.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**(7): 337-340.

Leroy G. 2011. Genetic diversity, inbreeding and breeding practices in dogs: results from pedigree analyses. *Vet J* **189**(2): 177-182.

Leutenegger AL, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* **73**(3): 516-523.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**(7357): 493-496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866): 1100-1104.

Liu YG, Xia XH. 2011. Molecular characterization, polymorphism and association with reproductive traits of porcine CDK20 gene. *Journal of Animal and Feed Sciences* **20**(4): 566-574.

Loftus SK, Larson DM, Baxter LL, Antonellis A, Chen Y, Wu X, Jiang Y, Bittner M, Hammer JA, Pavan WJ. 2002. Mutation of melanosome protein RAB38 in chocolate mice. *Proceedings of the National Academy of Sciences* **99**(7): 4471-4476.

Lohmueller KE, Albrechtsen A, Li YR, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N et al. 2011. Natural Selection Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the Human Genome. *Plos Genetics* **7**(10).

Long JC, Kittles RA. 2003. Human genetic diversity and the nonexistence of biological races. *Human Biology* **75**(4): 449-471.

Lopez S, Rousset F, Shaw FH, Shaw RG, Ronce O. 2009. Joint Effects of Inbreeding and Local Adaptation on the Evolution of Genetic Load after Fragmentation. *Conservation Biology* **23**(6): 1618-1627.

Lynch M. 1991. The Genetic Interpretation of Inbreeding Depression and Outbreeding Depression. *Evolution* **45**(3): 622-629.

Lynch M -. 2010. Evolution of the mutation rate. *Trends Genet* **26**(8): 345-352.

Lynch M, Conery J, Burger R. 1995. Mutation Accumulation and the Extinction of Small Populations. *Am Nat* **146**(4): 489-518.

MacLeod IM, Hayes BJ, Goddard ME. 2014. The Effects of Demography and Long Term Selection on the Accuracy of Genomic Prediction with Sequence Data. *Genetics*.

MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, Goddard ME. 2013. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Mol Biol Evol* **30**(9): 2209-2223.

MacLeod IM, Meuwissen THE, Hayes BJ, Goddard ME. 2009. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics Research* **91**(6): 413-426.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**(16): 3448-3449.

Maine GN, Burstein E. 2007. COMMD proteins: COMMing to the scene. *Cellular and molecular life sciences* **64**(15): 1997-2005.

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A et al. 2008. Runs of Homozygosity in European Populations (vol 83, pg no 359, 2008). *Am J Hum Genet* **83**(5): 658-658.

Meagher TR, Charlesworth D. 2003. Effects of inbreeding on the genetic diversity of populations - Discussion. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **358**(1434): 1070-1070.

Megens HJ, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, Vereijken A, Silva P, Muir WM, Cheng HH et al. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. *Bmc Genetics* **10**.

Megens HJ, Crooijmans RPMA, Cristobal MS, Hui X, Li N, Groenen MAM. 2008. Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genetics Selection Evolution* **40**(1): 103-128.

Megens HJ, Groenen MAM. 2012. Domestication: a long-term genetic experiment Domesticated species form a treasure-trove for molecular characterization of Mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. *Heredity* **109**(1): 1-3.

Meijaard E dHJ, Oliver WLR 2011. In*: Handbook of the Mammals of the World Vol 2*, (ed. MR Wilson DE), pp. 248-291. Lynx Edicions, Barcelona.

Merks J, Mathur P, Knol E. 2012. New phenotypes for new breeding goals in pigs. *animal* **6**(04): 535-543.

Merks JW. 2000. One century of genetic changes in pigs and the future needs. *BSAS occasional publication*: 8-19.

Meuwissen T, Hayes B, Goddard M. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819-1829.

Meuwissen TH. 1997. Maximizing the response of selection with a predefined rate of inbreeding. *J Anim Sci* **75**(4): 934-940.

Miao YW, Peng MS, Wu GS, Ouyang YN, Yang ZY, Yu N, Liang JP, Pianchou G, Beja-Pereira A, Mitra B et al. 2013. Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity* **110**(3): 277-282.

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao FQ, Kim HL, Burhans RC, Drautz DI, Wittekindt NE et al. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *P Natl Acad Sci USA* **109**(36): E2382-E2390.

Mona S, Randi E, Tommaseo-Ponzetta M. 2007. Evolutionary history of the genus Sus inferred from cytochrome b sequences. *Molecular Phylogenetics and Evolution* **45**(2): 757-762.

Mukai T, Chigusa SI, Mettler LE, Crow JF. 1972. Mutation rate and dominance of genes affecting viability in Drosophila melanogaster. *Genetics* **72**(2): 335-355.

Mulder HA LM, Strandén I, Mäntysaari EA, Pool MH, Veerkamp RF. 2012. *MiXBLUP Manual*. Animal Breeding and Genomics Centre, Lelystad, The Netherlands.

Murgiano L, D'Alessandro A, Egidi MG, Crisa A, Prosperini G, Timperio AM, Valentini A, Zolla L. 2010. Proteomics and transcriptomics investigation on longissimus muscles in Large White and Casertana pig breeds. *Journal of proteome research* **9**(12): 6450-6466.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746): 321-324.

Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Launer L, Hardy J, Singleton AB. 2009. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10**(3): 183-190.

Nei M, Li WH. 1979. Mathematical-Model for Studying Genetic-Variation in Terms of Restriction Endonucleases. *P Natl Acad Sci USA* **76**(10): 5269-5273.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res* **11**(5): 863-874.

Nie HS, Crooijmans RPMA, Lammers A, van Schothorst EM, Keijer J, Neerincx PBT, Leunissen JAM, Megens HJ, Groenen MAM. 2010. Gene Expression in Chicken Reveals Correlation with Structural Genomic Features and Conserved Patterns of Transcription in the Terrestrial Vertebrates. *Plos One* **5**(8).

Nosil P. 2008. Speciation with gene flow could be common. *Mol Ecol* **17**(9): 2103-2106.

Nosil P, Feder JL. 2012. Genomic divergence during speciation: causes and consequences Introduction. *Philos T R Soc B* **367**(1587): 332-342.

Nothnagel M, Lu TT, Kayser M, Krawczak M. 2010. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Human Molecular Genetics* **19**(15): 2927-2935.

Ojeda A, Huang LS, Ren J, Angiolillo A, Cho IC, Soto H, Lemus-Flores C, Makuza SM, Folch JM, Perez-Enciso M. 2008. Selection in the making: A worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2. *Genetics* **178**(3): 1639-1652.

Okumura N, Matsumoto T, Hamasima N, Awata T. 2008. Single nucleotide polymorphisms of the KIT and KITLG genes in pigs. *Animal Science Journal* **79**(3): 303-313.

Oliver W. 2008. Sus cebifrons. In IUCN 2013 IUCN Red List of Threatened Species Version 20131.

Onteru SK, Fan B, Du ZQ, Garrick DJ, Stalder KJ, Rothschild MF. 2012. A whole-genome association study for pig reproductive traits. *Animal Genetics* **43**(1): 18-26.

Orozco-Terwengel PA, Bruford MW. 2014. Mixed signals from hybrid genomes. *Mol Ecol* **23**(16): 3941-3943.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW. 2010. Conservation genetics in transition to conservation genomics. *Trends Genet* **26**(4): 177-187.

Paabo S. 2003. The mosaic that is our genome. *Nature* **421**(6921): 409-412.

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History (vol 91, pg 809, 2012). *Am J Hum Genet* **91**(6): 1150-1150.

Patel S, Schell T, Eifert C, Feldmeyer B and Pfenninger M. 2015. Characterizing a hybrid zone between a cryptic species pair of freshwater snails Mol. Ecol. In press. DOI: 10.1111/mec.13049

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**(7097): 1103-1108.

Paudel Y, Madsen O, Megens HJ, Frantz LA, Bosse M, Bastiaansen JW, Crooijmans RP, Groenen MA. 2013. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics* **14**: 449.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic Patterns of Homozygosity in Worldwide Human Populations. *Am J Hum Genet* **91**(2): 275-292.

Pocar P, Fischer B, Klonisch T, Hombach-Klonisch S. 2005. Molecular interactions of the aryl hydrocarbon receptor and its biological and toxicological relevance for reproduction. *Reproduction* **129**(4): 379-389.

Porter V. 1993. Pigs – a Handbook to the Breeds of the World. Helm Information, Mountfield, East Sussex.

Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* **11**(11): 800-805.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G et al. 2013. Great ape genetic diversity and population history. *Nature* **499**(7459): 471-475.

Price MV, Waser NM. 1979. Pollen Dispersal and Optimal Outcrossing in Delphinium-Nelsoni. *Nature* **277**(5694): 294-297.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**(2): 945-959.

Pryce JE, Hayes BJ, Goddard ME. 2012. Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *Journal of Dairy Science* **95**(1): 377-388.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Purfield DC, Berry DP, McParland S, Bradley DG. 2012. Runs of homozygosity and population history in cattle. *Bmc Genetics* **13**.

Ralls K, Ballou JD, Rideout BA, Frankham R. 2000. Genetic management of chondrodystrophy in California condors. *Anim Conserv* **3**: 145-153.

Ralph P, Coop G. 2013. The Geography of Recent Genetic Ancestry across Europe. *Plos Biol* **11**(5).

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *P Natl Acad Sci USA* **102**(44): 15942-15947.

Ramirez O, Ojeda A, Tomas A, Gallardo D, Huang LS, Folch JM, Clop A, Sanchez A, Badaoui B, Hanotte O et al. 2009. Integrating Y-Chromosome, Mitochondrial, and Autosomal Data to Analyze the Origin of Pig Breeds. *Mol Biol Evol* **26**(9): 2061-2072.

Ramírez O, Quintanilla R, Varona L, Gallardo D, Díaz I, Pena R, Amills M. 2014. DECR1 and ME1 genotypes are associated with lipid composition traits in Duroc pigs. *Journal of Animal Breeding and Genetics* **131**(1): 46-52.

Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P et al. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *Plos One* **4**(8): e6524.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R et al. 2001. Linkage disequilibrium in the human genome. *Nature* **411**(6834): 199-204.

Rheindt FE, Edwards SV. 2011. Genetic introgression: an integral but neglected component of speciation in birds. *The Auk* **128**(4): 620-632.

Rohrer GA, Alexander LJ, Hu ZL, Smith TPL, Keele JW, Beattie CW. 1996. A comprehensive map of the porcine genome. *Genome Res* **6**(5): 371-391.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**(7526): 261-U243.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12): 1572-1574.

Ropka-Molik K, Bereta A, Tyra M, Różycki M, Piórkowska K, Szyndler-Nędza M, Szmatoła T. 2014. Association of calpastatin gene polymorphisms and meat quality traits in pig. *Meat science* **97**(2): 143-150.

Rousset F. 2008. GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**(1): 103-106.

Rubin CJ, Megens HJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg O, Jern P, Jorgensen CB et al. 2012. Strong signatures of selection in the domestic pig genome. *P Natl Acad Sci USA* **109**(48): 19529-19536.

Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**(7288): 587-U145.

Ruegg K, Anderson EC, Boone J, Pouls J, Smith TB. 2014. A role for migration-linked genes and genomic islands in divergence of a songbird. *Mol Ecol* **23**(19): 4757-4769.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**(6909): 832-837.

SanCristobal M, Chevalet C, Haley CS, Joosten R, Rattink AP, Harlizius B, Groenen MAM, Amigues Y, Boscher MY, Russell G et al. 2006. Genetic diversity within and between European pig breeds using microsatellite markers. *Animal Genetics* **37**(3): 189-198.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**(7492): 354-+.

Scandura M, Iacolina L, Crestanello B, Pecchioli E, Di Benedetto MF, Russo V, Davoli R, Apollonio M, Bertorelle G. 2008. Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: are the effects of the last glaciation still detectable? *Mol Ecol* **17**(7): 1745-1762.

Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, Rimbault M, Decker B, Kidd JM, Sood R. 2012. Variation of BMP3 contributes to dog breed skull diversity. *PLoS genetics* **8**(8): e1002849.

Schwenk K, Brede N, Streit B. 2008. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philos T R Soc B* **363**(1505): 2805-2811.

Seo H, Kim M, Choi Y, Ka H. 2011. Salivary Lipocalin Is Uniquely Expressed in the Uterine Endometrial Glands at the Time of Conceptus Implantation and Induced by Interleukin 1Beta in Pigs. *Biology of Reproduction* **84**(2): 279-287.

Shaffer ML. 1981. Minimum Population Sizes for Species Conservation. *Bioscience* **31**(2): 131-134.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**(11): 2498-2504.

Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J. 2006. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *Plos Biol* **4**(12): 2227-2237.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH. 2008. Genomic selection using different marker types and densities. *J Anim Sci* **86**(10): 2447-2454.

Sonesson AK, Woolliams JA, Meuwissen TH. 2012. Genomic selection requires genomic control of inbreeding. *Genet Sel Evol* **44**: 27.

Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, Nachman MW, Kohn MH. 2011. Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Current Biology* **21**(15): 1296-1301.

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. 2012. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). *Plos Genetics* **8**(8).

Stratz P, Wellmann R, Preuss S, Wimmers K, Bennewitz J. 2014. Genome-wide association analysis for growth, muscularity and meat quality in Pietrain pigs. *Anim Genet* **45**(3): 350-356.

Sun HFS, Ernst CW, Yerle M, Pinton P, Rothschild MF, Chardon P, Rogel-Gaillard C, Tuggle CK. 1999. Human chromosome 3 and pig chromosome 13 show complete synteny conservation but extensive gene-order differences. *Cytogenetics and Cell Genetics* **85**(3-4): 273-278.

Sweeney MT, Thomson MJ, Cho YG, Park YJ, Williamson SH, Bustamante CD, McCouch SR. 2007. Global dissemination of a single mutation conferring white pericarp in rice. *Plos Genetics* **3**(8): 1418-1424.

Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zollner S, Rosenberg NA, Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet* **93**(1): 90-102.

Taberlet P, Valentini A, Rezaei HR, Naderi S, Pompanon F, Negrini R, Ajmone-Marsan P. 2008. Are cattle, sheep, and goats endangered species? *Mol Ecol* **17**(1): 275-284.

Tachibana M, Amato P, Sparman M, Woodward J, Sanchis DM, Ma H, Gutierrez NM, Tippner-Hedges R, Kang E, Lee HS et al. 2013. Towards germline gene therapy of inherited mitochondrial diseases. *Nature* **493**(7434): 627-631.

Tajima F. 1983. Evolutionary Relationship of DNA-Sequences in Finite Populations. *Genetics* **105**(2): 437-460.

Takada T, Ebata T, Noguchi H, Keane TM, Adams DJ, Narita T, Shin-I T, Fujisawa H, Toyoda A, Abe K et al. 2013. The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. *Genome Res* **23**(8): 1329-1338.

Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *Plos Biol* **5**(7): 1587-1602.

Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW, Tucker PK. 2010. The Variable Genomic Architecture of Isolation between Hybridizing Species of House Mice. *Evolution* **64**(2): 472-485.

Thomas MA, Roemer GW, Donlan CJ, Dickson BG, Matocq M, Malaney J. 2013. Gene tweaking for conservation. *Nature* **501**(7468): 485-486.

Tier B, Meyer K. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *Journal of Animal Breeding and Genetics* **121**(2): 77-89.

Toro MA, Villanueva B, Fernadez J. 2014. Genomics applied to management strategies in conservation programmes. *Livestock Science* **166**: 48-53.

Tortereau F, Servin B, Frantz L, Megens HJ, Milan D, Rohrer G, Wiedmann R, Beever J, Archibald AL, Schook LB 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.

Vidal O, Varona L, Oliver M, Noguera J, Sanchez A, Amills M. 2006. Malic enzyme 1 genotype is associated with backfat thickness and meat quality traits in pigs. *Animal genetics* **37**(1): 28-32.

Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Robinson M, Lawrence J, Anjorin A, Sklar P, Gurling HMD et al. 2009. No evidence for excess runs of homozygosity in bipolar disorder. *Psychiatric Genetics* **19**(4): 165-170.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *Plos Biol* **4**(3): e72.

vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jedrzejewski W, Jedrzejewska B et al. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Res* **21**(8): 1294-1305.

vonHoldt BM, Pollinger JP, Lohmueller KE, Han EJ, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A et al. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**(7290): 898-U109.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**(16): e164.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population-Structure. *Evolution* **38**(6): 1358-1370.

Wernersson R, Schierup MH, Jorgensen FG, Gorodkin J, Panitz F, Staerfeldt HH, Christensen OF, Mailund T, Hornshoj H, Klein A et al. 2005. Pigs in sequence space: A 0.66X coverage pig genome survey based on shotgun sequencing. *Bmc Genomics* **6**.

White S. 2011. From globalized pig breeds to capitalist pigs: a study in animal cultures and evolutionary history. *Environmental History* **16**(1): 94-120.

Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA. 2015. Genetic rescue to the rescue. In press. doi:10.1016/j.tree.2014.10.009

Wilkins AS, Wrangham RW, Fitch WT. 2014. The 'Domestication Syndrome' in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics (vol 197, pg 795, 2014). *Genetics* **198**(4): 1771-1771.

Wilkinson S, Lu ZH, Megens HJ, Archibald AL, Haley C, Jackson IJ, Groenen MAM, Crooijmans RPMA, Ogden R, Wiener P. 2013. Signatures of Diversifying Selection in European Pig Breeds. *Plos Genetics* **9**(4).

Witzenberger KA, Hochkirch A. 2011. Ex situ conservation genetics: a review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodiversity and Conservation* **20**(9): 1843-1861.

Wright S. 1921. Systems of Mating. II. the Effects of Inbreeding on the Genetic Composition of a Population. Genetics 6: 1240143.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J et al. 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat Biotechnol* **32**(7): 656-+.

Yong WJ, Jing L, Jiugang Z, Lei C, Yonggang L. 2012. A novel porcine gene, POT1, differentially expressed in the longissimus muscle tissues from Wujin and Large White pigs. *Cytokine* **59**(1): 22-26.

Summary

Summary

The genome is a mosaic, consisting of a collection of DNA segments that represent different ancestries. The work described in this thesis contributes to our understanding of the factors influencing genomic variation, using the pigs as a model. I analyzed re-sequence and genotype data from hundreds of pigs and wild boars to answer questions about the underlying mechanisms that shape genomic variation. With this thesis I provide a comprehensive overview of the effects of demography, hybridization and selection on diversity patterns in the genome.

In **chapter 2** I demonstrate that regions of homozygosity in individual genomes are shaped mainly by (local) recombination frequency and demographic history of the population. Also selection for particular haplotypes can result in Regions Of Homozygosity (ROH), often also visible in the population as a selective sweep. ROHs are more frequent in European pigs compared to Asian pigs, especially in European wild populations. We demonstrate that overall heterozygosity in a genome is not a direct predictor of ROH occurrence, since they refer to different time points in demographic history. The fundamental insight gained is, that effects of inbreeding are contingent on location in the genome. These findings underline the importance of not only the population dynamics, but also the genome architecture, of managed and wild populations, to understand inbreeding patterns.

Chapter 3 discusses the consequences of introgression of Asian haplotypes into European domesticated pig lines. I provide a detailed analysis of the influence of human induced hybridization in commercial pigs. The Asian haplotypes generally increase local nucleotide diversity in the genome, resulting in domesticated pigs with a higher level of variation than the source population that they were domesticated from (European wild boars). Length and abundance of these introgressed haplotypes correlate negatively with recombination rate, similar to the findings for ROHs in chapter 2. I also elaborate on the possible bias that can occur when analyzing the demographic history of a population based on hybrid genomes.

Chapter 4 provides an in-depth analysis of the patterns of variation on chromosome 1, with special emphasis on the contribution of Asian introgressed haplotypes to the level of variation in European commercial pigs. In this paper we use haplotype homozygosity on pig chromosome 1 as a model for genetic diversity in pigs. In this chapter I conclude that most variation within the species is captured by the Asian pigs and wild boars. I also show that it is highly unlikely that concurrent hybridization between European populations leads to the high level of variation that is observed in commercial breeds, and conclude that the Asian

introgression rather than mixing European wild populations causes the higher genetic diversity in European commercial pigs.

Adaptation, by means of hybridization, may play a major role in evolution. In **chapter 5** I demonstrate that adaptive introgression is not restricted to natural settings. I provide an example of human induced "adaptive" hybridization to improve a domestic species. I show that the introgression landscape is highly heterogeneous in the commercial Large White population and some regions in the genome are generally more Asian than European, suggesting selection for the Asian haplotypes. I further find a significant effect of the Asian haplotypes on litter size for the introgressed region at the *AHR* locus, supporting this hypothesis.

In **chapter 6** I explore how much the introgression landscape has been influenced by selection for commercial traits. The majority of regions that contain an excess of Asian haplotypes has an effect on backfat thickness in the Large White population. Moreover, I demonstrate an additive effect of Asian haplotypes on backfat over these introgressed regions. Together with almost no correlation between introgression signal and recombination frequency and gene content, this finding suggests that the persistence of introgressed haplotypes is mainly determined by selection.

In **chapter 7** I utilize the distribution of haplotypes in a commercial Pietrain population and a zoo population of the endangered *Sus cebifrons* to optimally manage each population. In this study I introduce a new concept in population management by implementing information derived from next-generation sequencing into breeding programs. I use the prediction of putative deleterious variants as a proxy for fitness and measure coancestry based on identity-by-descent (IBD) segments in the genome. I demonstrate that coancestry measures based on molecular markers outperform pedigree-based measurements, but also that optimizing variation does not necessarily result in the highest fitness. On top of that, these management strategies may influence ongoing selection by reducing the frequency of selected haplotypes.

Finally, in **chapter 8** I place my findings in a broader perspective. I discuss how fine-scale haplotype patterns in genomes are a treasure-trove of information about the demographic and selective history of the population. I discuss different populations under management, which types of genetic characterization we can apply and what the feasibility is of implementing genomics into the management plan. I conclude with some future perspectives on genomics in conservation. My findings demonstrate that a single genome contains information on the demographic history of a population, from ancient bottlenecks till recent inbreeding, hybridization and selection.

Samenvatting

Samenvatting

Het genoom is een mozaïek, bestaande uit een collectie van DNA-segmenten die verschillende herkomsten representeren. Het werk dat in deze thesis beschreven is, draagt bij aan ons begrip van de factoren die genomische variatie beïnvloeden, met het varken als modelorganisme. Ik heb re-sequence en genotype data geanalyseerd van honderden varkens en wilde zwijnen om vragen te beantwoorden over de onderliggende mechanismen die genomische variatie vormgeven.

In **hoofdstuk 2** toon ik aan dat regio's van homozygositeit (ROH) in individuele genomen voornamelijk bepaald worden door (lokale) recombinatie frequentie en demografische geschiedenis van de populatie. Ook selectie voor bepaalde haplotypen kan resulteren in ROHs, vaak ook zichtbaar in de populatie als een 'selective sweep'. Vergeleken met Aziatische varkens zijn ROHs frequenter aanwezig in Europese varkens, voornamelijk in Europese wilde populaties. We laten zien dat globale heterozygositeit in een genoom geen directe voorspeller is van het voorkomen van ROHs, aangezien ze refereren naar verschillende tijdspunten in de demografische geschiedenis. Het fundamentele inzicht dat hier is verkregen is dat de effecten van inteelt afhangen van de locatie in het genoom. Deze vindingen onderstrepen niet alleen het belang van de populatiedynamiek, maar ook de genetische architectuur van gemanagede en wilde populaties, om patronen van inteelt te begrijpen.

