Nucleotide variation and footprints of selection in the porcine and chicken genomes

Andreia J. Amaral

Thesis committee

Thesis supervisor

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

Thesis co-supervisors

Dr. Ir. Hendrik-Jan Megens
Researcher
Wageningen University

Dr. Ir. Henri C.M. Heuven
Assistant Professor
Faculty of Veterinary Medicine
Utrecht University

Other members

Prof. Dr. Larry B. Schook
University of Illinois (USA)

Prof. Dr. Richard G.F. Visser
Wageningen University

Dr. Greger Larson
Durham University (UK)

Prof. Dr. Ir. Francine P.M. Govers
Wageningen University

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

Nucleotide variation and footprints of selection in the porcine and chicken genomes

Andreia J. Amaral

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 28 May 2010

at 1:30 p.m. in the Aula.

Andreia J. Amaral

Nucleotide variation and footprints of selection in the porcine and chicken genomes, 164 pages.

PhD-thesis, Wageningen University, Wageningen, the Netherlands (2010) with summaries in English and Dutch.

ISBN 978-90-8585-655-9

Para o Filipe

"Lucky to be in love with my best friend,

Lucky to have been where I have been

Lucky to be coming home…."

(Jason Miraz)

Abstract

The aim of the research presented in this thesis was the characterization of genetic diversity of the porcine and chicken genomes and to infer footprints of artificial selection. To this purpose, the extent of linkage disequilibrium (LD) was evaluated in ten European pig breeds, in ten Chinese pig breeds and in the European wild boar (Chapter 2). Results showed a contrasting difference between European and Chinese breeds. Chinese breeds showed levels of extent of LD similar to human populations whereas European pig breeds and the European wild boar showed levels of LD similar to other domesticated species. These results furthermore showed that in order to characterize genetic diversity in European pig breeds a minimum of 30,000 informative SNPs would be required. A cost-effective assay was performed to make genome-wide identification of SNPs in the porcine genome using pooled DNA and massive parallel sequencing (chapter 3). First, this experiment allowed the identification of ~17,000 informative SNPs with a success rate of 96%. Second, it showed that this strategy could also be used to characterize nucleotide variation along whole genomes. In chapters 4 and 5, the variation of nucleotide diversity and the identification of footprints of selection were assessed in the porcine and chicken genomes respectively. Results show that selection has most affected different genes related with similar pathways. As expected, results suggest that pig breeds have been positively selected for growth and muscle development. Interestingly, results showed that genes encoding for olfactory receptors are over represented in regions putatively under balancing selection. In chicken lines, results suggest as expected, contrasting differences between layers (egg production) and broiler (meat production) lines. The layer lines showed the lowest levels of nucleotide diversity. Genes related with growth and muscle development were most represented in regions under positive selection in the broiler lines. Genes related with folate biosynthesis important for reproduction were mostly observed in the positively selected regions.

Contents

General Introduction 1

Domestication and subsequent improvement of animals and plants by exploiting existing variation through artificial selection has been a key element to the success of humans as a species, giving humans a leading role in the evolutionary process (Roots 2007). Manipulation of gene pools of domesticated species caused tremendous phenotypic changes and created a large variety of domesticated breeds based on selection of superior phenotypes. Consequently agricultural food production became more efficient, allowing the human population to grow from an estimated 10 million in Neolithic (Harris 1996) to 6.8 billion today and still expanding.

Over the last decades, advances in molecular biology and the rapid development of DNA marker technology have increased knowledge about the genetic variation of domesticated species, including the genetic basis of phenotypic differences (Andersson & Georges 2004). Recent developments in molecular biological techniques, in particular the development of next generation sequencing technologies will allow the analysis of variation at the level of the whole genome and the availability of the sequence of the entire genome of many individuals will be a reality in the near future. Why is there such an interest for the understanding of genetic variation in domesticated species? First of all, studies of genetic variation of domesticated genomes allow answering complex questions about origins and evolution of domesticated species (Götherström et al. 2005; Fang et al. 2009; Larson et al. 2007; Eriksson et al. 2008; Larson et al. 2005; Le Rouzic & Carlborg 2008). In addition, the availability of DNA marker data in the past two decades, has helped to develop tools to perform rational decisions regarding animal selection and breeding (Dekkers 2004), and regarding the conservation of domesticated breeds at risk of extinction (Toro et al. 2009). Finally, no other animals have had their phenotypes monitored as closely as the principal animal domesticated species, making them a prime model to understand the genetic basis of phenotypic diversity. This was noted already by Charles Darwin, who made observations and performed experiments on animal breeding (Darwin 1868).

This thesis focuses on the effect of domestication and selection on the genetic variation of important domesticated species such as chicken and pig. Chicken and pig are important food sources in the world and important animal models for the study of domestication because their wild counterparts still exist (Giuffra et al. 2000; Liu et al. 2006). Furthermore, chicken and pig are widely used as a model for biomedical research and in areas covering developmental biology, virology, and immunology (Burt 2005; Lunney 2007). Because of the similarity in size, physiology, and in organ development and disease progression (Lunney 2007), the pig has been used as a mammalian model for human as well as possible donor of replacement organs through xenotransplantation.

Domestication and selective breeding have created marked phenotypic changes and genetic adaptation to different environments. Domesticated genomes were enriched for mutations that affect specific phenotypic traits. Some of these traits, such as coat color, follow a simple mendelian inheritance model, but most, such as growth, fertility and behavior, have a complex genetic basis. Because selective pressures cause an aberrant spectrum of nucleotide diversity and affect haplotype diversity, this thesis focuses on the characterization of linkage disequilibrium and on the identification of outlier regions deviating from the average observed variation under neutrality. Such regions, once identified, may help to unravel the genetic basis of complex traits (Nielsen & Bustamante 2005; Voight et al. 2006).

Next a more detailed explanation about the history of the species in study is presented as well as a description about the methods used.

Pig domestication and breeding

European and Chinese pigs were domesticated independently from European and Asian subspecies of wild boar (Giuffra et al. 2000; Larson et al. 2005). The earliest pigs reared in Europe in the Neolithic, however, were derived from the pigs domesticated in Asia Minor, later to be replaced gradually – at least in part – by pigs derived from local wild boar (Larson et al. 2007). The domestication of the pig generated a wide diversity of breeds throughout Eurasia (70% of total pig breed diversity (Scherf 2000) with sizes, shapes and colors that vary significantly from the wild conspecifics (Porter & Tebbit 1993). Furthermore, the domestication of the pig was based on a complex process of selection leading to new genotypes that were suited to the different agricultural environments. Traits selected during domestication in pigs included coat color, which can be traced back to 5,000 years ago (Fang et al. 2009) and size (Diamond 2002). Behavior also would have been modified since life in captivity required changes in social structure and interactions with humans (Price 1999). According to the United Nations Food and Agriculture Organization (FAO) (Scherf 2000), there are currently 228 breeds in Europe, plus 105 that are now extinct; in China, 118 existing breeds are listed, plus 10 now extinct. In both China and Europe, the pig remains a major meat producing species. However, the development of pig breeding practices has been very different in these two areas over the last two centuries.

In Europe, with the development of selection programs, pig breeding became decreasingly local. Breeders started importing genetic material from elsewhere in Europe and even imported pigs from China (Giuffra et al. 2000). As a consequence many local or traditional breeds are now extinct. Current commercial breeds are derived mostly from a small number of breeds generated during the 18th century in North-Western Europe and the British islands (Porter & Tebbit 1993). Performance for growth, carcass traits and reproduction were drastically altered since the nineteen sixties.

The main commercialized breeds today are Pietrain, Large White, Landrace and Duroc. The Large White breed was developed in England in the 18th century and came into prominence in the 19th century. The breed can withstand a wide range of climatic conditions, have a large size of the eye muscle and grows rapidly. Large White sows are appreciated for their large litters, are robust and have a long productive life. The Landrace breed was developed in Denmark by crossing local pigs with the Large White. The Landrace was noted for its superior milk production, excellent maternal behavior reflected in the higher weight at weaning of their litters. The Pietrain breed is originally from Belgium and it is appreciated by the muscular legs and extremely high proportion of lean meat to fat. The exact origin of the Duroc breed is unknown but the modern Duroc is derived from crosses between 'red breeds' developed in the US around 1900 and it is valued for bacon production (Porter & Tebbit 1993).

Conversely, in China traditional local breeding has been applied until more recent times, although this has been changing rapidly in the last decades. New synthetic Chinese breeds have been created, both from local breeds as well as using Western commercial pigs. Nevertheless, China remains the country that has the largest diversity of pig breeds (118 described) (Scherf 2000).

Pig breeding had important developments since the introduction of DNA marker technology such as the discovery of the malignant mutation in the halothane gene and the consequent higher efficiency through the application of molecular testing in breeding practices (Fujii et al. 1991; Houde et al. 1993). The development of a pig linkage map (Yerle et al. 1995) allowed the development of more accurate linkage and association studies (Van Laere et al. 2003) which provided knowledge about the genetic basis of important traits. Currently a joined effort to generate the first draft of the pig genome (Schook et al. 2005), is in place. The future availability of a large set of molecular markers and the identification of genomic regions affected by selection would allow creating strategies for selective breeding that will incorporate genetic diversity into breeding lines, and to develop strategies to prevent the extinction of more local breeds.

Chicken domestication and breeding

Current domestic chicken diversity can be traced back to at least four independent domestication events in South and Southeast Asia (Liu et al. 2006). While all chicken appear to be derived from a single species (*Gallus gallus*, the Red Jungle Fowl – (Fumihito et al. 1996; Fumihito et al. 1994)), it has been shown that some traits – such as the yellow skin trait – can actually be traced back to a different species of Jungle Fowl (Eriksson et al. 2008). The chicken has unique characteristics which resulted from a process of natural selection and which favored its domestication about 5,000 years ago (Zeuner 1963). These characteristics include the social structure that allowed large social groups of males with multiple females; early maturity age, flexible dietary habits, limited agility and adaptation to a wide range of environments. Throughout the common path of humans and chicken, there is a long history of breeding chickens for sports and for religious ceremonials. Also a large variety of multi-purpose breeds was generated, mainly selected for growth, body weight and egg laying. However in the beginning of the 1950s modern poultry breeding emerged resulting in specialized industrial chicken breeds selected intensively for either meat-type (broiler) or egg-type (layer) chickens due to the negative pleiotropic effects between growth and fertility. Specialized egg layers are classified into white egg layers (WEL) and brown egg layers (BEL). WEL are derived from the White Leghorn breed whereas BEL were mostly derived from North American dual-purpose breeds, such as Rhode Island Red and White Plymouth Rock, which originated from crosses between European and Asian breeds (Muir et al. 2008). Today's commercial chicken lines have suffered from two bottlenecks that created a reduction of genetic diversity. The first resulted from the limited number of breeds used to create the current commercial broiler and layer lines. The second was due to the breeding structure and within line selection. Gene flow does not occur from non-commercial to commercial lines and the breeding systems are such that at the pure line level only a relatively small number of individuals is selected for the critical production traits. This means that a single nucleus of layers or broilers is expanded through three-or-four-way crosses to millions of offspring to produce eggs or meat. Inbreeding is often controlled by making crosses between elite lines ('open lines') for the same or different trait after genetic improvement for a trait within a line for three or four generations (Muir et al. 2008). However, as the number of distinct commercial elite lines has been in steady decline over the past decades due to consolidation of breeding companies, future chicken breeding can rely less and less on the practice of merging lines for the purpose of alleviating inbreeding levels.

Chicken genomics has seen several important landmarks in the last decade with the availability of a high quality draft genome sequence (International Chicken Genome Sequencing Consortium 2004), and a high-density SNP map (International Chicken polymorphism Map Consortium 2004) which makes it possible to study how selection has affected the genome since the time chickens were first domesticated. Examples are studies that show a significant loss of genetic diversity in commercial pure lines of chicken while compared with Red Jungle Fowl and non-commercial lines (Muir et al. 2008), the unexpected reduced variation on the Z chromosome (Sundstrom et al. 2004) and the yellow skin allele, a mutation only found in domesticated chickens (Eriksson et al. 2008).

Linkage disequilibrium

Linkage disequilibrium (LD) is defined as non-random association of alleles at two or more loci in a population (Falconer & Mackay 1996). The extent of LD in populations is shaped by several forces, in particular mutation and recombination. LD is created when a new mutation occurs in a chromosome that carries a particular allele at a nearby locus, and is gradually eroded by recombination. Other molecular, demographic and evolutionary forces have also a significant effect on LD. In general the increase of genetic drift, inbreeding, population structure, hitchhiking and selection all contribute to the increase of LD while population growth diminishes it. LD also differs along the genome, being higher in non-recombinant areas (Ardlie et al. 2002).

The extent of LD has been extensively studied in humans where it shows to vary accross human populations and genomic regions (Reich et al. 2001). The extent of LD has also been studied in domestic animals e.g. cattle (Farnir et al. 2000; The Bovine HapMap Consortium 2009), sheep (McRae et al. 2002), pigs (Nsengimana et al. 2004; Du et al. 2007) and chicken (Aerts et al. 2007; Megens et al. 2009) for which higher levels of LD have been found compared to humans.

Several measures have been developed to quantify LD (reviewed by Devlin & Risch 1995). The most used are *D′* (Lewontin 1964) and r^2 (Hill & Robertson 1968). The measure *D′* is strongly inflated in studies of small sample sizes and using SNPs with rare alleles. The measure r^2 has shown to be very useful for biallelic markers and compared to *D′* it is less inflated in small sample sizes (Weiss & Clark 2002). The extent of LD in pigs has never been characterized using a high density of SNP markers and a comparison between European and Chinese pig breeds is lacking. The possibility of using large numbers of single nucleotide polymorphisms (SNPs) could provide the sufficient resolution to detect islands of high or low LD. Such aberrant patterns may be associated with introgression and or phenotypic selection. Information about the extent of useful LD (Kruglyak 1999) could provide information about sample sizes and amount of markers required for all sorts of mapping studies, for instance to fine map genes responsible for common diseases and phenotypic traits (Zhang et al. 2002).

Characterization of nucleotide diversity and identification of outliers

Population genetics is concerned with estimating the amount of genetic variation in populations and explaining how selective processes change the patterns of genetic variation. In this thesis, data that allows the characterization of genetic variation within species at the sequence level was generated and inferences of non-neutral processes were made in order to identify regions with an aberrant level of variation. In the next section, a description of several statistics to characterize genetic variation is presented and forces that shape genetic variation are described.

Measuring genetic variation within populations

Three historically important summary statistics that measure genetic variation at the nucleotide level are: (1) the number of segregating sites in a sample, (2) the average number of pairwise differences and (3) the scaled mutation rate (θ) (Hamilton 2009). The simplest measure is the number of segregating sites (*S*). A segregating site is any of the L nucleotide sites that maintains two or more nucleotides within a population, such as sites 2,6 and 8 in Figure 1.

The total number of segregating sites can be expressed as the number of segregating sites per nucleotide, p_s, by dividing *S* by the total number of sites *L*. The frequency of DNA sequences within a given nucleotide at a site does not influence *S*, but *S* will increase as the number of individuals sampled increases, because DNA sequences with additional polymorphisms will be added to the sample. The average number of pairwise nucleotide differences between sequences,Π, can be estimated by;

$$\Pi = \frac{1}{[n(n-1)/2]} \sum_{i<j} \Pi_{ij}$$

where n is the number of sequences in the sample (so that n(n-1)/2 is the number of pairwise comparisons) and Π_{ij} is the difference between the *i*th and the *j*th sequences (see example in Figure 1). This number can be standardized by dividing by *L.* This new measure Π/L is also known as the nucleotide diversity or π. The third and one of the most important measures of genetic diversity within populations is given by θ, which is equal to $4N_e\mu$ (Watterson 1975), where Ne is the effective population size and μ is the mutation rate. The scaled mutation rate (θ) describes the expected amount of variation at each nucleotide site if evolution is entirely neutral and can actually be used to test the neutral theory. Under the infinite-site model and under random mating, Watterson (1975) showed that θ can be estimated from the number of segregating sites in a sample of DNA sequences. If we define a new variable, $a_i = \sum_{k=1}^{n-1} \frac{1}{k}$ then, $\theta = \frac{S}{a_i}$ using the absolute number of segregating sites, or $\theta = \frac{p_s}{a_i}$, using the number of segregating sites per nucleotide sampled. An example is shown in Figure 1.

The neutral theory of molecular evolution and the detection of outliers

The neutral theory of molecular evolution assumes that most if not all mutations at low frequency in the population, do not affect the fitness of individuals (Kimura 1985). Those mutations that are deleterious will largely be kept at low frequency or lost and the most common polymorphisms found in the population are selectively neutral. A second assertion of the neutral theory is that most evolutionary changes are the result of genetic drift acting on neutral alleles. According to the neutral theory of molecular evolution, both the variation within and between populations are mainly due to neutral mutations. Using the coalescent model, the observed nucleotide diversity may be compared with the expected nucleotide diversity under the neutral model of evolution, in order to identify outlier regions where aberrant levels of nucleotide diversity are observed (Wakeley 2008).

Nucleotide diversity (π):

1	A	A	T	G	T	C	A	A	C	G	d_{12}=0
2	A	A	T	G	T	C	A	A	C	G	
1	A	A	T	G	T	C	A	A	C	G	d_{13}=1
3	A	T	T	G	T	C	A	A	C	G	
1	A	A	T	G	T	C	A	A	C	G	d_{14}=3
4	A	T	T	G	T	G	A	T	C	G	
2	A	A	T	G	T	C	A	A	C	G	d_{23}=1
3	A	T	T	G	T	C	A	A	C	G	
2	A	A	T	G	T	C	A	A	C	G	d_{24}=3
4	A	T	T	G	T	G	A	T	C	G	
3	A	T	T	G	T	C	A	A	C	G	d_{34}=2
4	A	T	T	G	T	G	A	T	C	G	

$\sum d_{ij}=d_{12}+d_{13}+d_{14}+d_{23}+d_{24}+d_{34}$=0+1+3+1+3+2=10 differences

Number of pairs of sequences compared=[n(n-1)]/2=6

Π=10/6= 1.67 average pairwise differences

π=1.67/10=0.167 pairwise differences per site

Segregating sites (*S* and *p*$_s$)

Sites 2,6, and 8 have variable base pairs among the four sequences (columns in shade).

These are segregating sites. Therefore, for these sequences, S=3 and p_s =3/10=0.3 segregating sites per nucleotide site examined.

Scaled mutation rate (θ)

$\theta=0.3/((1/3)+(1/2)+(1/1))=0.164$

Figure 1- Estimation of genetic diversity. Practical example using a sample of four hypothetical DNA sequences with 10 nucleotides each adapted from (Hamilton 2009).

The unusual values of nucleotide diversity in outlier regions can be interpreted as footprints of selection. Selection can be classified in positive or directional selection, balancing selection and purifying or background selection (Figure 2) (Volis 2008). Positive selection produces a decrease of variability in the affected region (selective sweep). The sweep causes a higher frequency of derived alleles and extended haplotype homozygosity. With the increase in time, new mutations are generated, resulting in an excess of low frequency alleles (Figure 2b). Around the affected region, an increase in the level of linkage disequilibrium is observed. When balancing selection occurs many alleles are maintained in the population for long periods of time (Figure 2c), an increase of variability is observed along with an increase of intermediate allele frequencies. The level of linkage disequilibrium around the selected site also increases. Purifying selection occurs when weak deleterious mutations are eliminated from the populations (Figure 2c) through a long process, during which the allele frequency of the derived allele decreases, resulting in an excess of low frequency alleles and in a decrease of linkage disequilibrium. Nevertheless, the interpretation of results is not trivial, since demographic processes can mimic selection. The occurrence of population growth can also result in a high occurrence of low-frequency alleles and population subdivision also results in the high occurrence of intermediate-frequency alleles.

Another approach for the identification of outlier regions is estimating the degree of differentiation between sub-populations. Outlier regions can be identified in the extreme tails of the empirical distribution of the fixation index (F_{st}, Wright 1984) which, measures the degree of differentiation that has occurred among sub-populations (Volis 2008).

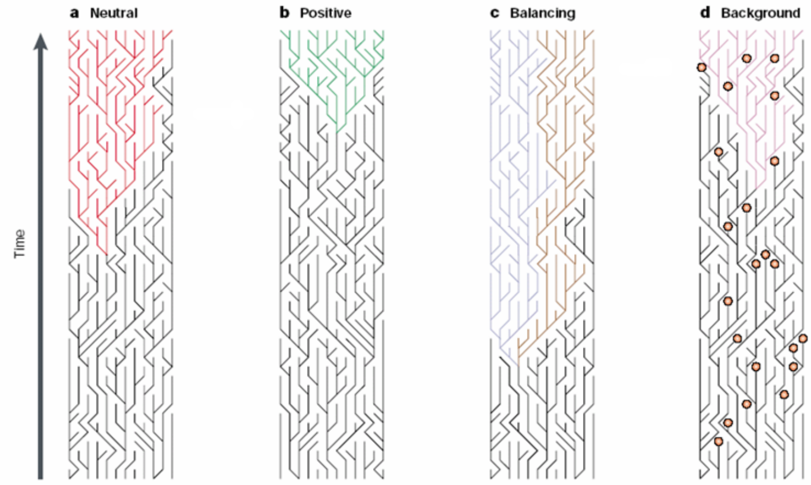


Figure 2 – Models of selection. (a) shows the evolution under the neutral model; (b) effect of positive selection; (c) effect of balancing selection; (d) effect of background or purifying selection (Adapted from Awadalla and Hobolth, 2009).

Massive parallel sequencing technologies

Many new sequencing technologies, called massive parallel sequencing (MPS) are currently available and are being used at present. MPS technologies outperform Sanger sequencing in terms of amount of generated sequence data per monetary unit, dramatically decreasing the sequencing costs of a complete genome (Bentley 2006). Several MPS platforms are available, which current output ranges from 50 to 400 bp. The Roche (454) platform (http://www.454.com/products-solutions/system-features.asp) uses pyrosequencing and generates 1 million sequences which length can reach 400bp. The Illumina system (http://www.illumina.com/downloads/SQ_GAIIx_spec_sheet2_04_09LR.pdf) uses sequencing by synthesis and generates ~200 million 75-100 bp sequences. The SOLID system (http://www3.appliedbiosystems.com/AB_Home/applicationtechnologies/SOLIDsystemSequencing/overviewofsolidsystem/index.htm), uses a ligation based sequencing method, generating 400 million 50 bp sequences. All the MPS platforms provide quality measures for each generated sequence. However, the length of the obtained sequences is still shorter than in Sanger sequencing, thus creating challenges in terms of downstream analysis of the data. Challenges arise at different levels of the analysis of sequences. (a) Sequence quality; error rates are still at least one magnitude higher compared to traditional Sanger sequencing. (b) Ambiguous alignments; due to the short length of the sequences, ambiguous alignments are frequent and cannot be considered for downstream analysis. As the read length is increasing and paired-end technologies are becoming commonplace, alignment ambiguity will decrease. (c) Identification of polymorphisms and assembly; several algorithms for MPS data have been developed, however different algorithms still provide different outcomes. Nevertheless, the potential for the exponential increase of information about genetic variation that can be obtained using of MPS data is immense.

Thesis Outline

This thesis focused in the characterization of genetic variation and detection of selection in chicken and pig. In Chapter 2, SNPs derived from BAC sequences were used to study the linkage disequilibrium in 10 European pig breeds, 10 Chinese pig breeds and European wild boar. European breeds and European wild boar, showed the highest values of linkage disequilibrium at larger genomic distances. In chapter 2 the amount of SNPs required to characterize genetic variation in European pigs on a genome-wide scale was estimated in ~30,000 SNPs. These results showed that the increase of available genetic markers was crucial for future studies. While MPS offers an unprecedented amount of sequence data and the power to rapidly increase the available information on genetic variation, it also brings an increase in bioinformatics challenges in order to extract relevant biological information. Bioinformatics challenges exist in several steps of the data analysis: data filtering, sequence alignment and extraction of relevant output information. As part of a sequence run there are sequences which cannot be mapped (which may be contaminants) and sequences with dubious quality (low quality scores and Ns, i.e. ambiguous bases, called in the sequence). The removal of such reads will decrease the noise during the alignment and will accelerate subsequent downstream analysis and therefore should be removed. In Chapter 3 different data filtering strategies of MPS data are compared in order to maximize the genome-wide identification of SNPs using an approach that combines DNA pooling with a reduced representation of the pig genome. The experimental approach presented here allowed the identification of thousands of SNPs in a cost-effective manner which can be applied to any species with a complete or partial reference genome. The results described in Chapters 2 and 3 contributed to the development of the large-scale project which generated the 60K SNP chip for pig (Ramos et al. 2009). The project of the porcine 60K SNP chip generated a large amount of MPS data which can be used to further address biological research questions. One interesting question, with a genome-wide sampling of sequences along the pig genome available, was the analysis of the variation of nucleotide diversity and the identification of footprints of selection.

The identification of footprints of selective events in the genome is relevant to the understanding of species evolution, specifically for the understanding of the genetic basis of complex traits and diseases. The identification of selection in the genome adds information to the annotation of the genome. Combined with association analyses a genome-wide map of footprints of selection can aid in explaining the genetic basis of complex diseases and traits. In Chapter 4, the first genome-wide map of the variation of nucleotide diversity in the pig genome, along with a preliminary identification of footprints of selection is presented showing how massive parallel sequencing can become the method of choice in population genomics. Moreover in Chapter 4 evidence for adaptive events that may have occurred during pig domestication and even during pig speciation which have affected specific biological pathways, are presented. In Chapter 5 the same approach is used to produce the first genome-wide map of selection in commercial chicken, showing that in spite of the intensive selection, commercial lines still hold high levels of genetic diversity. Results suggest that selection might have had an effect on genes related to specific biological pathways in different lines/breeds of chicken. Finally, in chapter 6 the results of this thesis are discussed and the implications of this work are explored.

References

Aerts, J. et al., 2007. Extent pf linkage disequilibrium in chicken. *Cytogenetic and Genome Research*, 117, 338-345.

Andersson, L. & Georges, M., 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet*, 5(3), 202-212.

Ardlie, K.G., Kruglyak, L. & Seielstad, M., 2002. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3(4), 299-309.

Awadalla, P., Hobolth, A,, 2009. Lecture notes from module 6: Coalescent theory. Summer Institute of Statistical Genetics, University of Liège, Belgium.

Bentley, D.R., 2006. Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, 16(6), 545-552.

Burt, D., 2005. Chicken genome: current status and future opportunities. *Genome Research*, 15, 1692-1698.

Darwin, C., 1868. *The variation of animals and plants under domestication*, John Murray.

Dekkers, J.C.M., 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *J. Anim Sci.*, 82(13_suppl), E313-328.

Devlin, B. & Risch, N., 1995. A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics*, 29(2), 311-322.

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

Du, F., Clutter, A. & Lohuis, M., 2007. Characterizing Linkage Disequilibrium in Pig Populations. *International Journal of Biological Sciences*, 3(3), 2007.

Eriksson, J. et al., 2008. Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. *PLoS Genet*, 4(2), e1000010.

Falconer, D.S. & Mackay, T.F.C., 1996. *Introduction to quantitative genetics*, Longman.

Fang, M. et al., 2009. Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs. *Plos Genetics*, 5(1), e1000341.

Farnir, F. et al., 2000. Extensive Genome-wide Linkage Disequilibrium in Cattle. *Genome Research*, 10, 220-227.

Fujii, J. et al., 1991. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science*, 253(5018), 448-451.

Fumihito, A. et al., 1994. One subspecies of the red junglefowl (Gallus gallus gallus) suffices as the matriarchic ancestor of all domestic breeds. *Proceedings of the National Academy of Sciences*, 91(26), 12505-12509.

Fumihito, A. et al., 1996. Monophyletic origin and unique dispersal patterns of domestic fowls. *Proceedings of the National Academy of Sciences*, 93(13), 6792-6795.

Giuffra, E. et al., 2000. The Origin of the Domestic Pig: Independent Domestication and Subsequent Introgression. *Genetics*, 154(4), 1785-1791.

Götherström, A. et al., 2005. Cattle Domestication in the Near East Was Followed by Hybridization with Aurochs Bulls in Europe. *Proceedings: Biological Sciences*, 272(1579), 2345-2350.

Hamilton, M., 2009. *Population Genetics*, Wiley-Blackwell.

Harris, D.R., 1996. *The origins and spread of agriculture and pastoralism in Eurasia*, Routledge.

Hill, W.G. & Robertson, A., 1968. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics*, 38(6), 226-231.

Houde, A., Pommier, S.A. & Roy, R., 1993. Detection of the ryanodine receptor mutation associated with malignant hyperthermia in purebred swine populations. *J. Anim Sci.*, 71(6), 1414-1418.

Kimura, M., 1985. *The neutral theory of molecular evolution*, Cambridge University Press.

Kruglyak, L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*, 22(2), 139-144.

Larson, G. et al., 2007. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences*, 104(39), 15276-15281.

Larson, G. et al., 2005. Worldwide Phylogeography of Wild Boar Reveals Multiple Centers of Pig Domestication. *Science*, 307(5715), 1618-1621.

Le Rouzic, A. & Carlborg, Ö., 2008. Evolutionary potential of hidden genetic variation. *Trends in Ecology & Evolution*, 23(1), 33-37.

Lewontin, R., 1964. The Interaction of Selection and Linkage. II. Optimum Models . *Genetics*, 50(4), 757-782.

Liu, Y. et al., 2006. Multiple maternal origins of chickens: Out of the Asian jungles. *Molecular Phylogenetics and Evolution*, 38(1), 12-19.

Lunney, J., 2007. Advances in Swine Biomedical Model Genomics. *Int. J. Biol. Sci.*, 3(3), 179-184.

McRae, A.F. et al., 2002. Linkage Disequilibrium in Domestic Sheep. *Genetics*, 160(3), 1113-

1122.

Megens, H. et al., 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. *BMC Genetics*, 10(1), 86.

Muir, W. et al., 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Sciences*, 105(45), 17312-17317.

Nielsen , R. & Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15, 1566-1575.

Nsengimana, J. et al., 2004. Linkage Disequilibrium in the Domesticated Pig. *Genetics*, 166(3), 1395-1404.

Porter, V. & Tebbit, J., 1993. *Pigs*, Helm Information.

Price, E.O., 1999. Behavioral development in animals undergoing domestication. *Applied Animal Behaviour Science*, 65(3), 245-271.

Ramos, A.M. 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, D.E. et al., 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834), 199-204.

Roots, C., 2007. *Domestication*, Greenwood Publishing Group.

Scherf, B., 2000. *World watch list for domestic animal diversity* 2nd ed., Rome: Food and Agricultural Organization.

Schook, L.B. et al., 2005. Swine Genome Sequencing Consortium (SGSC): A Strategic Roadmap for Sequencing The Pig Genome. *Comparative and Functional Genomics*, 6(4), 251-255.

Sundstrom, H., Webster, M.T. & Ellegren, H., 2004. Reduced Variation on the Chicken Z Chromosome. *Genetics*, 167(1), 377-385.

The Bovine HapMap Consortium, 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, 324(5926), 528-532.

Toro, M.A., Fernández, J. & Caballero, A., 2009. Molecular characterization of breeds and its use in conservation. *Livestock Science*, 120(3), 174-195.

Van Laere, A. et al., 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*, 425(6960), 832-836.

Voight, B.F. et al., 2006. A Map of Recent Positive Selection in the Human Genome. *Plos*

Biology, 4(3), e72.

Volis, 2008. Detection of signatures of positive selection in naturally occuring genetic variation. In *Population Genetics Research Progress*. Nova Science Publishers, Inc, pp. 279-310.

Wakeley, J., 2008. *Coalescent Theory*, Roberts & Co.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), 256-276.

Weiss, K.M. & Clark, A.G., 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, 18(1), 19-24.

Wright, S., 1984. *Evolution and the Genetics of Populations, Volume 2*, University of Chicago Press.

Yerle, M. et al., 1995. The PiGMaP consortium cytogenetic map of the domestic pig (Sus scrofa domestica). *Mammalian Genome*, 6(3), 176-186.

Zeuner, F.E., 1963. *A history of domesticated animals*, Harper & Row.

Zhang, K. et al., 2002. A dynamic programming algorithm for haplotype partitioning. *Proceedings of the National Academy of Sciences*, 99(11), 7335-7339.

2

Linkage disequilibrium decay and haplotype block structure in the pig

Andreia J. Amaral

Hendrik-Jan Megens

Richard. P.M. A. Crooijmans

Henri C. M. Heuven

Martien A. M. Groenen

Genetics (2008) 179:569-579

Abstract

Linkage disequilibrium (LD) may reveal much about domestication and breed history. An investigation was conducted, to analyze the extent of LD, haploblock partitioning and haplotype diversity within haploblocks across several pig breeds from China, Europe and in European wild boar. In total, 371 single-nucleotide-polymorphisms located in three genomic regions were genotyped.