Hoofdstuk 3 bediscussieert de consequenties van introgressie van Aziatische haplotypen in Europese gedomesticeerde varkenslijnen. Ik verstrek een gedetailleerde analyse van de invloed van humaan-geïnduceerde hybridisatie in commerciële varkens. In het algemeen verhogen de Aziatische haplotypen de lokale nucleotidevariatie in het genoom, wat resulteert in gedomesticeerde varkens met een hoger niveau van variatie dan de oorspronkelijke populatie waaruit ze gedomesticeerd zijn (Europese wilde zwijnen). De lengte en hoeveelheid van deze ingekruiste haplotypen correleren negatief met de recombinatiefrequentie, vergelijkbaar met de bevindingen voor ROHs in hoofdstuk 2. Ook wijd ik uit over de mogelijke bias die kan ontstaan wanneer de demografische geschiedenis van een populatie bepaald wordt op basis van hybride genomen.

Hoofdstuk 4 verschaft een diepteanalyse van de patronen van variatie op chromosoom 1, met speciale nadruk op de bijdrage van Aziatische ingekruiste haplotypen aan de niveaus van variatie in Europese commerciële varkens. In dit artikel gebruiken we haplotype homozygisiteit op het chromosoom 1 van varken als

een model voor algemene genetische variabiliteit in varkens. Ik concludeer in dit hoofdstuk dat de meeste variatie binnen de soort gevat wordt door de Aziatische varkens en wilde zwijnen. Ik toon eveneens aan dat het hoogst onwaarschijnlijk is dat herhaaldelijke hybridisatie tussen Europese populaties geleid kan hebben tot het hoge niveau van variatie dat geobserveerd is in commerciële varkensrassen. Hierbij concludeer ik dat voornamelijk de Aziatische introgressie ten opzichte van het mixen van Europese wilde populaties de hogere genetische diversiteit in Europese commerciële varkens heeft veroorzaakt.

Adaptatie, aan de hand van hybridisatie, zou een grote rol kunnen spelen in evolutie. In **hoofdstuk 5** toon ik aan dat adaptieve introgressie niet beperkt is tot natuurlijke omstandigheden. Ik geef een voorbeeld van door mensen teweeggebrachte 'adaptieve' hybridisatie om een gedomesticeerde soort te verbeteren. Ik laat zien dat het introgressielandschap uiterst heterogeen is in de commerciële Large White populatie, en dat sommige regio's in het genoom meer Aziatisch dan Europees zijn, wat selectie voor de Aziatische haplotypen suggereert. Verder vind ik een significant effect van de Aziatische haplotypen op de worpgrootte voor de introgressieregio op het AHR locus, wat deze hypothese ondersteunt.

In **hoofdstuk 6** onderzoek ik hoeveel het introgressie landschap in Europese varkens beïnvloed wordt door selectie voor commerciële kenmerken. Het merendeel van regio's die een overvloed aan Aziatische haplotypes hebben, heeft een effect op de dikte van rugspek in de Large White populatie. Bovendien demonstreer ik een additief effect van de Aziatische haplotypen op rugspek over al deze geïntegreerde regio's heen. Tezamen met haast geen correlatie tussen het introgressiesignaal en de recombinatiefrequentie en gendichtheid, suggereren deze resultaten dat de persistentie van geïntroduceerde haplotypen voornamelijk door selectie bepaald wordt.

In **hoofdstuk 7** gebruik ik de distributie van haplotypen in een commerciële Pietrainpopulatie en in een dierentuinpopulatie van het zwaar bedreigde *Sus cebifrons* zwijn om elke populatie optimaal te beheren. In dit onderzoek introduceer ik een nieuw concept in populatiemanagement door middel van de implementatie van informatie afkomstig van next-genetarion sequencing in fokprogramma's. Ik gebruik de voorspelling van vermeende schadelijke mutaties als een proxy voor fitness en ik meet gemeenschappelijke afstamming op basis van identity-by-descent (IBD) segmenten in het genoom. Ik toon aan dat metingen van gemeenschappelijke afstamming gebaseerd op moleculaire merkers nauwkeuriger zijn dan metingen op basis van stamboek gegevens, maar ook dat het optimaliseren van variatie niet noodzakelijkerwijs leidt tot de hoogste fitness. Deze

managementstrategieën kunnen bovendien mogelijk voortdurende selectie beïnvloeden doordat ze de frequentie van geselecteerde haplotypen reduceren.
Tenslotte plaats ik mijn resultaten in **hoofdstuk 8** in een breder perspectief. Ik bediscussieer hoe haplotype patronen op fijne schaal in het genoom een schat aan informatie bevatten over de demografische en selectie geschiedenis van een populatie. Ik behandel verschillende populaties onder management, welk type genetische karakterisering we hierop kunnen toepassen en wat de haalbaarheid is van het implementeren van genomics in het managementplan. Ik sluit af met een aantal toekomstperspectieven op het gebruik van genomics in natuurbehoud. Mijn bevindingen demonstreren dat een enkel genoom informatie bevat over de demografische geschiedenis van een populatie, van oeroude bottlenecks tot recente inteelt, hybridisatie en selectie.

Acknowledgements

Acknowledgements

Beste Martien. Yogesh, Laurent en ik hebben vaak tegen elkaar gezegd hoeveel geluk wij hebben gehad om onder zo'n fantastische inspirerende, gedreven en georganiseerde wetenschapper te mogen werken. We hadden het niet beter kunnen treffen. Bedankt voor de persoonlijke en wetenschappelijke begeleiding de afgelopen jaren, en dat ik aan dit geweldige project heb mogen deelnemen.

Hendrik-Jan, eigenlijk past kort en bondig mijn dank naar jou uitdrukken niet bij mij, laat staan bij jou.... :p Bedankt voor alle discussies, dat je altijd in mij bent blijven geloven en vooral voor jouw visie op genetica. Ik heb het enorm getroffen met een supervisor waaronder ik mijn eigen ideeën op deze manier heb kunnen ontwikkelen.

Dear Ole, grote smurf. Thank you for all the scientific guidance and personal support over the last 4 years. You always made me think about my research from a different perspective. I admire your personal style in science and your positive attitude in life. Oh, and finally... thank you for unplugging Laurents computer.

Laurent, thank you for all the great stuff we did together. Your scientific knowledge and drive in combination with your personality make you an excellent scientist who is not easily forgotten. You always push me a little bit further, in science as well as in bars, which is a gift that not many people have ;)

Yogesh, hoe gaat het? Thanks for your patience with teaching me the Linux basics and for all the great work we have accomplished. Thank you for our beautiful Wageningen journey together – from talking about fundamental biology and heredity via dictionaries and perl/python discussions to the more philosophical aspects of life – like one-sips.

Dear Juanma, I admire how you've grown over the last few years – and I obviously refer to your scientific career as well as your tennis skills ;) Thanks for our smooth collaboration and all valuable and nonsense discussions we had, and for adding your joy in life and science to the group.

Richard, Bert en Kimberley, bedankt voor het aanbrengen van structuur in de zee van samples, alle hulp bij analyses en sneeuwtjes en dergelijke.. dit werk was onmogelijk geweest zonder jullie.

Dat geldt ook absoluut voor het werk van de ladies van ons weergaloze secretariaat. Bedankt voor de ontelbare details en grotere uitdagingen die jullie zo moeiteloos signaleren en oplossen nog voordat ik überhaupt door heb dat er iets aan de hand is. Bedankt ook voor de warmte en menselijkheid die jullie toevoegen aan de groep.

Thank you all my molecular and quantitative genomics colleagues for providing this stimulating environment. I feel privileged that I have had the opportunity to be part of such a high-quality, vibrant research group.

Thanks PhD-group for sharing this difficult course of life with me. My roomies in the old and new Zodiac and now in Radix, thanks for the (mostly..) friendly faces in the morning, switching on the heater and all the stimulating discussions. Thanks all my ABG friends for making my life a bit easier by sharing supervisor frustrations (not you, Ole and H-J, of course.. ;)) organising dinners, painting nails, sharing cultures, sharing homes, having babies, going to festivals, playing sports and your overall friendship.

Dear pubquiz group, thank you for making me humble – how many times have I felt extremely stupid during the quiz when we did not know 1 answer for sure... Thanks for sharing this stupidity with me during the pleasure of some good beers.

Cher groupe de Paris, merci de m'inviter chez votre magnifique musée. My time in Paris was an unique experience and our collaboration made me into the independent researcher I am today.

Bedankt lieve tennisteamies, voor het zijn van de onverwachte stabiele factor in mijn leven de afgelopen ~ 10 jaar. Bedankt voor de nutteloze discussies over rare combinaties op de baan, en winning ugly. En natuurlijk voor de keiharde afspraken voor dit jaar: begeren, consumeren, solliciteren, presteren en....promoveren!

Bedankt leukste, beste, mooiste hockeyteam van Myra voor het accepteren van mijn afwezigheden op de training, het luisteren naar mijn knorrige verhalen en dat ik spits sta – een harde klap op doel (en soms een slap rolletje) heeft veel PhD-frustratie doen verdwijnen de afgelopen jaren!

Lieve oude vossen, bedankt dat jullie samen met mij de periode van mijn leven hebben doorlopen die me gevormd heeft als Vossiaan en persoon. Lieve maissies, ik ben ontzettend dankbaar voor een vriendschap die zo fundamenteel aanvoelt. Het is ongelooflijk hoe goed jullie mij kennen en kunnen steunen. Bedankt voor jullie gedeelde liefde voor de Beatles en jullie eeuwige enthousiasme en interesse voor wat ik doe.

Lieve Annie, Roos, Pax, Rijco, Rens, Silo, Thom en andere Gyrinusliefdes. Bij gebrek aan een echte Gyrinus-baby wil ik graag mijn knorwerk presenteren als een product van onze jarenlange ingewikkelde en bijzondere relatie. Dankzij onze ontelbare nachtelijke gesprekken, vergeten budelsdiscussies en alle fantastische ervaringen waar een (nog) ongeschreven boek gevuld mee kan worden ben ik de bioloog en persoon geworden om deze thesis te kunnen beginnen, en jullie vriendschap die stadsdeel- stads- land- en continents-grenzen overschrijdend is heeft geholpen hem af te maken.

Lieve opa en oma leeuw, ik ben enorm gelukkig met zulke levenslustige grootouders. Het is fantastisch om te zien hoe jullie moeiteloos meekomen zowel op wetenschappelijk niveau als met Lars en Rutgers voorliefde voor bepaalde lichaamsuitscheidingen. Bedankt voor jullie steun en liefde de afgelopen 32 jaar.

Lieve ConTRoL. Wat een enorme inspiratie zijn jullie voor mij geweest. Een voorbeeld nemen aan je grote broer is misschien cliché, maar ik bewonder enorm hoe jullie met z'n tweeën, drieën en zelfs vieren vol in het leven staan, en hoop daar altijd onderdeel van uit te blijven maken.

Joor en Har. Lief dynamisch duo. Ik heb de afgelopen jaren op veel verschillende plekken gezeten waar ik geprobeerd heb een thuis van te maken, maar mijn absolute basis en gevoel van thuiskomen is bij jullie. Bedankt voor alle steun, liefde en respect die altijd onvoorwaardelijk aanwezig is, en voor jullie geweldige levenshouding. Dankzij jullie heb ik de kracht om vast te houden aan mijn idealen en ben ik de optimistische Doch die jullie zo goed kennen.

Lieve Moes. Je kaartje hangt nog steeds boven m'n bureau: hup Doch, publiceeeern! Niemand kon en kan ooit zo enthousiast reageren op elk minisuccesje in m'n carrière als bioloog als jij. Aan niemand vertel ik liever wat ik nu weer heb gedaan als aan jou. Ik vraag me soms af waar ik de kracht vandaan heb gehaald om deze PhD zo te kunnen beëindigen. Dat weet ik nu. Dat komt door jou. Je bent en blijft mijn allergrootste steun, en bent onderdeel van mij geworden. Door jou ben ik wie ik nu ben. Bedankt daarvoor. Voor alles.

Lieve Bollie, bedankt dat je op mijn zeldzame Fanta-avond als enige keer tegen me gelogen hebt, anders had je hier misschien nooit gezeten ;) Ik kan nog steeds zo gelukkig worden als ik denk aan hoe wij stiekem tegelijk enorm verschillend en hetzelfde zijn. Je hebt me verrast met hoe ontzettend zorgzaam je kan zijn, hoe onvoorwaardelijk jouw steun en liefde voelt en hoe je weet wanneer je je onderbroek moet opsturen naar Bolivia om me op te vrolijken. Bedankt voor alle B. van de Water productions en dat je mijn enthousiasme voor de knorries met me gedeeld hebt, zowel op papier als de ontelbare keren dat we over hekjes zijn geklommen. Je bent preciieeess goed! Neufneuf.

Curriculum Vitae

About the author

Mirte Bosse was born on November 27th in Amstelveen, The Netherlands. After her graduation from the Vossius Gymnasium in Amsterdam, she spent some months volunteering in development work in South Africa. This experience strengthened her love for biology and she decided to obtain her BSc in Biology at the Vrije Universiteit Amsterdam. Societal relevance has always been an important motivation for her research. During her MSc Ecology at the VU she came into contact with DANTA: Association for Conservation of the Tropics for which she is now a board member. She deliberately has chosen a Masters education in which environmental changes are investigated on multiple levels because she believes in a multidisciplinary research environment. This has driven her through the Costa Rican and Bolivian jungle as well as through the extended climate chambers and molecular lab at the VU. During her MSc she became interested in the genomic approach of problems in ecology and evolution. Knowledge on fundamental questions in divergence, hybridization and variation (between, but also within species) can contribute to a better operation procedure when it comes to conservation efforts. In 2010 she joined the Animal Breeding and Genomics Centre in Wageningen to work as a PhD student on the ERC SelSweep project of Prof. Martien Groenen. Her PhD research focused on genomic applications in answering questions about (maintenance of) genetic diversity. The main goal of this research is to investigate genome-wide variation in wild boars and changes during domestication and artificial selection by commercial breeding in pigs (*Sus scrofa*). The pig is an excellent model for population genomics research. However, the fundamentals of this study exceed species boundaries and are applicable to a variety of (endangered) populations and species. Mirte is convinced that genomic tools will become of high importance for conservation efforts in the (near) future. She recently received a WIAS postdoctoral fellowship to write her own grant proposal so she can further develop her envisaged research line in conservation genomics.

Peer reviewed publications

1. **Bosse M**, Madsen O, Megens HJ, Frantz LAF, Paudel Y, Crooijmans RPMA and Groenen MAM (2015). Hybrid origin of European commercial pigs examined by an in-depth haplotype analysis on chromosome 1. Frontiers in Genetics 5:442
2. **Bosse M**, Megens HJ, Frantz LAF, Madsen O, Larson G, Paudel Y, Duijvesteijn N, Harlizius B, Hagemeijer Y, Crooijmans RPMA and Groenen MAM. (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature Communications 5:4392.
3. **Bosse M**, Megens HJ, Madsen O, Frantz LAF, Paudel Y, Crooijmans RPMA, and Groenen MAM. (2014) Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations. Molecular Ecology 23: 4089–4102.
4. **Bosse M**, Megens H-J, Madsen O, Paudel Y, Frantz LAF, Schook LB, Crooijmans RPMA and Groenen MAM. (2012) Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. PLoS Genetics 8(11): e1003100.
5. Frantz LAF., Madsen O, Megens H-J, **Bosse M**, Schraiber JG, Paudel Y, Crooijmans RPMA and Groenen MAM (2015) The evolution of Tibetan wild boars. Nature Genetics 47: 188-189.
6. Herrero-Medrano J, Megens HJ, Groenen MAM, **Bosse M,** Pérez-Enciso M and Crooijmans RPMA. (2014) Whole-genome sequence analysis reveals differences in population management and selection of European low-input breeds. BMC Genomics 15:601.
7. Herrero-Medrano J, Megens HJ, Groenen MAM, Ramis G, **Bosse M**, Pérez-Enciso M and Crooijmans RPMA. (2013) Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. BMC Genetics 14:106
8. Frantz LAF, Schraiber JG, Madsen O, Megens HJ, **Bosse M**, Paudel Y, Semiadi G, Meijaard E, Li N, Crooijmans RPMA, Archibald AL, Slatkin M, Schook LB, Larson G and Groenen MAM. (2013) Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biology 14:R107
9. Paudel Y, Madsen O, Megens HJ, Frantz LAF, **Bosse M**, Bastiaansen JWM, Crooijmans RPMA and Groenen MAM (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14:449

10. Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, Li S,Larkin DM, Kim H, Frantz LAF, Caccamo M, Ahn H, Aken BL, Anselmo A, Anthon C, Auvil L, Badaoui B, Beattie CW, Bendixen C, Berman D, Blecha F, Blomberg J, Bolund L, **Bosse M** et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–398

11. Nota B, **Bosse M**, Ylstra B, van Straalen NM and Roelofs D. (2009) "Transcriptomics reveals extensive inducible biotransformation in the soil-dwelling invertebrate Folsomia candida exposed to phenanthrene." BMC Genomics 10:236

12. **Bosse M**, Megens HJ, Madsen O, Crooijmans RPMA, Ryder OA, Austerlitz F, Groenen MAM, de Cara AMR (2015). Genomic Data in Population Management: Implications for Conservation and Selection Programmes. Submitted.

13. **Bosse M***, Lopes MS*, Madsen O, Megens HJ, Crooijmans RPMA, Frantz LAF, Harlizius B, Bastiaansen JWM, Groenen MAM (2015). Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. Submitted.

14. Paudel Y, Madsen O, Megens HJ, Frantz LAF, **Bosse M**, Crooijmans RPMA, Groenen MAM (2015) Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. BMC Genomics, in press.

Training and Supervision Plan

The Basic Package (2ECTS)	year	credits *
WIAS Introduction Course	2010	1.5
Scientific Exposure (16 ECTS)	year	credits
International conferences (5 ECTS)		
Benelux Congress of Zoology, Utrecht, The Netherlands 2nd-4th November 2011	2011	1.0
NBIC Integrative Bioinformatics, Wageningen, The Netherlands, March 21-23 2011	2011	0.9
SMBE Dublin, Ireland, 23rd to 26th June 2012	2012	1.2
ESEB Lisboa, Portugal, 19 to 24 August 2013	2013	1.2
SMBE San Juan, Puerto Rico, 8th-12th June 2014	2014	1.2
Seminars and workshops (1 ECTS)		
Genomics and Animal Breeding, Hof van Wageningen, 01-21-2011	2011	0.2
CBSG Technology Symposium "Advances in life-science Technologies"	2010	0.2
WIAS science day 2011, From DNA to daily practice, 02-03-2011	2011	0.3
Ensembl/Havana genome browser workshop, VU Amsterdam	2011	0.3
Presentations (10 ECTS)		
WIAS science day, Wageningen - poster	2012	1.0
Dutch annual ecology meeting NAEM, Lunteren – poster	2012	1.0
SMBE Dublin, Ireland – poster	2012	1.0
ESEB Lisboa, Portugal – poster	2013	1.0
Dutch annual ecology meeting NAEM, Lunteren – poster	2013	1.0
WIAS science day – poster (1st prize publication)	2013	1.0
PigBioDiv, Corsica, France – oral	2013	1.0
Symposium Dierentuin en wetenschap, Arnhem – oral	2014	1.0
SMBE, San Juan, Puerto Rico – poster	2014	1.0
Benelux Congress of Zoology, Liege, Belgium – oral	2014	1.0
In-Depth Studies (13 ECTS)	year	credits
Disciplinary and interdisciplinary courses (6 ECTS)		
Workshop on Molecular Evolution – University of Maryland, Colorado	2011	4.0
Livestock conservation genomics – University of Zagreb, Pag (travel Grant)	2012	2.0
PhD students' discussion groups (1 ECTS)		
PhD-paper discussion group, ABG	2011-12	1.0
MSc level courses (6 ECTS)		
Advanced Bioinformatics, WUR	2011	6.0
Professional Skills Support Courses (5 ECTS)	year	credits
Scientific Writing	2011	1.8
The essentials of writing and presenting a scientific paper	2013	1.2
Course Career Perspectives	2014	1.6
Research Skills Training (8 ECTS)	year	credits
Preparing own PhD research proposal	2011	6.0
External training period – Paris : Januari-April	2014	2.0

Didactic Skills Training (8 ECTS)	year	credits
Supervising practicals and excursions (3 ECTS)		
Genomics course – WUR	2011	1.0
Genomics course – WUR	2012	1.0
Genomics course – WUR	2013	1.0
Supervising theses (4 ECTS)		
Animal Breeding and Genomics Msc student Minor	2013	1.5
Animal Breeding and Genomics Msc student Major	2013	2.0
Tutorship (1 ECTS)		
RMC	2012	1.0

Management Skills Training (2 ECTS)	year	credits
Organisation of seminars and courses		
Wageningen Evolution and Ecology Seminars, organisation committee	2011/13	1.0
Membership of boards and committees		
Board of directors of DANTA, Association for Conservation of the Tropics	2011/14	1.0

Education and Training Total (54 ECTS)

Data availability and supplementary material

Data availability and supplementary material

All BAM files have been deposited in the European Nucleotide Archive (ENA) under the accession number ERP001813. Essential code for all chapters is available through the WUR Animal Breeding and Genomics Centre and can be accessed at http://git.wageningenur.nl/ABGC_Genomics/Hybrid_nature_of_pig_genomes.git.

The publications and supplementary materials for chapter 2,3,4 and 5 are available online through the journal websites:

Chapter 2:
http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003100
Chapter 3:
http://onlinelibrary.wiley.com/doi/10.1111/mec.12807/full
Chapter 4:
http://journal.frontiersin.org/Journal/10.3389/fgene.2014.00442/full
Chapter 5:
http://www.nature.com/ncomms/2014/140715/ncomms5392/full/ncomms5392.html

For completeness, the supplementary material for all chapters is also available in the supplements of this thesis.

Chapter 2

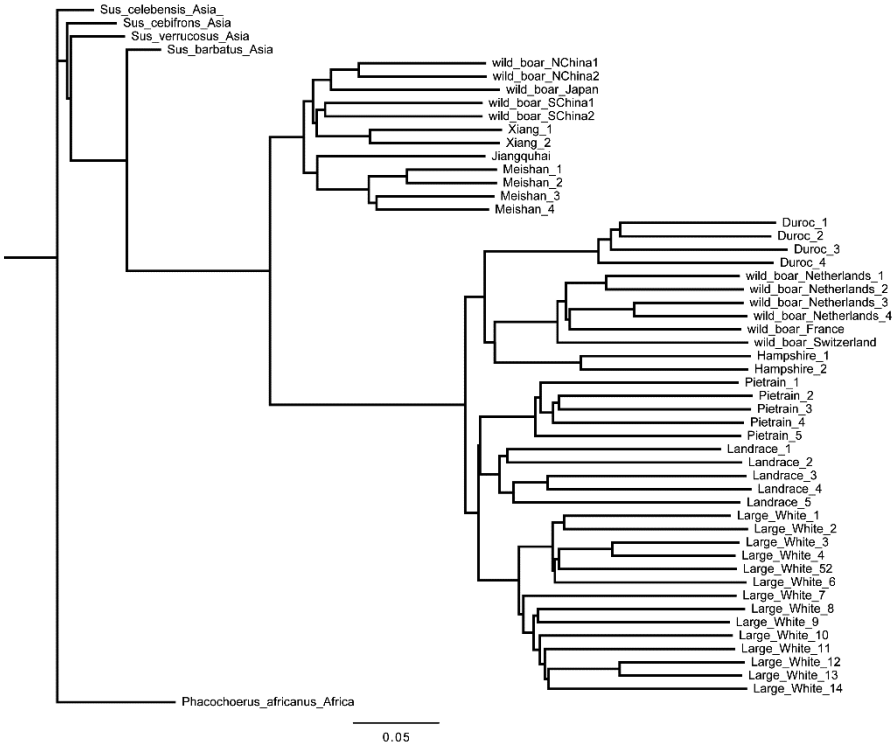


Figure S2.1 Phylogenetic tree for all 52 sequenced individuals. Distances are based on the genotypes on the Illumina Porcine 60K iSelect Beadchip. Three main clusters can be observed: The other *Sus* species originated from the South-East Asian Islands, The wild and domesticated Asian *Sus scrofa* and the European wild and domesticated *Sus scrofa*. Branch lengths may be affected because of the ascertainment bias introduced by the focus on variable sites in European pigs during SNP chip construction.

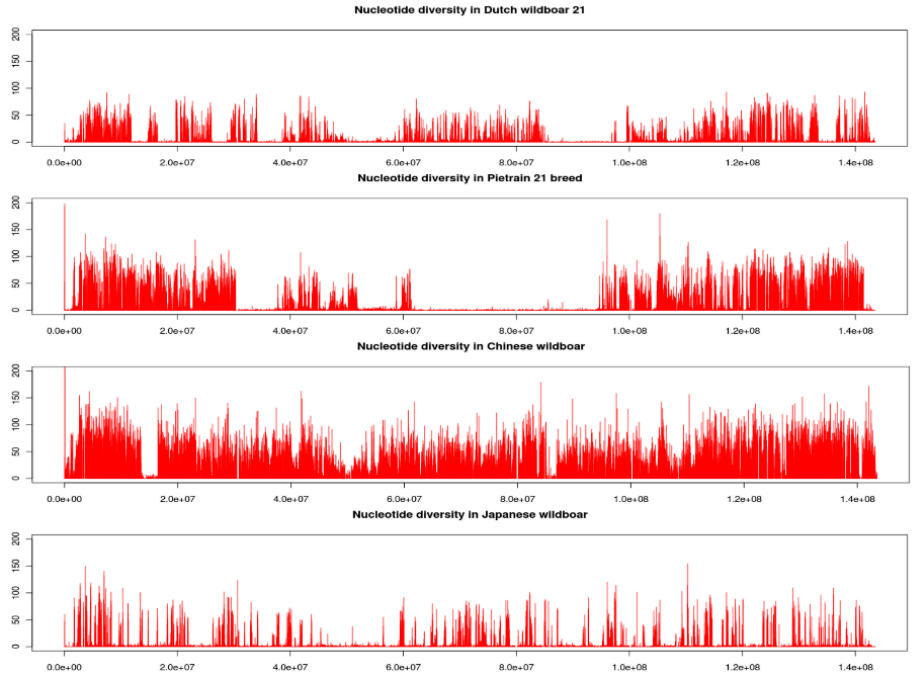


Figure S2.2 Distribution of nucleotide diversity over chromosome 1. The x-axis displays the physical position on the chromosome in bp and the y-axis shows the corrected number of SNPs that was called in bins of 10kbp. Data is shown for a Dutch wild boar from the Veluwe, for a pig from the European Pietrain breed, for a wild boar from North China and for a wild boar from a Japanese island.

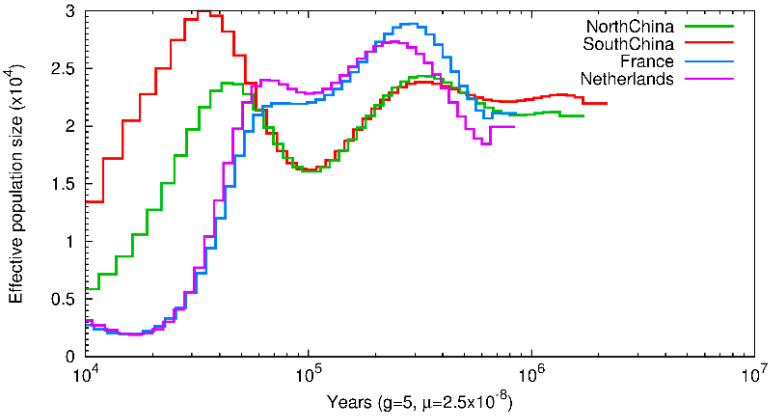


Figure S2.3 Estimation of demographic history and population size with the Pairwise Sequentially Markovian Coalescent (PSMC). The x-axis displays the years back in time, and the y-axis shows the estimated effective population size N. Data is shown for Two Asian wild boars from North (red) and South China (green) , and two European wild boars from the Netherlands (purple) and France (blue).

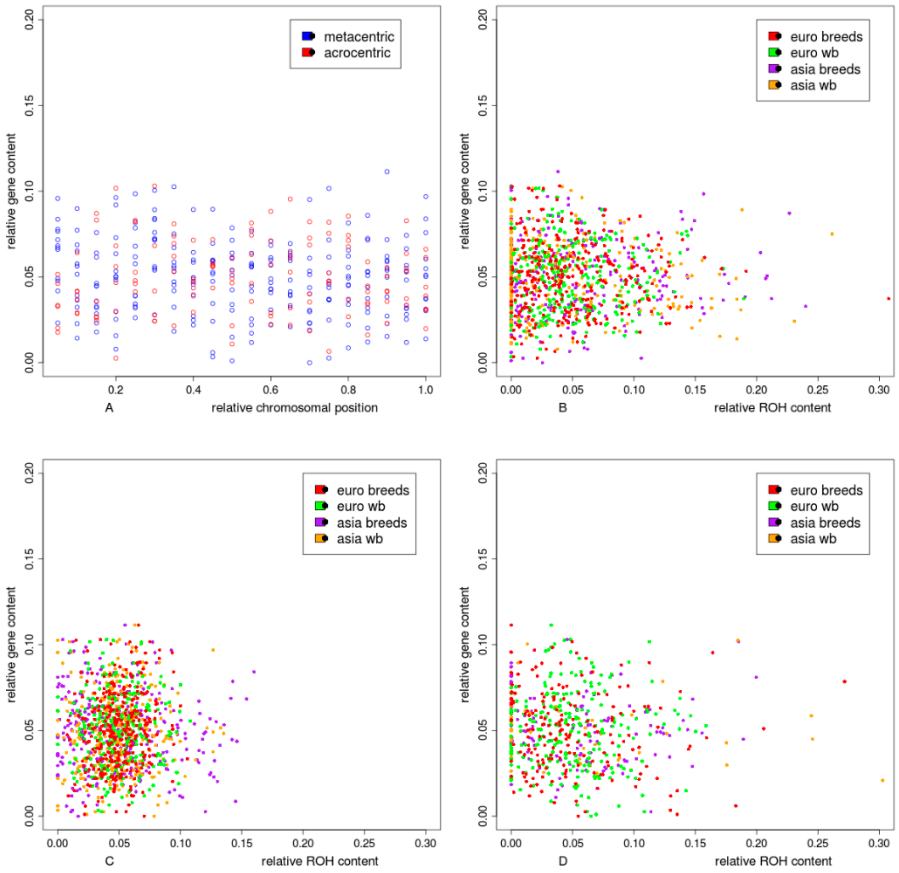


Figure S2.4 Correlation between genomic gene content and ROH frequency. **A**. Distribution of gene content over relative chromosomal position, plotted for all chromosomes separately. Metacentric chromosomes are displayed in blue and acrocentric chromosomes in red. Relative gene content plotted against ROH frequency for small (**B**), medium (**C**) and large size ROHs (**D**). ROH distribution is given for four groups: European breeds (red), European wild boars (green), Asian breeds (purple) and Asian wild boars (including the Japanese, orange).

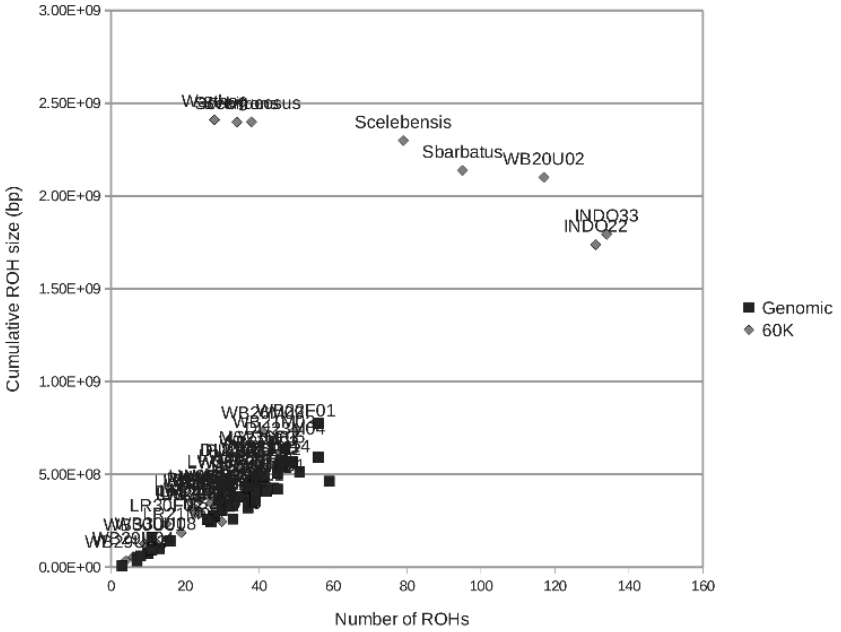


Figure S2.5 Comparison between the genomic data and 60K data on ROH number and cumulative size. The x-axis displays the number of ROHs that was counted for each individuals, and the y-axis shows the cumulative size of ROHs per individual. Data is shown for Genomic data (blue) and 60K data (red). All 52 sequenced animals are included in the analysis. Only for the 60K data the names of the individuals are included, showing that for the non Sus scrofa species (Warthog, *Sus verrucosus, Sus cebifrons, Sus celebensis* and *Sus barbatus*) and for the Japanese wild boar (WB20U02) the number and size of ROHs based on the 60K data are overestimated compared to the number and size of ROH based on the Genomic data. This is probably due to the ascertainment bias that is introduced to the 60K data because the chip is constructed based on polymorphisms that are found in European pig breeds

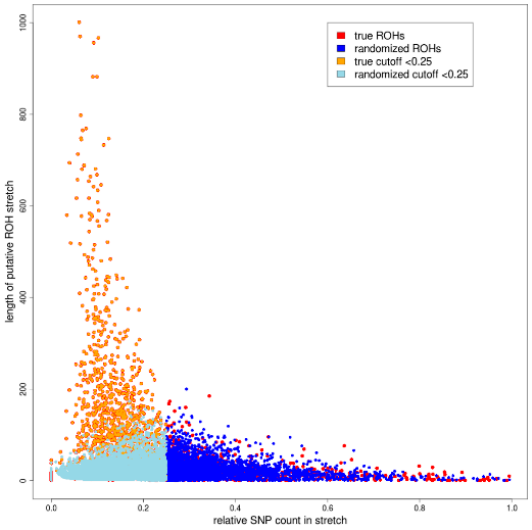


Figure S2.6 Example of ROH detection test where SNP distribution was randomized. The x-axis shows the number of SNPs, averaged over all bins within a ROH, relative to the genome-wide average number of SNPs in a bin. The length of the ROH in terms of consecutive bins is displayed on the y-axis. ROH calculation was executed as explained in the methods section, except for the cutoff of 0.25 times the genomic average. The red dots display the true distribution of ROH length and SNP count within an individual. The blue dots show the distribution after permutation. As can be seen in the plot, the true distribution and the distribution based on a randomized SNP distribution over the genome differ significantly below a relative SNP count of 0.25 times the genomic average per ROH. Values below the cutoff are shown in orange for the true distribution, and lightblue for the randomized distribution.

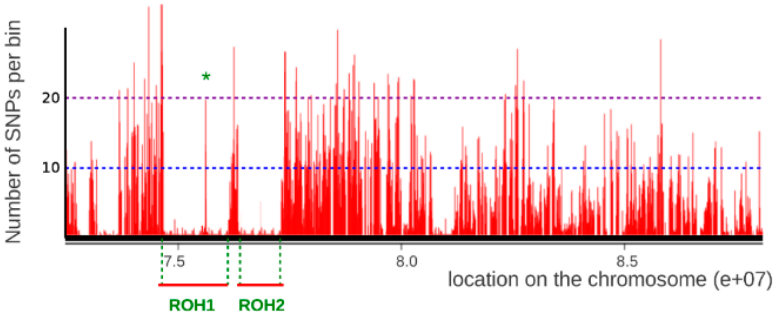


Figure S2.7 Example of ROH calculation. The x-axis represents the location on the chromosome and the y-axis shown the corrected number of SNPs that were counter per bin of 10Kbp. The blue dotted line represents the chromosomal average and the purple line 2* the average. The mutation rate $\mu = 2.5*10^{-8}$ (=0.0025 SNP per bin of 10kbp) and the false discovery rate is 0.0002 (2 SNPs per bin). The maximum SNP count in a ROHbin is in this case 0.25*10= 2.5, because (2+0.0025)< 2.5. The star indicates one bin within a ROH with SNP count 20. Because the local maximum does not exceed 2* the average (=20) and the maximum average of 10 surrounding bins (=(9*2.5 +20)/10) = 4.3) does not exceed 2/3 times the average (=6.67) the bin is included in ROH1. Because the bins between ROH1 and ROH2 locally do exceed this maximum, they are not considered as being part of a ROH.

Table S2.1 Summary statistics of all sequenced individuals. The first two columns "Background" and "Groups" define the background of the individuals. The individuals that are included in the European breeds belong to the Duroc, Hampshire, Landrace, Large White and Pietrain breeds. The Chinese breeds are Jhianquai, Meishan and Xiang. European wild boars are The Dutch, French and Italian individuals and the Asian wild boars come from Japan and China. The *Sus barbatus, Sus cebifrons, Sus celebensis, Sus verrucosus* and Warthog (*Phacochoerus africanus*) are clustered in the group other species. The column "Individual" displays the codes for each individual pig. The average ROH size in bp within an individual is shown in column 4. The total number of ROHs detected within an individual is displayed in column 5. Column 6 shows the average nucleotide diversity in the genome of an individual, outside ROHs. The seventh column displays the average coverage in read depth for each sequenced individual. The last column shows the relative coverage of the genome of each individual, for which each base has at least a read depth of 7 and a maximum read depth of 2 times the average coverage.

Origin	Groups	Individual	av ROH size (bp)	nr of ROHs	π-out	average coverage	Used coverage
Euro breed	Duroc	DU23M01	1082811,9	978	0,002652	10,65333712	0,491559
Euro breed	Duroc	DU23M02	1319945,5	734	0,002531	12,56444414	0,498404
Euro breed	Duroc	DU23M03	1057046,3	1036	0,00251	6,458237256	0,306876
Euro breed	Duroc	DU23M04	1293691,6	921	0,002662	8,337662427	0,407819
Euro breed	Hampshire	HA20U01	1062405,2	1002	0,002608	12,30343973	0,527492
Euro breed	Hampshire	HA20U02	1060795,6	993	0,00269	9,782284727	0,437132
Euro breed	Landrace1	LR21M03	938327,9	921	0,002859	10,396363	0,497363
Euro breed	Landrace2	LR24F01	1004319,4	933	0,002925	13,64539468	0,489508
Euro breed	Landrace2	LR24F08	1035941,6	924	0,002759	8,481167522	0,412339
Euro breed	Landrace3	LR30F02	908869,57	920	0,002647	7,755455501	0,406599
Euro breed	Landrace3	LR30F03	962127,45	1020	0,002721	7,76558895	0,426507
Euro breed	Large White 1	LW22M04	1015213,5	890	0,002892	9,764239246	0,48397
Euro breed	Large White 1	LW22M07	1079909,2	881	0,002582	10,76537049	0,426675
Euro breed	Large White 1	LW22F01	1147012,3	810	0,002744	5,655812236	0,244704
Euro breed	Large White 1	LW22F02	1087146	904	0,002997	10,41696187	0,557362
Euro breed	Large White 1	LW22F03	1058729,2	960	0,002943	10,50719157	0,554664
Euro breed	Large White 1	LW22F04	1033857,1	910	0,002929	10,41717463	0,557365
Euro breed	Large White 1	LW22F06	1101646,9	929	0,00288	9,457636043	0,533186
Euro breed	Large White 2	LW36F01	1137671,5	962	0,002975	10,00821606	0,552747
Euro breed	Large White 2	LW36F02	1248162,2	925	0,002798	8,614409192	0,475714
Euro breed	Large White 2	LW36F03	1162221,1	977	0,002729	9,223497025	0,48695
Euro breed	Large White 2	LW36F04	1085440,4	965	0,002738	9,860501303	0,511628
Euro breed	Large White 2	LW36F05	1290262,3	877	0,002745	8,649150887	0,479551
Euro breed	Large White 2	LW36F06	1173984,3	891	0,002807	8,818326266	0,49668
Euro breed	Pietrain	PI21F02	1051993,8	973	0,002569	10,83280419	0,492542
Euro breed	Pietrain	PI21F06	974808,1	938	0,002962	10,85788363	0,536574
Euro breed	Pietrain	PI21M17	1041031,3	960	0,002918	7,97346239	0,380582
Euro breed	Pietrain	PI21M20	1043203,8	849	0,002812	5,756139443	0,249433
Euro breed	Pietrain	PI21M21	1142233,2	909	0,002498	12,11234212	0,450057
Asia breed	Jianquhai	JQ01U02	1419959,9	499	0,003127	11,20020954	0,50287
Asia breed	Meishan1	MS20U10	1152198,9	523	0,003014	9,324629585	0,521456
Asia breed	Meishan1	MS20U11	1255590,9	440	0,003094	9,241385119	0,499541
Asia breed	Meishan2	MS21M07	1556150,7	491	0,003062	8,71423528	0,447726
Asia breed	Meishan2	MS21M14	1347531,5	555	0,003066	11,1966999	0,509636
Asia breed	Xiang	XI01U03	1992607,4	326	0,003056	9,382646677	0,491836
Asia breed	Xiang	XI01U04	1172509,5	263	0,003064	9,202472147	0,486174
Euro wild	Dutch wild1	WB21F05	748132,1	1408	0,001864	9,902107839	0,520055
Euro wild	Dutch wild1	WB21M03	959697,22	1189	0,001836	11,86069829	0,528753
Euro wild	Dutch wild2	WB22F01	1174204,4	861	0,001593	5,723752976	0,233363
Euro wild	Dutch wild2	WB22F02	886799,08	1309	0,001749	8,443029787	0,445864
Euro wild	French wild	WB25U11	783594,22	1316	0,001858	9,580001823	0,510147
Euro wild	Italian wild	WB26M09	1013249,6	1154	0,001651	15,13171624	0,47235
Asia wild	Japanese wild	WB20U02	1292395,5	1436	0,001818	11,92821661	0,486414
Asia wild	Chinese wild1	WB29U04	1184273,5	117	0,003448	5,302932939	0,214785
Asia wild	Chinese wild1	WB29U12	1023603,6	111	0,003392	10,50852102	0,519725
Asia wild	Chinese wild2	WB30U01	941789,47	190	0,003067	4,855867724	0,189313
Asia wild	Chinese wild2	WB30U08	902376,81	345	0,002895	10,67091624	0,506261
Species	Sbarbatus	Sbarbatus	78275,862	58	0,002641	7,104478282	0,318184
Species	Scebifrons	Scebifrons	3064761,9	63	0,001184	10,03646659	0,537281
Species	Scelebensis	Scelebensis	860517,24	116	0,00232	24,94750036	0,626745
Species	Sverrucosus	Sverrucosus	2070371,8	511	0,000626	13,401695	0,502614
Species	Warthog	Warthog	324353,45	232	0,001989	13,52509472	0,425011

Table S2.2 Summary statistics of all individuals genotyped on the Illumina Porcine 60K iSelect Beadchip. The first two columns "Background" and "Groups" define the background of the individuals. The total number of ROHs detected by PLINK is shown in the fourth column. The last column displays the total sum of ROHs in the genome.