The extent of LD differed significantly between European and Chinese breeds extending up to 2cM in Europe and up to 0.05cM in China. In European breeds, LD extended over large haploblocks up to 400kb, whereas in Chinese breeds the extent of LD was smaller and generally did not exceed 10kb. The European wild boar showed an intermediate level of LD between Chinese and European breeds. In Europe, the extent of LD also differed according to genomic region. Chinese breeds showed higher level of haplotype diversity and shared high frequent haplotypes with Large White, Landrace and Duroc.

The extent of LD differs between both centers of pig domestication, being higher in Europe. Two hypotheses can explain these findings. First, the European ancestral stock had higher level of LD. Second, modern breeding programs increased the extent of LD in Europe and caused differences of LD between genomic regions. Large White, Landrace and Duroc showed evidence of past introgression from Chinese breeds.

Introduction

Linkage Disequilibrium (LD), which refers to non-random association of alleles at different loci, has received increasing attention in recent years, and has gained unprecedented momentum as a result of the availability of genome sequences and large numbers of identified SNPs (single nucleotide polymorphisms). The Human HapMap project (The International Hapmap Consortium 2003, 2005) has revealed a large degree of variation of LD across the human genome and the intrinsic difficulty of analysis of genome-wide LD data (Reich *et al.* 2001). It also showed the presence of important differences in LD among human populations, which result from differences in population history and demography (Reich *et al.* 2001, Ardlie *et al.* 2002). Furthermore, the detailed information on genomic haplotype structure was shown to be of high utility for fine mapping of genes responsible for complex multfactorial diseases (Baessler *et al.* 2007, RIGBY *et al.* 2006; Wright *et al.* 2006; Welcome Trust Case Control Consortium, 2007).

Understanding the properties of LD in domesticated animals is important because it underlies all forms of genetic mapping (Nordborg and Tavare 2002). LD can reveal much about domestication and breed history because the distribution of LD is, in part, determined by population history and demography (Pritchard and Przeworski 2001, Tenesa *et al.*, 2007).

LD has been studied in a variety of domestic animal species, e.g. cattle (Farnir *et al.* 2000), sheep (Mcrae *et al.* 2002), pigs (Nsengimana *et al.* 2004), dog (Lindblad-Toh *et al.* 2005 and chicken (Aerts *et al.* 2007). In some of these species, a substantial extent of LD was found over several centimorgans and, exceeding the extent of LD found in humans (Reich *et al.* 2001). This larger extent of LD in animal species may be due to small effective population sizes in commercially held populations, and these may not be typical for the entire species. Dogs, for instance, show a large degree of variation in LD patterns, reflecting both high variability of the ancestor (wolf) and the result of low population sizes in breed formation and maintenance (Lindblad-Toh *et al.* 2005). In addition, most animal species are now known to have complex domestication histories (Bruford *et al.* 2003).

Pigs are among the most important domestic animals (Chen *et al.* 2007), being an important protein source. They are also an important animal model to study domestication because, Chinese and European pigs ancestors' still exist (Giuffra *et al.* 2000). European and Chinese pigs were domesticated independently from European and Asian subspecies of wild boar

(Giuffra *et al.* 2000, Larson *et al.* 2005). Studies on mitochondrial DNA, suggest the occurrence of introgression of Asian domestic pigs into European breeds after domestication (Giuffra *et al.* 2000, Fang and Andersson 2006). More recently, Larson *et al.* (2007) demonstrated that domestic pigs of Near Eastern ancestry were introduced into Europe during the Neolithic. European wild boar was also domesticated by this time, possibly as a direct consequence of the introduction of Near Eastern pigs.

The possibility of using large numbers of SNPs enables the detection of nuclear haplotypes which may be associated with introgression and or phenotypic selection which occurred during the domestication process. Analysis of extent of useful LD (Kruglyak 1999) could provide information about sample sizes and number of markers required to fine map genes responsible for common diseases and other phenotypic traits (Zhang *et al.* 2002).

Our aim was to investigate the extent of LD, LD haploblock partitioning and haplotype diversity within haploblocks across a total of 20 pig breeds in Europe and China and the ancestral European wild boar. With the commercial lines possibly containing the larger extent of LD in the species, we examined three genomic regions, each around 1 to 3 cM, at a higher SNP density than in previous studies (Nsengimana *et al.* 2004, Du *et al.* 2007). This study provides insight into the extent of useful LD across a wide range of breeds and the required sample sizes and number of markers for association studies.

Materials and Methods

DNA samples

DNA samples were obtained from 10 European and 10 Chinese pig breeds and from wild boar individuals from France which came from a single reserve, were certified as 2n=36, and are to the best of our knowledge unrelated. Sample size ranged from 15 to 25 individuals (Table 1). The material from these breeds was collected in the framework of PigBioDiv (Ollivier *et al.* 2005, Sancristobal *et al.* 2006) and PigBiodiv II (Blott *et al.* 2003) projects. European breeds were grouped by origin and history into local, international and commercial breeds. Chinese breeds were grouped by Lower Changjiang River Basin, Southwest China, Central China, North China, Plateau and South China (Zhang 1986; Fang *et al.* 2005, Megens *et al.* in 2007).

SNP development and selection

The NIH Intramural Sequencing Center (NISC; http://www.nisc.nih.gov) sequenced a large number of porcine BACs (supplementary Table). These BACs were derived from a pig BAC library developed using DNA of four crossbred male pigs (breed composition: 37.5% Yorkshire, 37.5% Landrace and 25% Meishan) (Fahrenkrug *et al.* 2001). NISC grouped these BACs by Targets. The Porcine sequences from Targets 1, 2 and 4 have previously been used to randomly identify SNPs (Jungerius *et al.* 2005). In our study additional SNPs were identified within these genomic regions by aligning sequences derived from overlapping BAC clones (supplemental Table 1). The list of identified SNPs and respective accession number is listed in supplemental Table 2.

SNP mapping

BAC sequences were masked for repeat motifs using RepeatMAsker v3.1.6 and RepBase 11.06 (http://repeatmasker.org) and aligned to porcine BAC end sequences (BES) available in GeneBank using Megablast v2.2.14 (Altschul *et al.* 1990). Positions of BES in pig genome are available in the FPC map (http://www.sanger.ac.uk/Projects/S_scrofa/). Hits with a bit score equal or over 1000 were therefore selected, and used to obtain BAC position on the FPC map (FPCmap of 08.10.06). The SNP position on the BAC was converted to a SNP position on the chromosome, using information on the BAC position and sequence length. SNP positions are in supplemental Table 2.

SNP genotyping

Because only small amounts of genomic DNA were available for each Chinese breeds except Meishan, whole genome amplification (WGA; Dean *et al.* 2002) was performed on these samples using the REPLI-g® kit from QIAGEN, with 50 – 150ng of input genomic DNA.

Genotyping was done in a 1536-plex format using the GoldenGate ™ assay and Sentrix ™ array matrices (Illumina Inc., San Diego, CA, Fan *et al.* 2003). Genotyping, including data editing, was performed by Illumina Inc. service facility. A total of the 1536 SNPs were genotyped with this procedure, but only 44 located in Target 1, 128 located in Target 2, and 199 located in Target 4 are described in this study.

Predicted decay of LD by breed

To measure LD, pairwirse r^2 was calculated using Haploview version 3.2 (Barrett *et al.* 2005). In this study, r^2 was chosen because it is very useful in the case of biallelic markers such as SNPs and, because it is independent from sample size (Devlin and Risch 1995). Further, Du *et al.* (2007) evaluated recently how r^2 and *D'* are affected by several levels of minor allele frequency (MAF). Their results suggest that *D'* is highly dependent on levels of MAF, whereas r^2 is less.

For SNPs genotyped for ≥75% of the total samples per breed in each genomic region within breed, tests for deviations from Hardy-Weinberg (H-W) equilibrium were performed and allele frequencies for all SNPs were estimated. SNPs in disequilibrium of H-W (*p*<0.001) and or with MAF smaller than 0.05 were excluded.

To assess the extent and decline of LD between breeds the equation was used (Sved 1971, Heifetz *et al.* 2005),

$$LD_{ijk} = \frac{1}{1+4\beta_{jk}d_{ijk}} + e_{ijk} \qquad (1)$$

where LD_{ijk} is the observed *LD* for marker pair *i* of breed *j* in region *k*, d_{ijk} is the distance in *bp* for marker pair *i* of breed *j* in genomic region k, β_{jk} is the coefficient that describes the decline of *LD* with distance for breed *j* in genomic region *k* and e_{ijk} is a random residual. For each genomic region within breed $\hat{L}D_{ijk}$, $\hat{\beta}_{jk}$ and $\hat{e}_{ijk}$ were estimated using the non linear fit function in the R environment (http://www.r-project.org/). Graphic displays of $\hat{L}D_{ijk}$ versus distance were produced.

Test for breed and genomic region effects in the extent of LD

Markers were not evenly distributed within genomic regions. This can have an effect in the evaluation of LD, since pairwise calculations are not assessed at equal distances and may cause a distortion in LD values. In order to test for breed and genomic region effects it was necessary to correct LD_{ijk} for differences in map distance when evaluating differences in LD between genomic regions. LDc_{ijk} is the distance corrected and variance stabilized LD for marker pair *i* in genomic region *k* and breed *j* and it was estimated using $\hat{\beta}_{jk}$ and $\hat{e}_{ijk}$ obtained with equation 1:

$$LDc_{ijk} = \frac{\hat{e}_{ijk}}{1 + 4\hat{\beta}_j d_{ijk}} \tag{2}$$

Differences in LD between genomic regions (Target 1, Target 2 and Target 4) and breeds were analyzed,

$$LDc_{ijk} = B_j + T_k + BT_{jk} + \varepsilon_{ijk} \tag{3}$$

where B_j is the fixed effect of breed *j*, T_k is the fixed effect of genomic region *k*, BT_{jk} is the fixed interaction effect and ε_{ijk} is the random residual. Equation 3 was fitted using linear model function in R environment (http://www.r-project.org/). Differences among all interaction levels were tested using the *lsmeans* function in SAS version 9.1 (SAS Institute, Inc., Cary, North Carolina).

LD haploblock partitioning and haplotype diversity

Due to the variation in local recombination rates, the breakdown of LD is often discontinuous and presents a haploblock-like structure (Daly *et al.* 2001; The International Hapmap Consortium, 2005). Therefore, it is important to analyse the haploblock structure and haplotypes which underlie LD. Analysis of haploblock partition defines the haploblock from the LD measure r^2 initiating and extending an haploblock according to the pairwise and grouped r^2 values (Gu *et al.* 2005). The algorithm starts an haploblock by selecting the pair of adjacent

SNPs with the highest r^2 (r^2>α) and extends the haploblock if the average r^2 between an adjacent site and current haploblock members is greater than β and each r^2 is greater than γ. Here, α > β > γ, and in this case were α=0.4, β=0.3 and γ=0.1 (Gu *et al.* 2005). After the first haploblock is identified, a new pair of adjacent SNPs with the highest r^2 (r^2≥α) is used to start a new haploblock accretion process.

Haplotypes within haploblocks were obtained using an accelerated EM algorithm, similar to the partition/ligation method of Qin *et al.* (2002) and implemented in Haploview version 3.2 (Barrett *et al.* 2005). The method creates highly accurate population frequency estimates of the phased haplotypes based on the maximum likelihood as determined from the unphased genotypes.

Plots of LD were generated using Haploview version 3.2 (Barrett *et al.* 2005). Frequencies of classes of haploblock sizes were calculated per breed. In this study, haplotype diversity was considered as the number of haplotypes found within an haploblock. Plots of haplotype diversity for each genomic region were produced.

Areas in the analyzed genomic regions that presented haploblock-like structure in all breeds were selected for study of haplotype frequency and of haplotype sharing between breeds. In these areas, a unique haploblock was forced for all breeds (supplemental Table 3 at http://www.genetics.org/supplemental/) and haplotypes were generated as described above. Median-joining networks (Bandelt *et al.* 1999) of these haplotypes were made using Network 4.2(http://www.fluxus-engineering.com).

Results

SNP identification and selection

Accurate estimation of the extent of LD within a selection of breeds from Europe and China required availability of genomic regions with high densities of identified SNPs. Because such information was not available at the start of the current study, we decided to analyze three genomic regions in pigs for which high quality sequences were available. The comparative vertebrate sequencing project of NISC (www.nisc.nih.gov) provided the necessary porcine genome sequences and NISC Target 1, Target 2 and Target 4 were chosen for the present study. Because the available sequenced BACs were derived from crossbred animals originally derived from three different breeds, it is likely that overlapping BACs are derived from different haplotypes. Consequently, these overlapping sequences provide a rich resource for the identification of SNPs. Alignment of these sequences identified several hundreds potential SNPs of which 371 were selected for genotyping. Of these SNPs 93% yielded genotyping results for over 75% of individuals for all genomic regions. In total, 40 SNPs in Target 1, 114 in Target 2 and 183 in Target 4 remained for further analysis.

Due to the absence of a genome sequence for the pig, positions of SNPs were based on alignment of BAC sequences with BES from clones located on the porcine BAC contig map (FPC map: http://www.sanger.ac.uk/Projects/S_scrofa/). This comparison indicates that Target 1 and Target 4 were located within q21 of SSC18 and Target 2 within q17 of SSC3 (Figure 1). The SNPs' distribution across each genomic region for SNPs with genotypes in at least 75% of the animals are in Figure 1. Available BACs are unevenly distributed along the chromosome and, consequently SNPs also are unevenly distributed across the different genomic regions.

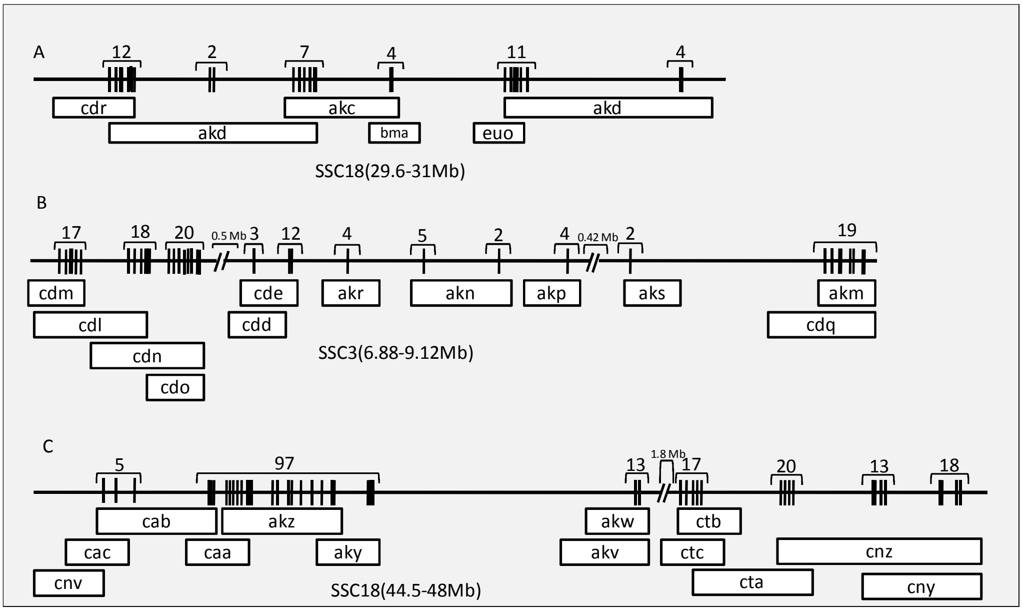


Figure 1 – SNPs' distribution by genomic region. (A) Target 1, (B) Target 2, (C) Target 4. SNP positions are indicated by vertical tick marks, numbers indicate total number of SNPs in a region. BACs are named according to NISC code and are indicated by horizontal bars. Accession numbers and clone names of the BACs are in Supplemental Table 1).

Predicted decay of LD per breed

Markers with departures from H-W were found at low frequencies in each genomic region and discarded. The proportion of markers with MAF<0.05 is higher for most European breeds especially in Target 1 and for Mangalitsa in Target 2. Predicted values of LD versus linkage distance per genomic region and per breed are in Figure 2. Most of tightly-linked SNP pairs have the highest r^2 and average r^2 rapidly decreases as linkage distance increases. Overall, there is a clear difference in the decay of LD between Chinese and European pig breeds for each of three genomic regions; r^2 decreases over short distances in Chinese breeds. This difference is most prominent in Target 1 and Target 4 (Figures 2A and 3).

In Target 1, observed r^2 is 1 for breeds Tamworth, Duroc, Middle White, British Saddleback and Large Black. Therefore predicted values were not estimated for these breeds and Figure 2B shows observed r^2.

For all genomic regions, LD decays more rapidly in Chinese breeds than in European breeds, indicating that in these breeds extent of LD is smaller than in European breeds. LD in the European wild boar (FrWb) is in-between European and Chinese breeds for Target 2 and Target 4 (Figures 2, C and D).

In European breeds, different patterns of decay of LD were observed across the three analyzed genomic regions between local, international and commercial groups of breeds (Table 1, Figure 2). International and commercial breeds present larger extent of LD than local European breeds, but there are some exceptions. For example, breed Landrace an international breed, which shows large extent of LD in Target 1 and in Target 4, presents a rapid LD decay in Target 2. These results showed that besides differences in the pattern of decay of LD between breeds, differences between genomic regions also exist.

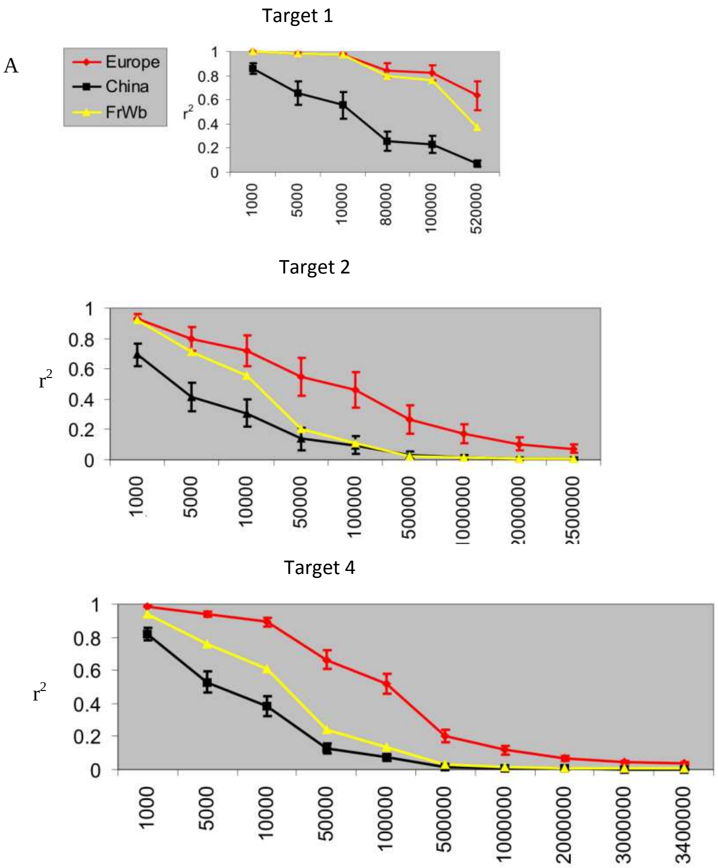


Figure 2- Predicted LD and physical distance in the three genomic regions. For each genomic region, the relationship between average predicted LD (*LD_{ij}*) and genomic distance (bp) is shown per biogeographical region for each genomic region (A), vertical bars represent the standard error. The relationship between predicted LD (*LD_{ij}*) vs distance (bp) is shown per breed and per genomic region, Target 1 (B), Target 2 (C) and Target 4 (D),. Chinese breeds are represented by dashed lines, European breeds are represented by solid lines and wild boar is represented by a thick line. In Target 1, observed r^2 is 1 for breeds Tamworth, Duroc, Middle White British Saddleback and Large Black. Therefore predicted values were not estimated for these breeds and Figure 2B shows observed r^2.

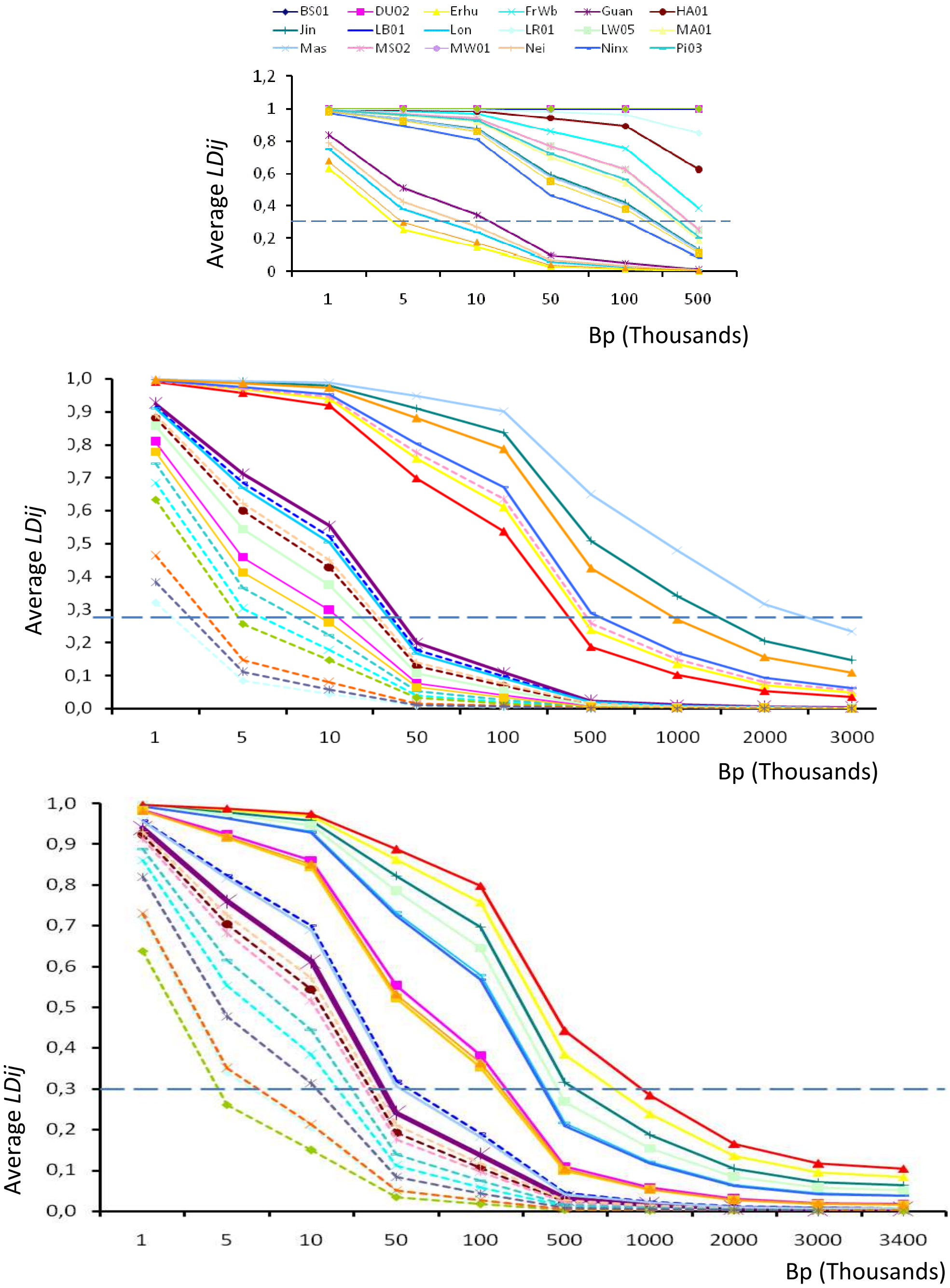


Figure 2 (Continued)

Differences were also observed among Chinese breeds. Breeds Ningxiang, Wuzishan, Tibetan and Neijang had the lowest levels of LD than other Chinese breeds in all genomic regions whereas, breeds Mashen, Meishan and Guanling had the highest levels of LD than other Chinese breeds. Differences in LD do not appear to be correlated to geographic area of origin.

Table 1- Sample size by breed.

Origin	Breed	Sample size	Target 1 SNPs (%)[a]	Target 2 SNPs (%)[a]	Target 4 SNPs (%)[a]
EUROPE					
Local	British Saddleback (BS01)	25	72.50	40.35	38.25
	Large Black (LB01)	25	75.00	62.28	38.25
	Mangalitsa (MA01)	23	42.50	78.95	52.46
	Middle White (MW01)	23	75.00	47.37	32.79
	Tamworth (TA01)	25	82.50	64.04	22.40
International	Landrace (LR01)	15	72.50	42.11	31.69
	Pietrain (Pi03)	24	47.50	67.54	23.50
Commercial	Duroc (DU02)	25	47.50	56.14	80.33
	Hampshire (HA01)	25	75.00	59.65	66.12
	Large White (LW05)	25	37.50	37.72	38.25
CHINA					
Lower Chagjiang River Basin	Erhulian (Erhu)	24	27.50	53.51	34.97
	Meishan (MS02)	25	25.00	30.70	51.37
Southwest China	Guanling (Guan)	24	32.50	50.00	49.18
	Neijiang (Nei)	25	42.50	50.88	61.20
Central China	Jinhua (Jin)	25	55.00	77.19	61.75
	Ningxiang (Ninx)	23	37.50	46.49	48.63
North China	Mashen (Mas)	24	17.50	64.04	38.80
Plateau	Tibetan (Tib)	24	40.00	52.63	27.87
South China	Longling (Lon)	25	17.50	40.35	20.22
	Wuzishan (Wuz)	25	22.50	41.23	25.68
FRENCH WILD BOAR (FrWb)		25	45.00	82.46	57.92

[a] Minor allele frequency ≤0.05.

Test for breed and genomic region effects in the extent of LD

Plotting predicted LD values versus genetic distance reveals clear differences between breeds as well as between genomic regions. Results of fitting the linear model (Equation 3) indicated that differences among breed and genomic region were significant (Table 2). Also, interaction between these two main effects was significant (Table 2). Differences among levels of interactions (breed *j vs* region *k)* were tested and *p* values matrix is shown in supplemental Table 3. The extent of LD was significantly different between Chinese and European breeds (*p*<0.001). Within the Chinese group, interaction effect between breed and genomic region was non significant (*p*>0.05) for most breeds. Most Chinese breeds and European wild boar do not have significantly different level of extent of LD. A clear exception is breed MS02 which showed to have an extent of LD significantly different from all the other Chinese breeds for each genomic region. Ningxiang and Tibetan showed significantly different levels of extent of LD in Target 1 (*p* <0.001).

Table 2- ANOVA of main effects and their interaction on LDc_{ijk}^{a}.

	d.f.	SS	MS	F value	*p* value
Breed	19	29.26	1.54	70.57	<0.001
Target	2	2.32	1.16	79.98	<0.001
Breed x Target	33	23.86	0.72	33.429	<0.001
Residuals	134373	2795.16	0.02		

[a] LDc_{ijk} is the distance corrected and variance stabilized LD for marker pair *i* in genomic region *k* and breed *j*.

Contrary to Chinese breeds, European breeds showed high levels of significant differences in the extent of LD within the group (*p*<0.01) across all genomic regions. For example, Mangalitsa a local breed, presented significant different level of extent of LD when compared with other European breeds and wild boar (*p*<0.001) and with Chinese breeds (*p*<0.05) in Target 2. In Target 4, Mangalitsa showed to be different from all breeds and European wild boar at a high significance level (*p*<0.001).

Haploblock partitioning

Haploblocks were defined using r^2 (Gu *et al.* 2005). The pattern of haploblock partitioning showed similarities across genomic regions for Chinese breeds (Figure 3 and supplemental Figure 1). A large number of haploblocks with size up to 10 kb were generally present (supplemental Figure 2). Among European breeds and wild boar, the overall pattern of haploblock partitioning consisted of SNPs grouped in lower number of haploblocks, ranging from 50 to more than 200 kb in Target 1 and Target 4 (Figure 3 and supplemental Figure 1). In Target 2, larger haploblocks were also present but the proportion of haploblocks up to 10 kb was higher (supplemental Figure 2).

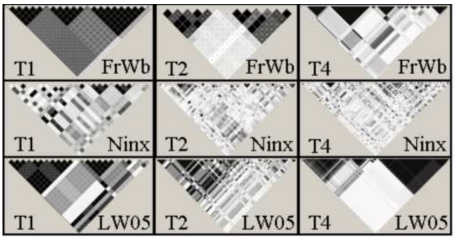


Figure 3 – Comparison of LD between genomic regions: (T1) Target 1; (T2) Target 2; (T4) Target 4. Pairwise LD (observed r^2) plots are shown for Ningxiang (Ninx) and Large White (LW05) as an example of the LD in the two main centers of pig domestication (Europe and China) and the European wild boar representing the ancestral European population: white, r^2=0; shades of grey, 0<r^2<1 and black, r^2=1.

The average number of haplotypes found in haploblocks with equal number of SNP's for Chinese and European breeds and wild boar by genomic region are in Figure 4. For haploblocks with an equal number of SNPs, Chinese breeds had higher number of haplotypes than wild boar and European breeds.

For each genomic region, areas characterized by the existence of larger haploblocks in one or several breeds and smaller haploblocks on the remaining breeds were selected to investigate haplotype sharing between European and Chinese breeds and European wild boar. SNPs included in this analysis are presented in supplemental Table 4. Supplemental Figure 3 shows the overall view of haploblock partitioning per genomic region. The selected areas were in Target 1, block1 (4kb) and block2 (149kb); in Target 2, block 1 (8kb) and block 2 (2kb) and in Target 4, block1 (293kb). In general, smaller haploblocks were found in Chinese breeds, resulting in larger number of haplotypes.

The haplotype diversity and respective frequencies are represented by a median-joining network for the 8 kb region of Target 2 (Figure 5), for the remaining genomic regions these figures are available as supplemental material (supplemental Figure 4).

Wild boar shared most haplotypes with European breeds, although one haplotype is also shared with Chinese breeds. Some haplotypes which occur at high frequency in Chinese Breeds are shared by Large White Duroc, Landrace, Pietrain and Mangalitsa. Meishan shares haplotypes which occur at high frequency in European breeds.

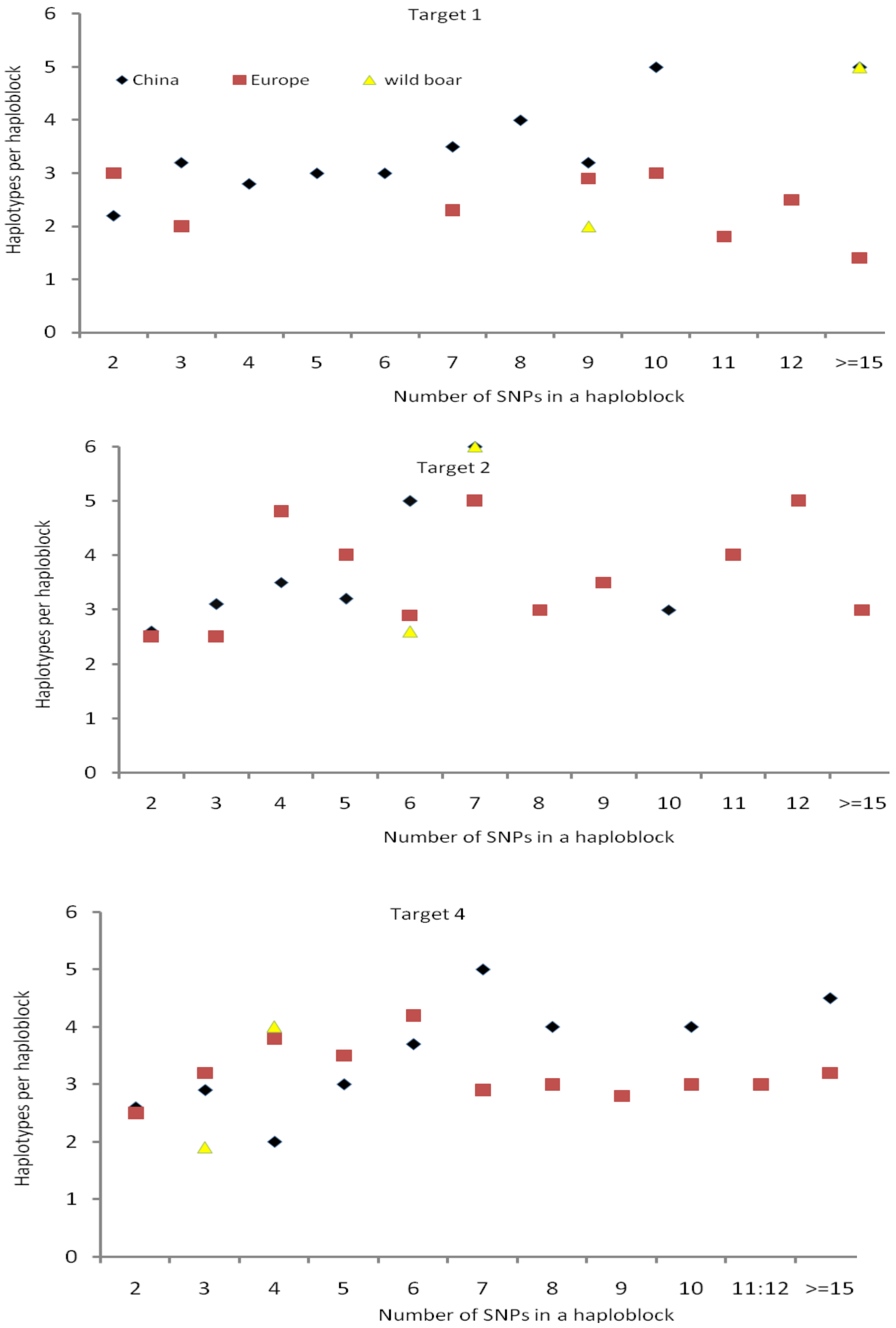


Figure 4– Comparison of haplotype diversity per genomic region. Number of haplotypes per haploblock plotted against the number of SNPs in each haploblock for each genomic region (Target1, Target 2 and Target 4) for Europe and China and the European wild boar.