Origin	Group	Individual	ROHnr	ROHsum (*1000)
Euro breed	Duroc	SSDU23M07	35	484920
Euro breed	Duroc	SSDU23M03	38	444361
Euro breed	Duroc	SSDU23M01	29	362212
Euro breed	Duroc	SSDU23M02	35	530795
Euro breed	Duroc	SSDU23M04	47	645843
Euro breed	Duroc	SSDU23M05	26	272860
Euro breed	Duroc	SSDU23M07	40	430827
Euro breed	Duroc	SSDU23M09	38	445964
Euro breed	Duroc	SSDU23M11	39	556849
Euro breed	Duroc	SSDU23M12	50	702951
Euro breed	Duroc	SSDU23M13	45	501453
Euro breed	Hampshire	SSHA20U02	40	461395
Euro breed	Hampshire	SSHA20U04	34	446081
Euro breed	Hampshire	SSHA20U06	47	552621
Euro breed	Hampshire	SSHA20U07	47	628043
Euro breed	Hampshire	SSHA20U08	36	520300
Euro breed	Hampshire	SSHA20U01	42	448931
Euro breed	Large white	SSLW22M01	20	280896
Euro breed	Large white	SSLW22M13	36	417902
Euro breed	Large white	SSLW22M06	22	231179
Euro breed	Large white	SSLW22M14	37	501068
Euro breed	Large white	SSLW22M07	29	346530
Euro breed	Large white	SSLW22M15	18	277113
Euro breed	Large white	SSLW22M08	30	433998
Euro breed	Large white	SSLW22M09	27	391959
Euro breed	Large white	SSLW22M17	21	298548
Euro breed	Large white	SSLW22M02	16	288808
Euro breed	Large white	SSLW22M10	30	415912
Euro breed	Large white	SSLW22M18	24	411695
Euro breed	Large white	SSLW22M03	24	316859
Euro breed	Large white	SSLW22M11	29	294926
Euro breed	Large white	SSLW22M19	23	257828
Euro breed	Large white	SSLW22M04	23	293901
Euro breed	Large white	SSLW22M12	24	337776
Euro breed	Large white	SSLW22M20	27	296589
Euro breed	Large white	SSLW22F01	29	382709
Euro breed	Large white	SSLW22F09	23	282564
Euro breed	Large white	SSLW22F02	26	389613
Euro breed	Large white	SSLW22F10	20	311824
Euro breed	Large white	SSLW22F03	26	326805
Euro breed	Large white	SSLW22F11	29	444565
Euro breed	Large white	SSLW22F04	24	299558
Euro breed	Large white	SSLW22F12	28	435309
Euro breed	Large white	SSLW22F13	30	360039
Euro breed	Large white	SSLW22F06	22	318209
Euro breed	Large white	SSLW22F14	29	307996
Euro breed	Large white	SSLW22F07	22	362203
Euro breed	Large white	SSLW22F08	29	291891
Euro breed	Large white 2	SSLW36F01	34	499209
Euro breed	Large white 3	SSLW36F02	41	534207
Euro breed	Large white 4	SSLW36F03	31	463725
Euro breed	Large white 5	SSLW36F04	30	393196
Euro breed	Large white 6	SSLW36F05	42	600841
Euro breed	Large white 7	SSLW36F06	31	471196
Euro breed	Large white 8	SSLW36F07	25	375956
Euro breed	Landrace	LR21M04	25	271664
Euro breed	Landrace	LR21M02	15	152549
Euro breed	Landrace	LR21M03	19	186672
Euro breed	Landrace	LR21M01	26	262664
Euro breed	Landrace	LR24M04	24	255782
Euro breed	Landrace	LR24M07	27	316131
Euro breed	Landrace	LR24M05	32	427195
Euro breed	Landrace	LR24F06	17	194190
Euro breed	Landrace	LR24F07	19	265526
Euro breed	Landrace	LR24F08	30	365931
Euro breed	Landrace	LR24M08	28	358442
Euro breed	Landrace	LR24F01	23	309972
Euro breed	Landrace	LR24M01	27	360939
Euro breed	Landrace	LR24M02	24	244554
Euro breed	Landrace	LR24M03	36	454776
Euro breed	Landrace	SSLR24M10	36	565133
Euro breed	Landrace	SSLR24M11	23	281442
Euro breed	Landrace	SSLR24M12	23	218435
Euro breed	Landrace	SSLR30F08	24	293387
Euro breed	Landrace	SSLR30F09	36	514610
Euro breed	Landrace	SSLR30F10	27	357868
Euro breed	Landrace	SSLR30F12	22	425654
Euro breed	Landrace	SSLR30F13	28	317395
Euro breed	Landrace	SSLR30F14	30	594172
Euro breed	Landrace	SSLR30F01	29	655506
Euro breed	Landrace	SSLR30F02	15	231735
Euro breed	Landrace	SSLR30F03	24	365117
Euro breed	Landrace	SSLR30F04	22	344500
Euro breed	Landrace	SSLR30F05	22	242496
Euro breed	Landrace	SSLR30F06	23	297540
Euro breed	Landrace	SSLR30F07	11	174030
Asia breed	Meishan	SSMS20U13	32	355688
Asia breed	Meishan	SSMS20M07	27	326196
Asia breed	Meishan	SSMS20M05	45	656286
Asia breed	Meishan	SSMS20M08	41	813044
Asia breed	Meishan	SSMS20M09	35	418275
Asia breed	Meishan	SSMS20M01	58	833783
Asia breed	Meishan	SSMS20M03	37	408747
Asia breed	Meishan	SSMS20U10	30	245787
Asia breed	Meishan	SSMS20U02	35	386805
Asia breed	Meishan	SSMS20U11	36	399964
Asia breed	Meishan	SSMS20M06	33	340642
Asia breed	Meishan	SSMS20M04	37	465444
Asia breed	Meishan	SSMS21M01	50	758374
Asia breed	Meishan	SSMS21M08	44	903016
Asia breed	Meishan	SSMS21M02	39	514009
Asia breed	Meishan	SSMS21M03	32	535858
Asia breed	Meishan	SSMS21M04	44	738630
Asia breed	Meishan	SSMS21M05	45	755144
Asia breed	Meishan	SSMS21M06	40	645505
Asia breed	Meishan	SSMS21M07	40	600244
Asia breed	Meishan	MS21M11	53	840168
Asia breed	Meishan	MS21M12	39	981188
Asia breed	Meishan	MS21M13	33	692950
Asia breed	Meishan	MS21M14	43	552202
Asia breed	Meishan	MS21M15	31	486749
Asia breed	Meishan	MS21M09	39	804758
Asia breed	Meishan	MS21M10	43	1033610
Asia breed	Jianquhai	SSJQ01U02	39	538795
Asia breed	Jianquhai	SSJQ01U05	22	252817
Asia breed	Jianquhai	SSJQ01U06	25	299915
Asia breed	Jianquhai	SSJQ01U09	15	272396
Asia breed	Jianquhai	SSJQ01U14	5	35444,8
Asia breed	Jianquhai	SSJQ01U15	19	167371
Asia breed	Jianquhai	SSJQ01U16	34	443529
Asia breed	Jianquhai	SSJQ01U17	21	293168
Asia breed	Jianquhai	SSJQ01U18	18	249496
Asia breed	Jianquhai	SSJQ01U19	16	297786
Asia breed	Jianquhai	SSJQ01U20	7	53240,9
Asia breed	Xiang	SSXI01U01	38	639966
Asia breed	Xiang	SSXI01U02	31	668579
Asia breed	Xiang	SSXI01U03	33	472811
Asia breed	Xiang	SSXI01U04	13	132990
Asia breed	Xiang	SSXI01U07	30	731657
Asia breed	Xiang	SSXI01U08	37	711834
Asia breed	Xiang	SSXI01U10	38	448432
Asia breed	Xiang	SSXI01U12	9	114781
Asia breed	Xiang	SSXI01U13	38	470380
Asia breed	Xiang	SSXI01U15	33	369331
Asia breed	Xiang	SSXI01U17	7	88456
Asia breed	Xiang	SSXI01U19	31	506703
Asia breed	Xiang	SSXI01U20	38	664852
Asia wild	WB20_Japan	SSWB20M01	69	740680
Asia wild	WB20_Japan	SSWB20M02	117	2093360
Asia wild	WB20_Japan	SSWB20U01	30	279918
Asia wild	WB20_Japan	SSWB20U02	56	686186
Asia wild	WB20_Japan	SSWB20U03	117	2099530
Asia wild	WB20_Japan	SSWB20M03	67	958699
Asia wild	WB20_Japan	SSWB20F01	53	868696
Asia wild	WB20_Japan	SSWB20F02	48	706314
Asia wild	WB20_Japan	SSWB20M04	41	619634
Asia wild	WB20_Japan	SSWB20U04	57	638226
Asia wild	WB20_Japan	SSWB20F03	111	2129230
Asia wild	WB20_Japan	SSWB20M06	74	1426540
Asia wild	WB20_Japan	SSWB20F04	117	1977990

Origin	Group	Individual	ROHnr	ROHsum (*1000)
Euro wild	WB21_Netherland	SSWB21M01	44	497911
Euro wild	WB21_Netherland	SSWB21M02	47	813281
Euro wild	WB21_Netherland	SSWB21M03	44	682264
Euro wild	WB21_Netherland	SSWB21M04	43	758303
Euro wild	WB21_Netherland	SSWB21F01	22	241753
Euro wild	WB21_Netherland	SSWB21F02	35	644028
Euro wild	WB21_Netherland	SSWB21F03	41	551080
Euro wild	WB21_Netherland	SSWB21M05	54	960247
Euro wild	WB21_Netherland	SSWB21F04	39	410332
Euro wild	WB21_Netherland	SSWB21F05	40	487304
Euro wild	WB21_Netherland	SSWB21M06	36	410316
Euro wild	WB21_Netherland	SSWB21M07	49	642302
Euro wild	WB21_Netherland	SSWB21U01	35	736498
Euro wild	WB21_Netherland	SSWB21U02	29	457652
Euro wild	WB21_Netherland	SSWB21U03	40	1422300
Euro wild	WB21_Netherland	SSWB21U04	28	435624
Euro wild	WB21_Netherland	SSWB21U05	35	653672
Euro wild	WB21_Netherland	SSWB21U06	36	522335
Euro wild	WB21_Netherland	SSWB21U07	26	326643
Euro wild	WB21_Netherland	SSWB21U08	30	331370
Euro wild	WB21_Netherland	SSWB21U09	35	711515
Euro wild	WB21_Netherland	SSWB21U10	30	625632
Euro wild	WB21_Netherland	SSWB21U11	44	741271
Euro wild	WB21_Netherland	SSWB21U12	35	654374
Euro wild	WB21_Netherland	SSWB21U13	33	435707
Euro wild	WB21_Netherland	SSWB21U14	39	635291
Euro wild	WB21_Netherland	SSWB21U15	34	465177
Euro wild	WB21_Netherland	SSWB21U16	27	412270
Euro wild	WB21_Netherland	SSWB21U17	21	394311
Euro wild	WB21_Netherland	SSWB21U18	31	357127
Euro wild	WB21_Netherland	SSWB21U19	40	649457
Euro wild	WB21_Netherland	SSWB21U21	26	534723
Euro wild	WB21_Netherland	SSWB21U22	42	914208
Euro wild	WB21_Netherland	SSWB21U23	29	458681
Euro wild	WB21_Netherland	SSWB21U24	28	401713
Euro wild	WB21_Netherland	SSWB21U55	31	596009
Euro wild	WB21_Netherland	SSWB21U56	40	537999
Euro wild	WB21_Netherland	SSWB21U57	41	759519
Euro wild	WB21_Netherland	SSWB21U58	15	156555
Euro wild	WB21_Netherland	SSWB21U59	47	671392
Euro wild	WB21_Netherland	SSWB21U60	48	628220
Euro wild	WB21_Netherland	SSWB21U61	45	564054
Euro wild	WB21_Netherland	SSWB21U62	42	521280
Euro wild	WB21_Netherland	SSWB21U63	35	588892
Euro wild	WB21_Netherland	SSWB21U64	45	526278
Euro wild	WB21_Netherland	SSWB21U65	39	635443
Euro wild	WB21_Netherland	SSWB21U66	58	853344
Euro wild	WB21_Netherland	SSWB21U67	36	566737
Euro wild	WB21_Netherland	SSWB21U68	45	568588
Euro wild	WB21_Netherland	SSWB21U69	46	595603
Euro wild	WB21_Netherland	SSWB21U70	49	637129
Euro wild	WB21_Netherland	SSWB21U72	5	41071
Euro wild	WB26_Italy	SSWB26F02	37	744051
Euro wild	WB26_Italy	SSWB26M10	28	464423
Euro wild	WB26_Italy	SSWB26F03	48	783238
Euro wild	WB26_Italy	SSWB26F04	41	796141
Euro wild	WB26_Italy	SSWB26M11	48	1088950
Euro wild	WB26_Italy	SSWB26F05	47	565939
Euro wild	WB26_Italy	SSWB26F06	34	482087
Euro wild	WB26_Italy	SSWB26M13	37	413697
Euro wild	WB26_Italy	SSWB26F07	19	237468
Euro wild	WB26_Italy	SSWB26M17	40	968748
Euro wild	WB26_Italy	SSWB26M19	43	641101
Euro wild	WB26_Italy	SSWB26F11	21	432806
Euro wild	WB26_Italy	SSWB26F12	42	881166
Euro wild	WB26_Italy	SSWB26F09	27	537643
Euro wild	WB26_Italy	SSWB26F13	35	836667
Euro wild	WB26_Italy	SSWB26F10	20	298058
Euro wild	WB26_Italy	SSWB26F14	38	923249
Euro wild	WB26_Italy	SSWB26M14	47	783556
Euro wild	WB26_Italy	SSWB26F15	32	542089
Euro wild	WB26_Italy	SSWB26M15	41	547140
Euro wild	WB26_Italy	SSWB26M18	26	276105
Euro wild	WB26_Italy	SSWB26M16	38	587944
Euro wild	WB26_Italy	SSWB26F16	39	544171
Asia wild	WB29_China	SSWB29U04	6	55914,8
Asia wild	WB29_China	SSWB29U12	4	35540,5
Asia wild	WB29_China	SSWB29U13	13	110249
Asia wild	WB29_China	SSWB29U14	7	57876,3
Asia wild	WB29_China	SSWB29U19	7	127349
Asia wild	WB30_China	SSWB30U01	9	127297
Asia wild	WB30_China	SSWB30U05	9	70588,6
Asia wild	WB30_China	SSWB30U07	11	199541
Asia wild	WB30_China	SSWB30U08	12	133819
Asia wild	WB30_China	SSWB30U09	6	48574,7

Table S2.3 List of GO-IDs that were overrepresented in the Large White ROHs. Genes from all genomic regions that were homozygous in all Large Whites were extracted. The GO-IDs in column 1 were overrepresented in these ROH regions (p-values in column B, Benjamini and Hochberg corrected p-values in column C). The description of the corresponding biological process is shown in column D and all genes that contributed to the GO group are listed in column E.

GO-ID	p-value	corr p-value	Description	Genes in test set
6334	5.3804E-10	1.0146E-6	nucleosome assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
31497	9.5623E-10	1.0146E-6	chromatin assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
65004	1.8844E-9	1.1392E-6	protein-DNA complex assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
34728	2.1475E-9	1.1392E-6	nucleosome organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
6323	3.6373E-8	1.5437E-5	DNA packaging	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
6333	8.5400E-8	3.0203E-5	chromatin assembly or disassembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
71103	2.4023E-7	7.2823E-5	DNA conformation change	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF
34622	1.5862E-6	4.2075E-4	cellular macromolecular complex assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|SNRPD1|RPS6|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|TUBA3C|TUBA3E|TGS1
34621	2.0902E-6	4.9283E-4	cellular macromolecular complex subunit organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|SNRPD1|RPS6|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|TUBA3C|TUBA3E|TGS1|GHSR
6325	1.7285E-5	3.6678E-3	chromatin organization	HIST1H2BB|SATB2|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|SOX2|HIST1H2BH|HIRA|HIST1H2BO|RNF8|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|JAK2|MLL2
65003	3.2334E-5	6.2375E-3	macromolecular complex assembly	CAPZA1|SNRPD1|TP63|POLR2B|HIST1H2BO|IGF1R|HIST1H2BN|HIST1H2BL|HIST1H2BI|TUBA3C|TUBA3E|TGS1|SLC1A1|HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|RPS6|P2RX6|HIST2H2BE|HIST2H2BF|KPNA3
43933	3.6434E-5	6.4428E-3	macromolecular complex subunit organization	CAPZA1|SNRPD1|TP63|POLR2B|HIST1H2BO|IGF1R|HIST1H2BN|HIST1H2BL|HIST1H2BI|TUBA3C|TUBA3E|TGS1|SLC1A1|HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|RPS6|P2RX6|HIST2H2BE|HIST2H2BF|GHSR|KPNA3
22607	4.5967E-5	7.5032E-3	cellular component assembly	DRD1|CAPZA1|SNRPD1|TP63|POLR2B|HIST1H2BO|IGF1R|HIST1H2BN|HIST1H2BL|HIST1H2BI|TUBA3C|NEFH|TUBA3E|TGS1|NKX2-5|SLC1A1|HIST1H2BB|HIST1H2BC|ACTC1|HIST1H2BD|ROCK1|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|RPS6|P2RX6|HIST2H2BE|HIST2H2BF|KPNA3
42509	7.7024E-5	1.1300E-2	regulation of tyrosine phosphorylation of STAT protein	NF2|LYN|TIFA|JAK2|HGF
44085	7.9878E-5	1.1300E-2	cellular component biogenesis	NAF1|DRD1|CAPZA1|SNRPD1|TP63|POLR2B|HIST1H2BO|IGF1R|HIST1H2BN|HIST1H2BL|HIST1H2BI|TUBA3C|NEFH|TUBA3E|TGS1|SLC1A1|NKX2-5|HIST1H2BB|HIST1H2BC|ACTC1|HIST1H2BD|ROCK1|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|RPS6|RCL1|P2RX6|HIST2H2BE|HIST2H2BF|KPNA3
6996	8.8106E-5	1.1685E-2	organelle organization	GRPEL1|CCDC99|USP8|SOX2|CAPZA1|LRRCC1|HIRA|HIST1H2BO|PFN3|HIST1H2BN|HIST1H2BL|HIST1H2BI|NEFH|NKX2-5|DNAJC19|MLL2|BOD1|HIST1H2BB|SATB2|HIST1H2BC|ACTC1|HIST1H2BD|NF2|ROCK1|HIST1H2BE|TRPM7|HIST1H2BF|HIST1H2BG|ERMN|HIST1H2BH|HGF|RPS6|RNF8|BNIP1|HIST2H2BE|HIST2H2BF|JAK2|SYNM|GHSR
42516	1.0152E-4	1.2673E-2	regulation of tyrosine phosphorylation of Stat3 protein	NF2|TIFA|JAK2|HGF
51276	1.4205E-4	1.4472E-2	chromosome organization	HIST1H2BB|SATB2|CCDC99|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|SOX2|HIST1H2BH|HIRA|HIST1H2BO|RNF8|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|JAK2|MLL2
9712	1.4322E-4	1.4472E-2	catechol metabolic process	DRD1|HDC|NR4A2|CHRNA7|COMT
6584	1.4322E-4	1.4472E-2	catecholamine metabolic process	DRD1|HDC|NR4A2|CHRNA7|COMT
34311	1.4322E-4	1.4472E-2	diol metabolic process	DRD1|HDC|NR4A2|CHRNA7|COMT
18958	1.6506E-4	1.5921E-2	phenol metabolic process	DRD1|HDC|NR4A2|CHRNA7|COMT
46425	1.8936E-4	1.7471E-2	regulation of JAK-STAT cascade	NF2|LYN|TIFA|JAK2|HGF
42417	1.9948E-4	1.7638E-2	dopamine metabolic process	DRD1|NR4A2|CHRNA7|COMT
60112	2.2514E-4	1.9110E-2	generation of ovulation cycle rhythm	CHRNA7|NPY5R
42053	2.6355E-4	2.1510E-2	regulation of dopamine metabolic process	DRD1|NR4A2|COMT
9653	3.1384E-4	2.4665E-2	anatomical structure morphogenesis	MRPL40|SOX2|HIRA|TP63|GREM1|IGF1R|FOXQ1|DGCR2|GATA6|FOXF2|ROBO4|ROBO3|NKX2-5|PLAG1|ACTC1|SATB2|NF2|LZTR1|ERMN|RTN4R|NR4A2|HGF|GSC2|RPS6|WNT2B|OSM|MAPK1|VEGFC|MIB1|CRKL|JAK2|CA2|FEZ1|ACVR1
42069	3.7233E-4	2.8217E-2	regulation of catecholamine metabolic process	DRD1|NR4A2|COMT
22414	4.7850E-4	3.5013E-2	reproductive process	ERMP1|DRD1|IGFBP7|SOX2|TP63|CHEK1|COMT|IGF1R|GATA6|WWP1|DDX25|FOXF2|SPATA4|CHRNA7|DNAJC19|PLAG1|RPS6|SPA17|NPY5R|WNT2B|CYLC2|TSSK2|TSSK1B|GHSR|ACVR1
3	5.0525E-4	3.5738E-2	reproduction	ERMP1|DRD1|IGFBP7|SOX2|TP63|CHEK1|COMT|IGF1R|GATA6|WWP1|DDX25|FOXF2|SPATA4|CHRNA7|DNAJC19|PLAG1|RPS6|SPA17|NPY5R|WNT2B|CYLC2|TSSK2|TSSK1B|GHSR|ACVR1
7267	5.4915E-4	3.7590E-2	cell-cell signaling	GJD2|SEPT5|DRD1|STC2|SOX2|PI4KA|TIFA|TP63|PANX3|COMT|HGF|GREM1|NPY5R|WNT2B|MAPK1|TNFSF10|GRM3|CHRNA7|GHSR|SLC1A1
10033	6.0931E-4	3.9227E-2	response to organic substance	ATP6V0E1|DRD1|RBM3|IGFBP7|SOX2|COMT|IGF1R|DGCR8|GATA6|CHRNA7|NKX2-5|MLL2|ACTC1|LYN|AK3|NR4A2|NPY1R|HGF|UBE2L3|MAPK1|TXNDC4|PCNA|JAK2|CA2|GHSR|FEZ1
16043	6.1003E-4	3.9227E-2	cellular component organization	GRPEL1|AP1B1|CAPZA1|HIRA|SNRPD1|TP63|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST1H2BI|FOXF2|CHRNA7|ROBO3|TGS1|SLC1A1|DNAJC19|MLL2|SATB2|ROCK1|TRPM7|ERMN|BNIP1|HIST2H2BE|HIST2H2BF|KPNA3|SNAP29|CCDC99|DRD1|USP8|SOX2|LRRCC1|POLR2B|IGF1R|PFN3|TUBA3C|NEFH|TUBA3E|NKX2-5|BOD1|HIST1H2BB|HIST1H2BC|ACTC1|HIST1H2BD|NF2|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|RTN4R|NR4A2|HGF|RPS6|RNF8|P2RX6|JAK2|SYNM|GHSR|FEZ1
32101	6.3302E-4	3.9417E-2	regulation of response to external stimulus	OSM|VEGFC|DRD1|JAK2|CHRNA7|HGF|GREM1|GHSR|USP14|NPY5R
42423	6.6758E-4	3.9417E-2	catecholamine biosynthetic process	HDC|NR4A2|CHRNA7
42508	6.6871E-4	3.9417E-2	tyrosine phosphorylation of Stat1 protein	OSM|JAK2

Table S2.4 List of GO-IDs that were overrepresented in the European wild boar ROHs. Genes from all genomic regions that were homozygous in all European wild boars were extracted. The GO-IDs in column 1 were overrepresented in these ROH regions (p-values in column B, Benjamini and Hochberg corrected p-values in column C). The description of the corresponding biological process is shown in column D and all genes that contributed to the GO group are listed in column E.