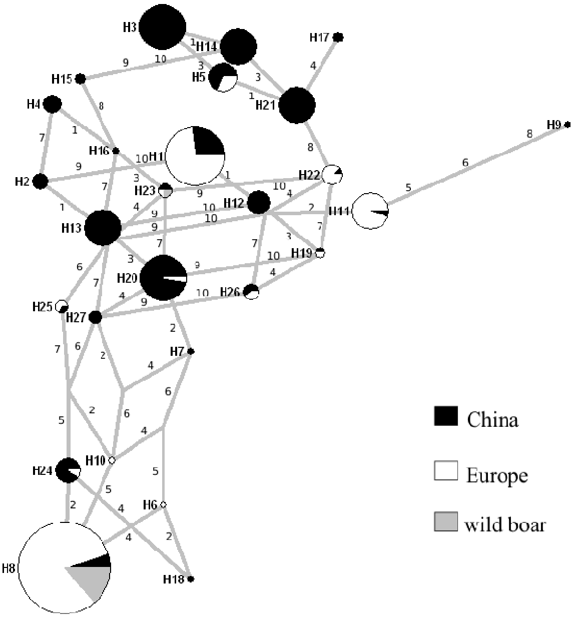


Figure 5 – Median-joining network of the SNP-based haplotypes from Target 2, block1. In each circle (identified in bold), haplotype frequencies in European and Chinese breeds and European wild boar are shown. The circle area is proportional to frequency and branch length proportional to the number of mutations. Each line is annotated with its corresponding mutational change (corresponding to the SNP ID in supplemental Table 2).

Discussion

Differences of LD between European and Chinese breeds

In this study, comparison between pig breeds from two major areas of domestication (China and Europe) across three genomic regions revealed large and significant differences in the extent and pattern of LD. The observed pattern of decay of LD is similar to what has been observed in previous studies in humans (Daly *et al.* 2001), cattle (Farnir *et al.* 2000), chicken (Aerts *et al.*, 2007) and pigs (Du *et al.* 2007).

In Chinese breeds, LD is mostly organized in haploblocks of up to 10 kb, while in European breeds LD haploblocks may be up to 400 kb in size. Haplotype diversity within the haploblocks was higher in Chinese breeds. Chinese breeds also showed a lower percentage of SNPs with MAF lower than 0.05. Higher haplotype diversity and lower proportion of fixed markers is an indication of higher genetic diversity in Chinese breeds. This has also been reported by Fang *et* *al.* (2005) and Megens *et al.* (2008) using microsatellites and by Fang and Andersson (2006) studying mitochondrial DNA.

European wild boar showed levels of LD and haplotype diversity in-between the values found in the European and Chinese breeds, partitioned in low number of haploblocks of up to 200 kb. Further analysis showed that European breeds share the most frequent haplotypes with the studied population of European wild boar.

Two hypotheses for the existence of significant differences of LD between European and Chinese pig breeds can be considered: a) differences are due to differences in ancestral stock; b) modern breeding practices in Europe resulted in small effective population size.

The first hypothesis is supported by the fact that the European wild boar had higher level of LD and lower genetic diversity compared to Chinese breeds. European and Chinese breeds were domesticated from different ancestors that might have had different levels of genetic diversity and LD. This hypothesis is also supported by the results obtained by Larson *et al.* (2005) which showed that European wild boars have lower genetic diversity in mitochondrial DNA compared to Asian wild boars. In addition, these authors suggest that domestication of pig breeds in Asia involved several lineages of wild boar, suggesting that the original gene pool for domestication was more diverse when compared to Europe. However, our findings regarding the level of LD in European wild boar cannot be generalized for the species since the European wild boar is largely distributed (from Western Europe and Mediterranean basin to Russia) and this study only analyzed one wild boar population from France. European wild boar populations suffered a serious decrease, resulting in the extinction in the British Isles and parts of Northern Europe. Only recently the population started to increase and, areas such as Sweden, Finland and Estonia were recolonized (Erkinaro *et al.* 1982, Leaper *et al.* 1999), suggesting that a recent bottleneck could have resulted in increased level of LD. To validate this hypothesis a wider geographic sampling of both European and Asian wild boar is necessary which was beyond the scope of this study.

Animal domestication is the process by which captive animals adapt to humans and to the captive environment. This involves directed changes in the gene pool occurring over generations during which, domesticated animals will differ from their wild counter parts (Price, 2002), diminishing effective population sizes, increasing inbreeding and consequently increasing LD (Lindblad-Toh *et al.* 2005). This takes us to our second hypothesis which relates the increase of LD to an effect of small population sizes and modern breeding practices. Domestication of pigs in Europe occurred independently from the pig domestication in Asia (Giuffra et al. 2000; Larson et al. 2005) but may have started due to the introduction of Near Eastern pigs during the Neolithic (Larson *et al.* 2007). Nowadays, Europe has 37% of world pig breeds (Scherf 2000). In this study, several local breeds, international and commercial lines were analyzed. Tamworth is listed as endangered breed (breeds with less than 1000 breeding females and or 20 or less breeding males). British Saddleback, Large Black, Mangalitsa and Middle White are listed as endangered maintained breeds (breeds with less than 1000 breeding females and or 20 or less breeding males but, are maintained by an active conservation program) (Scherf 2000). Therefore these local breeds are characterized by having small effective population sizes, which is affecting severely it's diversity, has shown by the high levels of LD and the high proportion of SNPs with fixed alleles. This study analyzed international breeds, Landrace and Pietrain and commercial lines of Duroc, Hampshire and Large White and these were the ones which showed the highest levels of LD. These are the breeds used by the pig industry to produce pork meat and related products. Modern breeding practices, started in the middle of the previous century and the introduction of BLUP selection after 1990 allowed an rapid increase of genetic gain (Merk 2000). At the same time it is likely that this had resulted in decreased effective population sizes by limiting the genetic inflow into commercial breeding lines (Jones, 1998).

Founder effect on LD of populations seems to be evident in the case of the Meishan line which was brought to Europe 25 years ago and started from a small number of individuals and has since been kept in a small population. The level of LD in this line is higher compared to the other Chinese breeds although, still lower than most European breeds.

Analysis of nuclear haplotypes using SNPs allows the detection of introgression and phenotype selection (Zhang *et al.* 2002). In this study, haplotype sharing between these European breeds (Large White, Landrace, Duroc) and Chinese breeds was found in all genomic regions. These results add support to the hybrid origin of these European breeds, reported by historical documentation (Porter 1993), and by studies of mitochondrial haplotypes (Giuffra *et al.* 2000, Fang and Andersson 2006). One of the shared haplotypes was also shared at a low frequency by European wild boar. This may be due to recent introgression of pig genes into European wild boar (Giuffra *et al.* 2000), however more samples of European wild boar should be analyzed.

Differences of LD between regions

The extent of LD and haploblock pattern varies significantly between genomic regions. Differences of extent of LD between genes located in different chromosomes were observed by Reich *et al.* (2001) in human populations. These authors found levels of LD extending up to 160 kb in some genomic regions while in others the LD only extended up to 40 kb.

Nsengimana *et al.* (2004) assessed LD in five populations of commercial pigs (Large White, Landrace, Duroc/Large White and Yorkshire/Large White) in two chromosomal regions, one on SCC4 (33 cM) and another on SCC7 (48 cM) using 15 microsatellites with an interval spacing of 5 cM. The region on SSC7 presented significant larger extent of LD compared to SSC4. Since SSC7 harbors QTLs associated to growth rate and fat deposition Nsengimana *et al.* (2004) suggested that these differences in the extent of LD were due to selection.

A likely cause to the observed differences in the extent of LD between genomic regions located in chromosomes SSC18 and SSC3 is selection. In region q21 of SSC18, previous studies identified a number of genes which are obvious candidates to be under selection (e.g. *INSIG1,* Qiu *et al.* 2005; *LEP,* Campbell *et al.* 2001 and *GHRHR,* Sun, *et al.* 1997). Further, several QTLs have been mapped to this region as well (pH, cook loss and feed-conversion ratio) (Hu *et al.* 2005). In contrast, region q17 on SSC3 mainly contains genes involved in general cellular processes such as DNA transcription, transduction and cell differentiation which are not likely candidate genes to be under selection and no major QTLs have been described for this region (Hu *et al.* 2005). This hypothesis is also supported by the similarity in the extent of LD across genomic regions in the European wild boar population. The effect of selection on the extent of LD in other domestic animals has been reported in other studies (cattle Farnir *et al.* 2000 and sheep Mcrae *et al.* 2002).

Assessing the extent of useful LD

The threshold for useful LD that was chosen in this study was the same as previously used in LD studies of pig populations using r^2 as a measure of LD (Jungerius *et al.* 2005, Du *et al.* 2007). With a threshold of 0.3, and considering that on average 1 cM is equivalent to 1 Mb, LD extended in the European breeds over 0.5 to 2 cM on SSC18 and 0.1 to 1 cM on SSC3. In the case of Chinese breeds LD ranged between 0.005 to 0.05 cM among the studied genomic regions.

This study shows that LD for European breeds is higher compared to human populations (Reich *et al.* 2001, Ardlie *et al.* 2002). Higher values for the extent of LD in domestic animals have been reported in previous studies (e.g. cattle Farnir *et al.* 2000, sheep Mcrae *et al.* 2002). Previous reports on the extent of LD in European pigs (Nsengimana *et al.* 2004, DU *et al.* 2007) also showed large levels of LD exceeding the values obtained in the current study. As described above, Nsengimana *et al.* (2004) studied the extent of LD using microsatellites, and LD was measured using *D*'. These authors concluded using the threshold for useful LD of 0.5 that, LD ranges from 3 to 10 cM. However, these conclusions were based on a low number of distantly spaced markers. Du *et al.* (2007) also studied LD using commercial lines (Pietrain, Duroc, Landrace and Large White). LD was assessed using SNPs across 18 chromosomes with an average of 330 markers per chromosome, with a maximum length of 100 cM and a marker spacing on average of 0.44 cM. LD was measured using r^2 and 0.3 as threshold for useful LD. These authors suggest that LD extends to 1-3 cM in these pig populations. This is only somewhat higher than values obtained in the current study, which could be due to the higher marker spacing used by Du *et al.* (2007).

By contrast, the present study aimed at assessing the extent of LD with high definition, and the trade-off was to have high SNP frequencies (0.02 cM) preferably covering regions with few cM's in size. The size of Target 1 however, was only 1cM and in the case of many European breeds this length was too small to observe the decay of the LD. In Targets 2 and 4 the decay of LD can be observed in detail with maxima ranging from 1 cM in Target 2 to 2 cM in Target 4. This experimental design allowed us to study LD in a large set of breeds, from Europe and China. The most important difference between our study and that of nsengimana *et al.* (2004) and DU *et al.* (2007) is, the high marker density used and wider sampling of pig populations. In Chinese breeds, the results of decay of LD would have been inconclusive if the same lower marker spacing would have been used as by Du *et al.* (2007). Our results indicate that we would have found a very steep decay in LD and would not have been able to precisely identify the point at which LD drops below 0.3. For Chinese breeds, this is the first study that aimed to assess LD in these breeds. The level of extent of LD is very small, 0.005 to 0.05 cM which is similar to the extent of LD observed in human populations (Reich *et al.* 2001, Ardlie *et al.* 2002, The International Hapmap Consortium, 2005). Populations with shorter extent of LD are more suitable for fine mapping of genes responsible for phenotypic traits (Ardlie *et al.* 2002).

Therefore, Chinese pig breeds may be useful to fine map QTLs, however, the QTL alleles which have an effect in the phenotypic trait have to be segregating in these breeds.

For future population-wide studies with a whole genome approach, our results indicate that assuming a threshold of 0.3 for r^2, the SNP spacing for European pig breeds should be of around 0.1 cM. This implies the use of 30.000 SNPs per individual using the same sample sizes as in this study and assuming that all SNPs are informative (with a MAF over 0.05). For Chinese breeds in a study with similar sample size a SNP spacing of 0.005 cM and the use of approximate 500.000 SNPs per individual would be required.

Aknowledgments

This work was funded by European Union grant QLRT-2001-01059. European pig DNA samples, other then provided by the authors, were provided by Agence de la Sélection Porcine, France, Georg-August University Göttingen, Germany, Rare Breeds Survival Trust, UK, Roslin Institute, UK, and PIC International Group, UK. European wild boar samples were provided by Alain Ducos (UMR INRA-ENVT 898, Cytogénétique des Populations Animales), Chinese breeds were provided by Ning Li (China Agricultural University, National Laboratories for Agrobiotechnology). This work was conducted as part of the SABRETRAIN Project, funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission.

References

Aerts, J., H. J. Megens, T. Veenendaal, I. Ovcharenko, R. Crooijmans *et al.*, 2007 Extent of linkage disequilibrium in chicken. Cytogenet *Genome Res* 117, 338-345.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215, 403-410.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilbrium in the human genome (vol 3, pg 299, 2002). *Nature Reviews Genetics* 3, 566-566.

Baessler, A., M. Fischer, B. Mayer, M. Koehler, S. Wiedmann *et al.*, 2007 Epistatic interaction between haplotypes of the ghrelin ligand and receptor genes influence susceptibility to myocardial infarction and coronary artery disease. *Human Molecular Genetics* 16, 887-899.

Bandelt, H. J., P. Forster and A. Rohl, 1999 Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* 16, 37-48.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly, 2005 Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-265.

Blott, S., L. Andersson, M. Groenen, M. Sancristobal, C. Chevalet *et al.*, 2003 Characterisation of genetic variation in the pig breeds of China and Europe-the pigbiodiv2 project. *Archivos de Zootecnia* 52, 207.

Bruford, M. W. M. W., D. G. D. G. Bradley and G. G. Luikart, 2003 DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics* 4, 900-910.

Campbell, E. M. G., S. C. Fahrenkrug, J. L. Vallet, T. P. L. Smith and G. A. Rohrer, 2001 An updated linkage and comparative map of porcine chromosome 18. *Animal Genetics* 32, 375-379.

Chen, K. K., T. T. Baxter, W. M. Muir, M. A. M. Groenen and L. B. Schook, 2007 Genetic resources, genome mapping and evolutionary genomics of the pig (*Sus scrofa*). *International journal of biological sciences* 3, 153-165.

Daly, M. J., J. D. Rioux, S. E. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. *Nature Genetics* 29, 229-232.

Dean, F. B., S. Hosono, L. H. Fang, X. H. Wu, A. F. Faruqi *et al.*, 2002 Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America 99, 5261-5266.

Devlin, B., and N. Risch, 1995 A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics* 29, 311-322.

Du, F. X., A. C. Clutter and M. M. Lohuis, 2007 Characterizing linkage disequilibrium in pig populations. *International Journal of Biological Sciences* 3, 166-178.

Erkinaro, E., K. Heikura, Pullianen E., and S. Sulkava, 1982 Occurrence and spread of the wild boar (*Sus scrofa*) in eastern Fennoscandia. *Memoranda Flora and Fauna Fennoscandia* 58, 39-47.

Fahrenkrug, S. C., G. A. Rohrer, B. A. Freking, T. P. L. Smith, K. Osoegawa *et al.*, 2001 A porcine BAC library with tenfold genome coverage: a resource for physical and genetic map integration*.* *Mammalian Genome* 12, 472-474.

Fan, J. B., A. Oliphant, R. Shen, B. G. Kermani, F. Garcia *et al.*, 2003 Highly parallel SNP genotyping. *Cold Spring Harbor Symposia on Quantitative Biology* 68, 69-78.

Fang, M., X. Hu, T. Jiang, M. Braunschweig, L. Hu *et al.*, 2005 The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Animal Genetics* 36, 7-13.

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

Farnir, F., W. Coppieters, J. J. Arranz, P. Berzi, N. Cambisano *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* 10, 220-227.

Gabriel, S. B., S. F. Schaffner, D. E. Reich, P. Sabeti, M. Freedman *et al.*, 2002 An extended analysis of haplotype structure in the human genome. *American Journal of Human Genetics* 71, 205-205.

Giuffra, E., J. M. H. Kijas, V. Amarger, O. Carlborg, J. T. Jeon *et al.*, 2000 The origin of the domestic pig: Independent domestication and subsequent introgression. *Genetics* 154, 1785-1791.

Gu, S., A. J. Pakstis and K. K. Kidd, 2005 HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* 21, 3938-3939.

Heifetz, E. M., J. E. Fulton, N. O'sullivan, H. Zhao, J. C. M. Dekkers *et al.*, 2005 Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* 171, 1173-1181.

Hu, Z. L. Z.-L., S. S. Dracheva, W. W. Jang, D. D. Maglott, J. J. Bastiaansen *et al.*, 2005 A QTL resource and comparison tool for pigs: PigQTLDB. *Mammalian genome* 16, 792-800.

Jones, G.F., 1998 Aspects of domestication, common breeds and their origin, pp.17-50 in *The genetics of the pig,* edited by M. F. Rothschild and A. Ruvinsky. CAB International, New York, USA.

Jungerius, B. J., J. J. Gu, R. Crooijmans, J. J. Van Der Poel, M. A. M. Groenen *et al.*, 2005 Estimation of the extent of linkage disequilibrium in seven regions of the porcine genome. *Animal Biotechnology* 16, 41-54.

Kanis, E., K. H. De Greef, A. Hiemstra and J. A. M. Van Arendonk, 2005 Breeding for societally important traits in pigs. *Journal of Animal Science* 83, 948-957.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 22, 139-144.

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

Larson, G., U. Albarella, K. Dobney, P. Rowley-Conwy, J. Schibler *et al.*, 2007 From the Cover: Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences* 104, 15276-15281.

Leaper, R., G. Massei, M. L. Gorman and R. Aspinall, 1999. The feasibility of reintroducing wild boar (*Sus scofa*) to Scotland. *Mammal Rev.* 29, 239-259.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803-819.

Mcrae, A. F., J. C. Mcewan, K. G. Dodds, T. Wilson, A. M. Crawford *et al.*, 2002 Linkage disequilibrium in domestic sheep. *Genetics* 160, 1113-1122.

Megens, H-J., R.P.M.A. Crooijmans, M. Sancristobal, X. Hui, N. Li, *et al.*, 2008 Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genetic Selection and Evolution* 40, 103-128.

Merks, J.W.M., 2000 One century of genetic changes in pigs and future needs, pp. 8-19 in the Challenge of Genetic Change in Animal Production. W.G.Hill, S.C. Bishop, B. McQuirck, J.C. McKay, G. Simm and A.J. Webb. Br. Soc. Anim. Sci. Occasional Publication. Edinburgh, U.K.

Nordborg, M., and S. Tavare, 2002 Linkage disequilibrium: what history has to tell us. *Trends in Genetics* 18, 83-90.

Nsengimana, J., P. Baret, C. S. Haley and P. M. Visscher, 2004 Linkage disequilibrium in the domesticated pig. *Genetics* 166, 1395-1404.

Ollivier, L., L. Alderson, G. C. Gandini, J. L. Foulley, C. S. Haley *et al.*, 2005 An assessment of European pig diversity using molecular markers: Partitioning of diversity among breeds. *Conservation Genetics* 6, 729-741.

Porter, V., 1993 *Pigs, a handbook to the breeds of the world.* Helm Information, London, UK.

Price, E. O., 2002 *Animal Domestication and Behavior*. CABI International, New York, USA.

Pritchard, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* 69, 1-14.

Qin, Z. H. S., T. H. Niu and J. S. Liu, 2002 Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics* 71, 1242-1247.

Qiu, H., T. Xia, X. D. Chen, L. Gan, S. Q. Feng *et al.*, 2005 Characterization of pig INSIG1 and assignment to SSC18. *Animal Genetics* 36, 284-286.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. *Nature* 411, 199-204.

Rigby, R. J., M. M. A. Fernando and T. J. Vyse, 2006 Mice, humans and haplotypes - the hunt for disease genes in SLE. *Rheumatology* 45, 1062-1067.

Sancristobal, M., C. Chevalet, C. S. Haley, R. Joosten, A. P. Rattink *et al.*, 2006 Genetic diversity within and between European pig breeds using microsatellite markers. *Animal Genetics* 37, 189-198.

Scherf, B. D., 2000 *World watch list for domestic animal diversity*.3rd Food and Agriculture Organization, Rome, Italy.

Sun, H. S., C. Taylor, L. Wang, M. F. Rothschild, C. K. Tuggle *et al.*, 1997 Mapping of growth hormone releasing hormone receptor to swine chromosome 18. *Animal Genetics* 28, 351-353.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations*. Theoretical Population Biology* 2, 125-141.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17, 520-526.

The International Hapmap Consortium, 2003 The International HapMap Project*. Nature* 426, 789-796.

The International Hapmap Consortium, 2005 A haplotype map of the human genome. *Nature* 437, 1299-1320.

WELLCOME TRUST CASE CONTROL CONSORTIUM, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-78

Wright, W. T., I. S. Young, D. P. Nicholls, C. Patterson, K. Lyttle *et al.*, 2006 SNPs at the APOA5 gene account for the strong association with hypertriglyceridaemia at the APOA5/A4/C3/A1 locus on chromosome 11q23 in the Northern Irish population. Atherosclerosis 185, 353-360.

Zhang, Z., 1986 *Pig breeds in China*. Shanghai Scientific and Technical Publishers, Shanghai, China.

Zhang, K., M. H. Deng, T. Chen, M. S. Waterman and F. Z. Sun, 2002 A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences* 99, 7335-7339.

3

Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome

Andreia J. Amaral

Hendrik-Jan Megens

Hindrik H.D. Kerstens

Henri C.M. Heuven

Bert Dibbits

Richard. P.M. A. Crooijmans

Johan T. den Dunnen

Martien A. M. Groenen

BMC Genomics (2009) 10:374

Abstract

Background

Although the Illumina 1G Genome Analyzer generates billions of base pairs of sequence data, challenges arise in sequence selection due to the varying sequence quality. Therefore, in the framework of the International Porcine SNP Chip Consortium, this pilot study aimed to evaluate the impact of the quality level of the sequenced bases on mapping quality and identification of true SNPs on a large scale.

Results

DNA pooled from five animals from a commercial boar line was digested with *Dra*I; 150-250-bp fragments were isolated and end-sequenced using the Illumina 1G Genome Analyzer, yielding 70,348,064 sequences 36-bp long. Rules were developed to select sequences, which were then aligned to unique positions in a reference genome. Sequences were selected based on quality, and three thresholds of sequence quality (SQ) were compared. The highest threshold of SQ allowed identification of a larger number of SNPs (17,489), distributed widely across the pig genome. In total, 3,142 SNPs were validated with a success rate of 96%. The correlation between estimated minor allele frequency (MAF) and genotyped MAF was moderate, and SNPs were highly polymorphic in other pig breeds. Lowering the SQ threshold and maintaining the same criteria for SNP identification resulted in the discovery of fewer SNPs (16,768), of which 259 were not identified using higher SQ levels. Validation of SNPs found exclusively in the lower SQ threshold had a success rate of 94% and a low correlation between estimated MAF and genotyped MAF. Base change analysis suggested that the rate of transitions in the pig genome is likely to be similar to that observed in humans. Chromosome X showed reduced nucleotide diversity relative to autosomes, as observed for other species.

Conclusion

Large numbers of SNPs can be identified reliably by creating strict rules for sequence selection, which simultaneously decreases sequence ambiguity. Selection of sequences using a higher SQ threshold leads to more reliable identification of SNPs. Lower SQ thresholds can be used to guarantee sufficient sequence coverage, resulting in high success rate but less reliable MAF estimation. Nucleotide diversity varies between porcine chromosomes, with the X chromosome showing less variation as observed in other species.

Background

The Sanger DNA sequencing technique has been and still is the method of choice for *de novo* sequencing of complete genomes (J. C. Venter et al. 2001; Schook et al. 2005). However, whole genome sequencing using the Sanger method is relatively expensive, labor intensive, and time consuming.

Several methods for ultra high-throughput DNA sequencing that reduce the cost and labor demands of Sanger sequencing are currently available (Ahmadian et al. 2006; Metzker 2005). The Illumina 1G Genome Analyzer (ILLUMINA, San Diego, CA, USA) uses a sequencing by synthesis method, during which millions of DNA fragments are sequenced in parallel (massive parallel sequencing). With this method, costly and often problematic procedures, such as cloning are eliminated. Another advantage is that accuracy is independent of sequence context because a discrete signal is generated per each base. Thus, this method is very accurate in cases of homopolymeric sequences and generates quality values that are analogous to Phred scores (Bentley 2006). Sequence lengths generally range from 25-50 bp (short sequences), which is sufficient for unique alignment to a reference genome (Whiteford et al. 2005). Because millions of fragments are sequenced in parallel, a fragment can be sequenced even if it exists in low abundance in the sample, thereby increasing sequencing depth and enabling identification of single nucleotide polymorphisms (SNPs) with high accuracy (Schendure et al. n.d.; Van Tassell et al. 2008; Hillier et al. 2008)

Sequencing of reduced representation libraries (RRLs), which are reproducible subsets of the genome, allows cost-effective genome-wide SNP discovery with accurate estimations of minor allele frequencies (MAF) (Altshuler et al. 2000). Because the cost of large-scale sequencing of RRLs is still prohibitive for individual samples, DNA samples can be pooled to further reduce sequencing costs and simultaneously infer the frequencies of two SNP alleles with high accuracy (Sham et al. 2002). Previous studies have shown that large-scale SNP discovery can be accurate using massive parallel sequences of RRLs prepared from pooled DNA (Van Tassell et al. 2008; Wiedmann et al. 2008).

Despite the efficiency of massive parallel sequencing for providing large amounts of sequencing data, a sequence selection stage is still required. Previous studies have applied various rules for sequence selection: sequences must start with a restriction motif (Van Tassell et al. 2008), sequences must be aligned to a unique location in the genome (Hillier et al. 2008), and sequences must have a minimum average sequence quality (SQ) score of 20 (Van Tassell et al. 2008; Hillier et al. 2008). An effective sequence selection stage can decrease noise in the data that can compromise alignment and SNP identification. Therefore, the effect of different levels of SQ in identifying SNPs needs to be evaluated.

With increasing attention being paid to genomic selection by animal breeders, there is a need for high-density SNP maps of the genomes of farm animals. Experimental evidence has shown that linkage disequilibrium extends from 0.1 to 2 cM in European commercial pig breeds (Amaral et al. 2008); thus, an SNP assay should contain a minimum of 30 k informative SNPs. To achieve this goal, we designed a cost-effective strategy for large scale identification of SNPs in the porcine genome that could be applied to other species.

In this study, an RRL generated from a DNA pool of a boar line was sequenced using the Illumina 1G Genome Analyzer. The two main goals of our study were (a) to develop rules for decreasing sequence ambiguity (sequence alignment to several locations in the genome), which would decrease noise and increase efficiency in sequence alignment and SNP identification, and (b) to evaluate the effects of different SQ filtering strategies for cost-effective, large-scale identification of SNPs.

Results and discussion

Sequencing and filtering the RRL

An RRL was produced from a DNA pool of five boars from a crossbred (Large White *vs.* Pietrain) commercial boar line (PW), using the restriction enzyme *Dra*I*,* which recognizes pattern “TTTAAA” and generates blunt-ended fragments starting with AAA. Fragments 150-250-bp long were selected and end-sequenced using the Illumina 1G Genome Analyzer. Short sequences will align to a unique genomic location (target), creating groups (clusters) with a number of sequences (target depth) sufficient for SNP identification. The *in silico* digest of *Sus scrofa* build 7 (pre-EnsEMBL Sus scrofa build 7) indicated an expected sequence coverage of ~1% of the reference genome, that is, 11,089,914 bp uniquely aligned to the porcine genome.

In total, 70,348,064 sequences were generated during three different runs. In addition to sequence information, this second generation platform generates quality scores that are analogous to Phred scores (which assign a probability to the four possible nucleotides for each base in the sequence) (Ewing et al. 1998). Levels of base quality varied between runs and along the sequence length, and decreased considerably at the 3'end [see Additional file 1]. This variation in base quality along the sequence has been reported in previous studies using short sequencing (Hillier et al. 2008). Base quality for the first three bases is poor, but increases before decreasing again at the end of the 31-bp sequence. Poor base quality at the first three bases is due to the properties of the algorithm implemented in the Illumina base caller BUSTARD® (ILLUMINA, San Diego, CA, USA). The quality score of a base is calculated by comparing the fluorescence signals of the previous and following bases. The algorithm does not expect a repeated motif in the beginning of a sequence (AAA) and therefore estimates poor quality scores. The severe decrease in base quality at the 3' end of the sequence indicates the existence of a higher level of sequencing errors at the 3' end. The proportion of unique sequences (sequences occurring only once) ranged from 35% when considering a sequence length of 29 bp to 55% when considering a sequence length of 35 bp, which again indicates an increase in sequencing errors with an increase in sequence length. Therefore, sequences were trimmed at 33 bp and filtering rules were applied.

We applied a number of rules to select sequences for further analysis: (a) screening for properties of the RRL (i.e., discarding sequences without the restriction motif "AAA" in the 5' end); (b) filtering for sequence ambiguity, and (c) filtering for SQ. Filtering for sequence ambiguity was performed by removing sequences with homopolymers and removing sequences with a high rate of re-sampling. Sequences with a re-sampling rate above an expected level were discarded as potentially paralogous sequences. Because pre-EnsEMBL *Sus scrofa* build 7 comprises approximately 70% of the porcine genome, unique alignment of reads does not guarantee that there is no other similar (duplicated) sequence in the remainder of the genome. Therefore, potential paralogous sequences should be eliminated from the data set to avoid identification of false-positive SNPs. This was done by estimating the ratio between the total number of fragments obtained after filtering for the restriction motif, unknown bases, homopolymers, and the numbers of fragments generated from an *in silico* digest. This ratio, the estimated average level of sequence re-sampling, was estimated at 35x, and sequences with a frequency approximately 3-fold greater (>100x) were removed.

The Illumina 1G Genome Analyzer® produces quality scores analogous to PHRED scores. SQ has been defined in previous studies as the average of the quality scores of all bases in a sequence, and the threshold has been set to a minimum of SQ = 20 (Van Tassell et al. 2008; Hillier et al. 2008), which implies that on average, 1 in 100 bases is wrongly identified. Applying this strict filtering rule left sufficient target coverage for SNP identification. In this study, we aimed to evaluate the impact of different thresholds of SQ on the identification of true SNPs. SQ was also evaluated by calculating the average of the base quality scores for all the bases of a given sequence. Three data sets with different SQ levels (12, 15, and 20) were generated and compared for SNP identification. These three different data sets are hereafter referred to as Data 12 for a quality level of 12, Data 15 for a quality level of 15, and Data 20 for a quality level of 20. The total number of sequences that remained after applying all of the filtering rules and that were used for alignment with the reference genome for Data 12, Data15, and Data 20 are shown in Table 1.

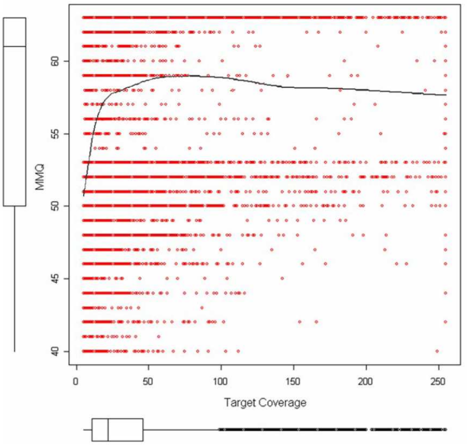


Figure 1 - Maximum mapping quality (MMQ) (mapping quality of the best mapped sequence of a cluster) on an SNP position versus target coverage. Box plots show the data distribution for each parameter. Red dots show MMQ values for the best mapped sequence on an SNP position versus target coverage. The black solid line shows the smooth-fit line.

Comparison of strategies for SNP identification

Sequence mapping was performed using an algorithm that calculates the probability that a sequence maps to a specific target in the genome (Li et al. 2008). Filtered sequences of Data12, Data15, and Data 20 were mapped to pre-EnsEMBL *Sus scrofa* build 7. Mapping quality (which is the probability with which sequences were aligned to a unique location in the genome) was very similar between the three strategies (approximately 60; Table 1). This value indicates an error in the mapping procedure of approximately 1/6000 sequences (Li et al. 2008). After mapping, consensus sequences were generated and SNPs were extracted, creating a large set of potential SNPs. At this stage, the algorithm identified 1,703,360 potential SNPs in Data 12, 1,541,991 potential SNPs in Data 15, and 1,193,814 potential SNPs in Data 20. Four filters were then applied to decrease the rate of false-positive SNPs: 1) SNPs were only accepted if they were identified in targets to which only non-ambiguous sequences were assigned; 2) the maximum mapping quality (mapping quality of the best mapped sequence of a cluster) of the target was larger than or equal to 40; 3) the minimum mapping quality (mapping quality of the sequence with the lowest mapping quality) of a target should be 10 or greater, and 4) the consensus quality (CQ), which measures the probability of the existence of a polymorphism, was 10 or greater (90% of the identified SNP are true positives). Figure 1 shows the relationship between target coverage and mapping quality. The smooth line shows a decrease after target coverage exceeds 100 sequences. This indicates that clusters with a level of target coverage above the expected number calculated from the *in silico* analysis have a lower mapping quality and are less reliable for SNP identification. Additional filters were used to further decrease the rate of false-positive SNPs: 1) occurrence of the minor allele in a minimum of three sequences (to increase the accuracy of detecting SNPs with high MAF), and 2) a maximum target coverage of 100 reads. Again, the restriction of maximum target coverage aims to decrease the rate of false-positive SNPs identified in potential paralogous regions that align to each other because the available assembly only comprises around 70% of the total pig genome. The results allowed us to identify a larger number of SNPs in Data 20 (Table 1) with a higher level of CQ, lower target coverage, and similar MAF values as compared to Data 12 and Data 15. Although a larger set of sequences was used in Data 12, resulting in a higher number of potential SNPs, the actual number of true SNPs was lower due to the removal of more false positives in the final round of filtering. This indicates that a large number of sequences from this data set were mapped ambiguously, introducing noise into the analysis, and shows that the application of filters for SNP selection is crucial for decreasing the rate of false positives. Because the DNA pool contained 10 genomes and the threshold for the minor allele count was three sequences, the observed MAFs are greater than 0.1 for all strategies analyzed [see Additional file 1] and quite adequate for the use of these SNPs in whole genome association and genomic selection studies. For a higher level of SQ (Data 20), more SNPs with MAFs between 0.1 and 0.2 were identified.

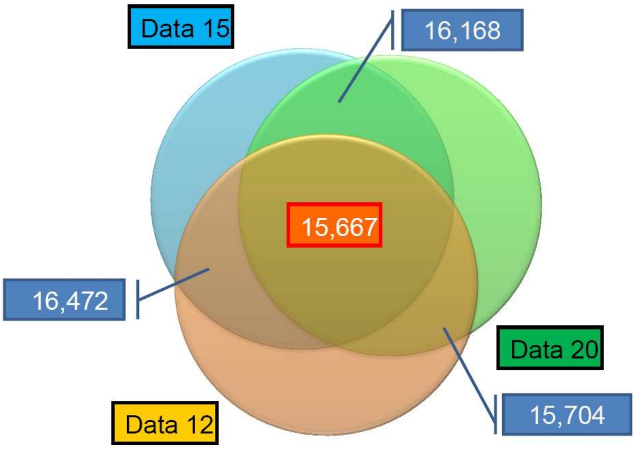


Figure 2 - Venn diagram showing the number of identical SNPs between the analyzed data sets with different levels of sequence quality.

This indicates that in SNP discovery studies aimed at identifying rare SNPs, greater sequence depth and higher levels of SQ are advisable. A large number of the SNPs identified using Data 12 were also identified using Data 15 and Data 20 (Figure 2) as a result of the CQ threshold used in the analysis (90% correct SNP calling rate). Our results indicate that in cases of lower target coverage, lowering the SQ threshold may increase the SNP discovery rate while keeping the rate of false-positive SNPs low.

Table 1 - Sequence production and filtering for the three strategies used to identify SNPs.

	Data 12	Data 15	Data 20
Total sequences after filtering	45,498,558	41,610,684	34,061,918
Total number of SNPs	16,768	17,047	17,489
Mapping quality[a]	61.76 (0.027)	61.78 (0.027)	62.02 (0.0237)
Consensus quality[a]	59.04 (0.257)	60.23 (0.259)	63.30 (0.263)
Target coverage[a]	29.37 (0.164)	29.19 (0.163)	28.60 (0.155)
MAF[a,b]	0.36 (0.0007)	0.36 (0.0007)	0.36 (0.0007)

[a] Mean (s.e.)
[b] MAF, minor allele frequency.

The decrease in quality at the 3’ end of the sequences affected the number of SNPs found per position in the sequence reads. Figure 3 shows that the number of SNPs identified decreased from the 5’ to the 3’ end, indicating that with the strict rules for SNP selection used in our study, base errors in the 3' end were not incorrectly identified as SNPs.

RRL sequence coverage along the pig genome

The reference genome was digested *in silico* and the predicted coverage compared to that of the aligned consensus sequences is shown in Table 2. The consensus sequences aligned evenly to all chromosomes, indicating that the obtained RRL represents a good random sample of the genome. Table 2 also shows that for all chromosomes, the total sequence length that uniquely aligned to the reference genome was somewhat greater than the value expected from the *in silico* digest, most obviously for SSC7.

Sequence polymorphism in the pig genome

Figure 4 shows the SNP map obtained for the SNP identification strategy with an SQ level of 20. A total of 17,489 SNPs were identified and, as expected, more SNPs were found on chromosomes for which more sequence was available in the build of the reference genome (pre-EnsEMBL Sus scrofa build 7). Therefore, chromosomes SSC1, SSC4, and SSC14 contain the largest number of SNPs identified. When analyzing the nature of the base changes, we found that transitions were more frequent than transversions and comprised 67.15% of the identified SNPs. This transition-to-transversion ratio is similar to the 2:1 ratio in observed in the human genome (D. G. Wang et al. 1998). As well, this frequency agrees with that reported in an earlier porcine SNP discovery study (Kerstens et al. 2009).

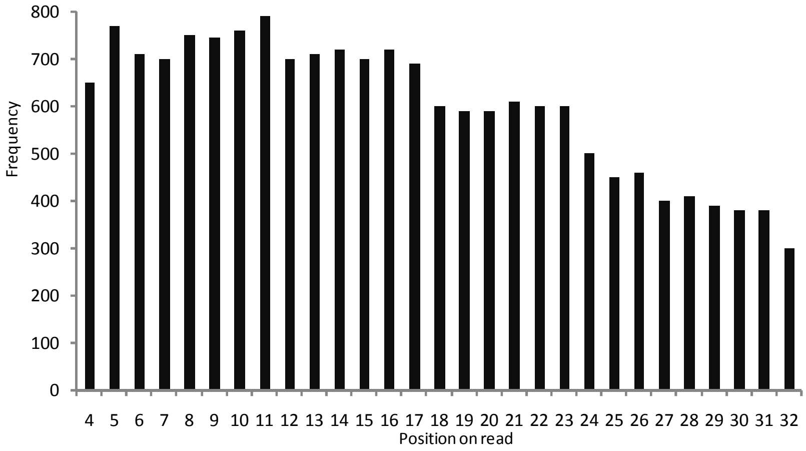


Figure 3 - Number of identified SNPs per position in a short read for Data 20.

This could have resulted from inadequate resolution of DNA fragments during electrophoresis, leading to selection of fragments larger than 150-250 bp and resulting in 13,376,663 bps of aligned sequences, significantly more than the 11,089,914 bps expected from the *in silico* digest of the genome assembly (pre-EnsEMBL Sus scrofa build 7).

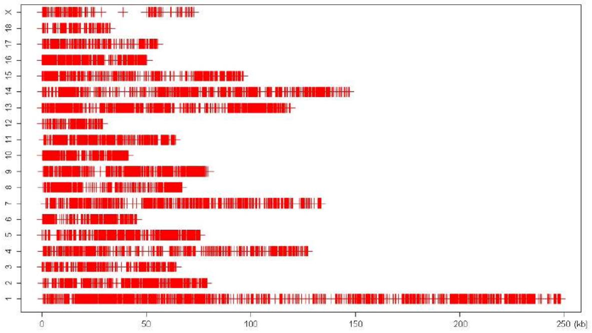


Figure 4 - SNP map of each chromosome based on Data 20. The colored vertical lines represent the location of each SNP.

Nucleotide diversity (Watterson, 1975) across all chromosomes was evaluated in 1-Mb windows based on the pre-EnsEMBL *Sus scrofa* build 7. This analysis showed that SSC 5, SSC10, and SSC12 have relatively greater nucleotide diversity, whereas SSC18 and SSCX have relatively lower nucleotide diversity (Table 2). Figure 5 illustrates the variation in nucleotide diversity and the length of sequence coverage for SSC1. Regions towards the telomeres have greater levels of nucleotide diversity and regions close to the centromere have the lowest levels of nucleotide diversity. The results for window 149, shown in Figure 5, also indicate that in windows of lower sequence coverage, nucleotide diversity may be overestimated. For some chromosomes (SSC4, SSC8, and SSC15), a correlation was observed between the level of GC content and nucleotide diversity, suggesting a relationship between GC content and polymorphism patterns for specific chromosomal regions. Previous studies have shown a relationship between GC content and polymorphism patterns in humans. Such GC-rich regions have been identified as regions of gene conversion and recombination hot spots (Galtier et al. 2001). Our results suggest that such relationships exist in many porcine chromosomes.

Although this study covered only ~1% of the porcine genome, using an RRL allowed estimation of genome-wide nucleotide diversity. Variation in nucleotide diversity and length of sequence coverage for the remaining chromosomes are shown in [see Additional file 1]. The pattern of variation in nucleotide diversity along chromosomes varies between chromosomes; SSC4, SSC8, SSC9, SSC10, SSC13, SSC15, and SSCX have higher levels of nucleotide diversity towards the telomeric regions and lower levels of nucleotide diversity in the centromeric region. On SSCX, large areas flanking the centromere were devoid of nucleotide diversity. Reduced variability in the X chromosome relative to the autosomes has been described for other species, including humans (Payseur & Nachman 2002;

Lu & Chung 2005), *Drosophila* (Kauer et al. 2002), and mice (Baines & Harr 2007), and is explained mainly by the fact that this chromosome has a lower mutation rate and a smaller effective population size (Betancourt et al. 2004).

SNP genotyping and validation

A SNP chip assay was conducted to validate a sample of 3,230 SNPs in the original SNP discovery panel. Of the 3,230 SNPs, 3% failed as a result of the assay design. The validation assay included 68 assayable SNPs exclusively found in Data 12, 147 assayable SNPs exclusively shared between Data 12 and Data 15, 48 assayable SNPs exclusively found in Data 20, and 2,879 assayable SNPs shared between Data 12, Data 15, and Data 20. The correlation between the estimated MAF (calculated from the analysis of short sequences) and the genotyped MAF in the animals used in the discovery panel was calculated for the 2,879 SNPs shared between Data 12, Data 15, and Data 20. The observed correlation of 0.32 was somewhat lower than that reported by Wiedmann et al. (2008). In order to investigate this result, MAF obtained from short sequence data and from genotyping were simulated and correlations were calculated.

Results from simulations showed that the correlation between MAF estimated from short sequences and MAF estimated from genotype data can range from 0.1 to 0.5. Of the total number of SNPs tested, 4% appeared to be monomorphic and 85% had an MAF above 0.2, showing that our strategy yielded a high proportion of informative SNPs useful for whole genome association assays. The fact that the correlation between estimated and observed MAFs was not high could restrict the type of measures used to evaluate genomic variation in population genomics studies using short sequences. For example, estimation of pairwise nucleotide diversity (π) requires accurate estimates of MAF, and a correlation of 0.3 shows that the π estimated from short sequences can be biased. Table 3 shows that assayed SNPs are also informative for other breeds. Large White and Pietrain, which are the breeds used in the cross of the animals analyzed in the discovery panel, show an average MAF closer to the average MAF of the discovery panel.

Duroc has one of the highest rates of monomorphic SNPs and the lowest average MAF. This is in agreement with other diversity studies (Megens et al. 2008) showing that Duroc is genetically more distant from Large White and Pietrain. The breed with highest rate of monomorphic SNPs, Hampshire, has also been reported by previous studies to present high level of genetic differentiation in comparison with Large White (Megens et al. 2008). The average MAF is higher for PW (the commercial boar line used in the discovery panel), probably due to the criteria used for SNP selection, which demanded observation of the minor allele in a minimum of three sequences. For the 68 SNPs identified exclusively in Data 12 and in Data 15, 7% were monomorphic and the correlation between the estimated MAF and the genotyped MAF was 0.08 in the animals used for the discovery panel. This indicates that highly accurate SNP identification is possible even when using a lower SQ threshold for SNP identification, although the cost is less precise estimation of the MAF.

Table 2 – Summary of *in silico* digest of reference genome and analysis of consensus sequences.

Chr	TSL[a] (bp)	TOSL[b] (bp)	TOSL/ 1Mb		Nucleotide diversity		GC content (%)	
			Mean	s.e.	Mean	s.e.	Mean	s.e.
1	1,942,512	2,167,060	8,738.15	204.29	0.0013	0.00007	31.72	0.11
2	564,894	704,052	8,912.05	430.80	0.0017	0.00015	31.82	0.21
3	396,990	510,151	7,971.11	337.50	0.0013	0.00011	32.60	0.22
4	931,194	988,690	7,724.14	286.62	0.0012	0.00009	32.26	0.16
5	541,200	589,399	7,805.12	417.18	0.0019	0.00012	32.29	0.24
6	263,538	354,332	7,874.04	508.06	0.0015	0.00013	32.65	0.24
7	263,538	831,673	6,253.18	283.95	0.0015	0.00010	32.62	0.17
8	590,700	732,338	10,930.42	338.24	0.0011	0.00013	31.36	0.22
9	582,384	763,287	9,541.09	406.33	0.0014	0.00014	32.06	0.18
10	292,842	367,327	8,959.20	380.01	0.0021	0.00019	32.44	0.21
11	527,274	584,831	9,137.98	370.76	0.0015	0.00013	31.85	0.30
12	135,894	174,581	6,020.03	462.41	0.0018	0.00015	33.71	0.36
13	924,396	1,177,791	9,897.40	293.47	0.0014	0.00009	31.64	0.14
14	874,500	952,048	6,476.52	265.24	0.0014	0.00008	32.83	0.17
15	822,822	974,572	10,185.12	326.63	0.0010	0.00009	31.47	0.14
16	402,270	500,390	10,007.80	481.24	0.0016	0.00013	31.97	0.26
17	280,434	303,111	5,511.11	367.98	0.0015	0.00011	33.22	0.24
18	256,806	314,098	9,518.12	593.62	0.0007	0.00010	32,45	0.29
X	495,726	386,932	5,300.44	181.11	0.0005	0.00009	31.77	0.20

[a]-*In silico* total sequence length.
[b]- Total observed sequence length.

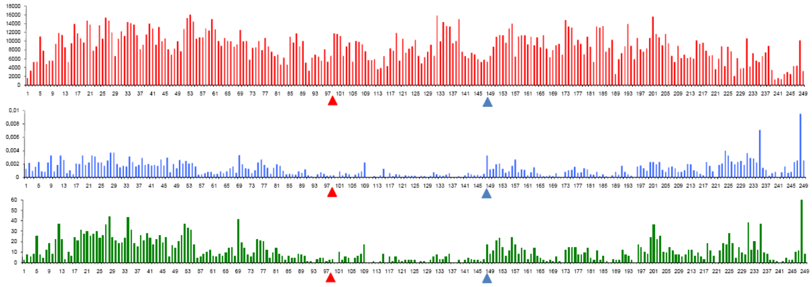


Figure 5 - Sequence coverage, nucleotide diversity, and SNP occurrence along chromosome 1. Each bar represents a window of 1 Mb. Red bars show the length of the aligned consensus sequence, blue bars show the estimated level of nucleotide diversity, and green bars show the number of SNPs found in each window. The red triangle designates the position of the centromere. The blue triangle designates a position where nucleotide diversity is high where coverage is low.

Table 3 - Percentage of monomorphic SNPs and average minor allele frequencies (MAF) by breed for 3,142 SNPs.

Breed	N	Data 12 # SNPs = 68		Data 12 & Data 15 # SNPs = 147		Data 20 # SNPs = 48		All* # SNPs = 2,879	
		M (%)	Avg MAF	M (%)	Avg MAF	M (%)	Avg MAF	M (%)	Avg MAF
Duroc	82	34	0.13	28	0.16	17	0.17	29	0.13
Large White	13	13	0.23	7	0.24	8	0.28	5	0.24
Landrace	80	12	0.22	12	0.20	17	0.20	11	0.21
Pietrain	90	16	0.19	12	0.22	10	0.26	12	0.22
Berkshire	67	32	0.14	31	0.14	21	0.18	28	0.14
Hampshire	59	28	0.15	28	0.16	23	0.17	27	0.16
Wild boar	20	34	0.17	25	0.19	17	0.20	23	0.18
PW	6	13	0.27	6	0.33	6	0.37	4	0.33

*SNPs identified in Data 12, Data 15, and Data 20.

Conclusion

We presented a strategy for using short sequences derived from second generation sequencing technology to efficiently identify large numbers of SNPs with MAF estimates at a low false discovery rate. These results show that by lowering the SQ it is possible to identify SNPs while still keeping the false discovery rate low, although the cost is a lower correlation between the estimated and true MAFs. Finally, our data show that nucleotide diversity is quite variable among porcine chromosomes and is particularly low on SSCX.

Methods

Library construction and sequencing

DNA was extracted from five individual boars produced from a cross between Landrace and Pietrain. Extracted DNA was pooled (60 ng) and digested with *Dra*I (100 units; New England Biolabs, Ipswich, MA, USA) at 37ºC for 16 hours. The fragments were separated by 1.2% agarose gel electrophoresis (2 hours; 60 volts), and 150-250 bp-fragments were eluted (yielding 1069 ng), end-repaired, and ligated to Illumina's oligonucleotide adapters. Fragments were end-sequenced using the Illumina 1G Genome Analyzer.

The Illumina 1G Genome Analyzer generates image information that is translated by BUSTARD® into sequences and base quality scores similar to Phred (Ewing et al. 1998). Using PERL scripts prepared by the authors, quality scores were converted into PHRED scores.

Sequence mapping and SNP identification

Filtering rules were applied to select sequences that were mapped to the pig genome (pre-EnsEMBL Sus scrofa build 7 n.d.) using MAQ 0.6.6 (Li et al. 2008). The algorithm implemented in MAQ calculates a mapping quality for each sequence that measures the probability that a sequence belongs to a specific target (Li et al. 2008). Mapping was performed allowing two mismatches and a mutation rate of 0.001. To generate consensus sequences, the algorithm implemented in MAQ estimated CQ, which is the value at which the probability of each genotype is maximized (Li et al. 2008). Consensus sequences were generated allowing a maximum of four mismatches since the expected SNP frequency in pigs is 1⁄336 bp (Jungerius et al. 2005); therefore, the percentage of clusters with more than one SNP should be low.

Nucleotide diversity analysis

The total aligned consensus sequence was obtained using option *cns.view* of MAQ (Li et al. 2008). Full output was filtered using the same rules applied to SNP identification. A PERL script was developed by the authors to calculate mapping density, nucleotide diversity (Watterson 1975), and GC content over 1-Mb windows.

Validation

A total of 3,142 SNPs located in *Sus scrofa* (pre-EnsEMBL Sus scrofa build 7) were validated using the Illumina Infinium® Genotyping assay on an Illumina® BeadStation. Oligonucleotides were designed, synthesized, and assembled into oligo-pooled assays by Illumina Inc.

Individual DNA samples from the animals used for the SNP identification panel (PW) were genotyped, plus one more animal from PW, 20 samples of Wild Boar, 136 samples of Large White, 80 samples of Landrace, 82 samples of Duroc, 90 samples of Pietrain, 67 samples of Berkshire, and 39 samples of Hampshire.

Authors' contributions

AJA designed and developed the PERL pipeline for sequence filtering, performed all bioinformatics and statistical analyses, and wrote the manuscript. H-JM was involved in discussions about all of the analyses performed, developed the PERL script to perform nucleotide diversity analysis, GC content, mapping density, and *in silico* digests. HHDK developed a PERL pipeline to parallelize sequence mapping analysis and to extract information for SNP statistical analysis. HCMH was involved in discussions about all of the analyses performed. BD and RPMA collected and prepared the samples for sequencing. MAMG coordinated and supervised the project. JD provided facilities for DNA sequencing. MAMG, H-JM, and HCMH assisted in manuscript preparation. All authors read and approved the final manuscript.

Acknowledgments

This work was conducted as part of the SABRETRAIN Project funded by the Marie Curie Host Fellowships for Early Stage Research Training and as part of the 6th Framework Programme of the European Commission. Sequencing costs were funded by the Institute for Pig Genetics, the Netherlands. The authors thank Luca Ferretti (Universidade Autonoma de Barcelona) for his help with the MAF simulations.

References

Ahmadian, A., Ehn, M. & Hober, S., 2006. Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta*, 363(1-2), 83-94.

Altshuler, D. et al., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803), 513-516.

Amaral, A.J. et al., 2008. Linkage Disequilibrium Decay and Haplotype Block Structure in the Pig. *Genetics*, 179(1), 569-579.

Baines, J.F. & Harr, B., 2007. Reduced X-Linked Diversity in Derived Populations of House Mice. *Genetics*, 175(4), 1911-1921.

Bentley, D., 2006. Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, 16(6), 545-552.

Betancourt, A.J., Kim, Y. & Allen Orr, H., 2004. A Pseudohitchhiking Model of X vs. Autosomal Diversity. *Genetics*, 168(4), 2261-2269.

Ewing, B. et al., 1998. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy�Assessment. *Genome Research*, 8, 175-85.

Galtier, N. et al., 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, 159(2), 907-911.

Hillier, L.W. et al., 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Meth*, 5(2), 183-188.

Jungerius, B. et al., 2005. Estimation of the extent of linkage disequilibrium in seven regions of the porcine genome. *Animal Biotechnology*, 16, 41-54.

Kauer, M. et al., 2002. Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of Drosophila melanogaster. *Genetics*, 160(1), 247-256.

Kerstens, H.H. et al., 2009. Mining for single nucleotide polymorphisms in pig genome sequence data. *BMC Genomics*, 10, 4.

Li, H., Ruan, J. & Richard, D., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851-1858.

Lu, J. & Chung-I, W., 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proceedings of the National Academy of Sciences*, 102(11), 4063-4067.

Megens, H. et al., 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), 26 pages.

Metzker, M.L., 2005. Emerging technologies in DNA sequencing. *Genome Research*, 15(12), 1767-1776.

Payseur, B.A. & Nachman, M.W., 2002. Natural selection at linked sites in humans. *Gene*, 300(1-2), 31-42.

pre-EnsEMBL Sus scrofa build 7, [ftp:ftp.sanger.ac.uk/pub/S_scrofa/assemblies/PreEnsembl_Sscrofa7].

Schendure, J. et al., Advanced sequencing technologies:Methods and goals. *Nature Reviews Genetics*, 5(5), 335-344.

Schook, L.B. et al., 2005. Swine Genome Sequencing Consortium (SGSC): A Strategic Roadmap for Sequencing The Pig Genome. *Comparative and Functional Genomics*, 6(4), 251-255.

Sham, P. et al., 2002. DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3(11), 862–871.

Van Tassell, C.P. et al., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Meth*, 5(3), 247-252.

Venter, J.C. et al., 2001. The Sequence of the Human Genome. *Science*, 291(5507), 1304-1351.

Wang, D.G. et al., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366), 1077.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), 256-276.

Whiteford, N. et al., 2005. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, 33(19), e171.

Wiedmann, R., Smith, T. & Nonneman, D., 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics*, 9(1), 81.

4

Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing

Andreia J. Amaral

Luca Ferretti

Hendrik-Jan Megens

Richard P.M. A. Crooijmans

Haisheng Nie

Henri C.M. Heuven

Sebatian Ramos-Onsins

Miguel Perez-Enciso

Larry B. Schook

Martien A.M. Groenen

In preparation

Abstract

Genomic footprints of pig domestication and selection were identified using GA sequences from pooled Reduced Representation Libraries of four pig breeds and the wild boar covering ~2% of the genome generated using massive parallel sequencing. We selected the domesticated breeds, Large White, Landrace, Duroc and Pietrain, which have been under strong selection for muscle development and growth, behavior and coat color. Using specifically developed statistical methods that accounted for DNA pooling, reduced read depths, and sequencing errors, we provide genome-wide estimates of nucleotide diversity and genetic differentiation in pig. In each breed and the wild boar widespread signals suggestive of positive and balancing selection were observed. Of the breeds studied, Pietrain is the breed most intensively selected for muscle development and this breed displayed the strongest signals of selection. Although, most signals were population-specific, affected genomic regions harbored genes for the same biological categories including coat color, brain development, muscle development, growth, metabolism, olfaction and immunity. We observed that in regions harboring genes related with muscle development and growth the genetic differentiation is higher between breeds than between a given breed and the wild boar. Results suggest that although domesticated breeds have experienced similar selective pressures, selection has acted upon different genes probably due to the occurrence of multiple domestication events in Europe or by the subsequent introgression of Asian alleles. This study suggests that selection events have affected about 7% of the porcine genome and that massive parallel sequencing of genomic pools is a cost-effective approach to identify footprints of selection.

Introduction

Animal domestication is a human-mediated process by which captive animals adapted to specific environments and selective pressures resulting in noticeable phenotypic changes compared to their wild counterparts. Adaptation to the captive environment and thousands of years of selective breeding have caused rapid genetic changes (Price 1984). Understanding how domestication has shaped the patterns of genetic variation is important, because domestication can be equaled to rapid evolution triggered by human-generated pressures. Previous studies have shown that domestication was associated with selective pressures on specific genes related to growth (Van Laere et al. 2003) and coat color (Moller et al. 1996; Fang et al. 2009). Artificial selection has reduced DNA sequence polymorphism in chicken (Muir et al. 2008) and may have contributed to the large extent of linkage disequilibrium in several domesticated species (Amaral et al. 2008; Farnir et al. 2000; McRae et al. 2002). However, the degree to which adaptive evolution has affected DNA polymorphism on a whole-genome scale has not been studied in much detail. Nor have any studies determined which type of selection is most prevalent in domesticated animals. Domestication of the pig, in particular, provides an opportunity to study how selection has affected DNA polymorphisms because, unlike other domesticated species, its wild ancestor still exists. The domestication of the European wild boar began approximately around 11,000 B.C. (Larson et al. 2007). The process of selection for different environments within the European continent resulted in the generation of a wide variety of pig breeds with quite divergent phenotypes (Porter & Tebbit 1993). Behavioral traits were also modified since life in captivity causes changes in social structure, reactions towards humans, reproduction, and intra specific relations (Price 1999). Thus, domestication of the pig can be defined as an admixture of artificial selection for favorable traits and natural selection for adaptation to captivity (Driscoll et al. 2009). Nowadays, four breeds dominate pig production worldwide on a large scale. These breeds, Large White, Landrace, Duroc and Pietrain have distinct phenotypes generated by human-mediated selection applied with different intensities according to the desired production aptitudes for each breed.

Selection generates footprints in DNA polymorphisms that can be detected through the study of allelic variation of SNPs (single nucleotide polymorphisms). In human populations, previous studies have revealed footprints of recent positive selection involving genes linked to response to malaria (Sabeti et al. 2002), dairy farming (Bersaglieri et al. 2004), and brain development (Evans et al. 2005). Recent studies have also detected signatures of recent balancing selection in the humans in response to disease (Meyer et al. 2006) and, in response to the need for individual recognition and survival (Alonso et al. 2008). However, those studies were based on SNP genotypes that had been initially identified by an ascertainment (or SNP discovery) process. Alternatively massive parallel sequencing (MPS) technologies provide an opportunity to obtain genome-wide data supporting the characterization of a species genetic diversity without ascertainment bias.

Within a large scale SNP discovery project (Ramos et al. 2009), a large data set of ~380 million Genome Analyzer (GA, Illumina) sequences was generated from pooled Reduced Representation Libraries (RRLs) of four commercial world-wide dominant pig breeds and the wild boar. Here the use of this MPS data for the evaluation of nucleotide diversity and genetic differentiation in the porcine genome and the identification of signatures of selection is demonstrated. Our results suggest a prevalence of positive selection for coat color, behavior, muscle development, and metabolism associated with domestication. Furthermore, these results similarly indicate that olfactory receptors and the immune system (MHC genes) have most likely undergone a process of balancing selection in both the domesticated European pig breeds and the wild boar.

Results

***Sequence* analysis**

We analyzed approximately 380 million GA sequences generated from pooled RRLs of four diverse domestic pig breeds (Duroc, Landrace, Large White, and Pietrain) with divergent phenotypes (Table S1) and the wild boar. The sampled animals are non-related and are representative of the global breed population. Raw GA sequence data was preprocessed to remove errors (see Methods), 200 million GA sequences remained that were aligned to the reference assembly. A total length of sampled sequences of approximately 2% of the porcine genome fitted alignment quality parameters (see Methods), and the average sequencing depth ranged from 7.5× for wild boar to 10× for Duroc (Table 1). SNPs observed included 70% transitions and 30% transversions (Fig.1 S2) and rare variants (SNPs observed in only one read) were nearly absent (Fig.2 S2). The correlation between the GC content and the total number of aligned bases per cluster was approximately 0.70 for all breeds and for wild boar.

Table 1 - Summary statistics of sequence filtering and alignment of overall chromosomes in pig breeds and wild boar.

Population	Total raw sequences	Total filtered sequences	Total aligned length	Average Sequence depth
Large White	81,501,174	45,233,951	40,580,40	7.8
Landrace	88,02,147	43,341,083	42,832,864	8.3
Pietrain	71,561104	42,242,606	51,566,246	9.7
Duroc	75,925,390	36,424,674	37,176,40	10
wild boar	65,053,290	28,723,892	29,229,683	7.5

Genome-wide estimates of nucleotide diversity in pig

Estimates of nucleotide diversity ($\hat{\theta}_W$) were estimated in windows of 500 kb for each chromosome and for each breed (Fig. S3). For each breed, the X chromosome showed the lowest nucleotide diversity (0.0008); the estimate of nucleotide diversity was between 0.001 and 0.0022 for the autosomes. In particular, the values of nucleotide diversity of Pietrain on SSC8, SSC15, and SSC18 were markedly lower than the nucleotide diversity observed in the other breeds and the wild boar. The Large White breed displayed the highest level of nucleotide diversity for most chromosomes. The value of $\hat{\theta}_W$ varied along the chromosomes with a similar pattern across breeds and in wild boar (Fig. S4). As expected from observations in other eukaryotic species, the level of nucleotide diversity was generally lower towards the centromere and increased towards the telomeres.

Genome-wide signals of recent selection

In order to identify outlier regions with unusual estimates of nucleotide diversity, and therefore, represent candidate regions that might be under selection, we obtained the 95% confidence interval (C.I.) of $\hat{\theta}_W$ by performing neutral coalescent simulations with recombination (see Methods). Genomic regions with values of $\hat{\theta}_W$ outside the boundaries established by the C.I. were classified as candidate regions for recent selection. The numbers of candidate regions for selection that were identified per breed are summarized in Table 2. Of the 4 breeds, Pietrain is the breed most intensely selected for muscle growth and this breed showed the highest number of regions with a significantly extreme value of $\hat{\theta}_W$ whereas the wild boar had the lowest. Most of the candidate regions that might be under selection (~70%)

potentially represent breed-specific selective events since they were found in only one of the five pig populations (Table 2).

Approximately 33% of the candidate regions with a value of $\hat{\theta}_W$ below the lowest C.I. boundary (LT) were shared among the white breeds, Pietrain, Large White, and Landrace. Only 10% were shared between each breed and wild boar and Duroc was the breed that shared fewest regions with the other breeds (Table 2). This is consistent with the pattern shown by the correspondence plot for these genomic regions, where a spatial distribution of the LT signals reflects how many LT regions are shared (Fig.1A).

Table 2- Number of regions per breed with significant values of nucleotide diversity ($\hat{\theta}_W$).Values shown above the diagonal are the number of regions with low $\hat{\theta}_W$ values that were shared between breeds (parentheses enclose the percentage, calculated over the minimum TWLT between the pairs compared); values shown below the diagonal are the number of regions with high $\hat{\theta}_W$ values that were shared between breeds (the percentage in parentheses is calculated over the maximum TWHT between the pairs compared).