GO-ID	p-value	corr p-value	Description	Genes in test set
6334	1.5020E-18	1.4737E-15	nucleosome assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
31497	2.9269E-18	1.4737E-15	chromatin assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
65004	6.4478E-18	1.8907E-15	protein-DNA complex assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
34728	7.5102E-18	1.8907E-15	nucleosome organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
6323	2.1119E-16	4.2533E-14	DNA packaging	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
6333	5.8654E-16	9.8442E-14	chromatin assembly or disassembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
71103	2.0457E-15	2.9428E-13	DNA conformation change	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
34622	1.2861E-10	1.6189E-8	cellular macromolecular complex assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
6325	3.9558E-10	4.4261E-8	chromatin organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|USP3|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
34621	6.0343E-10	6.0765E-8	cellular macromolecular complex subunit organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
51276	1.0164E-8	9.3046E-7	chromosome organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|USP3|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
65003	1.6789E-6	1.4089E-4	macromolecular complex assembly	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
22607	2.3444E-6	1.8160E-4	cellular component assembly	HIST1H2BB|DRD1|HIST1H2BC|E2F4|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
43933	3.7302E-6	2.6831E-4	macromolecular complex subunit organization	HIST1H2BB|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
6996	7.6753E-6	5.1527E-4	organelle organization	TPPP3|HIST1H2BB|FZR1|HIST1H2BC|HIST1H2BD|HIST1H2BE|USP3|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|DNA2|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX|FHOD1
44085	1.1601E-5	7.3016E-4	cellular component biogenesis	HIST1H2BB|DRD1|HIST1H2BC|E2F4|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|HIST1H2BH|HIST1H2BO|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|H2AFX
42742	1.9474E-5	1.1535E-3	defense response to bacterium	HIST1H2BC|HIST1H2BE|HIST1H2BF|HIST2H2BE|HIST1H2BG|HIST1H2BI
51707	2.1852E-4	1.2225E-2	response to other organism	HIST1H2BC|HIST1H2BE|HIST1H2BF|HIST2H2BE|IFNE|HIST1H2BG|HIST1H2BI|ITCH
9987	4.8327E-4	2.5613E-2	cellular process	E2F4|NDST2|HIST1H2BO|HIST1H2BN|CDKN2B|HIST1H2BL|HIST1H2BI|S1PR4|CNTNAP2|H2AFX|ITCH|ATOH7|MCRS1|CSNK1G1|FUT11|CA12|MECOM|ELMO3|DCAF6|ITGB1BP3|HIST2H2BE|HIST2H2BF|HSD11B2|KCNH3|SLC40A1|FHOD1|PRPF38A|TPPP3|DRD1|FZR1|MAMDC2|USP3|APH1B|ASNSD1|LRG1|APBA3|OXCT2|SPATA3|NKX2-3|ZBTB7A|HIST1H2BB|RPSA|HIST1H2BC|HIST1H2BD|HIST1H2BE|HIST1H2BF|HIST1H2BG|BCKDHB|HIST1H2BH|DNA2|HNRNPH3|PPIB|DOHH|PBX1|XCL1
16043	5.9500E-4	2.9958E-2	cellular component organization	HIST1H2BB|TPPP3|HIST1H2BC|FZR1|DRD1|HIST1H2BD|E2F4|HIST1H2BE|USP3|HIST1H2BF|HIST1H2BG|HIST1H2BH|ELMO3|HIST1H2BO|DNA2|HIST1H2BN|HIST1H2BL|HIST2H2BE|HIST1H2BI|HIST2H2BF|CNTNAP2|H2AFX|FHOD1
9617	6.7693E-4	3.2460E-2	response to bacterium	HIST1H2BC|HIST1H2BE|HIST1H2BF|HIST2H2BE|HIST1H2BG|HIST1H2BI
21756	7.1650E-4	3.2796E-2	striatum development	DRD1|CNTNAP2
9607	9.5573E-4	4.1844E-2	response to biotic stimulus	HIST1H2BC|HIST1H2BE|HIST1H2BF|HIST2H2BE|IFNE|HIST1H2BG|HIST1H2BI|ITCH
21544	1.0883E-3	4.5661E-2	subpallium development	DRD1|CNTNAP2

Chapter 3

for pairwise comparison:					
		MS20M03	LW22F02		
		MS20M05	LW22F04		
		MS20U10	LW22F06		
		MS20U11	LW22F07		
		MS20U13	LW22M07		
		MS21M05	LW36F01		
		MS21M07	LW36F03		
		MS21M08	LW36F05		
		MS21M14	LW36F06		
Individuals from 4 pig groups that the 2 selected breeds are compared to:		**8 ASWB**	**7 EUWB**	**6 ASDom**	**6 EUDom**
		WB20U02	WB21F05	MS20U11	DU23M01
		WB29U04	WB21M05	MS21M07	DU23M02
		WB29U12	WB22F02	MS21M14	HA20U01
		WB29U14	WB25U11	XI01U03	LW22F04
		WB29U16	WB26M09	XI01U04	LW36F03
		WB30U01	WB33U04	JQ01U02	LW36F05
		WB30U08	WB44U06		
		WB30U09			
Number of pairwise comparisons with:	**Meishans**	72	63	-	54
	Large whites	72	63	54	-

Figure S3.1 Schematic overview of all pairwise comparisons used in the IBD detection. All individuals from the LW breed and MS breed are shown in the top two columns. All individuals used for pairwise IBD detection are shown in the bottom 4 columns. Those individuals highlighted in yellow (LW) and blue (MS) are included in both the breed-group and the Domestic-group. When these two groups are compared, i.e. LW with EUDom, all pairwise comparisons are made except for those where individuals are compared to themselves (because this will result in a full IBD genome).

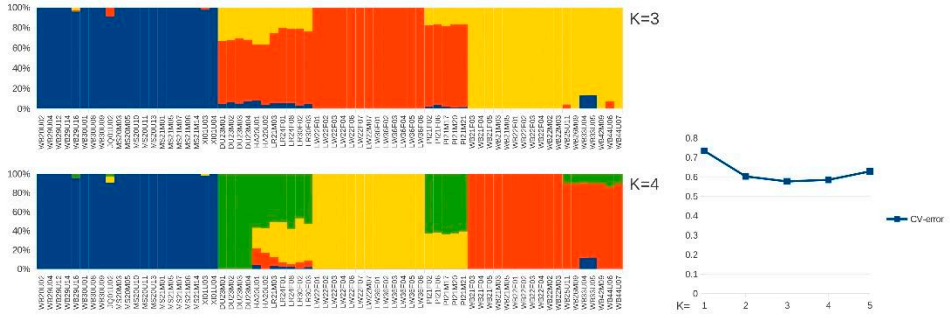


Figure S3.2 Admixture for all 70 pigs with K ranging from 1 to 5. The percentages of each assigned population contributing to the genetic variation in each individual is plotted for K=3 and K=4. Cross-validation errors are shown for K=1-5.

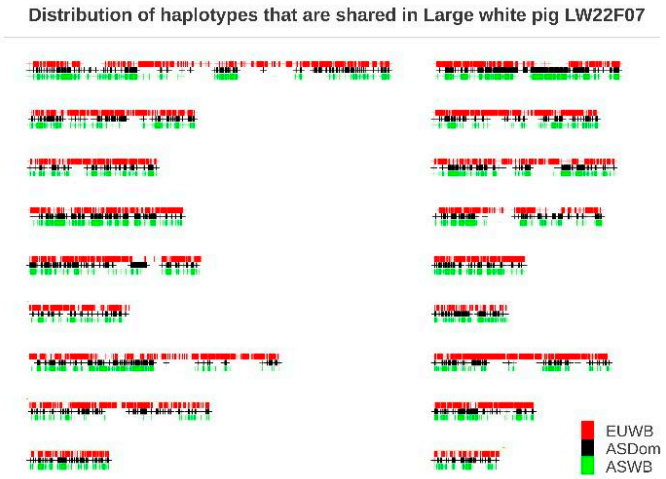


Figure S3.3 Example of the distribution of IBD haplotypes in the genome of a LW pig. Regions in the genome from one LW pig in bins of 10Kbp where one of the two haplotypes is IBD with EUDom (orange) and EUWB (red), ASDom (black) or ASWB (green). Bins are plotted on top of each other, so the same bin might contain a haplotype that is IBD with individuals from different pig groups simultaneously.

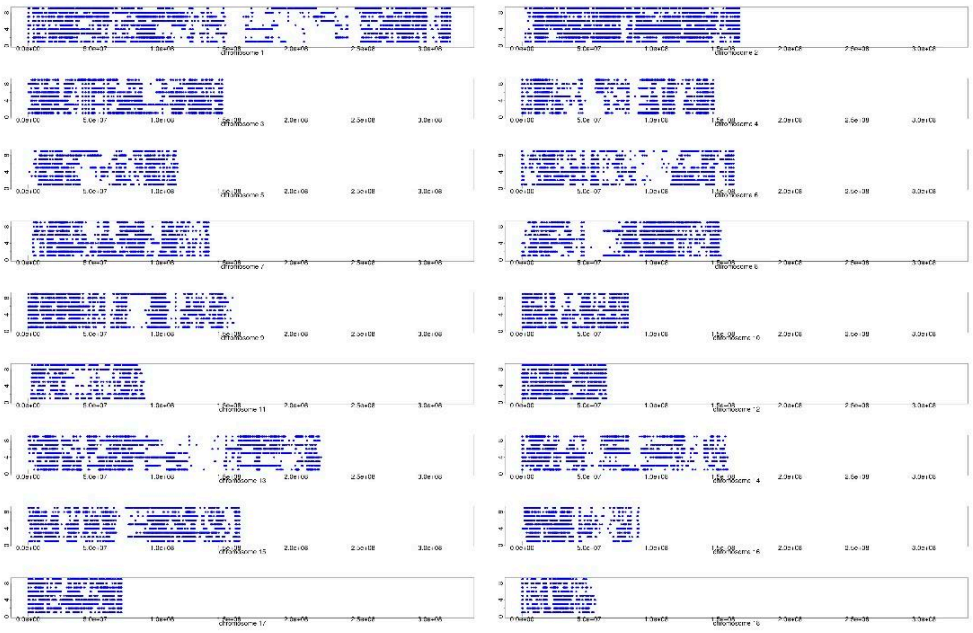


Figure S3.4 Distribution of haplotypes shared with ASDom over the genome of 9 LW pigs. The full length of all 18 autosomes is represented on the x-axis and each individual LW pig is listed on the y-axes (1 to 9). Regions of the genome in which a LW pig shares a haplotype with an ASDom pig are visualized in blue.

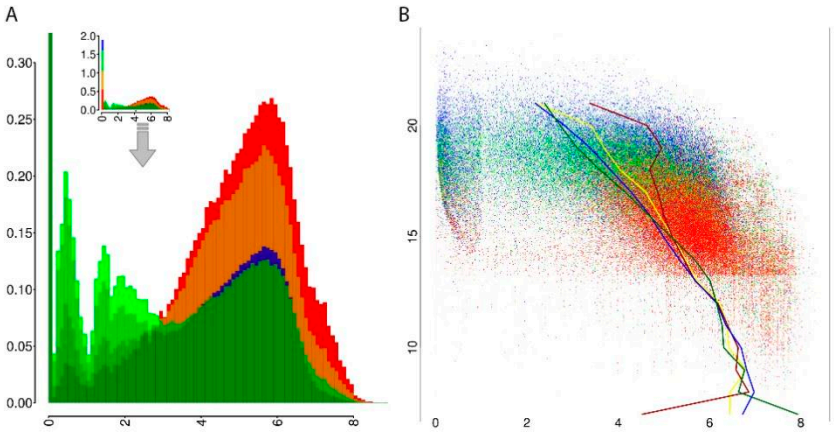


Figure S3.5 Distribution of nucleotide diversity between the two haplotypes in an individual. The x-axis displays the log-transformed nucleotide diversity (π) within a bin of 10Kbp in an individual, and the y-axis shows the relative frequency of that particular class of π in the full dataset. **A**. Relative distribution of π for regions in the genome where at least one of the two haplotypes in a LW individual is IBD with ASWB (red), ASDom (orange), EUDom (blue) and EUWB (green). The top left figure shows the full histogram, and the large graph zooms in on the bins>0. **B**. Size and nucleotide diversity of IBD fragments (both log-transformed) in LW pigs that are shared with EUDom (blue), EUWB (green), ASDom (orange), ASWB (red). Lines represent the average of each size-class.

Table 3.1 Overview of re-sequenced individuals that are used in the experiment. The origin of each sample is shown in the column "Origin" and the breed name or geographical region of the population is found in the column "Groups". The usage of each individual in the experiment is explained in the column "Usage" where "phasing" represents individuals that are used only for phasing the data, “analysis” indicates that these individuals are used in the LW or MS group to do the analyses on and "IBD" means that these individuals were also used for pairwise IBD comparisons.

total length	ungapped
2596639456	2323671356

Origin	Groups	Individual	total covered (bp)	average coverage	full_genome_coverage	usage
Euro breed	Duroc	DU23M01	2002023768	10,885	9,379	Phasing + IBD
Euro breed	Duroc	DU23M02	1990836233	11,776	10,089	Phasing + IBD
Euro breed	Duroc	DU23M03	1969535343	6,206	5,260	Phasing
Euro breed	Duroc	DU23M04	1978633263	7,759	6,607	Phasing
Euro breed	Hampshire	HA20U01	2002330006	11,646	10,035	Phasing + IBD
Euro breed	Hampshire	HA20U02	1993804314	10,299	8,837	Phasing
Euro breed	Landrace1	LR21M03	1988232860	9,548	8,170	Phasing
Euro breed	Landrace2	LR24F01	2006041898	13,917	12,014	Phasing
Euro breed	Landrace2	LR24F08	1984470378	9,213	7,868	Phasing
Euro breed	Landrace3	LR30F02	1986763206	7,604	6,502	Phasing
Euro breed	Landrace3	LR30F03	1988711923	7,782	6,660	Phasing
Euro breed	Large White 1	LW22M04	1996387069	9,771	8,394	Phasing
Euro breed	Large White 1	LW22M07	1953768579	10,642	8,948	Phasing + analysis
Euro breed	Large White 1	LW22F01	1900759489	5,888	4,817	Phasing
Euro breed	Large White 1	LW22F02	2010763103	10,318	8,928	Phasing + analysis
Euro breed	Large White 1	LW22F03	2008023120	10,316	8,914	Phasing
Euro breed	Large White 1	LW22F04	2011600059	10,338	8,950	Phasing + IBD + analysis
Euro breed	Large White 1	LW22F06	2007811588	9,547	8,250	Phasing + analysis
Euro breed	Large White 1	LW22F07	1996662559	11,790	10,131	Phasing + analysis
Euro breed	Large White 2	LW36F01	2012251940	10,070	8,720	Phasing + analysis
Euro breed	Large White 2	LW36F02	1995877141	8,633	7,415	Phasing
Euro breed	Large White 2	LW36F03	1996779275	9,004	7,737	Phasing + IBD + analysis
Euro breed	Large White 2	LW36F04	1998082648	9,576	8,234	Phasing
Euro breed	Large White 2	LW36F05	1997146423	8,737	7,509	Phasing + IBD + analysis
Euro breed	Large White 2	LW36F06	1997536328	8,820	7,582	Phasing + analysis
Euro breed	Pietrain	PI21F02	1984695950	10,760	9,191	Phasing
Euro breed	Pietrain	PI21F06	2000242335	10,783	9,282	Phasing
Euro breed	Pietrain	PI21M17	1957367968	8,671	7,304	Phasing
Euro breed	Pietrain	PI21M20	1879735507	5,931	4,798	Phasing
Euro breed	Pietrain	PI21M21	1973410348	11,435	9,712	Phasing
Asia breed	Jianquhai	JQ01U02	1988118084	10,746	9,194	Phasing + IBD + analysis
Asia breed	Meishan1	MS20M03	2010860374	10,331	8,940	Phasing + analysis
Asia breed	Meishan1	MS20M05	2016516860	10,457	9,075	Phasing + analysis
Asia breed	Meishan1	MS20U10	2002062901	9,410	8,107	Phasing + analysis
Asia breed	Meishan1	MS20U11	1997687711	9,398	8,080	Phasing + IBD + analysis
Asia breed	Meishan1	MS20U13	1997497029	7,782	6,689	Phasing + analysis
Asia breed	Meishan2	MS21M01	2006569137	8,653	7,473	Phasing
Asia breed	Meishan2	MS21M05	2014656185	10,610	9,199	Phasing + analysis
Asia breed	Meishan2	MS21M07	1973736597	8,995	7,640	Phasing + IBD + analysis
Asia breed	Meishan2	MS21M08	2009337184	9,013	7,794	Phasing + analysis
Asia breed	Meishan2	MS21M14	1990527039	10,429	8,933	Phasing + IBD + analysis
Asia breed	Xiang	XI01U03	1987624315	9,247	7,909	Phasing + IBD
Asia breed	Xiang	XI01U04	1988074028	9,082	7,770	Phasing + IBD
Euro wild	Dutch wild1	WB21F03	2017669054	10,559	9,168	Phasing
Euro wild	Dutch wild1	WB21F04	2024762016	12,613	10,991	Phasing
Euro wild	Dutch wild1	WB21F05	1995236506	9,406	8,076	Phasing + IBD
Euro wild	Dutch wild1	WB21M03	2003345400	11,663	10,056	Phasing
Euro wild	Dutch wild1	WB21M05	2025131146	15,245	13,287	Phasing + IBD
Euro wild	Dutch wild2	WB22F01	1893909559	5,725	4,666	Phasing
Euro wild	Dutch wild2	WB22F02	1989214850	8,167	6,991	Phasing + IBD
Euro wild	Dutch wild2	WB22F03	2012232772	9,429	8,165	Phasing
Euro wild	Dutch wild2	WB22F04	2010008029	9,022	7,804	Phasing
Euro wild	Dutch wild2	WB22M03	2021721345	11,907	10,359	Phasing
Euro wild	French wild	WB25U11	1988434379	9,622	8,234	Phasing + IBD
Euro wild	Swiss wild	WB26M09	1998495400	14,618	12,572	Phasing + IBD
Euro wild	Samos wild	WB33U04	2021164592	11,524	10,023	Phasing + IBD
Euro wild	Samos wild	WB33U05	2023128516	10,318	8,983	Phasing + IBD
Euro wild	Italian wild	WB42M09	2028924112	12,674	11,066	Phasing
Euro wild	Italian wild	WB44U06	2026387033	12,158	10,602	Phasing + IBD
Euro wild	Italian wild	WB44U07	2022238432	10,721	9,331	Phasing
Asia wild	Japanese wild	WB20U02	1985231894	11,315	9,667	Phasing + IBD
Asia wild	Chinese wild1	WB29U04	1851933086	5,492	4,377	Phasing + IBD
Asia wild	Chinese wild1	WB29U12	1992312707	10,385	8,904	Phasing + IBD
Asia wild	Chinese wild1	WB29U14	2011262234	9,035	7,820	Phasing + IBD
Asia wild	Chinese wild1	WB29U16	2023590514	12,772	11,122	Phasing + IBD
Asia wild	Chinese wild2	WB30U01	1795056199	5,405	4,175	Phasing + IBD
Asia wild	Chinese wild2	WB30U08	1987517519	10,092	8,632	Phasing + IBD
Asia wild	Chinese wild2	WB30U09	2023107155	11,846	10,314	Phasing + IBD
Outgroup	Sumatra wild	INDO22	1980223126	11,494	9,795	Phasing + root
Outgroup	Sumatra wild	INDO33	1975881286	11,155	9,485	Phasing + root
		Average	**1989738156,086**	**9,945**	**8,539**	

Chapter 4

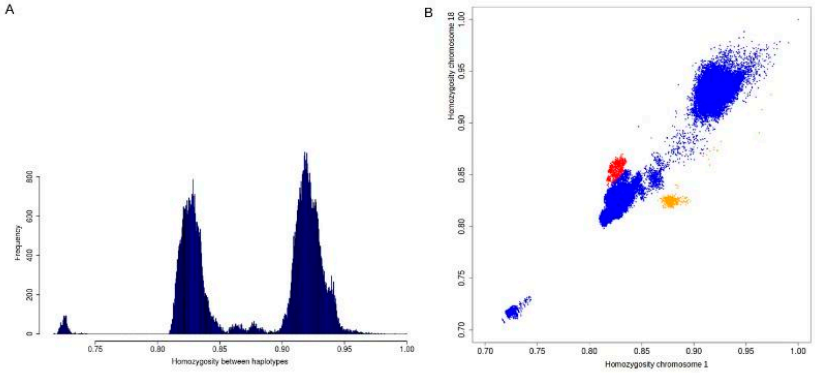


Figure S4.1 Haplotype homozygosity and consistency over chromosomes. A. Distribution of haplotype homozygosity between all possible pairs of haplotypes on chromosome 1. The first peak around 0.725 contains all haplotypes paired with a haplotype from Sumatra. The second peak at 0.825 represents all haplotypes paired with a Chinese wild or Chinese domesticated haplotype. The third peak round 0.92 shows all paired European haplotypes. **B**. Consistency over chromosomes. The x-axis displays homozygosity between haplotypes from chromosome 1, and the y-axis shows the homozygosity between the same pairs of haplotypes for chromosome 18.

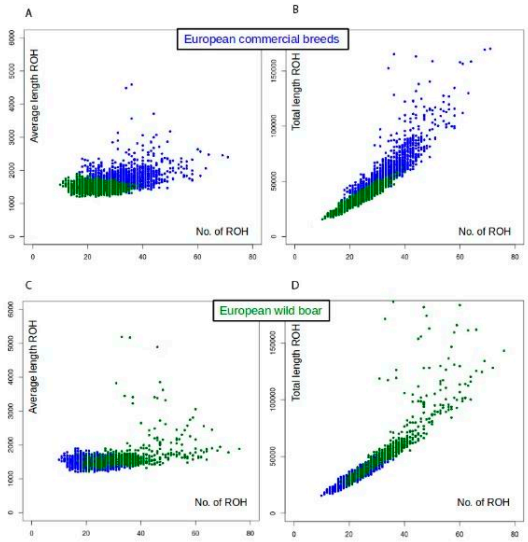


Figure S4.2 Runs of homozygosity between paired haplotypes. ROHs on chromosome 1 are recorded between pairs of haplotypes that belong to the European wild group (green) or the European commercial group (blue). **A**. Number of ROH and average ROH length when a haplotype is paired with a European domestic haplotype **B**. Number of ROH and total ROH length when a haplotype is paired with a European domestic haplotype **C**. Number of ROH and average ROH length when a haplotype is paired with a European wild haplotype **D**. Number of ROH and total ROH length when a haplotype is paired with a European wild haplotype.

Chapter 5

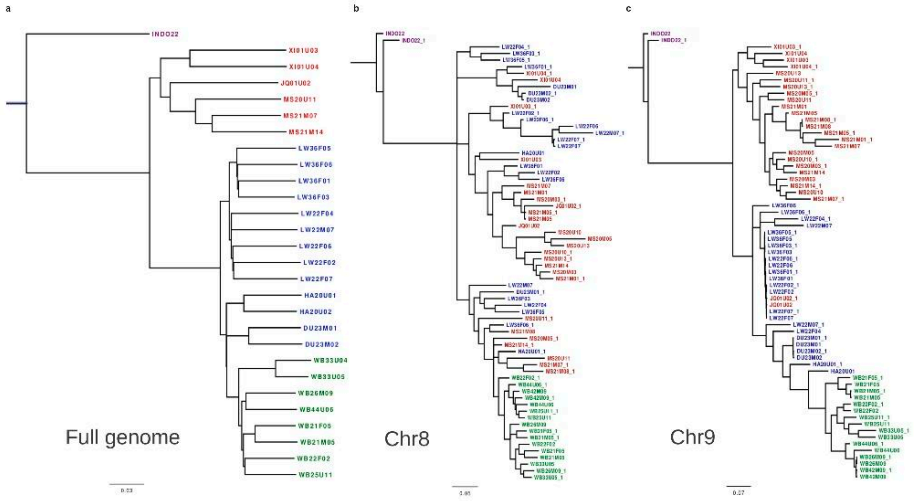


Figure S5.1 Bayesian trees for haplotypes covering the two largest regions of introgression. Coloration is based on the three different pig groups: Asian domesticated pigs (orange), European wild boars (green) and European domesticated pigs (blue). All individuals that are included in the pairwise comparisons as described in the methods section are included. **A**. Phylogenetic tree based on all markers in the dataset. **B**. Tree based on the introgression region on chromosome 8 (rIBD>0). Each individual has two haplotypes in the tree. **C**. Tree based on the introgression region on chromosome 9 (rIBD>0). Each individual has two haplotypes in the tree.