Breed	TWHT	TWLT	LW	LR	P	Du	WB
Large White (LW)	391	317		124 (39)	119 (38)	88 (28)	32 (10)
Landrace (LR)	446	354	155 (35)		122 (34)	87 (25)	34 (10)
Pietrain (P)	478	408	164 (34)	200 (42)		86 (21)	32 (8)
Duroc (Du)	421	331	137 (33)	150 (36)	147 (35)		27 (8)
wild boar (WB)	226	111	91 (40)	100 (44)	107 (47)	90 (40)	

TWHT- total number of windows with a significantly high value of $\hat{\theta}_W$. TWLT- total number of windows with a significantly low value of $\hat{\theta}_W$.

The genomic region containing the *KIT* gene on SSC18 is an example of a signal that was shared among the Large White, Landrace, and Pietrain breeds, but not by Duroc or wild boar (Fig. 2 A). The *KIT* gene, influences coat color and has been mapped to SSC8p12 (Van Laere et al. 2003). Several of the observed LT signals that were consistently shared between breeds but not by wild boar revealed to harbor genes related to brain and neuron function (Table S5) such as, for example, *PPP1R1B* located on SSC 12 and *LRRTM* located on SSC 2.

Approximately 30 to 50% of the candidate regions with values of $\hat{\theta}_W$ above the highest C.I. boundary (HT) were shared among the domesticated breeds but only 20% were shared with the wild boar. The correspondence analysis showed that, in these regions, the wild boar clustered together with the white breeds. Duroc, Landrace, and Pietrain were the domesticated breeds that mostly contributed to the spatial arrangement of the plot (Fig 1B).

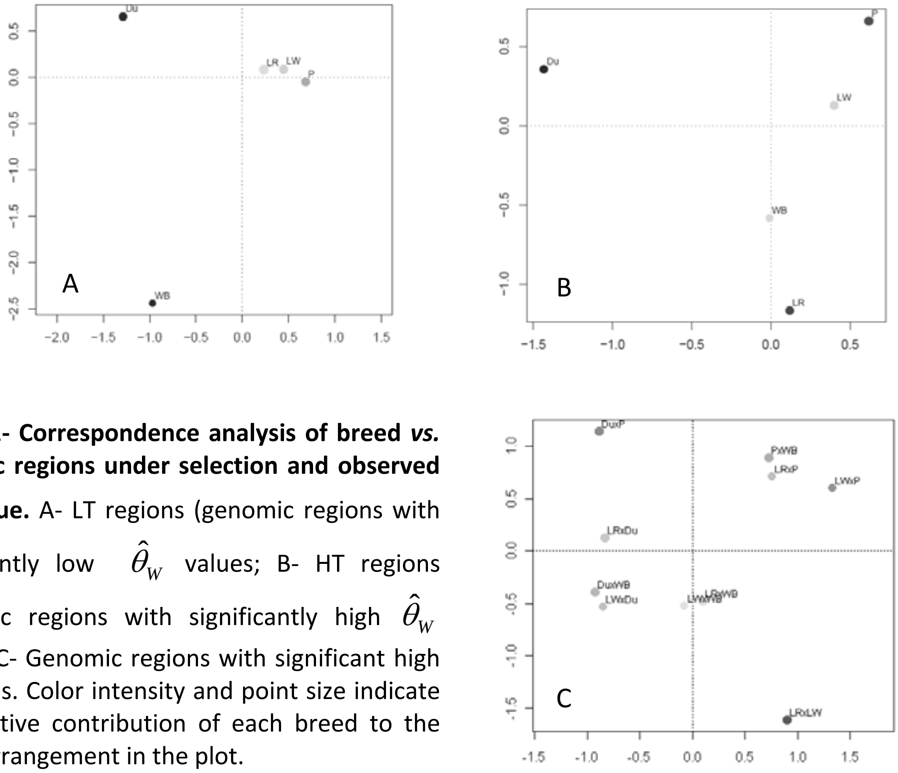


Figure 1- Correspondence analysis of breed *vs.* genomic regions under selection and observed $\hat{\theta}_W$ value. A- LT regions (genomic regions with significantly low $\hat{\theta}_W$ values; B- HT regions (genomic regions with significantly high $\hat{\theta}_W$ values; C- Genomic regions with significant high F_{st} values. Color intensity and point size indicate the relative contribution of each breed to the space arrangement in the plot.

Several HT regions were observed on SSC 7 (Fig. 2 B-C). The most significant regions corresponded to *TRIM26* (zinc finger protein 173) and to *OR4K13* (olfactory receptor 4K13). This significant signal of excess of nucleotide diversity was observed in all the domesticated breeds and in wild boar. The *TRIM26* gene overlaps the genomic region between 24-24.5 Mb and is one of the class I genes of the *SLA* locus (Renard et al. 2006), involved in immune responses. *OR4K13* is located within the genomic region between 86-86.5 Mb and belongs to a large family of genes involved in olfaction and in sensory transduction. The occurrence of misalignments had a small effect in the estimation of $\hat{\theta}_W$. After removing HT regions that deviated from the expected sequence coverage (see Methods), the $\hat{\theta}_W$ values decreased by 1% in most of the windows and did show a significant effect in the determination of HT regions.

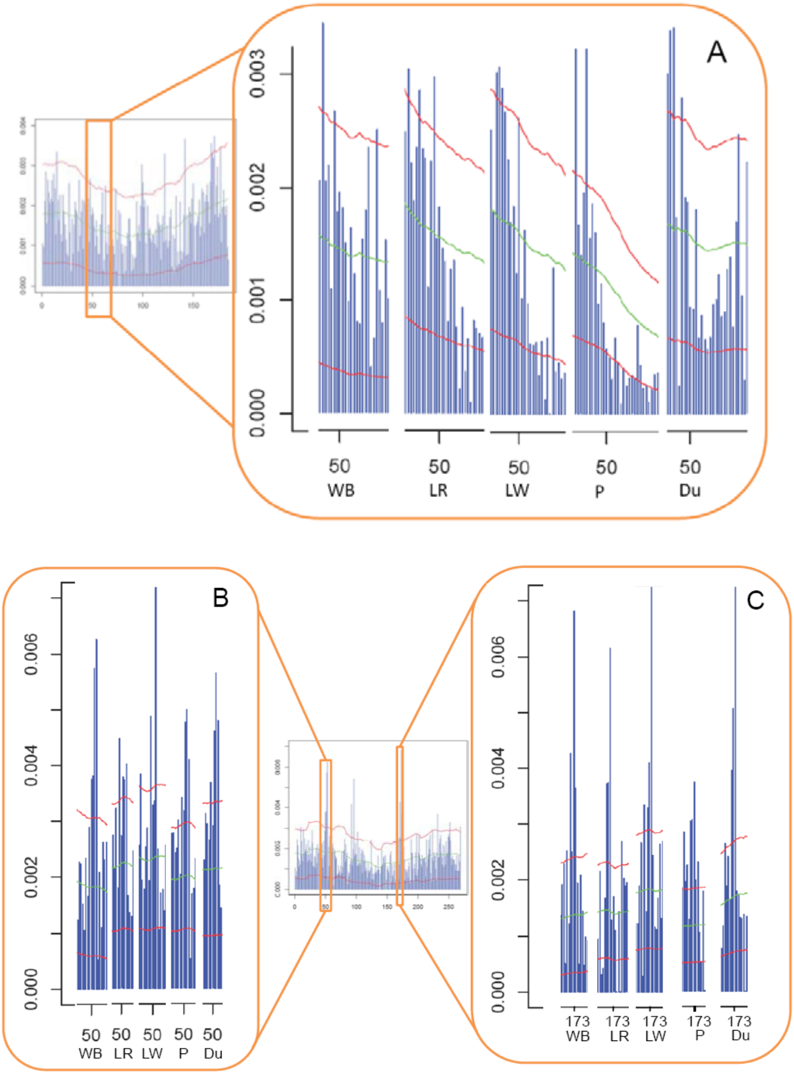


Figure 2- Variation of $\hat{\theta}_W$ along SSC8 and SSC7 for wild boar. Blue bars represent values for a 500 kb window. Red lines represent confidence interval limits with a significance level of 95%. Green lines represent the mean per window. The insets are enlargements of the orange boxes that show details in the variation of $\hat{\theta}_W$ in genomic regions that deviate from the standard neutral model in wild boar and domesticated pig breeds. A- Detail of genomic regions with a significantly low $\hat{\theta}_W$ and that potentially contains the *KIT* gene. B- Detail of genomic regions with a significantly high $\hat{\theta}_W$ and that potentially contains the *TRIM26* gene. C- Detail of genomic regions with a significantly low $\hat{\theta}_W$ and that potentially contains the *OR4K13* gene.

Widespread signals of breed differentiation

The measure of genetic differentiation (F_{ST}) varied along chromosomes (Fig.1 S6). Most F_{ST} values were significantly different from 0 (p<0.05) (Fig. 2 S6) displaying an overall mean of 0.122 with a standard deviation of 0.187. The F_{ST} values for all pair-wise breed comparisons are shown in Table 3. The Duroc breed displayed the highest F_{ST} value compared to the other breeds. The Pietrain and Duroc breeds showed the most differentiation. The Landrace and Large White breeds were the most similar to wild boar. Genomic regions that displayed high genetic differentiation among breeds (F_{ST} values in the 95% quartile with significant p-values (<0.05) were selected for further analysis. A correspondence analysis was performed on these genomic regions in order to investigate how breed pairs share these regions and the observed values of F_{ST} (Fig. 1C). Landrace *vs.* wild boar and Large White *vs.* wild boar were clustered together, indicating that the genetic differentiation between these breeds and wild boar mostly occurred within the same regions of the genome and with the same level of genetic differentiation. In contrast, Pietrain *vs.* Duroc clustered far apart from the other breeds *vs.* Pietrain, indicating that the genetic differentiation between Pietrain and Duroc mostly occurred in different genomic regions.

Table 3 – Genetic differentiation between breeds. Values represent the average genetic differentiation (F_{st}) between breed pairs.

	LW	LR	P	Du
Landrace (LR)	0.10			
Pietrain (P)	0.10	0.12		
Duroc (Du)	0.14	0.14	0.16	
wild boar (WB)	0.10	0.11	0.13	0.13

Types of biological processes under selection

The biological relevance of the genomic regions showing evidence for selection was addressed using Gene Ontology (GO) terms and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database (see Methods). Within the LT regions, several biological processes with a p-value <0.05 were detected but these no longer were significant after correcting for multiple testing. Nevertheless, results were suggestive of a common trend in domesticated pig breeds versus the wild boar. In the case of domesticated pig breeds, a high representation of specific genes involved in KEGG pathway classes that are related with growth and muscular tissue development (*MAPK* signaling pathway, *GnRH* signaling pathway, *p53* signaling pathway, and

VEGF signaling pathway), with metabolism (Arachidonic acid metabolism and the Selenoamino acid metabolism pathways), and melanogenesis was observed. In contrast, in the wild boar, a high representation of genes involved in KEGG pathway classes related with disease (pancreatic cancer and long term depression) was observed. These results are further supported by the observation of high representation of genes also involved in KEGG pathway classes related with growth and muscle development in regions with extreme genetic differentiation between the domesticated breeds and wild boar (Table S7).

Within the HT regions, a significant enrichment (p<0.0001, even after correcting for multiple testing) of genes related to the olfactory transduction pathway was observed in all domesticated breeds and in wild boar (Table 4). Also, a significant enrichment of GO terms related to olfaction and to sensorial abilities was found in all breeds and in wild boar (Table 4). Within genomic regions extremely differentiated between Duroc and Pietrain a significant enrichment of genes related to olfactory transduction (*p*=2x10^{-3}) was observed.

Table 4 - Enrichment of Gene Ontolgy (GO) categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways among HT regions. *p*-values ≤0.05 after multiple testing are shown in bold.

GO category	Landrace	Large White	Pietrain	Duroc	ALL breeds	Wild boar
Sensory perception of smell	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
G-protein, coupled receptor protein signaling pathway	0.03	0.04	<0.0001	0.0008	0.02	<0.0001
G-protein, coupled to cyclic nucleotide second messenger signaling	ns	0.05	<0.0001	ns	0.04	ns
neurological system process	0.009	0.003	0.01	0.0008	<0.0001	ns
sensory perception	0.004	0.001	0.02	0.002	<0.0001	<0.0001
response to stimulus	0.003	<0.0001	0.0009	0.002	<0.0001	0.0001
KEGG Pathway						
Olfactory transduction	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**

Discussion

This study showed that MPS of representative DNA pools is an easy, cost-effective approach that can be applied to any species, even in the absence of complete reference genomes. We identified candidate regions within the porcine genome that were putatively under selection by using a coalescent-based estimation of the uncertainty of the observed nucleotide diversity. Genomic regions that had a $\hat{\theta}_W$ value smaller than the lower limit of the confidence interval represent regions where the frequency of the favored allele in a haplotype with unusual low diversity could have increased very rapidly and thus, were considered as potential candidate regions of recent positive selection. In contrast, the interpretation of genomic regions that had a $\hat{\theta}_W$ value larger than the upper limit of the confidence interval is not so straightforward. These regions can either be considered candidate regions for the occurrence of balancing selection or regions with copy number polymorphisms. Nevertheless, this study provided the first genome-wide map of selection in the pig genome, or indeed, in the genome of any domesticated species. Our analyses clearly identified several regions of significant excess or lack of nucleotide diversity. These results, together with knowledge of the functional aspects of the region, provide new insights into the molecular nature of animal domestication and selection.

Variation of nucleotide* diversity *between pig breeds and wild boar

Overall, and in agreement with previous studies (L. Ollivier 2009), the Pietrain breed demonstrated the lowest nucleotide diversity which is indicative of a low effective population size compared to other breeds and a low contribution to the genetic diversity of the *Sus scrofa* species (Laval et al. 2000). It is worth noting that wild boar also revealed low levels of nucleotide diversity as compared to the Large White, Landrace, and Duroc breeds. This low level of nucleotide diversity in wild boar might be attributed to the severe decline of the European wild boar populations that over time might have lead to high levels of linkage disequilibrium and also to lower effective population size (Amaral et al. 2008). As reported previously for other species, low levels of nucleotide diversity in SSCX were observed, reflecting the small effective population size for this chromosome (J. Lu & Chung-I 2005) and nucleotide diversity was higher towards the telomeres and lower around the centromeric area what is consistent with previous studies in pigs (Amaral et al. 2009) and other mammals (Jensen-Seaman et al. 2004; The International SNP Map working group 2001).

Signatures of positive selection and genetic differentiation between breeds

Compared to the wild boar, in the domesticated breeds larger number of genomic regions that were putatively under positive selection were detected. Clustering results based on these regions suggested strong directional selection and are consistent with the history of pig populations. Landrace and Large White clustered together, sharing many of the regions putatively under positive selection and demonstrated the lowest levels of genetic differentiation between breeds. In fact, Landrace originated in Denmark from a cross between wild boar and Large White (Porter & Tebbit 1993). Duroc was the breed that shared less putative regions of positive selection with other breeds and was the most genetically differentiated breed. This result is consistent with results from previous studies (26) adding support to the hypothesis that the Duroc breed originated from European red breeds or the Red Guinea Hog. Our results showed that the genetic differentiation was higher between breeds than between a domesticated breed and its ancestor, the wild boar. These findings suggest that the occurrence of multiple domestication events in Europe resulted in the generation of several domestic pig breeds that are more similar to its ancestor and are highly differentiated in terms of phenotype and genotype. An alternative hypothesis is that the introgression of Asian germplasm in a post-domestication stage would have increased nucleotide diversity in domesticated breeds and the genetic differentiation between breeds. In fact, the use of a pig from Canton in China for the creation of the Large White pig is documented (Porter & Tebbit 1993), and the introgression of Asian alleles has been observed in genetic previous studies (Amaral et al. 2008; E. Giuffra et al. 2000) that may explain, the high levels of nucleotide diversity in Large White and Landrace as also the high genetic differentiation towards the Pietrain and Duroc breeds.

Footprints of selection reveal genetic differences in coat color, growth, and behavior

The biological functions of genes located within the genomic regions that were putatively under positive selection, suggest that a strong recent positive selection may have occurred in the domesticated breeds for coat color, behavior, growth, and muscle development. Concerning coat color, the genomic region containing the *KIT* gene was shown to have unusual low diversity in the white breeds Large White, Landrace, and Pietrain, but not in Duroc and wild boar. Consistent with this observation is the identification of a mutation in the *KIT* gene that was present in all white breeds and absent in breeds with colored coat (Moller et al. 1996).

The process of domestication also results in behavioral changes. Behaviors that are important for survival in nature, like finding food and predator avoidance, do not provide a significant advantage in humanized environments and in captive breeding (Price 1999) whereas it is expected that man would have selected for more docile animals. Indeed, previous studies have shown evidence that domesticated dogs have inferior observational learning skills compared to wolves (Frank 1980). More recently, analysis of gene expression patterns from three regions of the brain of dogs, gray wolves, and coyotes uncovered a unique pattern in the hypothalamus of dogs; in contrast, the amygdala and frontal cortex were less differentiated (Saetre et al. 2004). The hypothalamus controls specific emotional, endocrinological, and autonomic responses, particularly behaviors related to survival. Consequently, these altered expression patterns in the dog may represent adaptations to the novel challenges of human-altered environments (Saetre et al. 2004). Moreover, a recent study showed that human genes involved in brain development and function have been important targets of selection in recent human evolution (Voight et al. 2006). In this study, signatures of positive selection in genomic regions that contained genes related to brain and neuron functions were detected in the domesticated breeds, but not in the wild boar. For example, one region harbored the *PPP1R1B* gene which encodes for the dopaminergic neurotransmitter that is critical for motivated behavior, working memory, and reward-related learning (Meyer-Lindenberg et al. 2007).

Similar trends were found in regions with unusually low nucleotide diversity and in regions with high genetic differentiation that suggest positive selection of genes involved in growth rate, muscle development and melanogenesis in the domesticated breeds. Our results are consistent with previous findings that a mutation in the *IGF2* gene caused a causative major QTL for muscle growth in pig (Van Laere et al. 2003) and with the evidence that evolution for coat color is a consequence of domestication and not a consequence of natural selection or genetic drift . The observed suggestive evidence for selection of genes linked to metabolism, likely reflects selection due to adaptation to human-altered environments and feed. Interestingly, the wild boar showed evidence of positive selection in genes related to disease. This observation may be due to natural selection of individuals with better fitness for survival in the wild. Since the dataset we used comprised around 2% of the genome, it required usage of a relatively large window size of 500 Kb for analysis hence limiting the analyses. This limited the resolution of the gene enrichment analysis. Furthermore, incomplete annotation of the pig genome further limited the availability of GO terms and further reduced the sensitivity of the analysis. However, despite these limitations, the approach utilized provided suggestive evidence of a clear selection footprint of selection of behavioral, coat color, immunologic responsiveness, and growth related traits associated with the domestication of the pig.

Are the MHC and the olfactory receptors under the influence of balancing selection?

Several genes identified in this study, have been shown to be under the influence of balancing selection in other mammals. Our results indicate the maintenance of an unusually high nucleotide diversity in genes of the MHC genes of the porcine genome, suggesting an effect of balancing selection similarly to observations in other mammals such as e.g., dogs (Angles et al. 2005), cattle (Birch et al. 2006), sheep (Miltiadou et al. 2005), rat (Roos & Walter 2005), rhesus macaque (Otting et al. 2005), and humans (Meyer et al. 2006). The overrepresentation of olfactory receptor genes and of genes related to other sensory traits of the pig were very significant in our analysis (*p*<0.001). The maintenance of high variability in olfactory receptors has been observed in humans (Alonso et al. 2008), other primates (Gilad et al. 2005) and mouse (Niimura & Nei 2007). In humans, a model of overdominance has been proposed for the evolution of olfactory receptors (Alonso et al. 2008). Interestingly, our results are also consistent with results from low coverage sequencing of the Iberian pig (M. Pérez-Enciso, pers. comm.). Individuals that are heterozygous for olfactory receptors can potentially double the number of different odorant-binding sites encoded in the genome, thus allowing the individual to discriminate among closely related structural odorants (Lancet 1994). Our results suggest that individual recognition in pigs is crucial for survival not only for the wild boar reared in the wild but also for domesticated pigs in human-altered environments.

Conclusions

Artificial selection produced profound alterations in livestock environments, and undoubtedly left important selective footprints throughout domesticated genomes. To date, research of livestock has been centered in identifying individual genes as candidates for selection. This study provides the first genome-wide characterization of DNA polymorphism of a domesticated species and yields important insights into the types of biological processes that were targets of selection during pig domestication. The used coalescent simulations were relatively simple due to the lack of integration of physical recombination sites and to the incomplete functional annotations in the pig genome. This made the biological interpretation of our results challenging in situations when the target was clear, but the nature of its functionality was not. Therefore, the history of domestication continues to hold many challenging and complex questions that represent a promising area of important and fruitful research. Finally, this study showed that this research can progress rapidly due to the use of massive parallel sequencing proposing a top-down approach, where candidate genes can be identified in a whole-genome approach by using a representative sample of the genome.

Methods

Sequence analysis

We analyzed a total of 380 million 36 bp GA sequences from four pig breeds, including Duroc, Landrace, Large White, Pietrain, and from the wild boar (Table S1) (Ramos et al. 2009). GA sequences were generated from RRL libraries produced from DNA pools of each of the breeds. Details concerning DNA extraction, preparation of DNA pools and RRL libraries can be found in our previous publication (Ramos et al. 2009). Raw GA sequences were trimmed to 33 bp and filtered according to (Amaral et al. 2009). After filtering, the remaining GA sequences were aligned to *Sus scrofa* assembly 8 (ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/PreEnsembl_Sscrofa8), allowing two mismatches using MAQ (Heng Li et al. 2008). Information from MAQ output files was extracted with Perl scripts developed in house. Only unique alignments were considered, clusters were only selected if the read depth was 4-40×. An error model was applied to infer the rate of false SNPs. The error rate was minimized after performing several filtering steps (Text S8).

Modified Watterson estimator

For each cluster, nucleotide diversity ($\hat{\theta}_W$) was estimated according to equation (1), where *n_0* is the sample size as measured by the number of independent chromosomes, n_s (*i*) and *L(i)* are the read depth and length of the *ith* cluster, and *S* is the total number of segregating sites; *pSNP* is the consensus quality generated by MAQ (Li et al. 2008), a_j is the sum of *1/i* from *1* to *j-1*, p_c is the probability that a set of chromosomes randomly extracted (with repetitions) from n_o possible origins contains precisely *j* different chromosomes, and *k* is the derived allele frequency.

$$\hat{\theta}_W = \frac{S - \sum_{SNPs} 10^{-\frac{pSNP}{10}}}{\sum_i L(i)(\sum_{j=2}^{\min(n_s(i),n_o)} p_c(j \mid n_s(i), n_o) a_j - \sum_{k=1}^{n_o-1 n_s(i)} \frac{n_s(i)}{n_o} \left(\frac{k}{n_o}\right)^{n_s(i)-2}} \qquad (1)$$

A dynamic algorithm was programmed in *C* by the authors to estimate $\hat{\theta}_W$. R scripts were developed to perform the required statistical analysis (http://cran.r-project.org/).

Confidence intervals for $\hat{\theta}_W$

We quantified the statistical uncertainty (confidence intervals) of our estimates of $\hat{\theta}_W$ by performing neutral coalescent simulations with recombination using the MaCS (Markovian Coalescent Simulator) program (Chen et al. 2009), and details of the analysis are described in Text S9. The effect of misalignments on regions with high values of $\hat{\theta}_W$ was estimated by removing groups of consecutive clusters with a distribution of sequence coverage that deviated from the expected empirical distribution along the chromosome (p<0.05).

Modified estimator of genetic differentiation

The global statistic for multiples sites (F_{ST}) was defined previously by Nei (1973). We modified the definition in order to take into account pooled GA sequencing, sequence and consensus errors, and ascertainment bias towards singletons. For each population pair, we used Perl scripts to select aligned bases that commonly covered genomic regions in both breeds and a *C* script was used to estimate $\hat{F}_{ST}$ (Text S10).

Analysis of the types of genes involved in pig domestication and selection

We used the annotation for *Sus scrofa* assembly 8 available from pre-Ensembl (ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/PreEnsembl_Sscrofa8), generated by orthologous comparisons with human transcripts. Genes were considered to be overlapping with a genomic region of interest when the position of the gene was contained inside or partially inside the boundaries of the genomic region. The human Ensembl gene IDs were used to extract human Entrez gene IDs and protein family IDs by querying the Ensembl database (http://www.ensembl.org/) via the R package, biomaRt (Durinck et al. 2005). Using AnnotationDbi, we built a customized annotation R package with using the Entrez gene IDs (Pages et al. 2008). GOstats package (Falcon & Gentleman 2007) was used to analyze enrichment in GO terms and KEGG pathways (http://www.geneontology.org/; http://www.genome.jp/kegg/pathway.html). Within the GOstats package we applied a conditional hypergeometric test algorithm (Benjamini-Hochberg procedure), for multiple testing correction. The conditional hypergeometric test determined whether a GO term/KEGG pathway was significant when there was evidence beyond that provided by its significant children. Only the enriched GO term/KEGG pathway with raw *p*-values < 0.05 were used for further interpretation in this study.

Acknowledgments

We thank to Giuseppe Jacopo Guidi who provided the server for the simulations. This work was conducted as part of the SABRETRAIN Project funded by the Marie Curie Host Fellowships for Early Stage Research Training and as part of the 6th Framework Programme of the European Commission. This work was supported by USDA AG 2008-34480-19328, USDA-ARS 538 AG58-5438-7-3171, USDA AG 2007-04315 and EU SABRE project FOOD-CT-2006-01625. MPE is funded by a MICNN grant AGC2007-65563-C02-01/GAN (Spain). SERO is recipient of a *Ramon y Cajal* position and is funded by a MICNN grant CGL2009-09346/BOS (Spain).

References

Alonso, S. et al., 2008. Overdominance in the Human Genome and Olfactory Receptor Activity. *Mol Biol Evol*, 25(5), 997-1001.

Amaral, A.J. et al., 2008. Linkage Disequilibrium Decay and Haplotype Block Structure in the Pig. *Genetics*, 179(1), 569-579.

Amaral, A.J. et al., 2009. Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *B.M.C. Genomics*, 10, 374.

Angles, J., L. J. Kennedy & N. C. Pedersen, 2005. Frequency and distribution of alleles of canine MHC-II DLA-DQB1, DLA-DQA1 and DLA-DRB1 in 25 representative American Kennel Club breeds. *Tissue Antigens*, 66(3), 173-184.

Bersaglieri, T. et al., 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, 74(6), 1111-1120.

Birch, J. et al., 2006. Generation and maintenance of diversity in the cattle MHC class I region. *Immunogenetics*, 58(8), 670-679.

Chen, G.K., Marjoram, P. & Wall, J.D., 2009. Fast and flexible simulation of DNA sequence data. *Genome Research*, 19, 136-142.

Driscoll, C.A., Macdonald, D.W. & O'Brien, S.J., 2009. From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences of the United States of America*, 106 Suppl 1, 9971-9978.

Durinck, S. et al., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439-3440.

Evans, P.D. et al., 2005. Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans. *Science*, 309(5741), 1717-1720.

Falcon, S. & Gentleman, R., 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2), 257-258.

Fang, M. et al., 2009. Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs. *Plos Genetics*, 5(1), e1000341.

Farnir, F. et al., 2000. Extensive Genome-wide Linkage Disequilibrium in Cattle. *Genome Research*, 10, 220-227.

Frank, H., 1980. Evolution of canine information procession under conditions of natural and artificial selection. *Zeitschrift für Tierpsychologie*, 5, 389-399.

Gilad, Y., Man, O. & Glusman, G., 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Research*, 15(2), 224-230.

Giuffra, E. et al., 2000. The Origin of the Domestic Pig: Independent Domestication and Subsequent Introgression. *Genetics*, 154(4), 1785-1791.

Jensen-Seaman, M. et al., 2004. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Research*, 14, 528-538.

Lancet, D., 1994. Olfaction. Exclusive receptors. *Nature*, 372(6504), 321-322.

Larson, G. et al., 2007. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences*, 104(39), 15276-15281.

Laval, G. et al., 2000. Genetic diversity of eleven European pig breeds. *Genetics Selection Evolution*, 32(2), 17 pages.

Li, H., Ruan, J. & Richard, D., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851-1858.

Lu, J. & Chung-I, W., 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proceedings of the National Academy of Sciences*, 102(11), 4063-4067.

McRae, A. et al., 2002. Linkage disequilibrium in domestic sheep. *Genetics*, 160(3), 1113-1122.

Meyer, D. et al., 2006. Signatures of Demographic History and Natural Selection in the Human Major Histocompatibility Complex Loci. *Genetics*, 173(4), 2121-2142.

Meyer-Lindenberg, A. et al., 2007. Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. *Journal pf Clinic Investigation*, 117(3), 672-682.

Miltiadou, D. et al., 2005. Haplotype characterization of transcribed ovine major histocompatibility complex (MHC) class I genes. *Immunogenetics*, 57(7), 499-509.

Moller, M. et al., 1996. Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. *Mammalian Genome*, 7(11), 822-830.

Muir, W. et al., 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Sciences*, 105(45), 17312-17317.

Nei, M., 1973. Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12), 3321-3323.

Niimura, Y. & Nei, M., 2007. Extensive Gains and Losses of Olfactory Receptor Genes in Mammalian Evolution. *PLoS ONE*, 2(8), e708.

Ollivier, L., 2009. European Pig Genetic Diversity: A Minireview. *animal*, 3(07), 915-924.

Otting, N. et al., 2005. Unparalleled complexity of the MHC class I region in rhesus macaques. *Proceedings of the National Academy of Sciences*, 102(5), 1626-1631.

Pages, H. et al., 2008. *AnnotationDbi: Annotation Database Interface. R package version 1.6.0*,

Porter, V. & Tebbit, J., 1993. *Pigs*, Helm Information.

Price, E.O., 1984. Behavioral Aspects of Animal Domestication. *The Quarterly Review of Biology*, 59(1), 1-32.

Price, E.O., 1999. Behavioral development in animals undergoing domestication. *Applied Animal Behaviour Science*, 65(3), 245-271.

Ramos, A.M. 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.

Renard, C. et al., 2006. The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics*, 88(1), 96-110.

Roos, C. & Walter, L., 2005. Considerable haplotypic diversity in the RT1-CE class I gene region of the rat major histocompatibility complex. *Immunogenetics*, 56(10), 773-777.

Sabeti, P.C. et al., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909), 832-837.

Saetre, P. et al., 2004. From wild wolf to domestic dog: gene expression changes in the brain. *Molecular Brain Research*, 126(2), 198-206.

The International SNP Map working group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822), 928-933.

Van Laere, A. et al., 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature*, 425(6960), 832-836.

Voight, B.F. et al., 2006. A Map of Recent Positive Selection in the Human Genome. *Plos Biology*, 4(3), e72.

5

Localizing recent selection in the chicken genome from DNA pools

Andreia J. Amaral
Hendrik-Jan Megens
Yalda Zare
Richard P.M. A. Crooijmans
Haisheng Nie
Henri C.M. Heuven
Luca Ferretti
Sebatian Onsin-Ramos
Miguel Perez-Enciso
Addie Vereijken
Sean MacEachern
William M. Muir
Hans H. Cheng
Martien A.M. Groenen

In preparation

Abstract

Genome-wide assessment of nucleotide diversity enables a comprehensive and systematic understanding of chicken domestication and selective breeding. Here we present a genome-wide nucleotide diversity map of four populations of chicken based on Massively Parallel Sequencing of DNA pools. The two meat-type (broiler) populations displayed a higher degree of nucleotide diversity compared to the commercial egg layers, in line with known history and demography of these breeds. Nucleotide diversity was lower at the largest chromosomes (macrochromosomes) than at the smallest chromosomes (microchromsomes), which appears mostly related to differences in recombination rate. Outlier regions with unusual nucleotide variation, indicative of selection, were identified by comparing to expectations under a neutral model with recombination. The location of outlier regions in the genome was different between broiler and layer lines which, likely reflects the different breeding goals. Functional analysis of genes underlying the regions putatively under selection suggested involvement of breeding-goal specific pathways, demonstrating the value of screening for signatures of selection to enhance the understanding the genetic basis of phenotypic change.

Introduction

Describing how domestication and selective breeding shaped patterns of genetic variation in domesticated species is critical to a general understanding of evolution.

Selective breeding can be regarded as a long-term human experiment to alter the phenotypes of species that – while having been conducted 'blind' with regard to the underlying mechanisms is expected to have left a signature in the genomes of domesticated species. Such signatures can be present as aberrations in the variation spectrum, for instance unusually low nucleotide diversity or the presence of exceptionally long haplotypes. A global characterization – i.e. throughout the genome and for many breeds and populations – of these signatures of selection along with the functional knowledge of the region can reveal which genes are linked to traits or diseases with a complex genetic basis (Nielsen and Bustamante 2005). The variation in domesticated species can thus be mined not only to gain a better understanding of consequences of selective breeding, but can aid in elucidating fundamental biological and molecular pathways.