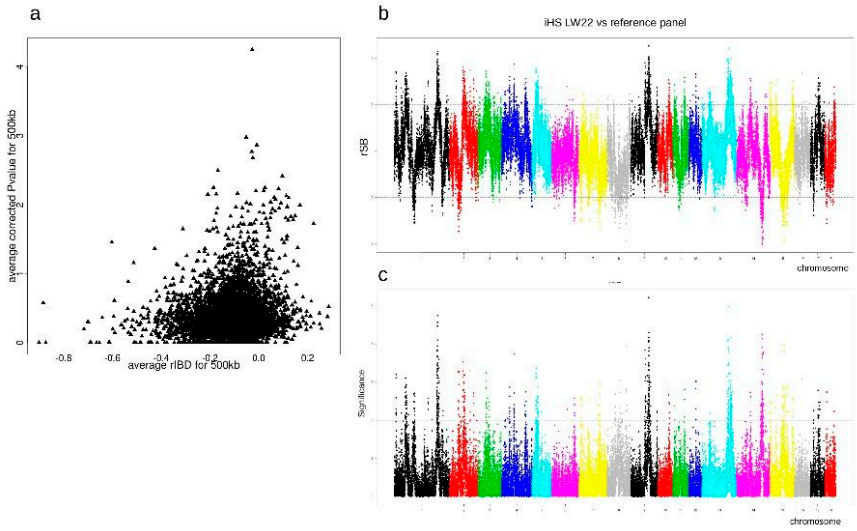


Figure S5.2 Signatures of selection based on EHH in Large White. Regions of extended haplotype homozygosity within 56 LW individuals are identified for each chromosome. **A**. Correlation between iHS

significance signals (y-axis) and rIBD signals (x-axis) in LW, averaged over 500kb. **B**. The iHS signals in LW are polarized with iHS signals in a reference panel of 4 other European breeds, resulting in rSB (standardized ratio of iES). **C**. Significance of the rSB signals in LW for each chromosome.

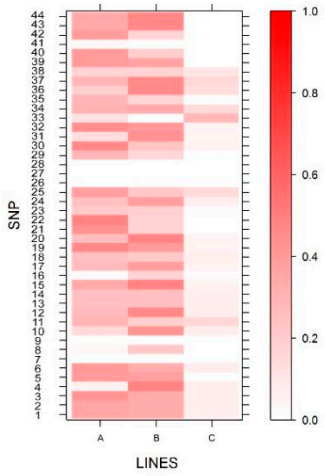


Figure S5.3 Allele frequencies for markers surrounding the introgression region on chromosome 8. Allele frequencies are displayed for each 60K marker (44 total) that lies within the longest introgression region on chromosome 8. Frequencies are calculate for two reproduction-associated lines (line A, N=1053 and Line B, N=568) and one growth-associated line (Line C, N=965).

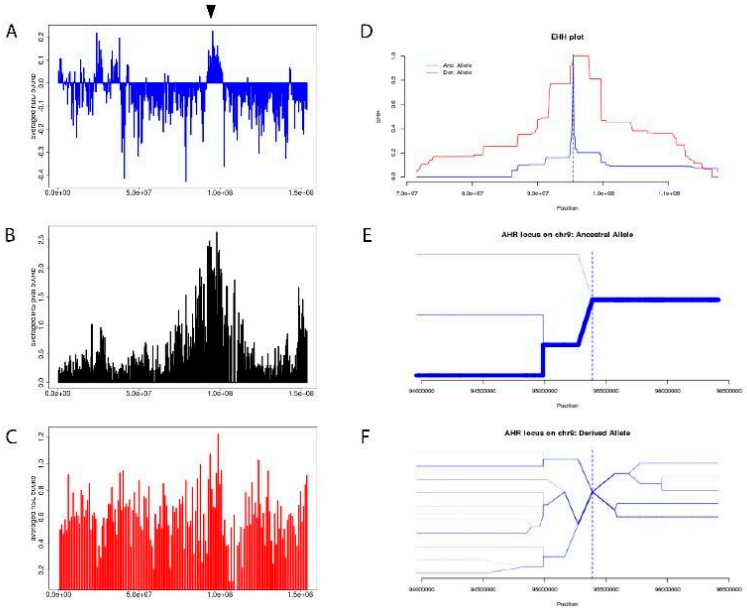


Figure S5.4 Extended haplotype homozygosity in Large White at the AHR locus. Regions of extended haplotype homozygosity within 56 Large White individuals are identified for chromosome 9. **A**. The

distribution of rIBD in LW over the full length of chromosome 9. The AHR locus is indicated with an arrow, resulting in the highest peak of rIBD. **B**. Significance of the iHS signals in LW for chromosome 9, averaged over 500kb. **C**. Significance of the nSL signals in LW for chromosome 9, averaged over 500kb. **D**. Extended haplotype homozygosity of the ancestral allele in LW at the AHR locus. **E-F**. Bifurcation diagrams of the breakdown of LD at increasing distance from the AHR locus.

Table S5.1 Overview of re-sequenced individuals that are used in the experiment. The origin of each sample is shown in the column "Origin" and the breed name or geographical region of the population is found in the column "Groups". The usage of each individual in the experiment is explained in the column "Usage" where "phasing" represents individuals that are used only for phasing the data, and "analysis" means that they were also used for pairwise IBD comparisons.

Origin	Groups	Individual	average coverage	usage
Euro breed	Duroc	DU23M01	10.885	Phasing
Euro breed	Duroc	DU23M02	11.776	Phasing
Euro breed	Duroc	DU23M03	6.206	Phasing
Euro breed	Duroc	DU23M04	7.759	Phasing
Euro breed	Hampshire	HA20U01	11.646	Phasing
Euro breed	Hampshire	HA20U02	10.299	Phasing
Euro breed	Landrace1	LR21M03	9.548	Phasing
Euro breed	Landrace2	LR24F01	13.917	Phasing
Euro breed	Landrace2	LR24F08	9.213	Phasing
Euro breed	Landrace3	LR30F02	7.604	Phasing
Euro breed	Landrace3	LR30F03	7.782	Phasing
Euro breed	Large White 1	LW22M04	9.771	Phasing + analysis
Euro breed	Large White 1	LW22M07	10.642	Phasing + analysis
Euro breed	Large White 1	LW22F01	5.888	Phasing + analysis
Euro breed	Large White 1	LW22F02	10.318	Phasing + analysis
Euro breed	Large White 1	LW22F03	10.316	Phasing + analysis
Euro breed	Large White 1	LW22F04	10.338	Phasing + analysis
Euro breed	Large White 1	LW22F06	9.547	Phasing + analysis
Euro breed	Large White 1	LW22F07	11.790	Phasing + analysis
Euro breed	Large White 2	LW36F01	10.070	Phasing + analysis
Euro breed	Large White 2	LW36F02	8.633	Phasing + analysis
Euro breed	Large White 2	LW36F03	9.004	Phasing + analysis
Euro breed	Large White 2	LW36F04	9.576	Phasing + analysis
Euro breed	Large White 2	LW36F05	8.737	Phasing + analysis
Euro breed	Large White 2	LW36F06	8.820	Phasing + analysis
Euro breed	Pietrain	PI21F02	10.760	Phasing
Euro breed	Pietrain	PI21F06	10.783	Phasing
Euro breed	Pietrain	PI21M17	8.671	Phasing
Euro breed	Pietrain	PI21M20	5.931	Phasing
Euro breed	Pietrain	PI21M21	11.435	Phasing
Asia breed	Jianquhai	JQ01U02	10.746	Phasing + analysis
Asia breed	Meishan1	MS20M03	10.331	Phasing
Asia breed	Meishan1	MS20M05	10.457	Phasing
Asia breed	Meishan1	MS20U10	9.410	Phasing
Asia breed	Meishan1	MS20U11	9.398	Phasing + analysis
Asia breed	Meishan1	MS20U13	7.782	Phasing
Asia breed	Meishan2	MS21M01	8.653	Phasing
Asia breed	Meishan2	MS21M05	10.610	Phasing
Asia breed	Meishan2	MS21M07	8.995	Phasing + analysis
Asia breed	Meishan2	MS21M08	9.013	Phasing
Asia breed	Meishan2	MS21M14	10.429	Phasing + analysis
Asia breed	Xiang	XI01U03	9.247	Phasing + analysis
Asia breed	Xiang	XI01U04	9.082	Phasing + analysis
Euro wild	Dutch wild1	WB21F03	10.559	Phasing
Euro wild	Dutch wild1	WB21F04	12.613	Phasing
Euro wild	Dutch wild1	WB21F05	9.406	Phasing + analysis
Euro wild	Dutch wild1	WB21M03	11.663	Phasing
Euro wild	Dutch wild1	WB21M05	15.245	Phasing + analysis
Euro wild	Dutch wild2	WB22F01	5.725	Phasing
Euro wild	Dutch wild2	WB22F02	8.167	Phasing + analysis
Euro wild	Dutch wild2	WB22F03	9.429	Phasing
Euro wild	Dutch wild2	WB22F04	9.022	Phasing
Euro wild	Dutch wild2	WB22M03	11.907	Phasing
Euro wild	French wild	WB25U11	9.622	Phasing + analysis
Euro wild	Swiss wild	WB26M09	14.618	Phasing + analysis
Euro wild	Samos wild	WB33U04	11.524	Phasing + analysis
Euro wild	Samos wild	WB33U05	10.318	Phasing + analysis
Euro wild	Italian wild	WB42U06	12.674	Phasing
Euro wild	Italian wild	WB44U06	12.158	Phasing + analysis
Euro wild	Italian wild	WB44U07	10.721	Phasing
Asia wild	Japanese wild	WB20U02	11.315	Phasing
Asia wild	Chinese wild1	WB29U04	5.492	Phasing
Asia wild	Chinese wild1	WB29U12	10.385	Phasing
Asia wild	Chinese wild1	WB29U14	9.035	Phasing
Asia wild	Chinese wild1	WB29U16	12.772	Phasing
Asia wild	Chinese wild2	WB30U01	5.405	Phasing
Asia wild	Chinese wild2	WB30U08	10.092	Phasing
Asia wild	Chinese wild2	WB30U09	11.846	Phasing
Outgroup	Sumatra wild	INDO22	11.494	Phasing + outgroup
Outgroup	Sumatra wild	INDO33	11.155	Phasing

Table S5.2 Overview of introgressed regions in the Large white genome. Regions in the genome where Asian introgression is most pronounced (ZrIBD >2) and the genes within these regions are listed. "Start" represents the startbin of 10.000bp and length indicates how long the stretch within the region of inferred introgression is.

Chromosome	start	stop	Length (ZrIBD>2)	genes			Chromosome	start	stop	Length (ZrIBD>2)	genes		
Ssc10_2_1	4090000	4100000	10000				Ssc10_2_13	61610000	61660000	50000			
Ssc10_2_1	6840000	6860000	20000				Ssc10_2_13	158690000	158700000	10000			
Ssc10_2_1	7550000	7640000	90000				Ssc10_2_13	170850000	170860000	10000			
Ssc10_2_1	7710000	7720000	10000				Ssc10_2_14	130000	150000	20000	CDK20		
Ssc10_2_1	27170000	27190000	20000				Ssc10_2_14	180000	220000	40000			
Ssc10_2_1	28540000	28700000	160000	REPS1	C8ORF115		Ssc10_2_14	240000	250000	10000			
Ssc10_2_1	29650000	29660000	10000				Ssc10_2_14	9450000	9470000	20000			
Ssc10_2_1	29670000	29700000	30000				Ssc10_2_14	9510000	9520000	10000			
Ssc10_2_1	30780000	30800000	20000				Ssc10_2_14	11830000	11840000	10000			
Ssc10_2_1	30810000	30860000	50000				Ssc10_2_14	13330000	13380000	50000	RPLP1		
Ssc10_2_1	30880000	30970000	90000				Ssc10_2_14	19320000	19340000	20000			
Ssc10_2_1	31090000	31130000	40000				Ssc10_2_14	25070000	25090000	20000			
Ssc10_2_1	31190000	31200000	10000				Ssc10_2_14	90970000	91000000	30000			
Ssc10_2_1	31790000	31840000	50000				Ssc10_2_15	7130000	7220000	90000			
Ssc10_2_1	32020000	32310000	290000	MYB	AHI1		Ssc10_2_15	7690000	7750000	60000			
Ssc10_2_1	32410000	32440000	30000				Ssc10_2_15	9330000	9500000	170000			
Ssc10_2_1	32620000	32630000	10000				Ssc10_2_15	30010000	30230000	220000			
Ssc10_2_1	52110000	52280000	170000	EYS			Ssc10_2_15	33360000	33380000	20000			
Ssc10_2_1	53390000	53760000	370000				Ssc10_2_15	33390000	33450000	60000			
Ssc10_2_1	54040000	54080000	40000				Ssc10_2_15	59950000	60170000	220000	NRG1		
Ssc10_2_1	56830000	57020000	190000	C6orf57	FAM135A		Ssc10_2_15	60290000	60300000	10000			
Ssc10_2_1	71440000	71470000	30000				Ssc10_2_15	97710000	97860000	150000	FRZB	DNAJC10	
Ssc10_2_1	71480000	71510000	30000				Ssc10_2_15	97960000	98380000	480000	NUP35	NCKAP1	DUSP19
Ssc10_2_1	71530000	71540000	10000				Ssc10_2_15	119770000	119780000	10000			
Ssc10_2_1	71610000	71640000	30000				Ssc10_2_15	119790000	119800000	10000	VCL		
Ssc10_2_1	74270000	74390000	120000	POU3F2			Ssc10_2_15	120330000	120510000	180000			
Ssc10_2_1	77060000	77160000	100000	GRIK2			Ssc10_2_15	123770000	123880000	110000			
Ssc10_2_1	77170000	77210000	40000				Ssc10_2_15	124040000	124200000	160000	MAP2		
Ssc10_2_1	81970000	82020000	50000	AIM1			Ssc10_2_15	136610000	136620000	10000			
Ssc10_2_1	83100000	83110000	10000	SCML4			Ssc10_2_15	140000000	140010000	10000			
Ssc10_2_1	84510000	84520000	10000				Ssc10_2_15	140790000	140800000	10000			
Ssc10_2_1	84550000	84800000	250000	CEP57L1			Ssc10_2_16	3500000	3510000	10000			
Ssc10_2_1	84870000	84890000	20000	AKD1			Ssc10_2_16	6130000	6140000	10000			
Ssc10_2_1	84970000	84990000	20000	AKD1			Ssc10_2_16	23150000	23160000	10000			
Ssc10_2_1	85040000	85060000	20000				Ssc10_2_16	23170000	23220000	50000			
Ssc10_2_1	85520000	85940000	120000	WASF1			Ssc10_2_16	24940000	25000000	60000			
Ssc10_2_1	86090000	86120000	30000				Ssc10_2_17	35900000	36040000	140000	SMOX		
Ssc10_2_1	86140000	86910000	770000	GTF3C6	CDK19	RPF2	Ssc10_2_17	41500000	41530000	30000	BPIFB4		
				AMD1	REV3L	KIAA1919	Ssc10_2_17	58300000	58340000	40000			
				SLC16A10			Ssc10_2_17	66810000	66820000	10000			
Ssc10_2_1	87040000	87060000	20000	TRAF3IP2			Ssc10_2_18	40000	50000	10000			
Ssc10_2_1	87400000	87530000	130000				Ssc10_2_18	17560000	17580000	20000			
Ssc10_2_1	87540000	87570000	30000				Ssc10_2_18	18370000	18410000	40000			
Ssc10_2_1	90300000	90340000	40000				Ssc10_2_18	23950000	24210000	260000			
Ssc10_2_1	90640000	90730000	90000				Ssc10_2_18	24630000	24800000	170000	GPR37		
Ssc10_2_1	91830000	91880000	50000	NT5DC1			Ssc10_2_18	24920000	24940000	20000			
Ssc10_2_1	92080000	92240000	160000				Ssc10_2_18	25070000	25100000	30000			
Ssc10_2_1	92460000	92540000	80000	SULT2B1			Ssc10_2_18	25130000	25140000	10000			
Ssc10_2_1	92840000	92890000	50000				Ssc10_2_2	11330000	11370000	40000	OSBP		
Ssc10_2_1	92900000	93420000	520000	RWDD2A	PRSS35	ME1	Ssc10_2_2	11390000	11410000	20000			
				PGM3	DOPEY1		Ssc10_2_2	11440000	11450000	10000	OR4D10		
Ssc10_2_1	93450000	93540000	90000	UBE3D	TSPAN13		Ssc10_2_2	11470000	11490000	20000	OR5A2		
Ssc10_2_1	94010000	94270000	260000				Ssc10_2_2	12620000	12640000	20000	LDHB		
Ssc10_2_1	97970000	97990000	20000				Ssc10_2_2	12990000	13140000	150000	OR4K15	OR4K1	OR4L1
Ssc10_2_1	98040000	98090000	50000	IRAK1BP1			Ssc10_2_2	13320000	13340000	20000	PRG3	PRG3	
Ssc10_2_1	103820000	103930000	110000				Ssc10_2_2	24410000	24450000	40000			
Ssc10_2_1	104740000	104870000	130000				Ssc10_2_2	24460000	24570000	110000			
Ssc10_2_1	107740000	107880000	140000				Ssc10_2_2	33760000	33810000	50000			
Ssc10_2_1	107910000	107920000	10000				Ssc10_2_2	33880000	33920000	40000			
Ssc10_2_1	108000000	108020000	20000				Ssc10_2_2	33990000	34000000	10000			
Ssc10_2_1	177280000	177340000	60000	CDH20			Ssc10_2_2	34210000	34280000	70000			
Ssc10_2_1	178880000	178890000	10000				Ssc10_2_2	34550000	34930000	380000	KIF18A	METTSD1	
Ssc10_2_1	178700000	178730000	30000				Ssc10_2_2	35310000	35320000	10000			
Ssc10_2_1	230010000	230030000	20000				Ssc10_2_2	36610000	36730000	120000			
Ssc10_2_1	281820000	281830000	10000				Ssc10_2_2	40640000	40650000	10000			
Ssc10_2_1	284430000	284510000	80000	SAL1			Ssc10_2_2	42060000	42080000	20000	NAV2		
Ssc10_2_1	312000000	312010000	10000				Ssc10_2_2	42110000	42120000	10000	DBX1		
Ssc10_2_10	410000	420000	10000				Ssc10_2_2	42180000	42190000	10000			
Ssc10_2_10	440000	450000	10000				Ssc10_2_2	50200000	50210000	10000			
Ssc10_2_10	460000	480000	20000				Ssc10_2_2	67840000	67860000	20000			
Ssc10_2_10	2290000	2310000	20000	B3GALT2			Ssc10_2_2	68110000	68120000	10000			
Ssc10_2_10	2320000	2330000	10000				Ssc10_2_2	68840000	68870000	30000	OR7D4	OR7D4	ZNF846
Ssc10_2_10	4200000	4210000	10000				Ssc10_2_2	69110000	69360000	250000	P2RY11	COL5A3	RDH8
Ssc10_2_10	4360000	4400000	40000								C3	C19ORF66	ANGPTL6
Ssc10_2_10	4410000	4420000	10000								PPAN	EIF3G	DNMT1
Ssc10_2_10	4430000	4440000	10000				Ssc10_2_2	69410000	69560000	150000	ZGLP1	ICAM1	ICAM5
Ssc10_2_10	4480000	4780000	300000								ICAM6	ZGLP1	FDX1L
Ssc10_2_10	23740000	23750000	10000								RAVER1	RAVER1	ICAM-3
Ssc10_2_10	47780000	47770000	10000								TYK2	CDC37	
Ssc10_2_10	48260000	48310000	50000				Ssc10_2_2	69570000	69610000	40000	PDE4A		
Ssc10_2_10	48370000	48420000	50000				Ssc10_2_2	70060000	70090000	30000	C19ORF52	YIPF2	
Ssc10_2_10	48550000	48630000	80000	PTPLA			Ssc10_2_2	86740000	86760000	20000			
Ssc10_2_10	50420000	50440000	20000				Ssc10_2_2	86850000	86870000	20000			
Ssc10_2_10	50530000	50550000	20000				Ssc10_2_2	86880000	86900000	20000			
Ssc10_2_11	8620000	8650000	30000	FRY			Ssc10_2_2	86910000	87010000	100000	IQGAP2		
Ssc10_2_11	8700000	8720000	20000	N4BP2L2			Ssc10_2_2	88570000	88780000	210000	AP3B1		
Ssc10_2_11	25780000	25800000	20000				Ssc10_2_2	88910000	88930000	20000			
Ssc10_2_11	25900000	25930000	30000				Ssc10_2_2	89280000	89300000	20000			
Ssc10_2_12	680000	720000	40000				Ssc10_2_2	89320000	89370000	50000	ARSB		
Ssc10_2_12	990000	1020000	30000				Ssc10_2_2	89500000	89530000	30000			
Ssc10_2_12	1030000	1180000	150000	NPLOC4	FASN	C17orf70	Ssc10_2_2	90480000	90490000	10000			
				FSCN2	BETA-ACTIN		Ssc10_2_2	90500000	90520000	20000	THBS4		
Ssc10_2_12	1230000	1380000	150000	SLC38A10			Ssc10_2_2	90580000	90620000	40000			
Ssc10_2_12	1400000	1410000	10000				Ssc10_2_2	90890000	91020000	130000	FAM151B	ZFYVE16	
Ssc10_2_12	1420000	1500000	80000	AATK	C12H17orf56	AZI1	Ssc10_2_2	93160000	93260000	100000			
Ssc10_2_12	1520000	1650000	130000	CHMP6	RPTOR		Ssc10_2_2	94090000	94110000	20000			
Ssc10_2_12	1660000	1720000	60000				Ssc10_2_2	97000000	97080000	80000			
Ssc10_2_12	21660000	21670000	10000	KRTAP1-3			Ssc10_2_2	101260000	101330000	70000			
Ssc10_2_12	21680000	21740000	60000	KRT39	KRTAP3-1		Ssc10_2_2	101590000	101640000	50000			
Ssc10_2_12	25130000	25220000	90000	IGF2BP1			Ssc10_2_2	106910000	106960000	50000	CAST		
Ssc10_2_12	25300000	25350000	50000				Ssc10_2_2	107030000	107330000	300000	ERAP2	ERAP1	
Ssc10_2_12	26800000	26820000	20000	ACSF2			Ssc10_2_2	107370000	107440000	70000	IRAP		
Ssc10_2_12	28440000	28460000	20000				Ssc10_2_2	107450000	107480000	30000			
Ssc10_2_12	41520000	41550000	30000	NLE1			Ssc10_2_2	107490000	107560000	70000	LIX1		
Ssc10_2_12	46340000	46380000	40000				Ssc10_2_2	107950000	107980000	30000			
Ssc10_2_12	53520000	53530000	10000				Ssc10_2_2	108680000	108850000	170000			
Ssc10_2_13	14220000	14290000	70000	NGLY1			Ssc10_2_2	110280000	110340000	60000			
Ssc10_2_13	15470000	15520000	50000	SLC4A7			Ssc10_2_2	110950000	111000000	50000			
Ssc10_2_13	16170000	16290000	120000				Ssc10_2_2	111340000	111370000	30000			
Ssc10_2_13	16670000	16720000	50000				Ssc10_2_2	111380000	111560000	180000			
Ssc10_2_13	29740000	29750000	10000				Ssc10_2_2	115770000	115810000	40000			
Ssc10_2_13	29800000	29810000	10000				Ssc10_2_2	115980000	115990000	10000			
Ssc10_2_13	45170000	45200000	30000				Ssc10_2_2	123230000	123240000	10000			
Ssc10_2_13	45210000	45240000	30000				Ssc10_2_2	125130000	125360000	230000	COMMD10		
Ssc10_2_13	53960000	54000000	40000				Ssc10_2_2	125390000	125470000	80000	SEMA6A		
Ssc10_2_13	54030000	54050000	20000				Ssc10_2_2	125550000	125560000	10000			
							Ssc10_2_2	125710000	125830000	120000			