One of the prominent domesticated species is the chicken. Domestication of this species occurred well over 5000 year ago in the Indus Valley (Zeuner 1963) and at least four independent domestication events throughout South and Southeast Asia explain current diversity in domesticated chicken (Liu et al. 2006). During the 20th century, the introduction of genetic evaluations in industrial breeding programs has produced cumulative change towards specific breeding goals, e.g. selection for increasing egg production, making chicken the leading domesticated avian species. Nowadays, industrial production replaced dual-purpose breeds by stocks which are intensively selected specifically for either meat or egg production. These commercial lines are the result of two main tiers of genetic diversity reduction. The first tier of genetic diversity reduction is characterized by the reduced number of breeds which were used to create commercial lines. The second tier of genetic diversity reduction is ongoing and due to the commercial breeding structure based on within-line selection whereas introgression from noncommercial birds is rarely used (Muir et al. 2008).

Given the availability of a reference genome (International Chicken Genome Sequencing Consortium 2004) and of a high density SNP map (International Chicken polymorphism Map Consortium 2004) the chicken is a prime model to study how domestication has affected genomic variation. To date, most examples of selection in domesticated species have been discovered in studies of candidate genes where there was a prior hypothesis of selection (Pollinger et al. 2005). Recently, massive parallel sequencing was shown to offer the opportunity to perform genome-wide scans to identify signatures of selection and to obtain information about the kind of genes most involved in pig domestication (chapter 4).

For the current study, we used Illumina Genome Analyzer (GA) sequence data to create a map of selection across the chicken genome by comparing lines bred specifically for meat (broilers) and for eggs (layers). Our search is aimed at finding loci where there is a strong selection in favor of alleles that have not yet reached fixation and of regions where a more even allele frequency distribution than expected under neutrality is observed. By doing so, we aim to provide preliminary answers about the nature and extent of adaptation that occurred during chicken domestication and selection. The loci that we identify will start to fill in the details about the selective pressures humans applied to chicken, the most important domesticated bird, and about the differences between domestication of birds and mammals.

Results and Discussion

Genome-wide estimates of nucleotide diversity in chicken

We analyzed genome-wide Illumina GA sequence data from chicken SNP chip project (Groenen et al., unpublished results).These data consist of 121,528,957 quality approved GA sequences (see Material & Methods) obtained from reduced representation libraries (RRLs) of pooled DNA of each of four chicken lines, two broiler lines and two egg layer lines. Since different fragment sizes were selected to produce the RRL libraries in broilers and layers (see Material and Methods), the portion of the genome covered was different between populations. The genome coverage was 2.85% and 4.64% in broiler A and broiler B respectively. In layers, GA sequencing covered 1.47% of the genome in brown egg layers (BEL) and 1.24% in white egg layers (WEL). In order to estimate the nucleotide diversity we used a modified version of Watterson's classical estimator (ϑ_W) applicable to GA sequences obtained from pooled RRL libraries, over windows of 500 kb (chapter 4). The average estimates of ϑ_W *are* shown in Table 1 for each of the four lines.

Our observations show that nucleotide diversity is highly increased only in the extreme regions of the telomeres (Supplemental Figure 1), similar as observed for the recombination rate (Groenen et al. 2008). This pattern is different from the observed in other species e.g. human

(Lercher & Hurst 2002) and pig (chapter 4), where nucleotide diversity progressively increases from the centromere towards the telomere.

A higher degree of nucleotide diversity was observed in broilers compared to layers with a broiler/layer ratio for ϑ_W of 1.27—1.77. This is consistent with the results presented by the International Chicken polymorphism Map Consortium (2004) that reported a higher degree of nucleotide diversity in broiler lines versus layer lines (broiler/layer ratio for ϑ_W of 1.15). This lower level of ϑ_W might be explained by a higher level of inbreeding in the population of the WEL due to intense within-line selection and complex background breeding structure. Our results suggest distinct differences in the levels of nucleotide diversity between the layer lines, BEL and WEL, which are consistent with documented breed origin and selection history. All commercial WEL lines are based on the White Leghorn breed, whereas the BEL lines were initially selected from North American dual-purpose breeds (selected both for meat and egg qualities), such as Rhode Island Red and White Plymouth Rock, which originated from Asian and European breeds (Muir et al. 2008).

Nucleotide diversity vs chromosome size and recombination rate

Average nucleotide diversity is shown by grouping chicken autosomes into classes of size, five large macrochromosomes (*G. gallus* (GGA)1-5), five intermediate chromosomes (GGA6-10) and 16 microchromosomes (GGA 11-28) (Table 1, Figure 1), GGA16 and GGA22 were excluded from this study due to insufficient sequence coverage.

The average of the observed nucleotide diversity was significantly higher on microchromosomes and on intermediate chromosomes than on macrochomosomes in all the lines studied except for the WEL (Table 1). This indicates that for broilers and BEL nucleotide diversity is dependent of chromosome size, with an upward trend in nucleotide diversity from macrochromosomes towards microchromosomes (Figure 1) as previously described by (Fang et al. 2008; Megens et al. 2009). The results of the WEL are consistent with previous results (International Chicken polymorphism Map Consortium 2004) but are mostly an effect of reduced sequence coverage in microchromosomes combined with the reduced number of observed variations per window. We observed a significant threefold reduction of nucleotide diversity on the Z chromosome (GGAZ) compared to the autosomes (Table 1). These results are not consistent with the results reported by the International Chicken polymorphism Map Consortium (2004), which by normalizing the SNP rate by the effective length of the alignment over the chromosome show only a smaller reduction of the SNP rate on GGAZ. Our results are consistent with Sundstrom et al. (2004), who studied 13 loci on GGAZ and 14 loci on autosomes and found a ratio for ϑ_W between autosomes and GGAZ of 4.1. The observed lower nucleotide diversity on GGAZ is the opposite of what would be expected. One would expect the GGAZ chromosome to have an elevated mutation rate caused by mutation bias because the GGAZ chromosome spends two-thirds of its time in males, where rates of germline cell division are high (Nanda et al. 2002). These expectations are in line with the estimations of Axelsson et al. (2004) which based on the analysis of 33 loci on autosomes and 28 loci on GGAZ obtained rates of 3.6 x 10^{-9} for the autosomes and 3.9 x 10^{-9} for GGAZ respectively. Reduced diversity on GGAZ may be related to a reduced effective population size. In chicken this may be more pronounced, as male population size is much smaller than female population size due to the social structure in chicken, and two-thirds of Z-chromosome haplotypes come via the male line. This effect can be further amplified in current elite lines, where the offspring of a single bird can exceed 200 and as a result, genetic superiority of a single prime layer or broiler can be expanded – through the four-way cross system – to more than a million times (Muir et al. 2008).

Table 1- The table lists the P-value of the test of differences between the average observed estimates of ϑ_W in macrochromosomes (Macro - GGA1-5), intermediate chromosomes (Int - GGA6-10), microchromosomes (Micro - GGA11-28) and Z chromosome in broiler and layer lines.

Line	Macro vs Micro	Micro vs Int	Macro vs Int	Z vs Macro	Z vs Micro	Z vs Int
Broiler A	0.001544	0.9275	0.009365	<2.2 x 10^{-16}	<2.2 x 10^{-16}	<2.2 x 10^{-16}
Broiler B	0.0003713	0.7649	0.05418	<2.2 x 10^{-16}	<2.2 x 10^{-16}	<2.2 x 10^{-16}
BEL	0.006258	0.4364	0.005593	<2.2 x 10^{-16}	<2.2 x 10^{-16}	<2.2 x 10^{-16}
WEL	0.9478	0.9154	0.9586	<2.2 x 10^{-16}	<2.2 x 10^{-16}	<2.2 x 10^{-16}

We estimated the recombination rate within windows of 500 kb based on data available from (Groenen 2009). The correlation between the recombination rate and the nucleotide diversity was estimated for each chicken population and for each chromosome. Results per group of chromosomes are shown in Table 2 and the results per individual chromosome are shown in Supplemental Table S1. In broiler B and the BEL a significant correlation was found between nucleotide diversity and recombination rate in microchromosomes and intermediate chromosomes. These results are in agreement with previous studies in chicken (Fang et al. 2008) and in other species (Nachman 1997; Lercher & Hurst 2002; Mouse Genome Sequencing

Consortium 2002), where shorter chromosomes have higher recombination rates. In WEL we did not find a correlation between recombination rate and nucleotide diversity. Due to the reduced level of nucleotide diversity observed in this line, the window size did not allow to observe a minimum number of segregating sites.

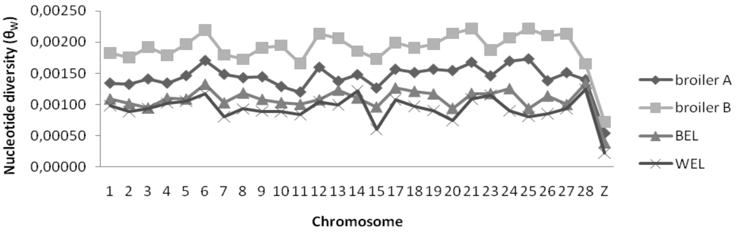


Figure 1- Average of observed nucleotide diversity (ϑ_W) in the chicken chromosomes for the studied egg and meat lines.

Table2- The table lists the correlation coefficient and the corresponding p-value between recombination rate and nucleotide diversity along windows of 500 kb. Values in bold indicate significant correlations (p-value ≤0.05).

	Broiler A r (*p*-value)	Broiler B r (*p*-value)	BEL r (p-value)	WEL r (p-value)
Macrochromosomes	0.02 (0.41)	0.04 (0.22)	0.04 (0.17)	-0.01 (0.85)
Microchromosomes	**0.24** **(0.0)**	**0.28** **(0.0)**	**0.17** **(0.0012)**	0.06 (0.27)
Intermediate chromosomes	0.08 (0.1823)	**0.13** **(0.02)**	**0.14** **(0.01)**	-0.04 (0.50)
All	**0.10** **(0.0)**	**0.11** **(0.0)**	**0.10** **(0.0)**	0.0 (0.93)

Identifying outliers

The goal of our study was to identify regions which may have been under positive or balancing selection. As a consequence of positive selection the favored allele increases in frequency very fast, and the frequency of the entire haplotype it is residing in may also increase. This results in lower genetic variation in adjacent regions and the observed ϑ_W is lower than the expected (ϑ_E) under neutrality. In case of balancing selection, a more even allele frequency distribution than that expected under neutrality (ϑ_E), is observed along a genomic region. In order to identify the HT regions and the LT regions we conducted coalescent simulations along non-overlapping windows of 500 kb to determine the deviations from ϑ_E under 95% confidence. Regions with ϑ_W larger than ϑ_E (with 95% confidence – called HT regions) may be indicative of balancing selection. Similarly, regions with ϑ_W smaller than ϑ_E (with 95% confidence – called LT regions) may be indicative of positive selection. In the WEL line we found a higher number of regions putatively under selection in comparison with all other lines, while the BEL line shows the lowest number of regions putatively under selection (Table 3, Figure 2A; Figure 2B). The broiler lines share more of the regions under selection than layer lines (Figure 2A; Figure 2B). The WEL shares higher number of LT and HT regions with the broiler lines than with the BEL (Figure 2A; Figure 2B). This is an unexpected result because the broiler lines are originated from crosses between male lines, derived from Cornish stock, selected primarily for growth traits and have higher proportion of breast muscle and female lines originated from the same dual purpose breeds as the BEL (Muir et al. 2008). However the WEL shows the highest number of LT and HT regions, meaning that there are more opportunities to find regions in common with broilers. Again the reduced sequence coverage and the reduced number of observed variations in a window could have created a bias and resulted in overestimation of HT and LT regions in WEL. We identified regions putatively under selection which are shared between broilers but not with layers that may reflect the different selection goals for meat and egg lines. For instance, LT signals on GGAZ are absent in layers and so is the LT signal overlapping the *IGF1* gene (Figure 2C). The occurrence of several LT signals on GGAZ for the broiler lines, suggests that reduced variability on this chromosome may be due to positive selection, possibly affecting sex-linked loci related with the improvement of male fitness. The 10 regions with the most significant HT regions are given in Table 4 and the 10 regions with the most significant LT regions are given in Table 5. WEL exhibited the strongest signals of potential positive selection whereas broiler B exhibited the strongest potential signals of balancing selection. We compared our results with previous published QTL analysis and studies focusing on the characterization of nucleotide diversity of a few genes. Considering the available published QTL studies, we only considered the published QTLs with a p-value<0.05 and with a maximum span of 1,000,000 bp. QTLs observed in regions overlapping LT and HT signals (Supplemental Table S2), were mainly related to body weight, Marek's disease and abdominal fat .

Table 3-Number of non-overlapping 500kb windows putatively under selection per breeding line.

	broiler A	broiler B	BEL	WEL
LT regions	183	94	49	219
HT regions	171	94	66	241

Characterizing biological functions in LT and HT regions

Many different genes may underlie a certain phenotype. As a consequence, selection on such a phenotype may result in selection on many different genes. These genes may be dispersed across the genome, but are expected to overlap in biological function. A genome-wide analysis of selection therefore is expected to result in many different regions under selection, but these regions are expected to harbor genes involved in the trait and should result in over-representation of certain functional annotations such as specific Gene Ontology terms (results not shown) or specific KEGG pathway annotations.

For each LT and HT region we determined which genes are overlapping and which biological functions are most represented. The KEGG ontology database provides a classification of genes into pathway categories according to biological process. We tested whether any of these categories showed an enrichment of genes and we show the results for all the categories which presented an uncorrected *p*-value <0.05 (Table 6 and Table 7). The expected counts for many categories are low and therefore, after correcting for multiple testing, we could not detect significant signals of enrichment.

Table 4- Summary of some of the strongest Genome-wide HT signals.

Chr	HT region (Mb)	Genes (HUGO symbol)	Line	θ_W	Mean θ_W	Low CI	High CI
GGA 2	87.5-88	BAG1, CCDC127, CHMP5,FH, ZNF830	WEL	0,0033	0,0009	0,0002	0,0016
GGA 8	2.0-2.5	ATP6V1G3, LHX9,C1orf5, DENND1B, PTPRC,NEK7	BEL	0,0035	0,0012	0,0004	0,0020
GGA 11	6.5-7	PAPD5, ADCY7, NKD1	WEL	0,0021	0,0008	0,0008	0,0009
GGA Z	50.5-51	NANS,CLTA, SHB,LNPEP, TRIM14, RNF38	Broiler A	0,0017	0,0005	0,0005	0,0005
GGA 11	16.5-17	C16orf61, ATMIN, CDYL2, CENPN, DYNLRB2	WEL	0,0020	0,0008	0,0008	0,0009
GGA 1	25-25.5	ING3, C7orf58, FAM3C,TSPAN12, WNT16	WEL	0,0028	0,0010	0,0003	0,0017
GGA 14	1.5-2	FAM20C	WEL	0,0026	0,0012	0,0009	0,0016
GGA 11	8.5-9	C19orf12, VSTM2B, UQCRFS1, UQCRFSL1	Broiler A	0,0025	0,0012	0,0010	0,0015
GGA 20	11.5-12	RAE1,AURKA, CASS4, C20orf43, SPO1,GCNT7, CSTF1, RBM38, TFAP2C, BMP7,PCK1	WEL	0,0017	0,0008	0,0008	0,0007
GGA 2	150.5-151	KCNK9	WEL	0,0027	0,0010	0,0003	0,0017

Table 5- Summary of some of the strongest Genome-wide LT signals.

Chr	LT region (Mb)	Genes (HUGO symbol)	Line	θ_W	Mean θ_W	Low CI	High CI
GGA 13	18-18.5	KCTD16,SLC4A9, SERPIND1,GLP2R, C20orf24,ZBTB26, HIC2,NFATC3	Broiler B	0,0002	0,002	0,0017	0,0026
GGA 13	1-1.5	SLC4A9,HBEGF, NRG,PFDN1, PURA,PSD2	Broiler B	0,0002	0,002	0,0018	0,0026
GGA 15	0-0.5	SERPIND1,UBE2L3, YDJC,CLTCL1, HIC2,SLC25A1, SNAP29,SDF2L1 ,PPM1F	Broiler B	0,0002	0,002	0,0016	0,0019
GGA 18	0-0.5	GLP2R,GAS7, MYH13, DHRS7C	Broiler B	0,0003	0,002	0,0017	0,0023
GGA 20	0-0.5	C20orf24,GGT7, C20orf4,DLGAP4, SFRS6,TGIF2, L3MBTL,ZBTB26	Broiler B	0,0005	0,002	0,0018	0,0026
GGA 17	9.5-10	ZBTB26,CRB2, DENND1A, RABGAP1, ZBTB6,LHX2, RC3H2,STRBP	Broiler B	0,0005	0,002	0,0017	0,0024
GGA 15	0-0.5	HIC2,SLC25A1, SNAP29,PPM1F, SERPIND1,SDF2L1, UBE2L3,YDJC, CLTCL1	Broiler A	0.0000	0,001	0,0012	0,0014
GGA 11	3-3.5	NFATC3,DUS2L, ESRP2,DDX56	Broiler B	0,0001	0,002	0,0012	0,0021
GGA 18	0.5-1	MAP2K4,MYOCD, SCO1	Broiler B	0,0005	0,002	0,0017	0,0023
GGA 11	2.5-3	NECAB2,MBTPS1, RIPK3,SLC38A8, OSGIN1	Broiler B	0,0001	0,002	0,0012	0,0021

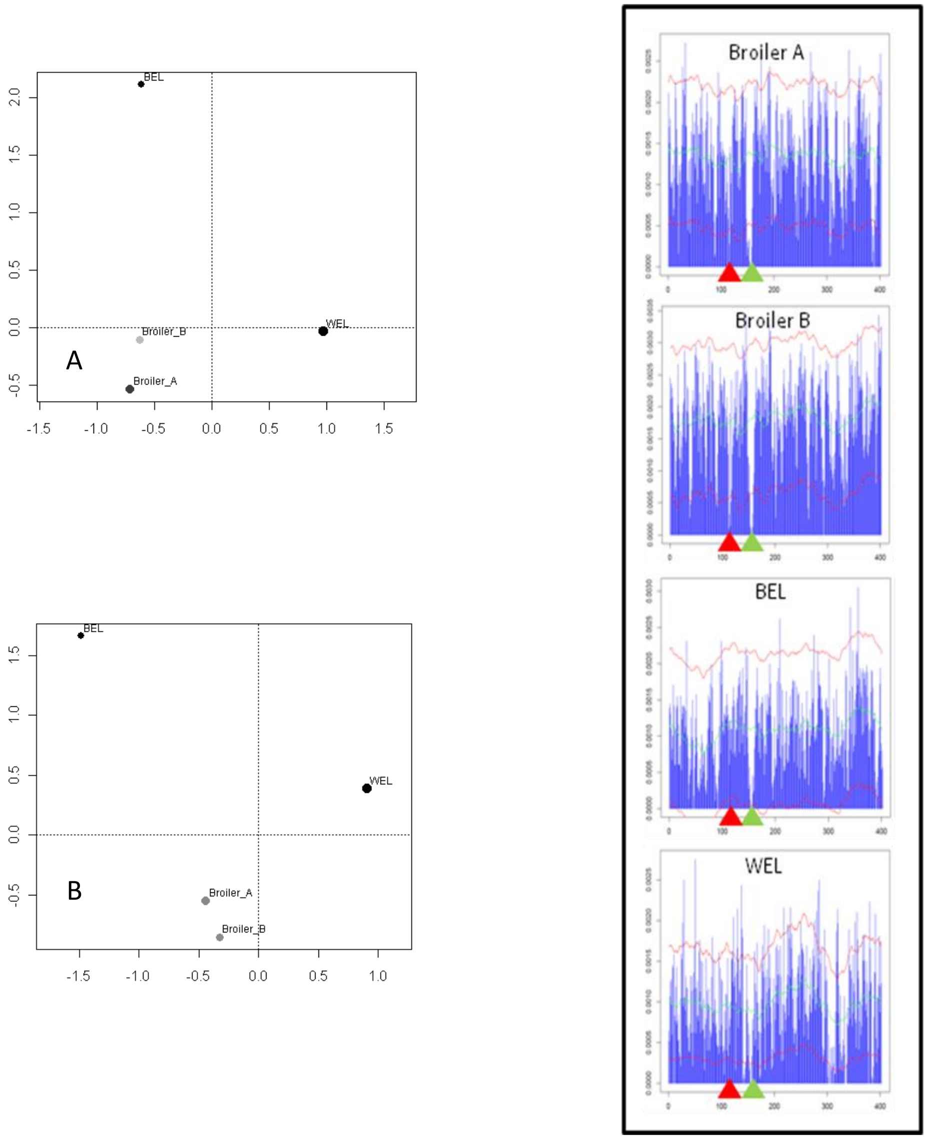


Figure 2 - A-Sharing of HT signals between breeding lines; B- Sharing of LT signals between breeding lines. Color intensity and point size indicate the relative contribution of each breed to the space arrangement in the plot. C- Plot of observed values of nucleotide diversity (θ_W) in GGA1. Vertical blue bars indicate the observed values θ_W for windows of 500 Kb. Green line indicates the value of local mean of θ_W , red lines represent the confidence intervals. The green triangle represents the position of the centromere, and the red triangle shows the position of a LT region, in the region containing the gene *IGF1*. Values of θ_W ranged from 0 to 0.003.

Nevertheless, consistent and biologically meaningful differences between broilers and layers were observed. In HT regions, we found in broilers a high representation of genes related with protein synthesis and sugar metabolism, whereas in layers we found a smaller number of categories for which there is an overrepresentation of genes in these regions related to growth (Table 6). In LT regions, again we observed consistent differences between broilers and layers, in terms of highly represented KEGG categories. Folate biosynthesis was highly represented in layers, which is important in reproduction, whereas in the broilers, genes related with muscle and growth were highly represented. Other KEGG categories related to metabolism were highly represented in LT regions of both layers and broilers (Table 7), possibly as a consequence of selection for adaptation for a wide range of environments. The lack of a significant gene set enrichment either in HT or LT regions may indicate that selection on chicken may have affected genes related to different biological functions or traits under selection such as growth, increased egg production and increased muscle growth are affected by a small number of genes. Moreover for many regions a biological link was not established what can also explain the lack of significance in the enrichment of specific biological pathways. This can be a consequence of the lack of information on gene annotation. The absence of genes in outlier regions can also occur and in this case, meaning the outlier region might be in linkage disequilibrium with loci which have an effect in important traits. Nevertheless, we could observe a distinct pattern between broiler and layer lines that is likely due to differences in selection goals.

Table 6 – Uncorrected p-values for enrichment of KEGG categories among genes overlapping HT regions in broiler and layer lines.

KEGG	KEGG category	Broilers	Layers
4020	Calcium signaling pathway	0.009	-
4120	Ubiquitin mediated proteolysis	0.013	-
602	Glycosphingolipid biosynthesis neo-lactoseries	0.025	-
4070	Phosphatidylinositol signaling system	0.04	-
4910	Insulin signaling pathway	0.047	0.013
530	Aminosugars metabolism	0.049	-
4742	Taste transduction	0.049	-
4012	ErbB signaling pathway	-	0.014
4664	Fc epsilon RI signaling pathway	-	0.023
1040	Biosynthesis of unsaturated fatty acids	-	0.047

Conclusion

This study presents a nucleotide diversity map for four populations of chicken, allowing a detailed and virtually unascertained comparison between populations but also between different regions in the genome. The meat-type chicken (broilers) have the highest nucleotide diversity and the white egg commercial layer the lowest, as expected from known breed history and demography. Furthermore, this study confirms the differences in nucleotide diversity between the macro and micro chromosomes – where macrochromosomes have lower nucleotide diversity – that are highly correlated with differences in recombination rate. On a finer scale, our study has identified chromosomal regions with either a higher or lower degree of polymorphisms, possibly related to selection. Functional analysis of genes in these regions suggests that a number of biological pathways may have been under selection. The pathways identified this way aid in better understanding functional genomic aspects of phenotypic change, and aid in identifying candidate genes involved in important commercial traits.

Table 7 – Uncorrected *p*-values for enrichment of KEGG categories among genes overlapping LT regions in broiler and layer lines.

KEGG	KEGG category	Broilers	Layers
4612	Antigen processing and presentation	0.007	-
4620	Toll-like receptor signaling pathway	0.026	-
530	Aminosugars metabolism	0.053	-
790	Folate biosynthesis	-	0.005553
3410	Base excision repair	-	0.0126
4912	GnRH signaling pathway	-	0.014
910	Nitrogen metabolism	-	0.020
5110	Vibrio cholerae infection	-	0.027
500	Starch and sucrose metabolism	-	0.039
620	Pyruvate metabolism	-	0.039
	Calcium signaling pathway	-	0.057

Material and Methods

Sequence analysis

We analyzed a total of about 121 million 36 bp Illumina GA sequences obtained from RRL libraries of pooled DNA from four commercial lines, including BEL (N=25), WEL (N=25), broiler

A (N=25) and broiler B (N=25). Filters were applied to select higher quality GA sequences: presence of the restriction motif CT in the 5' end; continuous repeats of the same nucleotide lower than 18; average quality score lower than 12. GA sequences were mapped in the chicken genome (assembly WASHUC2, May 2006, available from www.ensembl.org) using MAQ 0.6.6 (Li et al. 2008) using default parameters. SNP identification was performed using MAQ 0.6.6.

Table 8 – Recombination estimates used in coalescent simulations per breed and per chromosome.

Chr	WEL (Ne=50)	WEL (Ne=100)	BEL (Ne=100)	BEL (Ne=150)	broilers (Ne=200)	broilers (Ne=700)
1	0,0005	0,0010	0,0010	0,0015	0,0019	0,0068
2	0,0004	0,0008	0,0008	0,0012	0,0016	0,0057
3	0,0005	0,0009	0,0009	0,0014	0,0019	0,0066
4	0,0004	0,0009	0,0009	0,0013	0,0017	0,0060
5	0,0005	0,0010	0,0010	0,0015	0,0020	0,0071
6	0,0006	0,0012	0,0012	0,0018	0,0024	0,0083
7	0,0006	0,0012	0,0012	0,0018	0,0024	0,0083
8	0,0006	0,0012	0,0012	0,0018	0,0024	0,0084
9	0,0007	0,0014	0,0014	0,0021	0,0028	0,0098
10	0,0008	0,0016	0,0016	0,0024	0,0032	0,0112
11	0,0006	0,0013	0,0013	0,0019	0,0025	0,0089
12	0,0007	0,0014	0,0014	0,0022	0,0029	0,0101
13	0,0006	0,0013	0,0013	0,0019	0,0026	0,0090
14	0,0009	0,0017	0,0017	0,0026	0,0034	0,0120
15	0,0009	0,0017	0,0017	0,0026	0,0034	0,0119
17	0,0010	0,0019	0,0019	0,0029	0,0038	0,0134
18	0,0010	0,0019	0,0019	0,0029	0,0038	0,0134
19	0,0011	0,0021	0,0021	0,0032	0,0043	0,0150
20	0,0008	0,0015	0,0015	0,0023	0,0030	0,0105
21	0,0016	0,0031	0,0031	0,0047	0,0063	0,0219
23	0,0015	0,0030	0,0030	0,0045	0,0061	0,0212
24	0,0015	0,0030	0,0030	0,0045	0,0060	0,0209
25	0,0066	0,0132	0,0132	0,0198	0,0263	0,0922
26	0,0019	0,0037	0,0037	0,0056	0,0074	0,0259
27	0,0023	0,0046	0,0046	0,0068	0,0091	0,0320
28	0,0023	0,0047	0,0047	0,0070	0,0093	0,0327
Z	0,0007	0,0014	0,0014	0,0021	0,0029	0,0100

Thresholds were established to select reliable SNPs. MAQ's minimal map quality for the read, minimal consensus quality and minimal map quality of the best mapping read for each predicted SNP position were set at 10, 5, and 10 respectively. The required read depth ranged from minimum 6 to maximum 180.

Estimation of nucleotide diversity and outlier identification

We used a modification of Watterson's classical estimator (chapter 4) applicable to GA sequences obtained from RRL libraries generated from pooled DNA. The existence of differences in the levels of nucleotide diversity between chromosomes was tested using a two-sample t-test as implemented in package stats in R (http://cran.r-project.org/). The estimation of confidence intervals was based on simulation data, taking into account local recombination rates, sequencing errors and the number of reads sampled for genomic segments of 500 kb. To this end we used MACS (Chen et al. 2009) to perform 1000 coalescent simulations under the standard neutral model with recombination. Since effective population size (Ne) of the studied lines was unknown, we used the estimates of Ne provided for broiler and layer lines from Megens et al. (2009). 1000 simulations were performed by normalizing the estimates of recombination (Groenen et al. 2009) by the lowest possible value of Ne and 1000 simulations by normalizing the estimates of recombination (Groenen et al. 2009) by the highest possible value of Ne (described in Table 8). The values of mutation rate used for the simulations were obtained based on estimates of θ_W (see Table 9). Simulated reads were sampled from simulated 500 kb sequences in exactly the same way as they were obtained in the GA sequence data and θ_W was estimated as described in (chapter 4). Regional 95% confidence intervals were estimated by linear interpolation of the variance between the two simulations (Table 8) and by centering the confidence intervals for each window around a local average of θ_W, taking into account different levels of variability associated with regions of high and low recombination. Genomic regions with a value of θ_W higher than the θ_E under 95% confidence were identified as HT regions. Genomic regions with a value of θ_W lower than the θ_E under 95% confidence were identified as LT regions.

Estimation of correlations between nucleotide diversity and recombination rate

Data available from the most recent linkage map in chicken (Groenen et al. 2009) was used to estimate recombination rates in 500 kb windows. Correlation between nucleotide diversity and recombination rate was estimated using Spearman correlation test implemented in package Hmisc in R (http://cran.r-project.org/).

Table 9- Values of θ used in the coalescent simulations per breed line and chromosome size.

	broiler A		broiler B		BEL		WEL	
	ϑ_W $(x10^{-3})$ Ne=200	ϑ_W $(x10^{-3})$ Ne=700	ϑ_W $(x10^{-3})$ Ne=200	ϑ_W $(x10^{-3})$ Ne=700	ϑ_W $(x10^{-3})$ Ne=100	ϑ_W $(x10^{-3})$ Ne=150	ϑ_W $(x10^{-3})$ Ne=50	ϑ_W $(x10^{-3})$ Ne=100
Macro	1.345	1.708	1.750	1.966	1.087	1.324	0.889	1.5
Int.	1.290	1.487	1.795	1.946	1.026	1.83	0.810	1.174
Micro and Z	0.718	1.680	1.142	2.214	0.945	1.338	0.422	1.243

Analysis of the types of QTLs overlapping HT and LT regions

The QTL databse was downloaded from (http://www.animalgenome.org/cgi-bin/QTLdb/GG/download?file=gbp) and QTLs were considered if were inside or partially inside the boundaries of the HT and LT regions, if the *p*-value <0.05 and the span was smaller or equal than 1,000,000 bp.

Analysis of the types of genes overlapping HT and LT regions

All locuslink genes overlapping the identified HT and LT regions were identified using the annotation of *G. gallus* (ggallus version WASHUC2), Ensembl gene IDs of the chicken genome were extracted using the biomaRt (Durinck et al. 2005) package as were the Ensembl gene IDs of human orthologues. Only chicken genes with a location on genome build 2.1 were considered for this study. Only chicken-to-human one-to-one or many-to-one orthologues were considered for further analysis. Only genes that were inside or partially inside the boundaries of the HT and LT regions were considered. The GOstats package (Falcon & Gentleman 2007) was used to analyze enrichment in GO terms and KEGG pathways (http://www.geneontology.org/; http://www.genome.jp/kegg/pathway.html). Within the GOstats package we applied a conditional hypergeometric test algorithm (Benjamini-Hochberg procedure). The conditional hypergeometric test determined whether a GO term/KEGG pathway was significant when there was evidence beyond that provided by its significant children. Only the enriched GO term/KEGG pathway with uncorrected *p*-values < 0.05 were used for further interpretation in this study.

Aknowledgments

This work was conducted as part of the SABRETRAIN Project funded by the Marie Curie Host Fellowships for Early Stage Research Training and as part of the 6th Framework Programme of the European Commission.

References

Axelsson, E. et al., 2004. Male-Biased Mutation Rate and Divergence in Autosomal, Z-Linked and W-Linked Introns of Chicken and Turkey. *Mol Biol Evol*, 21(8), 1538-1547.

Chen, G.K., Marjoram, P. & Wall, J.D., 2009. Fast and flexible simulation of DNA sequence data. *Genome Research*, 19, 136-142.

Durinck, S. et al., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439-3440.

Falcon, S. & Gentleman, R., 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2), 257-258.

Fang, L. et al., 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.

Groenen, M., 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Available at: http://genome.cshlp.org.ezproxy.library.wur.nl/content/19/3/510 [Accessed December 22, 2009].

International Chicken Genome Sequencing Consortium, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018), 695-716.

International Chicken polymorphism Map Consortium, 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, 432(7018), 717-722.

Lercher, M.J. & Hurst, L.D., 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, 18(7), 337-340.

Li, H., Ruan, J. & Richard, D., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851-1858.

Liu, Y. et al., 2006. Multiple maternal origins of chickens: Out of the Asian jungles. *Molecular Phylogenetics and Evolution*, 38(1), 12-19.

Megens, H. et al., 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. *BMC Genetics*, 10(1), 86.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), 520-562.

Muir, W. et al., 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Sciences*, 105(45), 17312-17317.

Nachman, M.W., 1997. Patterns of DNA Variability at X-Linked Loci in Mus domesticus. *Genetics*, 147(3), 1303-1316.

Nanda, I. et al., 2002. Comparative mapping Z-orthologous genes in vertebrates: implications for the evolution of avian sex chromosomes. *Cytogenetic and Genome Research*, 99, 178-184.

Nielsen R. Bustamante C., 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *Plos Biology*, 3(6), e170.

Pollinger, J. et al., 2005. Selective sweep mapping of genes with large phenotypic effects. *Genome Research*, 15(12), 1809-1819.

Sundstrom, H., Webster, M.T. & Ellegren, H., 2004. Reduced Variation on the Chicken Z Chromosome. *Genetics*, 167(1), 377-385.

Zeuner, F.E., 1963. *A history of domesticated animals*, Harper & Row.

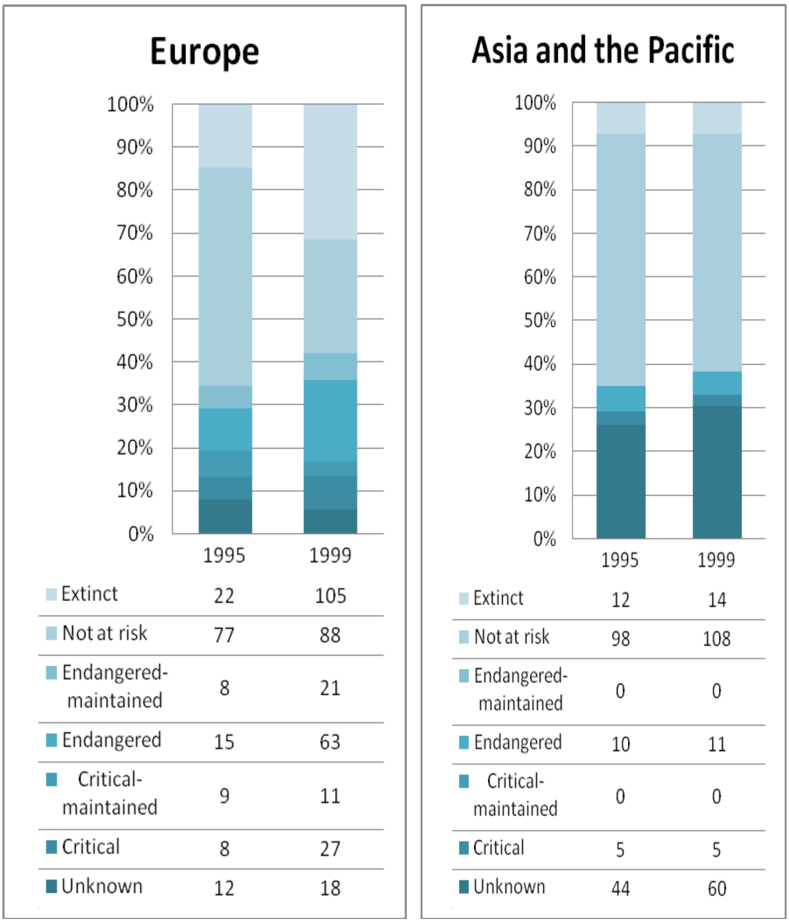
6

General discussion

One of the paths that provide insight about the genetic basis of phenotypic traits is the full characterization of the genome structure of a species. However, currently this is not realistic for any species because it would require the complete genomic sequence and spatial location of every individual within a species (Gillespie 2004). In spite of the absence of complete genome information at the individual level, the development of genomic resources such as linkage maps and reference genome sequences were crucial for a more detailed understanding at the molecular level of complex traits in two of the most important livestock species, the pig and the chicken. However the identification of the genetic basis of a complex trait or disease is not trivial. Complex traits and diseases are determined by the effects of several genes, where a few genes which have strong effect can be identified through classical QTL studies, but the genes with small effects are rarely detected. The research presented in this thesis aimed at characterizing and understanding the forces behind the genetic variation in pig and chicken. Domesticated species such as chicken and pig have a well recorded history of selection (at least since the 19th century) within confined environments in contrast to other species which followed a more complex history of selection. Furthermore, a reference genome is available for both species. For these reasons, these species are excellent models to evaluate footprints of selection and to identify the genetic basis of complex traits and diseases. in order to achieve these goals, an investigation to provide insights in the extent of LD was conducted in different pig breeds, covering the most common commercial breeds, European and Chinese traditional breeds and the European wild boar (chapter 2). Information about SNP markers in pig was lacking and therefore an investigation which showed how massive parallel sequencing (MPS) technologies can be used in a cost-effective manner to characterize within species diversity was presented in chapter 3. Finally in chapters 4 and 5 we present the first maps of selection footprints in two of the most important domesticated species, pig and chicken. Below I will discuss further considerations about the achieved results and future perspectives. Data analysis and management issues related with MPS data will also be briefly discussed.

Genetic diversity in the pig and future research

Pig was domesticated independently in Asia and Europe, the origins of the largest diversity of pig breeds in the world. Europe is the origin for 37% of the worlds' listed pig breeds where the majority of the pigs belong to a small number of selected breeds of international status (e.g. Large White, Landrace, Pietrain) (Figure 1), and Local breeds are mostly rare (many are at risk of extinction or even are already extinct). In Asia, the origin for 46% of the total list of pig breeds in the world, the situation is very different. The current state is characterized by the existence of many local breeds, from which only very few are known to be extinct. The largest proportion is not at risk, although the current state for many local breeds is unknown and has not been assessed (Figure 1).



Europe	1995	1999
Extinct	22	105
Not at risk	77	88
Endangered-maintained	8	21
Endangered	15	63
Critical-maintained	9	11
Critical	8	27
Unknown	12	18

Asia and the Pacific	1995	1999
Extinct	12	14
Not at risk	98	108
Endangered-maintained	0	0
Endangered	10	11
Critical-maintained	0	0
Critical	5	5
Unknown	44	60

Figure 1- Summary statistics of the FAO evaluation of the state of the pig breeds in the world (Scherf 2000)**.** In brief, besides the international breeds the samples included a large diversity of traditional breeds from Celtic to Iberian types. These projects allowed the development of microsatellite (MS) and AFLP markers, which were used to assess the within breeds diversity. Results showed that lower within-breed diversity was observed in local breeds and commercial lines as compared with international breeds (SanCristobal et al. 2006a; SanCristobal et al. 2006b) while the local breeds have the largest contribution (~56%) to the overall genetic diversity (Ollivier et al. 2005).

The endangered status of many European pig breeds shows that it is important to assess genetic variation. Collaborative programs were launched with the support of the European Commission (EC) such as the PigBioDiv I and the PigBioDiv II during which a large set of European pig breeds was sampled. A detailed list of all the breeds sampled in PigBioDiv I was reviewed by (Ollivier et al. 2005).

In China traditional local pig breeding has been applied until more recent times. Recently breeding lines are being developed with introgression of modern commercial breeding lines, mostly descendent from European breeds. Although many Chinese local breeds are not considered at risk by FAO, many have a small population size and are considered rare (Yang et al. 2003), showing a need for the evaluation of genetic diversity. In recent years studies such as the PigBioDiv II project, have contributed to the characterization of genetic diversity of Chinese breeds. Chinese breeds hold high levels of genetic diversity (Li et al. 2004; Yang et al. 2003; Fang et al. 2005) denoting a high level of population structure (Li et al. 2004; Fang et al. 2005; Megens et al. 2008). Compared to European breeds, Chinese breeds showed higher levels of heterozygosity and genetic distance (Megens et al. 2008).

The differences found in genetic diversity and population structure between European and Chinese breeds may be explained by several factors including (1) a smaller effective population size in the current European breeds and in their ancestral population, (2) a higher intensity of selection in the European breeds and (3) a lower genetic diversity in its ancestor the European wild boar. The characterization of linkage disequilibrium (LD) in pig breeds enables the reconstruction of population history and is of high utility for fine mapping of genes responsible for complex traits and diseases. In chapter 2, the levels of LD in 10 European pig breeds, 10 Chinese pig breeds and European wild boar were assessed using a total of 371 SNPs distributed along three genomic regions, two of which are located on SSC18 and one on SSC3. The results showed that the Chinese breeds presented much higher levels of genetic diversity than the European breeds and the European wild boar. We observed high values of r^2 at long distances (1Mb) in European pig populations, which suggests reduced effective sizes of the ancestral populations (Figure 2). We also observed that commercial breeds presented the highest values of r^2 from all the breeds studied. The extent of LD in Chinese breeds was similar to that observed for human populations (Ardlie et al. 2002) and we observed high values of r^2 at short distances. The estimated size of the ancestral population obtained from the estimates of linkage disequilibrium (Figure 2) suggest the existence of larger ancestral populations for some of Chinese breeds such as was found for the indicine cattle (The Bovine HapMap Consortium 2009).

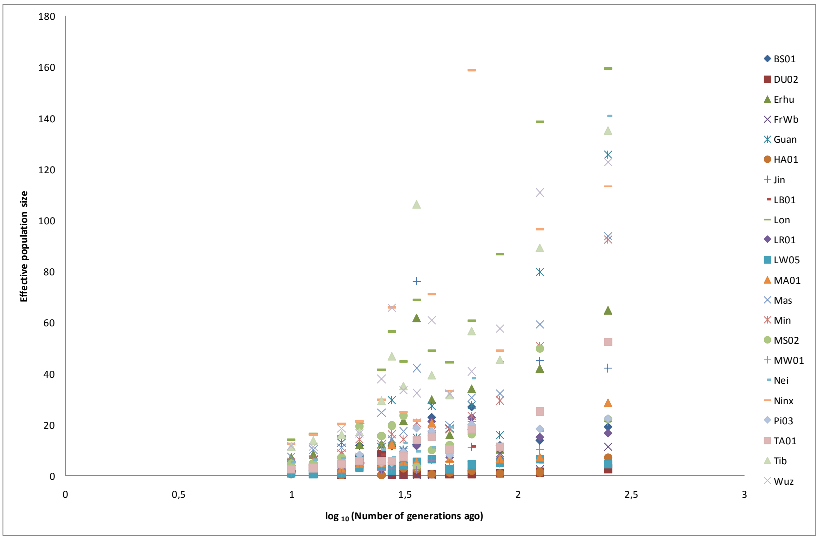


Figure 2 – Effective population size in the past estimated from linkage disequilibrium data in target 4 (see chapter 2).

The estimates of LD suggest that for a genome-wide characterization of the genetic diversity of pig breeds a set of 30,000 SNPs per individual in European breeds and a set of 500,000 SNPs in Chinese breeds would be required. The short extent of LD found in the Chinese breeds makes these breeds optimal models for the identification of genes with effect in complex traits and diseases in pig, because such short extent of LD allows narrowing QTLs confidence intervals (Ardlie et al. 2002).

Our results also show variability in the pattern of decay of LD in the Chinese breeds. Most outstanding is the case of the Meishan breed, which was sampled from a population established in Europe 25 years ago and which started from a small number of individuals. The level of LD in Meishan is higher than in the other Chinese breeds, most likely due to the effect of the small population size. This result shows the impact within a few generations of small effective population size on LD and on the decrease of genetic diversity of populations. The Chinese breed Jinhua, also showed higher levels of LD. Previous studies have shown a lower level of genetic diversity within this breed compared to other Chinese breeds. The breed Jinhua has been under high selection pressure to produce the Jinhua ham (Li et al. 2004). Due to the low extent of LD, Chinese breeds allow a higher resolution for gene mapping and therefore will be ideal models for the identification of candidate genes related to complex traits and diseases. Furthermore, the Chinese pig breeds represent an important reservoir of genetic diversity for the species which is important to preserve. The European wild boar population studied, showed a prevalence of high values of LD at long distances, although smaller than what is observed in the European traditional and commercial breeds. The estimated size of the ancestral population indicates the existence of a small ancestral population for the European wild boar probably the result of a bottleneck in the past (Figure 2).

In chapter 4 the genome-wide pattern of nucleotide diversity was evaluated in several European commercial breeds and in European wild boar. In this study the samples of European wild boar were representative of the species distribution in Europe. It was found that the genetic diversity in wild boar was lower compared to the European commercial pig breeds. The results reinforce the hypothesis of the existence of a bottleneck in the European wild boar population and that pig domestication in Europe is based on a gene pool with lower diversity compared to the domestication in Asia. In fact it is estimated that domestication of the pig in Europe occurred during the Neolithic about 9,000 B.C., after the last glacial period (~110,000-10,000 B.C.). During the last glacial period, Northern Europe was covered by ice and Central Europe was covered by Steppe-Tundra, being the Southern Peninsulas of Europe the only refugees where we could find the lowest temperatures and temperate climate. In Asia, in particular the region which today is China, the climate was not so severe, mainly with dry-steppe and forest-steppe ecosystems. These different climate conditions in both domestication centers must have shaped different gene pools for domestication of pig in Europe and Asia. The last glacial period has likely caused a severe bottleneck in the European wild boar populations and consequently domestication of pig in Europe could have started from a less diverse gene pool than in Asia.

Our results indicate that the largest reservoir of genetic diversity of European pigs is not in its wild relatives, as observed for most other species, but in the breeds. These findings emphasize the need to preserve local pig breeds in Europe. These results also raise another question: how did the European pig breeds become more genetically diverse than wild boar? Historical records show that during the 19th century pigs were imported from China (V. Porter & Tebbit 1993). Moreover, the results described in chapter 2, the occurrence of highly frequent haplotypes in Chinese breeds in some Europe pig breeds, clearly show the introgression of

Chinese pig breeds into European Pig breeds, as previously suggested by studies on mtDNA (Giuffra et al. 2000). Another source of variation could be the first domesticated pigs originated from the Near East which were introduced in central Europe during the Neolithic. However the Near Eastern haplotype is only observed in wild boars from Italy (Larson et al. 2007) and this most likely has a poor contribution to the variation found in Central and Northern European breeds.

Future studies should address the following questions: (1) what was the effective size of ancestral populations of European pig breeds, Chinese pig breeds, European wild boar and Asian wild boar? (2) how did genetic diversity change during domestication of pig?

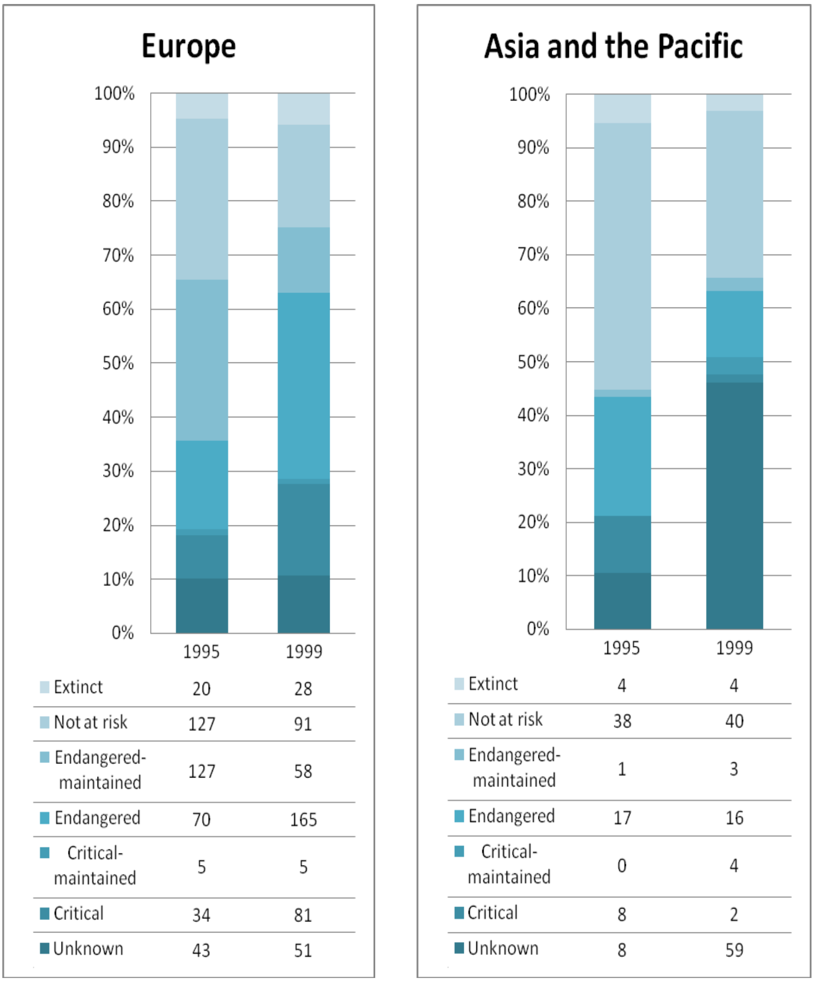
A genome-wide scale evaluation of the extent of LD based in sequence data, would allow estimating with more accuracy the effective size of the ancestral population. As shown in chapter 2, the extent of LD varies along different genomic regions. This means that measuring LD in a few and small genomic regions can create sampling biases in the estimates, which could be overcome by an estimate on a genome-wide level. Results from such assessment could allow investigation of the size of the effective population many generations ago (~4,000) and to verify: (1) the hypothesis for the occurrence of a bottleneck after the last glacial period in the population of European wild boar; (2) whether the domestication of pig in Europe is based on a less diverse gene pool compared to domestication of the pig in Asia and (3) which haplotypes found in European pig breeds resulted from introgression of Asian pig breeds.

A genome-wide scale of the genetic diversity based on sequence data of all the sampled breeds of PigBioDiv I & II, plus European and Asian wild boars and closely related species such as *Sus verrucosus* or *Sus barbatus* would allow an unbiased estimate of genetic diversity. All previous studies have been based on markers which have been developed based in the study of European pig breeds creating a so called ascertainment bias. Rare SNPs are more likely to be identified in larger samples than in smaller samples. Therefore, variant detection processes which involve an initial small discovery panel will produce an excess of high-frequency alleles in the genotyped sample. As a consequence, the data will be different from what is assumed in standard population genetic models with respect to allele frequency distribution (e.g., Nielsen 2000; Nielsen & Signorovitch 2003), patterns of linkage disequilibrium (Nielsen & Signorovitch 2003), and level of population subdivision (Nielsen et al. 2004). This ascertainment bias toward high-frequency alleles can have serious consequences when estimating genetic variation.

Estimates based on sequence data from whole-genome resequencing or resequencing of large parts of the genome could circumvent this bias and therefore would enable the assessment of which breeds contribute most to the genetic diversity of the species. Furthermore, the use of *Sus verrusocus* or *Sus barbatus* as outgroup species, allows the identification of the ancestral alleles enabling reconstructing the allele frequencies for the hypothetical ancestral population and the proportion of missing alleles.

Genetic diversity in chicken

Chicken was first domesticated in Asia about 5,000 years ago (Zeuner 1963) leading to a process which has generated 734 breeds worldwide from which 32 are extinct, 201 are at risk, 101 have critical status of conservation and the status of 141 is unknown. Europe is the location of the highest diversity of chicken breeds accounting for 64.2% of the total number, whereas populations of chicken in Asia and the Pacific account for 45.4% of the total. Many breeds have become extinct in Europe and many are endangered (Figure 3). In Asia and the Pacific, fewer breeds are extinct but endangered breeds account for 20% of the total (Figure 3). Furthermore, since the start of the 20th century, chicken genetic diversity started to be partitioned among relatively few specialized lines of production (meat or eggs) which were created from a few European local breeds. As a consequence and as shown in Figure 3, many chicken breeds, especially dual-purpose breeds (meat and eggs), have been extinct or are endangered. The situation shows that the assessment of genetic variation is crucial. In order to provide recommendations regarding future management or conservation, EC funded collaborative programs like AVIANDIV, which aimed to assess the diversity of a wide range of populations, including commercial lines, traditional breeds, experimental lines and the ancestor the Red Jungle Fowl. Results from this program show that the Red Jungle Fowl has the highest level of genetic diversity, being an important reservoir of chicken polymorphism, followed by traditional breeds, where an extensive diversity has accumulated during domestication. Commercial lines had the lowest levels of genetic diversity and differences were observed between broiler and layer lines. Genetic diversity was slightly higher in broiler lines, and within layer lines, the brown egg layers showed the highest degree of genetic diversity (Hillel et al. 2003). Furthermore, (Muir et al. 2008) showed that commercial lines (broiler and layers) are missing significant genetic diversity which is found in noncommercial chickens. In chapter 5 an assessment of the genome-wide distribution of genetic diversity in commercial chicken lines was presented.



Europe	1995	1999
Extinct	20	28
Not at risk	127	91
Endangered-maintained	127	58
Endangered	70	165
Critical-maintained	5	5
Critical	34	81
Unknown	43	51

Asia and the Pacific	1995	1999
Extinct	4	4
Not at risk	38	40
Endangered-maintained	1	3
Endangered	17	16
Critical-maintained	0	4
Critical	8	2
Unknown	8	59

Figure 3- Summary statistics of the FAO evaluation of the state of the chicken breeds in Europe, Asia and the Pacific (Scherf 2000)**.**

Results confirm previous assessments (Muir et al. 2008; International Chicken polymorphism Map Consortium 2004), showing that broiler lines had higher levels of genetic diversity than layer lines and that brown egg layers have a slightly higher level of genetic diversity than white egg layers. A genome-wide scale comparison of the genetic diversity in chicken breeds, commercial lines and Red Jungle Fowl is thus urgent in order to establish appropriate measures of breed management and conservation.

Identifying candidate genes using a population genomics approach and applications

Understanding the genetic basis of complex traits and diseases has been and continues to be a central goal for the scientists working with genetics of domesticated animals. Identifying the genes affecting phenotypic traits will allow answering important questions in genetics and evolution: how many genes influence complex traits and diseases and what are their relative effect sizes? Do these genes show evidence of non-neutral evolution at the sequence level? What forces maintain or decrease variation at these loci? Which loci were affected during domestication? These questions can be answered through quantitative traits loci (QTL) mapping and population genomics approaches.

QTL mapping has been by far the most used method in pig and poultry genetics to answer these questions, with the ultimate goal to identify specific genes affecting these complex traits. QTL mapping strategies have identified genomic regions that affect numerous phenotypes in pig (Max F. Rothschild et al. 2007) and chicken (Nadaf et al. 2007; Nadaf et al. 2009; Mignon et al. 2009). However, QTL mapping requires knowledge of phenotype and pedigree, the development of controlled crosses to generate a large number of progeny, genotyping of all the offspring and progenitors and the development of linkage maps to finally produce a statistical association of phenotypes with genetic markers. This makes experiments expensive and time consuming. Furthermore, the mapping resolution of QTL linage studies is low; most often only identifies the genes with the main effects or results in very wide confidence intervals with hundreds of genes.

The population genomics approach described in Figure 4 allows the identification of genes which have been involved in processes of recent selection. This approach requires the sampling of many individuals, genotyping or sequencing the sampled population at many loci and the estimation of statistics which allows the detection of selection effects on specific locus. Such an approach is not dependent on the availability of a pedigree, nor does it require any investment to produce specific experimental crosses. The accuracy depends on the genotyping or sequencing a large number of loci and on reliable statistics to detect outlier loci that may indicate regions that have been under recent selection. Because positive selection should have locus-specific effects by reducing genetic variability within populations and by increasing the differentiation between populations, loci that are outliers for these characteristics are strong candidates for selective sweeps.

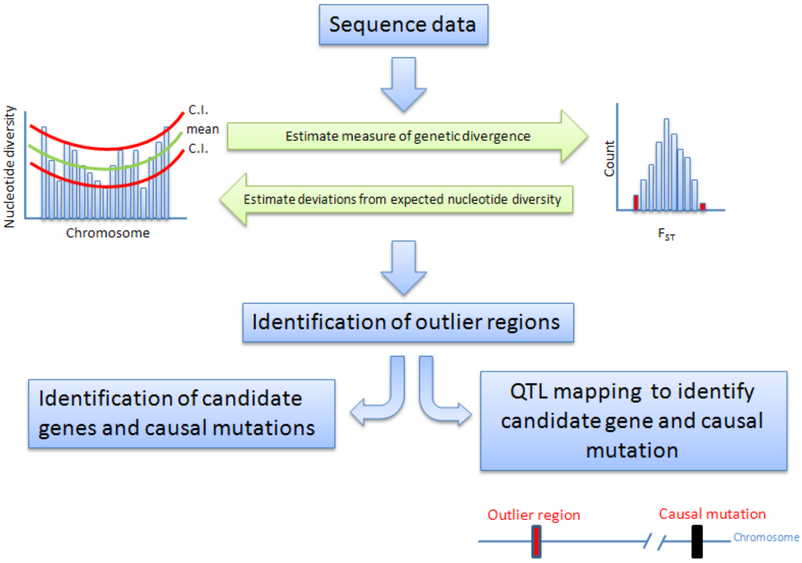

Sequence data
Nucleotide diversity
C.I.
mean
C.I.
Chromosome
Estimate measure of genetic divergence
Estimate deviations from expected nucleotide diversity
Count
F_{ST}
Identification of outlier regions
Identification of candidate genes and causal mutations
QTL mapping to identify candidate gene and causal mutation
Outlier region
Causal mutation
Chromosome

Figure 4 – Model for a cost-effective approach to identify genes affecting a trait. From sequence data regions deviating from neutrality can be identified either by (a) identifying outlier regions with values of nucleotide diversity outside the range of the confidence intervals (CI) or by (b) identifying regions with extreme values of genetic divergence (red bars). These regions harbor candidate genes or are in linkage disequilibrium with regions containing the genetic mutation. In the second case, QTL mapping studies should be pursued.

In contrast, balancing selection should have locus-specific effects, resulting in increasing genetic variability within populations. Loci that are outliers for these characteristics are strong candidates for co-dominance. The application of this approach allowed the identification of loci that have been under selection during human evolution and in response to dairy farming (Sabeti et al. 2002) and malaria (Bersaglieri et al. 2004). The population genomics approach identified loci under positive selection during maize domestication (Vigouroux et al. 2002) and genes for coat color and shortened limbs during dog breed formation (Pollinger et al. 2005). This approach allows the identification of genes with similarities in the histories of selection and with closely related functions. Previous studies have found regions under positive selection in mammals which represent whole pathways (Kosiol et al. 2008), allowing the identification of several genes involved in phenotypes with a complex genetic basis.

As already mentioned, the population genomics approach requires the availability of a large amount of marker information on loci across the complete genome, a requirement that thus far has limited its application in domesticated animal species. The use of MPS now allows the development of population genomic studies with high power of resolution in domesticated animals and the identification of outlier loci by studying genetic variation at the sequence level. Nevertheless, individual sequencing using MPS is still very expensive and DNA pooling presents an alternative to decrease costs while still allowing capturing information at the population level.

Classical statistics to identify outlier loci rely on the comparison of sequences or loci from different individuals. In chapter 4 an approach to identify outlier loci in the pig genome using MPS data obtained from DNA pools was presented. A coalescent based frequency test using a modified Watterson estimator which accounted for sequencing errors and DNA pooling was developed and allowed the creation of a genome-wide map of potential outlier regions along the pig genome. Outlier regions were also identified by measuring the genetic differentiation between breeds. Four international breeds were studied (Pietrain, Large White, Landrace and Duroc) along with the European wild boar. We used the *KIT* gene locus as a positive control showing that our approach could detect a previously established selective sweep (Moller et al. 1996). We observed an outlier region in the white breeds which overlapped perfectly the previously defined sweep. This result is in line with the study of (Moller et al. 1996) which identified a mutation in the sequence of the *KIT* gene in white breeds which was absent in colored breeds and in wild boar. We found in the outlier regions genes with closely related functions and which had the same KEGG categories. Regions putatively under the effect of positive selection and region with higher genetic differentiation were especially rich in genes related with growth, muscle development, behavior and metabolism in the pig breeds. Regions which putatively show evidence of the possible occurrence of balancing selection were enriched with genes of the MHC complex and of the Olfactory Receptor complex in the pig breeds and in the European wild boar. This result is in line with observation in other species, e.g. humans, for which a model of balancing selection for the MHC complex (Meyer et al. 2006), and for the olfactory receptor complex (Alonso et al. 2008) has been proposed. In chapter 5 the same approach was applied to generate a genome-wide map of potential outlier regions in chicken. Two lines of broiler (meat) chickens and two lines of egg layer chickens were studied. Again we used a positive control, the *IGF1* gene, to determine the sensitivity of our method to detect previously detected selective sweeps (Megens et al., unpublished results). The results indicate the occurrence of positive selection in the genomic region around the *IGF1* gene in broiler lines but not in the layers. For many regions was difficult to establish a link with the biological function. This may be due to the fact much more information concerning gene functionality is still to be discovered or due to the fact that the outlier region does not hold the causal locus itself, but is either physically linked or in linkage disequilibrium with the selected site (Nordborg & Tavaré 2002). An example of the second case is the identification of a deletion which was found in a QTL with effect in coronary artery disease in mice which affects the transcription of nearby genes in the heart tissue (Visel et al. 2010). The development of a QTL mapping experiment may allow the identification of the causal mutation which is in linkage disequilibrium with the outlier regions and has an effect on the phenotype (Figure 4).

Many outlier regions overlap with genes for which evolutionary models have been proposed in the species of study or in related species. However, the approach developed, relies solely on the accurate determination of outlier loci by determining what is significantly different from the neutral expectations. The method developed, does not distinguish between the effects of demography and selection in shaping genetic variation. Many demographic factors can affect patterns of nucleotide diversity in a way similar to the effects of selection. Population growth for example, can also result in the observed unusual low nucleotide diversity. Another example is the occurrence of population subdivision that results in the observation of unusual high nucleotide diversity. Future studies therefore, should: (1) estimate the expected pattern of nucleotide diversity under models of selection and demography; (2) complement the population genomics approach with QTL mapping approaches. The use of an outgroup population in future studies will increase the accuracy of the estimates of outlier regions because it will allow better fitting of the observed variation with the variation under a selection model (Nielsen & Bustamante 2005), or the evolution under a demographic model (Stajich & Hahn 2005). Including an outgroup population will allow inference of changes or losses of functionality during domestication and selective breeding. Nevertheless, the results presented, provide a valuable starting point for experimental follow-up. Simultaneous progress in gene annotation and in the development of more robust statistics will allow establishing strong connections between outlier regions and molecular function.

Genetic diversity and patterns of selection in chicken and pig

Domesticated species have different histories of domestication and selection. Results, presented in chapters 4 and 5, show that the different histories of domestication and selection leave footprints in the patterns of genetic variation along the genome. Results presented in chapter 4 suggest that in Europe, the largest reservoir of genetic diversity is found in the domesticated breeds and not in their wild relatives. In contrast in chicken the largest reservoir of genetic diversity is still the wild species, the Red Jungle Fowl.

Differences were also found in the pattern of variation of nucleotide diversity along the genome. In pig we observed (chapter 4) that nucleotide diversity increases towards the telomeres in metacentric as well as in submetacentric chromosomes. Chromosome X was the chromosome which exhibited a significant lower level of diversity. In chicken, we observed (chapter 5) that nucleotide diversity is only higher at the extreme end of the telomeres, presenting a pattern very different from that observed in pig. Furthermore, significant differences were found in the levels of nucleotide diversity between chromosomes. In most of the chicken populations studied, macrochromosomes show lower levels of genetic diversity compared to intermediate chromosomes and microchromosomes. Are these differences due to the impact of selection? Are these differences due to differences in recombination rates between mammals and birds? The answer of these questions through a comparative analysis of the pattern of variation between these two model domesticated species will lead us in the future to the understanding of evolutionary variation across species and will allow revealing the genetic basis of phenotypic variation. In fact (Kosiol et al. 2008) showed that the power to ascertain rates of protein evolution increased when the set of species under consideration increased considerably. In chapter 4 and chapter 5 we found some common trends in the enrichment of genes found in regions putatively under positive selection. In both species we found enrichment of genes related to growth, and metabolism, which however after correcting for multiple testing lost their significance. Nevertheless, the pathways found are strongly related with the history of selection of both species, what indicates that the increase of number of observations could increase the power of the analysis. With the same reasoning future studies should focus on comparing the patterns of selection in pig with other mammals and on comparing the chicken with other domesticated birds. The results obtained in pig, lead to the same conclusions obtained with other mammals, although with lower statistical significance. Therefore I am convinced that an increase of the number of species under study would increase the power of the analysis considerably.