Chromosome	start	stop	Length (ZrIBD>2)	genes			Chromosome	start	stop	Length (ZrIBD>2)	genes		
Ssc10_2_2	125860000	125890000	30000				Ssc10_2_7	24760000	24780000	20000	AFP		
Ssc10_2_2	126050000	126190000	140000				Ssc10_2_7	24810000	25230000	420000	RNF39	TRIM15	TRIM10
Ssc10_2_2	126970000	127010000	40000								TRIM40	TRIM31	PPP1R11
Ssc10_2_2	127040000	127090000	50000								ZNRD1	C7H6ORF12	KRAB
Ssc10_2_2	127130000	127190000	60000								MOG	RNF39	ZNRD1
Ssc10_2_2	127340000	127460000	120000								KRAB	GABBR1	
Ssc10_2_2	127480000	127540000	60000				Ssc10_2_7	25320000	25350000	30000	UBD		
Ssc10_2_2	127590000	127610000	20000				Ssc10_2_7	25360000	25390000	30000	OR2G6	OR2G3	OR6Q1
Ssc10_2_2	127780000	127850000	70000				Ssc10_2_7	25450000	25700000	250000	OR14J1	OR12D2	OR12D2
Ssc10_2_2	127860000	127910000	50000								OR5V1	OR14J1	
Ssc10_2_2	128260000	128270000	10000				Ssc10_2_7	25740000	25750000	10000	OR12D2		
Ssc10_2_2	129780000	129930000	150000	PRR16			Ssc10_2_7	25760000	25780000	20000	OR12D3	OR5V1	
Ssc10_2_2	129970000	130000000	30000				Ssc10_2_7	26670000	26930000	260000	C7H6orf136	SLA-11	TRIM39
Ssc10_2_2	130120000	130130000	10000								RPP21	GNL1	PRR3
Ssc10_2_2	130140000	130160000	20000								ABCF1	PPP1R10	MRPS18B
Ssc10_2_2	130620000	130660000	40000								ATAT1	DHX16	
Ssc10_2_2	130960000	131000000	40000				Ssc10_2_7	28400000	28430000	30000	C6orf10		
Ssc10_2_2	132080000	132250000	170000	CEP120			Ssc10_2_7	28530000	28540000	10000			
Ssc10_2_2	132340000	132580000	240000	CSNK1G3			Ssc10_2_7	28660000	28680000	20000			
Ssc10_2_2	132590000	132720000	130000				Ssc10_2_7	132350000	132380000	30000			
Ssc10_2_2	133010000	133020000	10000				Ssc10_2_7	132390000	132500000	110000			
Ssc10_2_2	133030000	133080000	50000				Ssc10_2_7	132640000	132680000	40000	OR4F3	OR4F16	OR4F29
Ssc10_2_2	133120000	133170000	50000				Ssc10_2_8	8850000	8870000	20000			
Ssc10_2_2	135910000	135920000	10000				Ssc10_2_8	9770000	9790000	20000			
Ssc10_2_2	137320000	137330000	10000				Ssc10_2_8	10340000	10350000	10000			
Ssc10_2_3	9960000	10150000	190000	HSPB1			Ssc10_2_8	13890000	13900000	10000			
Ssc10_2_3	12150000	12170000	20000				Ssc10_2_8	14070000	14080000	10000			
Ssc10_2_3	14360000	14410000	50000				Ssc10_2_8	14200000	14250000	50000			
Ssc10_2_3	14420000	14550000	130000				Ssc10_2_8	14620000	14830000	10000			
Ssc10_2_3	22300000	22310000	10000				Ssc10_2_8	16550000	16620000	70000			
Ssc10_2_3	29500000	29510000	10000				Ssc10_2_8	25020000	25030000	10000			
Ssc10_2_3	29650000	29660000	10000				Ssc10_2_8	29850000	29880000	30000			
Ssc10_2_3	29750000	29760000	10000				Ssc10_2_8	30220000	30230000	10000			
Ssc10_2_3	29830000	29840000	10000				Ssc10_2_8	92960000	92980000	20000			
Ssc10_2_3	30940000	30960000	20000				Ssc10_2_8	98780000	98800000	20000			
Ssc10_2_3	31200000	31260000	60000	SHISA9			Ssc10_2_8	103030000	103040000	10000			
Ssc10_2_3	31660000	32110000	450000				Ssc10_2_8	103300000	103850000	550000	LARP1B	PGRMC2	C4ORF29
Ssc10_2_3	59750000	59870000	120000	RPIA	IGKV1D-43						MFSD8		
Ssc10_2_3	59940000	59980000	40000				Ssc10_2_8	104050000	104060000	10000			
Ssc10_2_3	111220000	111240000	20000				Ssc10_2_8	104070000	104560000	490000			
Ssc10_2_3	111850000	111860000	10000				Ssc10_2_8	105170000	105180000	10000			
Ssc10_2_3	111870000	111900000	30000				Ssc10_2_8	105210000	105220000	10000			
Ssc10_2_3	111930000	111960000	30000				Ssc10_2_8	105250000	105390000	140000			
Ssc10_2_3	113210000	113250000	40000	LTBP1			Ssc10_2_8	107310000	107320000	10000			
Ssc10_2_3	113520000	113530000	10000				Ssc10_2_8	109050000	109080000	30000			
Ssc10_2_3	113770000	113850000	80000	TTC27			Ssc10_2_8	120090000	120140000	50000	ELOVL6		
Ssc10_2_3	120350000	120390000	40000				Ssc10_2_8	144280000	144290000	10000			
Ssc10_2_3	120400000	120600000	200000	DNMT3A	DTNB		Ssc10_2_8	144350000	144380000	30000	HPSE	HELQ	
Ssc10_2_3	121580000	121590000	10000				Ssc10_2_8	145150000	145200000	50000	HNRPDL		
Ssc10_2_3	130800000	130830000	30000				Ssc10_2_8	145210000	145230000	20000	HNRNPD		
Ssc10_2_4	60000	200000	140000	ZNF250	COMMD5	C8ORF33	Ssc10_2_8	145870000	145900000	30000			
				ZNF7			Ssc10_2_8	146150000	146240000	90000	BMP3		
Ssc10_2_4	16550000	16570000	20000				Ssc10_2_8	146660000	146690000	30000	FGF5		
Ssc10_2_4	43410000	43430000	20000				Ssc10_2_9	90000	110000	20000			
Ssc10_2_4	43460000	43590000	100000	MTERFD1	PTDSS1		Ssc10_2_9	230000	290000	60000	TMEM41B		
Ssc10_2_4	97150000	97170000	20000	DEDD	UFC1		Ssc10_2_9	1650000	1730000	80000	OR5R1	OR5M10	OR10A3
Ssc10_2_4	105310000	105330000	20000								OR10A6		
Ssc10_2_4	110240000	110250000	10000				Ssc10_2_9	2360000	2370000	10000	OR5P3		
Ssc10_2_4	110800000	110830000	30000				Ssc10_2_9	3710000	3720000	10000	URGCP		
Ssc10_2_4	111060000	111080000	20000				Ssc10_2_9	5210000	5230000	20000			
Ssc10_2_4	111230000	111240000	10000	REG4			Ssc10_2_9	5550000	5560000	10000	OR52A5		
Ssc10_2_4	130770000	130790000	20000				Ssc10_2_9	5890000	5910000	20000	OR51H1P		
Ssc10_2_4	130980000	130990000	10000				Ssc10_2_9	6560000	6570000	10000			
Ssc10_2_4	131000000	131020000	20000				Ssc10_2_9	6580000	6610000	30000	OR52B4		
Ssc10_2_4	132700000	132720000	20000				Ssc10_2_9	6630000	6660000	30000	SSU72	OR52B4	
Ssc10_2_4	132770000	132800000	30000				Ssc10_2_9	6730000	6740000	10000			
Ssc10_2_4	135800000	135990000	190000	FAM69A			Ssc10_2_9	13530000	13570000	40000	KCTD14		
Ssc10_2_4	138410000	138460000	50000				Ssc10_2_9	13610000	13620000	10000			
Ssc10_2_5	21250000	21260000	10000	OR6C6			Ssc10_2_9	13670000	13680000	10000			
Ssc10_2_5	24950000	25030000	80000				Ssc10_2_9	15740000	15750000	10000			
Ssc10_2_5	25060000	25070000	10000				Ssc10_2_9	15950000	16000000	50000			
Ssc10_2_5	25130000	25320000	190000				Ssc10_2_9	19310000	19330000	20000	CCDC90B		
Ssc10_2_5	35880000	35890000	10000	MDM2			Ssc10_2_9	23450000	23750000	300000			
Ssc10_2_5	35900000	35930000	30000				Ssc10_2_9	23910000	23950000	40000			
Ssc10_2_5	93150000	93240000	90000				Ssc10_2_9	24860000	24980000	120000	GRM5		
Ssc10_2_5	93330000	93380000	50000				Ssc10_2_9	25210000	25400000	190000	NOX4		
Ssc10_2_6	30000	40000	10000				Ssc10_2_9	25410000	25430000	20000			
Ssc10_2_6	16030000	16040000	10000				Ssc10_2_9	25510000	25540000	30000			
Ssc10_2_6	16050000	16190000	140000	NOB1			Ssc10_2_9	25560000	25570000	10000			
Ssc10_2_6	16730000	16760000	30000	CIRH1A			Ssc10_2_9	26190000	26380000	190000			
Ssc10_2_6	16930000	16970000	40000				Ssc10_2_9	26440000	26490000	50000			
Ssc10_2_6	19010000	19030000	20000				Ssc10_2_9	26500000	26580000	80000			
Ssc10_2_6	20230000	20260000	30000				Ssc10_2_9	32480000	32520000	40000			
Ssc10_2_6	21210000	21230000	20000				Ssc10_2_9	32530000	32560000	30000			
Ssc10_2_6	22850000	23020000	170000				Ssc10_2_9	32580000	32590000	10000			
Ssc10_2_6	26420000	26460000	40000	MT2A	BBS2		Ssc10_2_9	32780000	32860000	80000			
Ssc10_2_6	38340000	38430000	90000				Ssc10_2_9	33240000	33330000	90000	ENY2		
Ssc10_2_6	39160000	39300000	140000				Ssc10_2_9	33340000	33440000	100000			
Ssc10_2_6	39410000	39440000	30000				Ssc10_2_9	33800000	33830000	30000			
Ssc10_2_6	46230000	46370000	140000	PLAUR	PHLDB3	LYPD3	Ssc10_2_9	34110000	34140000	30000			
Ssc10_2_6	52330000	52440000	110000	DPRX	SLC7A3		Ssc10_2_9	34560000	34570000	10000			
Ssc10_2_6	52750000	52810000	60000	SLC7A3			Ssc10_2_9	35080000	35150000	70000	CNTN5		
Ssc10_2_6	52820000	52830000	10000				Ssc10_2_9	35660000	35670000	10000			
Ssc10_2_6	52840000	52860000	20000				Ssc10_2_9	35930000	35980000	50000			
Ssc10_2_6	54270000	54290000	20000	SYT5	DNAAF3		Ssc10_2_9	35990000	36020000	30000			
Ssc10_2_6	56310000	56400000	90000	OR5R1			Ssc10_2_9	37510000	37860000	350000	DYNC2H1	DCUN1D5	
Ssc10_2_6	56410000	56430000	20000				Ssc10_2_9	43410000	43450000	40000			
Ssc10_2_6	112900000	112920000	20000	FHOD3			Ssc10_2_9	92780000	92900000	20000			
Ssc10_2_6	112940000	112950000	10000				Ssc10_2_9	94860000	94950000	90000	AGR3		
Ssc10_2_6	119350000	119370000	20000				Ssc10_2_9	94960000	94990000	30000			
Ssc10_2_6	156300000	156310000	10000				Ssc10_2_9	95150000	95570000	420000	AHR		
Ssc10_2_7	15550000	15560000	10000				Ssc10_2_9	95830000	95870000	40000			
Ssc10_2_7	15640000	15660000	20000				Ssc10_2_9	96020000	96120000	100000			
Ssc10_2_7	15690000	15700000	10000				Ssc10_2_9	96220000	96380000	160000	ACBD5		
Ssc10_2_7	15840000	15850000	10000				Ssc10_2_9	97150000	97160000	10000			
Ssc10_2_7	18890000	18900000	10000				Ssc10_2_9	97820000	97860000	40000			
Ssc10_2_7	22950000	23000000	50000				Ssc10_2_9	97890000	98210000	320000	TMEM196	TWISTNB	
Ssc10_2_7	23010000	23060000	50000				Ssc10_2_9	98220000	98240000	20000			
Ssc10_2_7	23660000	23710000	50000	OR1F1	OR2G6		Ssc10_2_9	99880000	99940000	60000	DNAH11		
Ssc10_2_7	23870000	23990000	120000	ZNF167	ZNF192	ZNF389	Ssc10_2_9	100050000	100070000	20000			
				ZNF165	ZSCAN16		Ssc10_2_9	119600000	119610000	10000			
Ssc10_2_7	24000000	24260000	260000	ZNF323	ZNF193	NKAPL	Ssc10_2_9	125520000	125530000	10000			
				ZNF397	ZKSCAN4	PGBD1	Ssc10_2_9	142330000	142350000	20000			
				ZSCAN12	OR2B11	GPX5	Ssc10_2_9	142680000	142690000	10000			
				GPX6			Ssc10_2_9	143230000	143260000	30000			
Ssc10_2_7	24270000	24360000	90000	ZNF232									
Ssc10_2_7	24630000	24650000	20000	SLA-3									

Chapter 6

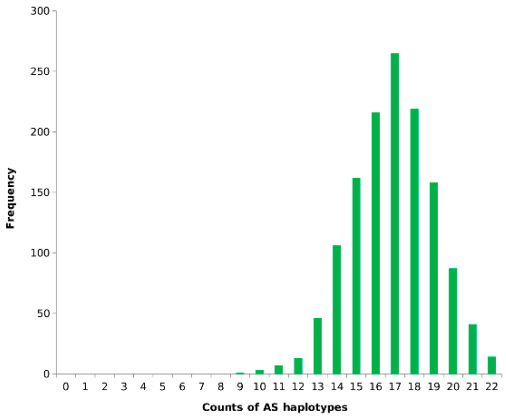


Figure S6.1 Frequency of Asian haplotypes in introgressed regions per individual. The x-axis displays the sum of "C" alleles (=number of Asian haplotypes) that are observed for an individual, summed over all 11 regions of introgression. The y-axis contains the frequency of individuals in the Large White population that are observed to have the associated number of Asian haplotypes.

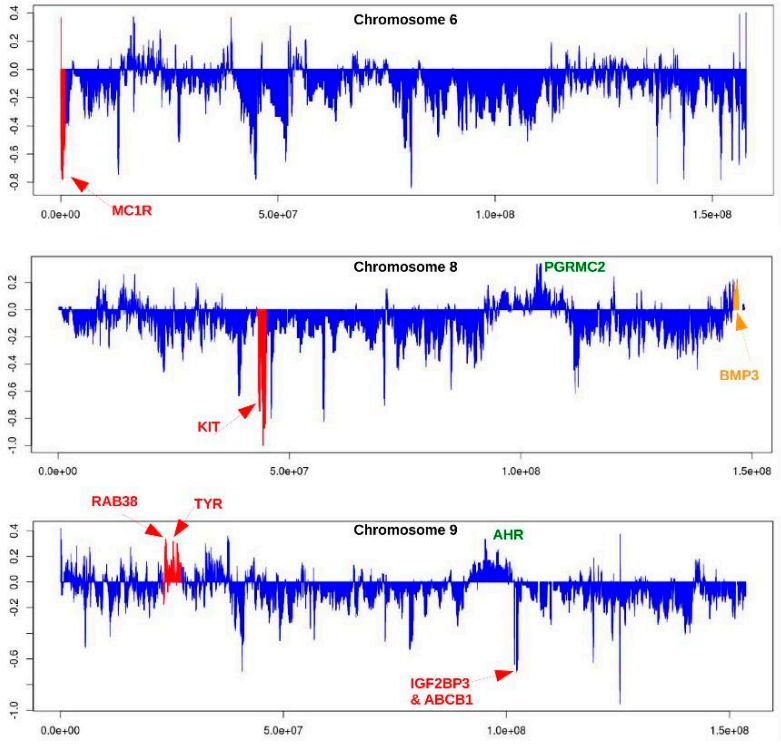


Figure S6.2 Detailed introgression signals of chromosome 6, 8 and 9.

Table S6.1 Regions with average introgression signal ZrIBD >2. Regions with average introgression signal ZrIBD >2. "E" allele stands for a European haplotype and "C" allele stands for an Asian (Chinese) haplotype. Regions indicated in red are discarded based on their allele frequencies. Italicized regions are merged with the neighboring region and regions in bold pased the hardy-weinberg threshold.

region	chromosome	startCore	stopCore	startEx	stopEx	lengthCore	lengthEx	nr_Markers	p	q	o_EE	o_EC	o_CC	e_EE	e_EC	e_CC	chi_sqr
1_1	Ssc10_2_1	32,02	32,31	31,7	32,55	0,29	0,85	29	0,18	0,82	0	631	1132	56	518	1188	84
1_2	Ssc10_2_1	52,11	52,28	52,05	52,35	0,17	0,3	8	0,25	0,75	91	949	1261	139	853	1309	29
1_3	Ssc10_2_1	53,39	53,76	53,15	54,01	0,37	0,86	21	0,24	0,76	29	1016	1224	127	819	1322	130
1_4	Ssc10_2_1	84,55	84,6	84,49	84,65	0,05	0,16	7	0,46	0,54	447	1249	615	497	1149	665	18
1_5	Ssc10_2_1	86,14	86,91	85,97	87,17	0,77	1,2	23	0,35	0,65	131	1182	775	249	944	894	132
1_6	Ssc10_2_1	92,9	93,42	92,36	94,4	0,52	2,04	41	0,60	0,40	448	1055	125	585	781	261	199
15_1	**Ssc10_2_15**	**30,01**	**30,23**	**29,97**	**30,4**	**0,22**	**0,43**	**5**	**0,11**	**0,89**	**13**	**471**	**1792**	**27**	**442**	**1806**	**10**
15_2	Ssc10_2_15	59,95	60,17	59,88	60,25	0,22	0,37	6	0,51	0,49	582	1202	525	606	1154	549	4
15_3	Ssc10_2_15	97,9	98,38	97,39	98,45	0,48	1,06	13	0,41	0,59	139	1582	562	379	1102	802	433
15_4	Ssc10_2_15	124,04	124,2	123,93	124,25	0,16	0,32	7	0,30	0,70	130	1107	1073	202	962	1145	52
18_1	**Ssc10_2_18**	**23,95**	**24,21**	**23,78**	**24,56**	**0,26**	**0,78**	**5**	**0,39**	**0,61**	**329**	**1085**	**830**	**338**	**1066**	**839**	**1**
2_1	Ssc10_2_2	34,55	34,93	34,47	35,05	0,38	0,58	11	0,41	0,59	322	1188	729	375	1082	782	21
2_2	**Ssc10_2_2**	**69,11**	**69,36**	**69,04**	**70,13**	**0,25**	**1,09**	**16**	**0,44**	**0,56**	**386**	**1192**	**676**	**428**	**1108**	**718**	**13**
2_3	Ssc10_2_2	88,57	88,78	88,44	88,92	0,21	0,48	14	0,56	0,44	640	1019	392	644	1011	396	0
2_4	**Ssc10_2_2**	**90,89**	**91,02**	**90,85**	**91,07**	**0,13**	**0,22**	**7**	**0,20**	**0,80**	**74**	**779**	**1447**	**93**	**740**	**1466**	**7**
2_5	*Ssc10_2_2*	*107,03*	*107,33*	*106,97*	*107,4*	*0,3*	*0,43*	*15*	*0,24*	*0,76*	*124*	*877*	*1305*	*137*	*850*	*1318*	*2*
2_6	**Ssc10_2_2**	**107,37**	**107,44**	**107,11**	**107,71**	**0,07**	**0,6**	**19**	**0,32**	**0,68**	**230**	**1027**	**1049**	**240**	**1007**	**1058**	**1**
2_7	**Ssc10_2_2**	**125,13**	**125,36**	**125,12**	**125,59**	**0,23**	**0,47**	**15**	**0,38**	**0,62**	**276**	**1054**	**763**	**308**	**989**	**795**	**9**
2_8	Ssc10_2_2	132,34	132,58	131,74	132,86	0,24	1,12	18	0,41	0,59	157	1329	495	340	961	678	290
3_1	**Ssc10_2_3**	**31,66**	**32,11**	**31,6**	**32,18**	**0,45**	**0,58**	**15**	**0,25**	**0,75**	**130**	**798**	**1189**	**132**	**794**	**1191**	**0**
3_2	Ssc10_2_3	120,4	120,6	120,4	120,66	0,2	0,26	6	0,31	0,69	48	1300	914	215	965	1081	273
6_1	**Ssc10_2_6**	**16,05**	**16,19**	**15,97**	**16,28**	**0,14**	**0,31**	**7**	**0,01**	**0,99**	**0**	**44**	**2198**	**0**	**43**	**2198**	**0**
7_1	Ssc10_2_7	24	24,26	23,36	24,74	0,26	1,38	22	0,60	0,40	734	1027	307	752	989	325	3
7_2	Ssc10_2_7	26,67	26,93	26,62	27	0,26	0,38	10	0,24	0,76	0	1122	1185	136	849	1321	238
8_1	Ssc10_2_8	103,3	103,85	102,63	106,39	0,55	3,76	49	0,52	0,48	76	1946	7	542	1013	473	1722
9_1	**Ssc10_2_9**	**23,45**	**23,75**	**23,4**	**24,07**	**0,3**	**0,67**	**14**	**0,03**	**0,97**	**2**	**105**	**1934**	**1**	**105**	**1933**	**0**
9_2	Ssc10_2_9	37,51	37,86	37,44	37,89	0,35	0,45	62	0,57	0,43	722	1175	389	750	1118	417	6
9_3	**Ssc10_2_9**	**95,15**	**95,57**	**94,47**	**96,26**	**0,42**	**1,79**	**30**	**0,26**	**0,74**	**134**	**818**	**1101**	**143**	**798**	**1110**	**1**
9_4	*Ssc10_2_9*	*96,22*	*96,38*	*96,17*	*96,44*	*0,16*	*0,27*	*5*	*0,28*	*0,72*	*196*	*924*	*1190*	*187*	*941*	*1181*	*1*
9_5	**Ssc10_2_9**	**97,89**	**98,21**	**97,69**	**98,42**	**0,32**	**0,73**	**6**	**0,28**	**0,72**	**197**	**918**	**1189**	**187**	**938**	**1179**	**1**

Table S6.2 Extreme 1% tails of the introgression distribution, based on 1Mb bins over all autosomes.

Lowest 1%						Top 1%					
chr	Mbp	nr genes	rIBD	Region	specialties	chr	Mbp	nr genes	rIBD	Region	specialties
1	134	6	-0,522	*NA*	GLDN, DMXL2	1	54	0	0,141	1_3	*no genes*
1	314	22	-0,417	*NA*	GO: fatty acid synthesis	1	87	11	0,267	1_5	AMD1
2	58	13	-0,426	*NA*	ZNF496, NLRP3	1	94	6	0,234	1_6	ME1
2	59	21	-0,465	*NA*	ZNF496, NLRP3	2	35	1	0,126	2_1	KIF18A; Rubin 2012**
2	161	5	-0,528	*NA*	Groenen 2012* (olfactory?)	2	70	42	0,159	2_2	MANY genes
2	162	14	-0,415	*NA*	Groenen 2012* (olfactory?)	2	90	7	0,118	*(2_3, 2_4)*	Wilkinson 2013***
3	41	43	-0,563	*NA*	Groenen 2012*	2	108	6	0,137	2_5	IRAP
3	42	58	-0,868	*NA*	Groenen 2012*	2	126	3	0,159	2_6	COMMD10
3	43	44	-0,774	*NA*	Groenen 2012*	2	128	0	0,152	*NA*	*no genes*
4	49	0	-0,483	*NA*	*no genes*	2	133	1	0,137	2_7	CSNK1G3,CEP120
4	50	3	-0,445	*NA*		3	32	1	0,126	3_1	SNX29
6	1	22	-0,561	*NA*	MC1R	7	24	128	0,142	7_1	MHC
6	45	30	-0,510	*NA*	Many genes (CYP2 (3x))	7	25	32	0,144	*NA*	MHC
6	52	39	-0,440	*NA*	Many genes	8	104	7	0,173	8_1	PGRMC2
7	86	39	-0,441	*NA*	olfactory?	8	105	0	0,204	8_1	*no genes*
8	45	10	-0,669	*NA*	KIT close	8	106	0	0,123	8_1	*no genes*
13	101	2	-0,655	*NA*	MBNL1	8	147	10	0,127	*NA*	ANTXR2,FGF5,BMP3; Wilkinson 2013***
13	102	1	-0,518	*NA*	MBNL1	9	26	13	0,125	*(9_1, 9_2)*	TYR, NOX4,RAB38 ; Wilkinson 2013***
13	127	10	-0,415	*NA*	PIK3Ca	9	27	0	0,122	*NA*	*no genes*
14	57	0	-0,478	*NA*	*no genes*	9	96	3	0,188	9_3	AHR,SNX13
15	37	7	-0,462	*NA*	PTPN4	9	97	1	0,133	9_4	AHR,SNX13
15	45	5	-0,420	*NA*	ASB5 , spata4	9	99	3	0,136	9_5	TMEM196, TWISTNB
15	65	0	-0,498	*NA*	*no genes*	12	2	21	0,170	*NA*	Many (FASN)
17	15	10	-0,469	*NA*	SLC23A2, ANKRD26	15	31	1	0,138	15_1	ELfe3
18	1	13	-0,441	*NA*	VIPR2	18	25	0	0,157	18_1	POT1

* Groenen et al. 2012 indicate that this region has also been found to be selected in European wild boar

** Rubin et al. 2012 report this region as being under selection in European domestic pigs

*** This region was found to be introgressed and selected in Wilkinson et al. 2013

Chapter 7

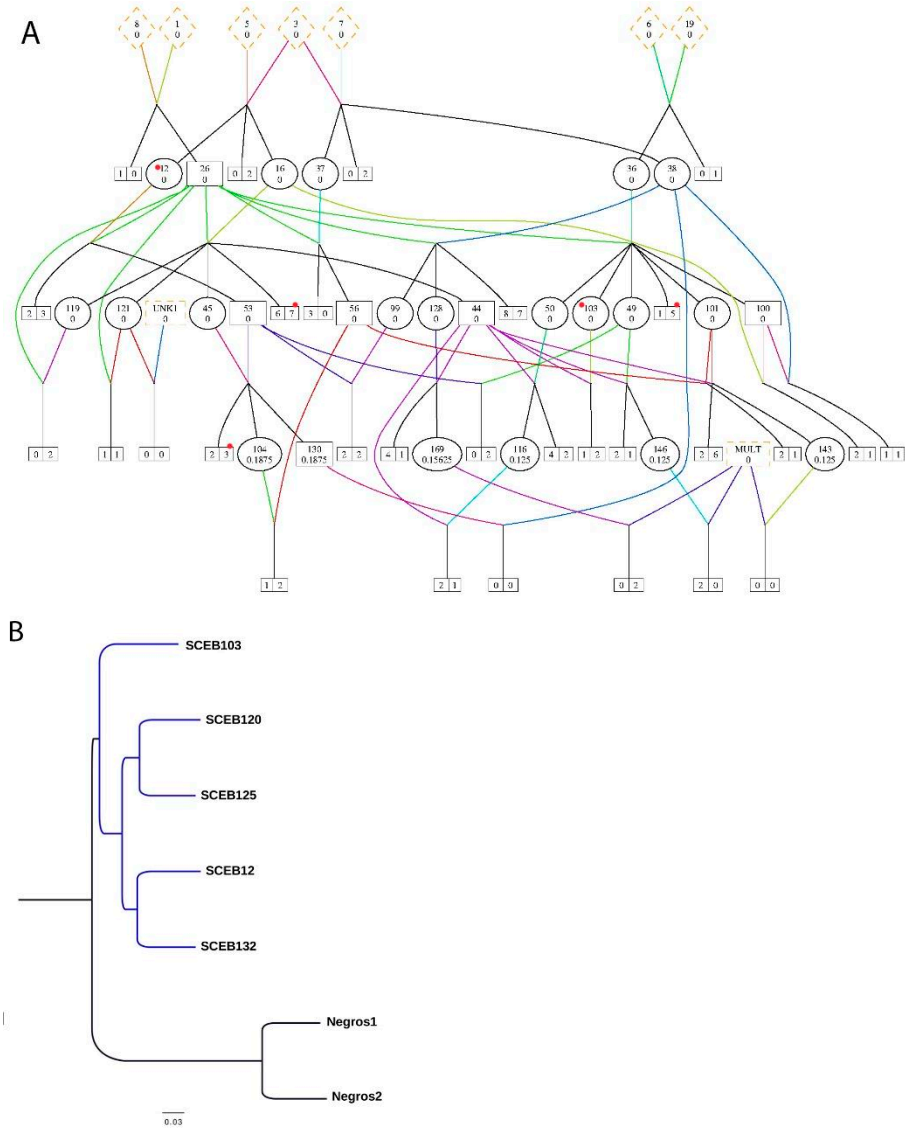


Figure S7.1 Relationship between the S. cebifrons individuals. A. Pedigree of the captive zoo population that the 5 re-sequenced individuals are sampled from. Inbreeding coefficients for breeding individuals are displayed below the number of the individual, and the number of non-breeding offspring from a particular breeding couple is shown within squared boxes. Sampled individuals are indicated with a red dot. **B**. Neighbor-joining phylogenetic tree of the 7 re-sequenced *S. cebifrons* individuals. Individuals highlighted in blue are from the San Diego Zoo and are used for the in silico management.

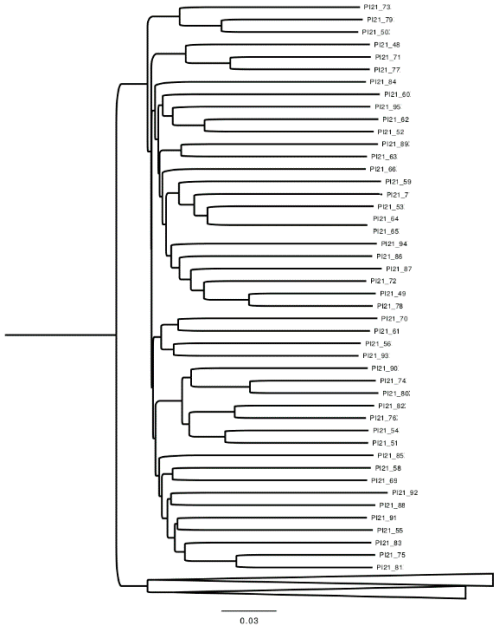


Figure S7.2 Relationship between all Pietrain individuals. Neighbor-joining phylogenetic tree of all Pietrain individuals that were used for the in silico management. We used Large White and Landrace pigs as outgroup.

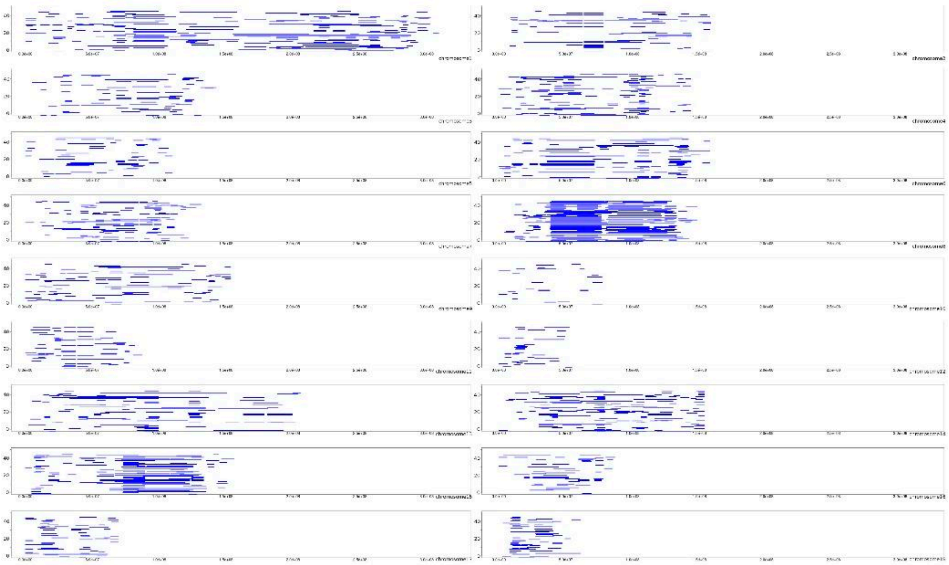


Figure S7.3. Runs of homozygosity in the Pietrain population. ROHs within individual genomes in the Pietrain population before the management are displayed per chromosome. The x-axis displays the full length of each chromosome in bp and individuals are listed on the y-axis so that each line represents the genome of one individual. ROHs are indicated in blue.

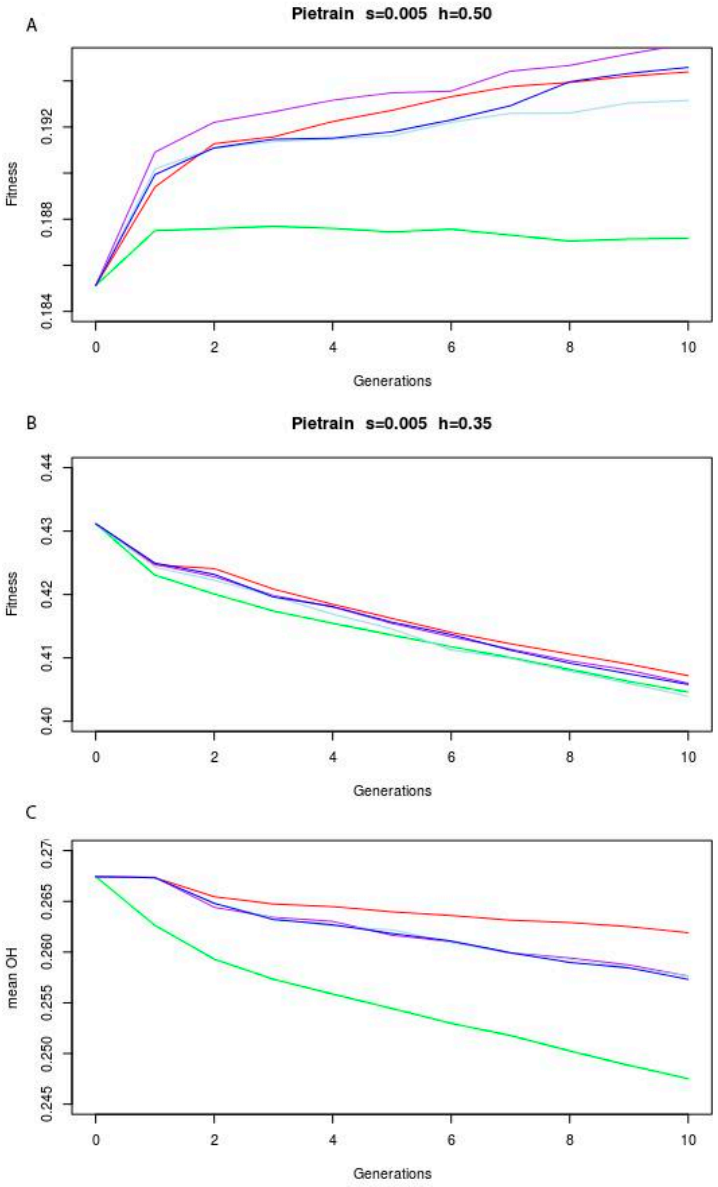


Figure S7.4 Fitness and diversity during management of the Pietrain population. The change in fitness and observed heterozygosity (OH) during 10 generations of management is displayed for 5 different management strategies. **A**. Fitness change over 10 generations of management when a dominance coefficient of 0.5 and selection coefficient of 0.005 is applied. **B**. Fitness change over 10 generations of management when a dominance coefficient of 0.35 and selection coefficient of 0.005 is applied. **C**. Observed heterozygosity during 10 generations of management.

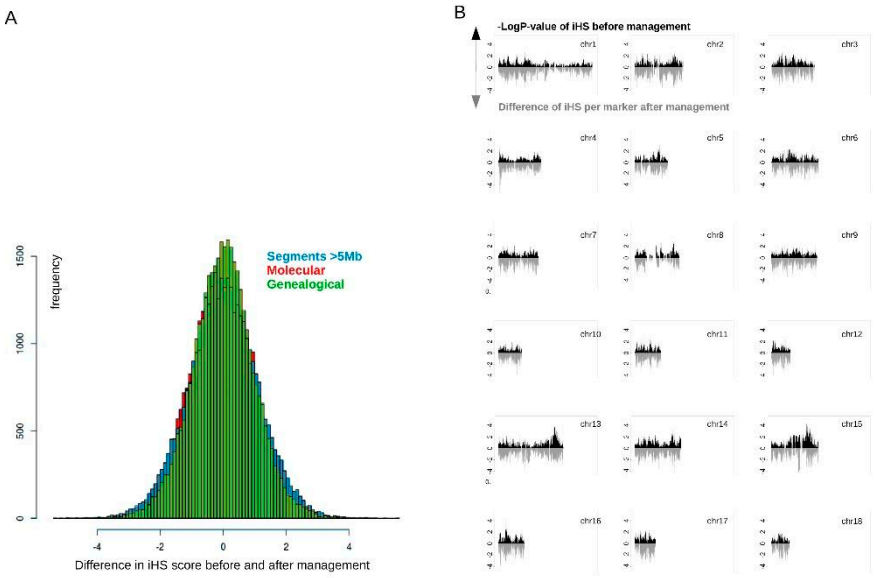


Figure S7.5 Effect of management on selective sweeps in the Pietrain population. Signatures of selection are measured as extended haplotype homozygosity (iHS signal) in the Pietrain population before and after management. **A** displays a histogram of the difference in iHS signal for each marker before and after 5Mb segment-based, molecular-based and genealogical-based management. **B** Genome-wide view of the correlation between the p-value of the iHS signal before management, and the magnitude of difference between iHS signal before and after management. The physical location on the chromosome in bp for each marker is indicated on the x-axis. The significance levels of the iHS signal before management are indicated in black, and displayed on the y-axis (-logp) and range between 0 and 4, so that markers with a signal >2 are considered to be significant. Differences in iHS signal per marker are indicated in grey and range from 0 to -5, with a strong negative number indicating a large difference.

Table S7.1 GO-enrichment analysis of genes containing deleterious variants. List of gene-ontology enrichment terms for Biological Process, Molecular Function and Protein Class. Those GO-terms are listed that were over- or under-represented in the list of genes that contained deleterious variants in the 5 re-sequenced S. cebifrons pigs and in the 11 re-sequenced Pietrain pigs.

Cebifrons					
Biological Process	**REFLIST**	**observed**	**expected**	**over/under**	**P-value**
RNA metabolic process	2008	190	252.93	-	0.00187
cell adhesion	735	131	92.58	+	0.017
biological adhesion	735	131	92.58	+	0.017
nucleobase-containing compound metabolic process	2982	311	375.62	-	0.0257
primary metabolic process	6302	711	793.82	-	0.0301
transcription, DNA-dependent	1579	151	198.9	-	0.0325
metabolic process	7531	864	948.63	-	0.0365
regulation of biological process	1760	172	221.7	-	0.0399
transcription from RNA polymerase II promoter	1569	151	197.64	-	0.046
response to external stimulus	201	44	25.32	+	0.11
regulation of transcription from RNA polymerase II promoter	1222	117	153.93	-	0.201
fertilization	89	23	11.21	+	0.321
anatomical structure morphogenesis	601	103	75.7	+	0.343
regulation of nucleobase-containing compound metabolic process	1319	130	166.15	-	0.376
sensory perception of sound	64	18	8.06	+	0.423
regulation of catalytic activity	983	157	123.82	+	0.434
cellular component morphogenesis	560	96	70.54	+	0.486
regulation of molecular function	993	158	125.08	+	0.486
blood coagulation	162	35	20.41	+	0.488
cell-matrix adhesion	104	25	13.1	+	0.533
Unclassified	6756	917	851.01	+	0.553
lipid transport	264	51	33.25	+	0.588
regulation of translation	119	5	14.99	-	0.687
nucleobase-containing compound transport	104	4	13.1	-	0.853
defense response to bacterium	24	9	3.02	+	0.991

Pietrain					
Biological Process	**REFLIST**	**observed**	**expected**	**over/under**	**P-value**
transcription from RNA polymerase II promoter	1579	125	174.1	-	0.00693
transcription, DNA-dependent	1589	126	175.2	-	0.00705
regulation of biological process	1786	145	196.92	-	0.00724
primary metabolic process	6371	624	702.46	-	0.0272
RNA metabolic process	2019	172	222.61	-	0.0284
nucleobase-containing compound metabolic process	3003	271	331.11	-	0.0293
metabolic process	7612	759	839.29	-	0.0326
skeletal system development	301	14	33.19	-	0.0331
Unclassified	6802	823	749.98	+	0.102
regulation of transcription from RNA polymerase II promoter	1229	99	135.51	-	0.104
regulation of nucleobase-containing compound metabolic process	1330	109	146.64	-	0.115
pattern specification process	210	9	23.15	-	0.173
antigen processing and presentation	58	16	6.4	+	0.241
segment specification	146	5	16.1	-	0.312
cellular component morphogenesis	576	88	63.51	+	0.441
anatomical structure morphogenesis	617	93	68.03	+	0.489
cell adhesion	770	112	84.9	+	0.57
biological adhesion	770	112	84.9	+	0.57

Cebifrons

Molecular Function	REFLIST	observed	expected	over/under	P-value
nucleic acid binding	2812	285	354.21	-	0.00346
binding	5102	568	642.66	-	0.036
Unclassified	7852	1071	989.06	+	0.0419
DNA binding	1666	165	209.85	-	0.0722
RNA binding	533	42	67.14	-	0.0927
GTPase activity	230	13	28.97	-	0.112
motor activity	85	22	10.71	+	0.254
translation factor activity, nucleic acid binding	122	5	15.37	-	0.337
nucleic acid binding transcription factor activity	1410	142	177.61	-	0.378
enzyme regulator activity	956	152	120.42	+	0.386
translation initiation factor activity	94	3	11.84	-	0.407
structural constituent of cytoskeleton	711	117	89.56	+	0.412
RNA helicase activity	79	2	9.95	-	0.455
sequence-specific DNA binding transcription factor activity	1401	142	176.47	-	0.492
translation regulator activity	117	5	14.74	-	0.524
lipid transporter activity	63	17	7.94	+	0.54
pyrophosphatase activity	213	42	26.83	+	0.611
serine-type peptidase activity	240	46	30.23	+	0.683

Pietrain

Molecular Function	REFLIST	observed	expected	over/under	P-value
DNA binding	1666	122	185.88	-	0.000019
nucleic acid binding	2812	240	313.74	-	0.000251
nucleic acid binding transcription factor activity	1410	107	157.31	-	0.00104
sequence-specific DNA binding transcription factor activity	1401	107	156.31	-	0.00149
Unclassified	7852	953	876.05	+	0.0433
receptor activity	1340	191	149.5	+	0.0606
binding	5102	507	569.23	-	0.155
serine-type endopeptidase inhibitor activity	82	19	9.15	+	0.451
structural molecule activity	1088	152	121.39	+	0.498
peptidase inhibitor activity	171	32	19.08	+	0.65
voltage-gated ion channel activity	137	6	15.29	-	0.995

Pietrain

PANTHER Protein Class	REFLIST	observed	expected	over/under	P-value
defense/immunity protein	457	93	50.99	+	0.00000943
nucleic acid binding	2248	185	250.81	-	0.000396
transcription factor	1427	109	159.21	-	0.00138
homeobox transcription factor	189	6	21.09	-	0.0197
helix-turn-helix transcription factor	189	6	21.09	-	0.0197
cell adhesion molecule	445	77	49.65	+	0.0285
DNA binding protein	768	55	85.69	-	0.0363
Unclassified	7069	864	788.69	+	0.0552
immunoglobulin receptor superfamily	127	29	14.17	+	0.0614
receptor	1354	192	151.07	+	0.0843
small GTPase	120	3	13.39	-	0.135
cytokine receptor	209	40	23.32	+	0.175
histone	53	0	5.91	-	0.482
extracellular matrix linker protein	21	8	2.34	+	0.517
antibacterial response protein	102	22	11.38	+	0.585
protease inhibitor	171	32	19.08	+	0.731
voltage-gated ion channel	142	6	15.84	-	0.778

Cebifrons

PANTHER Protein Class	REFLIST	observed	expected	over/under	P-value
cell adhesion molecule	445	96	56.05	+	0.0000969
nucleic acid binding	2248	213	283.17	-	0.000384
small GTPase	120	2	15.12	-	0.00615
extracellular matrix glycoprotein	115	30	14.49	+	0.0407
G-protein	197	9	24.81	-	0.0425
RNA binding protein	879	79	110.72	-	0.131
intermediate filament	69	19	8.69	+	0.292
transcription factor	1427	143	179.75	-	0.332
Unclassified	7069	958	890.43	+	0.343
cytoskeletal protein	727	120	91.58	+	0.373
homeobox transcription factor	189	11	23.81	-	0.488
helix-turn-helix transcription factor	189	11	23.81	-	0.488
RNA helicase	79	2	9.95	-	0.511
actin binding motor protein	42	13	5.29	+	0.576
extracellular matrix protein	394	70	49.63	+	0.599
ATP-binding cassette (ABC) transporter	58	16	7.31	+	0.64
DNA binding protein	768	72	96.74	-	0.801

Colophon

This work was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) / ERC Grant agreement no 249894 (SelSweep project).

The cover of this thesis was designed by Mirte Bosse and Dennis T.S. van de Water (dvdwphotography.com).

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