Future challenges in the development of experiments using MPS in a population genomics context

MPS technologies are boosting the era of genomics. With the application of MPS in the future, large genomic regions or even whole-genomes will be resequenced at the individual level. At the start of the PhD research project described in this thesis many research centers working in genomics had limited accessibility to MPS sequencers. After three years, many MPS technologies are provided by a large number of commercial sequence providers, and are accessible to a much larger group of the scientific community.

MPS is creating unprecedented opportunities to make novel biological observations and address novel research questions. In the previous section of this discussion the main results were summarized and questions for future research were raised. Next, I will present some considerations about the developments of experiments to address these research questions using MPS technologies.

Genome-wide identification of variants

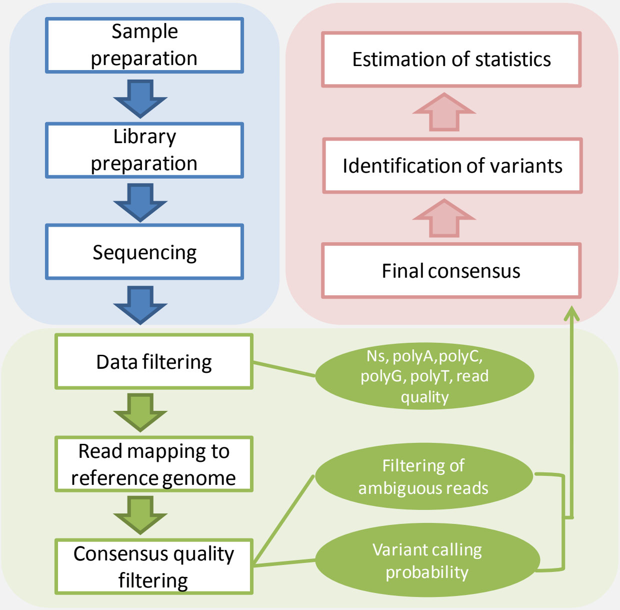
Using traditional Sanger sequencing for the identification of variants is time consuming and expensive. Approaches to reduce time and cost included sequencing a panel of individuals to identify variants followed by genotyping all the individuals in the study using e.g. SNAPshot and SNP array methods for the previously identified variants. Sequences stored in shared public databases could be another source to identify variants, examples are the strategies described in chapter 2, where SNPs were identified from BAC sequences of pigs available from the comparative vertebrate sequencing project of NISC (http://www.nisc.nih.gov) and the strategy followed by (Kerstens et al. 2009) who identified thousands of SNPs by analyzing sequences generated in the Sino-Danish Pig Genome Project. These approaches, although strongly decreasing experimental costs, created biases since the variation existent within the population of study could never be screened. The genomes of the individuals of the population to be analyzed were only interrogated for the existence of variation in the same location as the individuals used for the sequencing panel. MPS technologies allow the unbiased identification of variants simply because it is cost-effective to obtain sequences from all the individuals in the studied population. Therefore, MPS allows capturing the genetic variation within and between populations. Examples of such applications are the whole-genome SNP discovery in pigs (chapter 2; (Ramos et al. 2009), in *C. elegans* (Hillier et al. 2008) and cattle (Van Tassell et al. 2008). Population genomics studies aiming at the identification of regions under selection need to identify all variants on a genome-wide scale. MPS technologies still present challenges concerning the identification of variants. The first challenge concerns the quality of the data (see BOX). In chapter 3 we showed that the removal of low quality sequences is not arbitrary. If low quality reads are mapped to the reference genome, these can create noise and decrease accuracy of the correct identification of variants. The second challenge concerns sequence mapping. Next-generation aligners allow fast processing of sequence mapping, however, algorithms still have problems in dealing with repetitive sequences. The simplest approach and which was followed in the analysis presented in chapters 3, 4 and 5, was simply to not consider sequences that had ambiguous matches. However this decreases the amount of usable data and does not allow identification of variants in repetitive regions. Currently all MPS technologies are capable of generating “paired end” data. When one of the paired sequences maps to a highly repetitive element in the genome but the second does not, it allows both ends to be mapped unambiguously. Furthermore, for all MPS technologies the length of the reads is increasing, which will also contribute to decrease the effect of ambiguous sequence mapping. The third challenge concerns consensus quality filtering (see BOX). In chapters 4 and 5 the low sequencing depth was not sufficient to identify rare variants (variants observed in only one sequence). Furthermore, in chapter 4 difficulties in accurately identifying variants were higher towards the 3’ end of the sequences. In fact, it was observed that variant detection was dependent on the position in the sequence, something that should not happen. A bias in the quality scores associated with low sequencing depth was the reason for this phenomenon. In order to overcome this problem, a filter over the consensus quality was applied, where the threshold increased towards the 3’end of the sequences. This resulted in the decrease of false positive variants and in the almost absence of rare variants. Higher sequencing depth would have allowed higher accuracy and identification of rare variants. MPS technologies are evolving, the amount of generated sequences as also its quality is increasing making costs more manageable. Future studies should aim at experimental designs with larger sequencing depths and paired-end sequencing.

BOX | **MPS data handling, management and analysis**

The figure on the right represents a workflow of variant identification using MPS. A typical analysis of a biological sample is shown with the various sections color codes has wet-lab work (blue), and the dry-lab work corresponding to the creation of the filtered data set (green) and specific data analysis (pink).

Data filtering

MPS technologies produce percentage of sequences with dubious quality (low quality scores and/or Ns called in the sequence) or which could be contaminants and that should be removed.



Sample preparation
Library preparation
Sequencing
Data filtering
Ns, polyA,polyC, polyG, polyT, read quality
Read mapping to reference genome
Consensus quality filtering
Filtering of ambiguous reads
Variant calling probability
Final consensus
Identification of variants
Estimation of statistics

BOX | **MPS data handling, management and analysis - continued**

Sequence mapping

Once MPS data have been filtered to remove aberrant sequences, the next step is to match the sequences back to the reference genome. Alignment itself is the process of determining the most likely source within the genome sequence for the observed DNA sequencing read, given the knowledge of which species the sequence comes from. While this task is trivial on an individual basis, the large amount of sequences generated by MPS presents a challenge. Early-generation sequence aligners like MEGABLAST, BLAST or BLAT require a few days for while next generation aligners (MAQ (Li et al. 2008a), MOSAIK(http://bioinformatics.bc.edu/marthlad/Mosaik), SOAP (Li et al. 2008b) freely available and allow faster processing than early-generation aligners. This is mainly due to the fact that while early-generation aligners required alignments of protein sequences and searching through large databases to find homologous sequences, next-generation aligners are prepared to align a DNA sequence to the reference genomes of the species of interest. Although this difference is subtle, it allows letting the assumptions about number of expected mismatches be driven by the species polymorphism rate rather than by considerations of evolutionary substitutions.

Consensus quality filtering

After positioning the sequences in the reference genome, the next step is to define a consensus in order to extract the variant sites. The consensus analysis aims to determine the probability of occurrence of a variant on each position based on the probability of the sequence being aligned to a position, on the probability of the nucleotide given by its quality score and on the number of observations in that position. Next-generation aligners like MAQ (Li et al. 2009), SOAP (Li et al. 2008) and VarScan (Koboldt et al. 2009) make variant detection from the alignment output however, each algorithm results in a different set of variants.

Reduced representation libraries or whole genome resequencing

Whole-genome sequencing with classical methods such as Sanger sequencing are very expensive and time consuming. This is the reason why until recently a genome reference was only available for a few species. Projects such as the Human HapMap (The International SNP Map working group 2001) which aimed to characterize the variation of human populations by genotyping several individuals for thousands of SNPs, showed how important it is to characterize variation within a species to understand what a “normal” genome is. Again, projects like HapMap had the problem of ascertainment bias because most of the genotyped individuals were not part of the SNP discovery panel.

The use of reduced representation libraries (RRLS) allows performing a representative sampling of a species genome on a genome-wide scale. A restriction enzyme is used to digest the DNA and digested fragments of DNA are extracted and purified for sequencing. It may allow characterization of variation within a population without ascertainment bias issues because RRLS can be produced for all the sampled individuals, removing the need for an ascertainment process based on a discovery panel. RRLs have been used in combination with Sanger sequencing (Altshuler et al. 2000) and in chapters 3, 4 and 5, RRLs were used in combination with MPS for whole-genome identification of variants and characterization of nucleotide diversity.

Like any sampling procedure, the use of RRLs presents limitations. Patterns of sampling are different within a population as a consequence of the existence of mutations, insertions and deletions. This means that a comparison is not possible for all sites sampled. This effect is even higher when different populations are compared. In chapters 4 and 5, different populations of pig and chicken were compared to study the variation within the species. Several RRL libraries were created using the same restriction enzymes for all the populations. Results of the sequence alignment showed that only 30% of the sequence coverage was in common between two populations and only 10% was in common when considering all the studied populations. This means that for analyses demanding a site-to-site comparison, the use of RRLs results in a reduction of usable data. An example is the analysis of genetic differentiation described in chapter 4 where in a comparison of pairs of breeds usable sequence coverage was reduced to ~30%. The costs of MPS are decreasing, making whole-genome resequencing more affordable. Future studies should choose between sampling many individuals and perform sequencing of

RRLs or sample fewer individuals and perform whole-genome resequencing. For highly substructured populations, a common feature of domesticated animals such as pig and chicken, sampling bias that is a consequence of using RRLs can limit the interpretation and usability of data. Therefore, the second option, i.e. sequencing entire genomes, should be pursued in future studies.

DNA pooling versus individual sequencing

In order to reduce sequencing costs DNA pooling was used in several studies aimed at characterizing variation at the population level (Docherty et al. 2007; Wolford et al. 2000). In chapters 3, 4 and 5 DNA pooling was combined with MPS to characterize variation on a genome-wide scale in pig and chicken. The preparation of the pooled libraries needs to assure that each pool has the same amount of DNA from each individual, because biases created at this stage cannot be corrected as the identification of the origin of the sequences is no longer possible. The error associated with a bias in the preparation of pooled DNA libraries is higher in experiments using a smaller number of individuals to create the pool. In fact the correlation between the allele frequencies estimated from the sequencing of pools and the allele frequencies estimated from the individual genotyping of the same sampled individuals, was smaller when using a sample size of 5 individuals (chapter 2) than in a sample size of 25 (Ramos et al. 2009). Furthermore it can happen that sequences of one individual are observed more often than the sequences of other individuals in the pool due to stochastic effects of the sequencing method. Statistics can be developed to overcome the effect of DNA pooling (as in the case of chapter 4 and 5), however it involves making more complex models often resulting in a decrease in power of the analysis. Further limitations exist in the possibilities to address many other important research questions. For example, in the outlier regions identified in the pig populations (see chapter 4) and in the outlier regions identified in the chicken populations (see chapter 5) the analysis of the extent of linkage disequilibrium would have been useful in order to decrease the rate of false positives for effect of positive or balancing selection. In regions of positive selection an increase of linkage disequilibrium would have likely been observed, whereas in regions of balancing selection a decrease of linkage disequilibrium would have been expected. Because the use of DNA pooling does not allow the identification of the genome of origin of the sequence linkage disequilibrium cannot be assessed. The use of DNA pooling has to be carefully considered in any experimental design. DNA pooling is an option which is often preferred to reduce costs, but it does not allow answering several biological questions which require knowing how alleles co-segregate such as linkage disequilibrium, inbreeding and association between allele frequencies and phenotype. This means that DNA pooling can be very useful for a preliminary screen, but individual sequencing is essential in subsequent studies.

References

Alonso, S. et al., 2008. Overdominance in the Human Genome and Olfactory Receptor Activity. *Mol Biol Evol*, 25(5), 997-1001.

Altshuler, D. et al., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803), 513-516.

Ardlie, K.G., Kruglyak, L. & Seielstad, M., 2002. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3(4), 299-309.

Bersaglieri, T. et al., 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, 74(6), 1111-1120.

Docherty, S. et al., 2007. Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genomewide association studies. *BMC Genomics*, 8(1), 214.

Fang, M. et al., 2005. The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Animal Genetics*, 36(1), 7-13.

Gillespie, J.H., 2004. *Population genetics*, JHU Press.

Giuffra, E. et al., 2000. The Origin of the Domestic Pig: Independent Domestication and Subsequent Introgression. *Genetics*, 154(4), 1785-1791.

Hillel, J. et al., 2003. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genetics Selection Evolution*, 35(5), 25 pages.

Hillier, L.W. et al., 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Meth*, 5(2), 183-188.

International Chicken polymorphism Map Consortium, 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature*, 432(7018), 717-722.

Kerstens, H.H. et al., 2009. Mining for single nucleotide polymorphisms in pig genome sequence data. *BMC Genomics*, 10, 4.

Kosiol, C. et al., 2008. Patterns of Positive Selection in Six Mammalian Genomes. *PLoS Genetics*, 4(8).

Larson, G. et al., 2007. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences*, 104(39), 15276-15281.

Li, S. et al., 2004. Genetic diversity analyses of 10 indigenous Chinese pig populations based on 20 microsatellites. *J. Anim Sci.*, 82(2), 368-374.

Li, H., Ruan, J. & Richard, D., 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851-1858.

Li, R. et al., 2008b. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), 713-714.

Max F. Rothschild, Zhi-liang, Hu & Zhihua, Jiang, 2007. Advances in QTL Mapping in Pigs. *Int. J. Biol. Sci.*, 3(3), 192-197.

Megens, H. et al., 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), 26 pages.

Meyer, D. et al., 2006. Signatures of Demographic History and Natural Selection in the Human Major Histocompatibility Complex Loci. *Genetics*, 173(4), 2121-2142.

Mignon, G.L. et al., 2009. A comprehensive analysis of QTL for abdominal fat and breast muscle weights on chicken chromosome 5 using a multivariate approach. *Animal Genetics*, 40(2), 157-164.

Moller, M. et al., 1996. Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. *Mammalian Genome*, 7(11), 822-830.

Muir, W. et al., 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Sciences*, 105(45), 17312-17317.

Nadaf, J. et al., 2007. Identification of QTL controlling meat quality traits in an F2 cross between two chicken lines selected for either low or

Nadaf, J. et al., 2009. QTL for several metabolic traits map to loci controlling growth and body composition in an F2 intercross between high- and low-growth chicken lines. *Physiol. Genomics*, 38(3), 241-249.

Nielsen , R. & Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15, 1566-1575.

Nielsen, R., 2000. Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms. *Genetics*, 154(2), 931-942.

Nielsen, R., Hubisz, M.J. & Clark, A.G., 2004. Reconstituting the Frequency Spectrum of Ascertained Single-Nucleotide Polymorphism Data. *Genetics*, 168(4), 2373-2382.

Nielsen, R. & Signorovitch, J., 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, 63(3), 245-255.

Nordborg, M. & Tavaré, S., 2002. Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, 18(2), 83-90.

Ollivier, L. et al., 2005. An assessment of European pig diversity using molecular markers: Partitioning of diversity among breeds. *Conservation Genetics*, 6(5), 729-741.

Pollinger, J. et al., 2005. Selective sweep mapping of genes with large phenotypic effects. *Genome Research*, 15(12), 1809-1819.

Porter, V. & Tebbit, J., 1993. *Pigs*, Helm Information.

Ramos, A.M. 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.

Sabeti, P.C. et al., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909), 832-837.

SanCristobal, M., Chevalet, C., Haley, C.S. et al., 2006a. Genetic diversity within and between European pig breeds using microsatellite markers. *Animal Genetics*, 37(3), 189-198.

SanCristobal, M., Chevalet, C., Peleman, J. et al., 2006b. Genetic diversity in European pigs utilizing amplified fragment length polymorphism markers. *Animal Genetics*, 37(3), 232-238.

Scherf, B., 2000. *World watch list for domestic animal diversity* 2nd ed., Rome: Food and Agricultural Organization.

Stajich, J.E. & Hahn, M.W., 2005. Disentangling the Effects of Demography and Selection in Human History. *Mol Biol Evol*, 22(1), 63-73.

The Bovine HapMap Consortium, 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science*, 324(5926), 528-532.

The International SNP Map working group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822), 928-933.

Van Tassell, C.P. et al., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Meth*, 5(3), 247-252.

Vigouroux, Y. et al., 2002. Identifying Genes of Agronomic Importance in Maize by Screening Microsatellites for Evidence of Selection during Domestication. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15), 9650-9655.

Visel, A. et al., 2010. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature*, advance online publication. Available at: http://dx.doi.org/10.1038/nature08801 [Accessed February 27, 2010].

Wolford, J. et al., 2000. High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC). *Human Genetics*, 107(5), 483-487.

Yang, J. et al., 2003. Genetic Diversity Present Within the Near-Complete mtDNA Genome of 17 Breeds of Indigenous Chinese Pigs. *J Hered*, 94(5), 381-385.

Zeuner, F.E., 1963. *A history of domesticated animals*, Harper & Row.

Summary

Summary

Since 12,000 years ago humankind has exploited the genetic diversity of animals and plants through domestication, giving humans the leading role in evolutionary processes. Artificial selection is a process which only occurs in domestication and implies the selection of animals and plants with phenotypic superiority in appreciated traits. During this process, a valuable agricultural patrimony emerged from the available genetic diversity existent in nature. Chicken and pig are part of this valuable patrimony and are the main sources of protein in the diet of humans from most parts of the world. Furthermore, these two species are important models for studies in biomedical research.

Domestication of the pig occurred in two different geographical regions, Europe and Asia. It is estimated that pigs were domesticated during the Neolithic. The earliest pigs which reared in Europe were derived from pigs domesticated in Asia Minor, which were later replaced by pigs derived from the European wild boar. The domestication of the pig generated a wide diversity of pig breeds from which 70% were originated in Eurasia. Traits selected during domestication included coat color, growth and muscle development and litter size. With the development of industrialized animal production, pig breeding in Europe became decreasingly local, while in China, the country of the world with the highest number of pig breeds, local pig breeding has been applied until more recent times. A few breeds were selected for the pig breeding industry and gained a status of international breeds, e.g. Pietrain, Large White, Landrace and Duroc.

Nowadays most European local pig breeds are endangered or in critical state and have suffered a severe decrease of the effective population size. In contrast in China, the number of pig breeds at risk of extinction represents a small proportion of the breeds, although information about many breeds is not available. As a consequence of the awareness of the critical state of many local pig breeds, research programs were launched which aimed to characterize genetic variation of large panel of pig breeds. Results showed that European local pig breeds contribute the most for the genetic diversity in Europe and Chinese breeds have the highest levels of genetic diversity. These differences may be due to: (1) small effective population sizes; (3) different intensity of selection; (4) lower genetic variation in the ancestral gene pool.

Chicken was domesticated in Asia about 5,000 years ago. Europe and Asia are the location where most of the genetic diversity can be found. In Europe we can find the largest number of breeds, whereas in Asia we can find the largest populations of chicken. Nowadays many of the local chicken breeds – mostly dual purpose (meat and eggs) breeds - are extinct or are at risk of extinction. Commercial poultry breeding created specialized lines for meat or eggs, for which a reduced number of traditional breeds was used to create the elite lines. The structure of commercial poultry breeding rarely incorporates genetic variation from non-commercial lines in elite commercial lines. Genetic variation is maintained through crosses of elite lines. Studies have shown that commercial lines are significantly missing important genetic diversity.

Thesis research focused in the characterization of genetic diversity of pig and chicken. In Chapters 2, 3 and 4 research which aimed to understand the patterns of variation in pig is presented. In Chapter 5 patterns of variation of genetic diversity along the chicken genome are presented.

The extent of the linkage disequilibrium (LD) allows inferring reconstructing the history of a species. In Chapter 2, the extent of LD in European and Chinese pig breeds and in European wild boar was investigated using 371 SNPs located along three genomic regions. The extent of LD differed significantly between European and Chinese breeds, extending up to 2 cM in Europe pig breeds and up to 0.05 cM in Chinese pig breeds. The European wild boar showed an intermediate level of LD between Eurpean and Chines pig breeds. The extent of LD also differed according to genomic region in European pig breeds. Chinese pig breeds showed higher level of haplotype diversity and shared some of the most frequent haplotypes with Large White, Landrace and Duroc. Results also indicated that in order to characterize the genetic variation on a genome-wide scale, at least 30,000 informative SNPs would be required for European pig breeds, whereas, 500,000 informative SNPs would be required in the case of Chinese pig breeds. In chapter 6, an estimation of the size of the ancestral population for each breed based in the LD estimates, is shown. Results suggest that Chinese pig breeds had a larger ancestral population. This indicates that Chinese pig breeds might have been originated from a most diverse gene pool than European pig breeds. Furthermore, the high levels of LD found in European wild boar, suggest the occurrence of a bottleneck, however in this study, only animals from a nature reserve in France were sampled. The results in this chapter showed

the need to increase the amount of information about genetic variation along the genome of the pig.

In Chapter 3 an assay which aimed to identify SNPs in a genome-wide scale in the pig is presented. In this assay, an approach which combined DNA pooling of five animals from a crossbred line, construction of reduced representation libraries (RRLs) and massive parallel sequencing was followed. Massive parallel sequencing allows larger sequencing output than Sanger sequencing, but the error rates are higher. As a consequence the analysis of MPS data has challenges at the level of data filtering and analysis. In Chapter 3 different data filtering strategies of MPS data were compared in order to maximize the genome-wide identification of SNPs. The experimental approach allowed the identification of thousands of SNPs. Results show that the removal of low quality sequences is not arbitrary. If low quality sequences are mapped to the reference genome, these can create noise and decrease accuracy of the correct identification of variants. The patterns of nucleotide diversity along the pig genome were also investigated. Results show a trend of increasing levels of nucleotide diversity towards the telomeres and decreasing levels of nucleotide diversity towards the centromere in metacentric and submetacentric chromosomes. The level of nucleotide diversity observed in the X chromosome is smaller than in autosomes. Overall, this assay showed that this approach is a cost-effective manner to study genome-wide variation and which can be applied to any species with a complete or partial reference genome. Therefore in the context of continuing increasing the amount of information about genetic variation in the pig genome, the same approach was followed for a wider sampling of breeds. In Chapter 4, results of the genome-wide characterization of genetic variation in international pig breeds (Pietrain, Large White, Landrace and Duroc) and European wild are presented. The collected samples were representative of the studied breeds and the sample of European wild boar included animals from different locations in Europe. Results showed that Pietrain had the lowest levels of nucleotide diversity, followed by European wild boar. These results are in accordance with previous studies which showed that European wild boar contributes less to the genetic diversity of the species than the local pig breeds. These results are also in accordance with studies showing the reduced genetic variation of Pietrain while compared with other international pig breeds. The observed patterns of nucleotide diversity were very similar to the patterns reported in Chapter 3, with some particular differences between the pig breeds and the European wild boar.

Artificial selection has been performed through the long history of domestication mostly without the knowledge of the genetic basis of the phenotype. In fact, the genetic basis of phenotype has not yet been revealed for most production traits. Most production traits have a complex genetic basis. A complex genetic basis means that the phenotype is determined by the effect of several genes. Some genes can have higher effects, but a great proportion of the genetic effect is divided by many genes. As a consequence the identification of genes affecting complex traits is not trivial. Most research developed in this field has followed QTL mapping approaches which have been in some cases successful in identifying genes with major effects. Population genomics approaches may allow identifying the genetic basis of complex diseases. The rationale of this approach is that loci which have a pattern of genetic variation which deviated from the expected under neutrality are strong candidates of being under selection. In, fact it is expected that genes involved in the same biological pathways and which affect a complex trait to have similar evolutionary histories. Such population genomic approaches require the analysis of genetic variation of a large number of loci. The application of this approach in the study of the genetic basis of complex traits in domesticated animals has been limited by the limited availability of information about genetic variation. In Chapter 4, a comprehensive map of regions which have been putatively under the effect of selection in the pig genome is presented. The pattern of the location of the identified outlier regions reflected the history of selection of the breeds. Large White and Landrace were the breeds with more regions commonly with an unusual decrease of nucleotide diversity. In contrast Duroc was the breed which had patterns of signals of unusual decreases of nucleotide diversity more different. Although, most signals were population-specific, many signals affected genomic regions harbored genes related with the same biological pathways. In regions with an unusual decrease of nucleotide diversity, a high representation of genes related with biological pathways responsible for coat color, brain development, muscle development, growth, metabolism were found in the pig breeds. Whereas the European wild boar presented a high representation of genes related with metabolism and immunity. Genes related with olfaction were over-represented in regions with unusual increase of nucleotide diversity and in regions with high genetic differentiation between breeds. Similarly to other mammals, some genes related with olfaction a perception of smell seem to follow a model of balancing selection whether others seem to follow a model of positive selection. Although the research presented in Chapter 4 represents a preliminary map of regions with potential interest for the study of selection, results were encouraging because we could detect outlier regions which overlap with previously identified signals of positive selection such as the *KIT* gene.

In Chapter 5 a similar approach was followed to study the patterns of genetic variation in commercial lines of chicken. Two lines of broilers and two lines of egg layers (one brown egg layer and one white egg layer). We found that patterns of genetic variation in chickens are dependent on chromosome size, with an upward trend towards the microchromosomes. We also found a correlation between the patterns of nucleotide diversity and recombination rate. Even though a large portion of the genome was covered, in the white egg layers, the reduced levels of nucleotide diversity conducted to a decrease in the power of our analysis. Nevertheless, we found outlier regions overlapping previously described sweeps such as in regions harboring the *IGF1* gene. Results are in accordance with the selection history of the commercial lines. We found that as expected different biological pathways have been affected by a reduction of genetic diversity in broiler and layer lines. Whereas regions with unusual reduction of genetic diversity in egg layers, contained a high representation of genes related with folate synthesis which is important for reproduction. Broilers showed a high representation of genes related with protein synthesis and sugar metabolism in regions with an unusual increase of nucleotide diversity.

In Chapter 6 important points of the thesis are highlighted and discussed. It is argued that in Europe the largest reservoir of genetic diversity is not in the wild boar but in the pig breeds. Even though in Chapter 2 it is shown that some European breeds share haplotypes which have high frequency in Chinese breeds, the source of the higher genetic variation found in European breeds has not been disclosed as also for the differences found between European and Chinese pig breeds. Future studies should therefore include Asian and European wild boar besides the European and Chinese pig breeds, in order to understand how different were the gene pools in the pig domestication centers and how the populations have evolved. Closely related species should also be included in order to ascertain the proportion of missing alleles during artificial selection. It is also argued how population genomics approaches can be used in future studies to uncover the genetic basis of complex traits. Although, the methods followed in Chapters 4 and 5 require future improvements, results showed how the differential selection applied to commercial lines of chicken and to international pig breeds can be detected. In the final part it is discussed how massive parallel sequencing provides the necessary means in order to apply this approach in the study of complex traits in domesticated species.

Samenvatting

Samenvatting

Al sinds 12000 jaar benut de mens de genetische diversiteit van dieren en planten waardoor ze een dominantie rol kan spelen in evolutie. De benutting van deze diversiteit, ook wel domesticatie genoemd, wordt gekenmerkt door kunstmatige selectie. Dit is de drijvende kracht die samen met natuurlijke selectie zorgt voor het succes van de mens. Kunstmatige selectie richt zich op het verbeteren van kenmerken die voordelige zijn voor de mens. Door dit proces is de huidige diversiteit in planten en dieren vorm gegeven. Varkens en kippen zijn voorbeelden daarvan en ze vormen een belangrijke eiwitbron voor een groot deel van de wereldbevolking. Daarnaast zijn deze soorten belangrijke modeldieren in biomedisch onderzoek.

De domesticatie van het varken vond onafhankelijk van elkaar plaats in Europa en in Azië gedurende het neolithische tijdperk. De eerste varkens die in Europa werden gehouden waren waarschijnlijk afkomstig uit klein Azië, maar werden vervolgens goeddeels vervangen door gedomesticeerde Europese wilde zwijnen. Door de domesticatie ontstonden veel verschillende rassen en de schatting is dat 70% van deze rassen in Eurazië zijn ontstaan. De kenmerken waarop deze rassen geselecteerd zijn waren o.a. huid- en vachtkleur, groei- en spieraanzetvermogen en toomgrootte. Met de toename van de industriële productie veranderde ook de fokkerij van locaal georganiseerde naar internationale schaal. In China daarentegen, het land met de meeste varkensrassen, is de fokkerij nog lang lokaal georganiseerd geweest. In de huidige varkenshouderij worden slechts een beperkt aantal internationale rassen gebruikt, namelijk: Pietrain, Large White, Landras en Duroc. Dit heeft tot gevolg dat de locale Europese rassen met uitsterven bedreigd worden en nog maar een kleine effectieve populatieomvang hebben. In China daarentegen, zijn er nog maar weinig rassen die met uitsterven bedreigd worden, maar gedetailleerde gegevens hierover zijn schaars. Vanwege de toenemende zeldzaamheid van veel varkensrassen is er onderzoek gestart op de genetische variatie in kaart te brengen. Hieruit bleek dat er veel unieke variatie is in Europese locale rassen en dat er nog meer genetische variatie voor komt tussen de Chinese rassen. De verschillen tussen Europese en Chinese diversiteit is mogelijk het gevolg van (1) kleine effectieve fokpopulaties; (2) verschillende fokdoelen; (3) verschillen in selectie intensiteit; (4) verschillen in genetische variatie van de oorspronkelijk gedomesticeerde varkens.

Kippen zijn ongeveer 500 jaar geleden gedomesticeerd en de meeste genetische diversiteit bevindt zich in Europa en Azië. In Europa zijn de meeste kippenrassen te vinden terwijl de grootste populaties zich in Azië bevinden. Veel locale rassen, meestal dubbeldoel (vlees en eieren) rassen, zijn uitgestorven of worden daarmee bedreigd. Commerciële fokkerij organisaties hebben gespecialiseerde elite lijnen voor ei- en vleesproductie gecreëerd uit een beperkt aantal van deze traditionele rassen. Introductie van genetische variatie uit de locale rassen in deze elite lijnen komt zelden voor. De genetische variatie in de elite lijnen wordt in stand gehouden door het inkruisen van andere elite lijnen en uit studies is gebleken dat een belangrijk deel van de totale genetische variatie ontbreekt in deze elite lijnen.

Het onderzoek dat beschreven staat in dit proefschrift heeft zich gericht op het karakteriseren van genetische variatie in kippen en varkens. In hoofdstuk 2, 3, 4 komt de genetische variatie van het varken aan de orde terwijl in hoofdstuk 5 ingegaan wordt op de genetische variatie bij de kip.

De mate waarin allelen (vormen van genen) gezamenlijk overerven heet Linkage disequilibrium (LD). Genen op een deel van het chromosoom erven gezamenlijk over maar de grootte van dit stuk kan variëren. De hoogte van de LD (grootte van dit stuk) kan gebruikt worden om de geschiedenis van een populatie te beschrijven. In hoofdstuk 2 is de mate van LD bestudeerd met behulp van 371 SNPs (single nucleotide polymorphisms, genetische merkers) binnen 3 chromosomale regio's van Europese en Chinese rassen. De mate van LD verschilde significant tussen de Europese en Chinese rassen. De lengte van de chromosoomstukken bereikte waarden tot wel 2 cM in Europese rassen terwijl in de Chinese rassen dit maar 0.5 cM bedroeg. De LD in de Europese rassen bleek ook te variëren tussen de regio's die werden bestudeerd. Chinese rassen vertoonden meer haplotypen in vergelijk met de Europese rassen en enkele Chinese haplotypen bleken ook voor te komen bij Large White, Landras en Duroc (Westerse rassen). Op grond van deze 3 regio's is berekend dat de genetische variatie goed in kaart te brengen is als het gehele genoom wordt gescreend met behulp van 30.000 SNPs voor Europese rassen en 500.000 SNPs voor Chinese rassen. In hoofdstuk 6 is aan de hand van de variatie in de 3 regio's berekend hoe groot de oorspronkelijke populaties waren voor de verschillende rassen. Hieruit bleek dat de Chinese rassen oorspronkelijk groter waren, wat zou kunnen betekenen dat de Chinese genen-pool meer divers is geweest. De hoge mate van LD in het Europese wilde zwijn suggereert dat er sprake kan zijn geweest van 'bottle necks' in het ontstaan van deze soort. De wilde zwijnen zijn echter alleen bemonsterd in een Frans reservaat. Hoofdstuk 2 benadrukt de noodzaak voor het vergroten van de informatie over genetische variatie in het genoom van het varken.

In hoofdstuk 3 is een methode beschreven om genoomwijd SNPs te identificeren in varkens. De methode behelst het combineren van DNA-pooling van 5 dieren, het maken van 'reduced representation libraries' (RRLs) en het op grote schaal parallel sequensen van korte DNA-fragmenten (massive parallel sequencing: MPS). MPS geeft veel meer informatie dan de traditionele manier van sequencen (Sanger sequencen), maar bevat wel meer fouten. Hierdoor is de analyse van MPS data een grote uitdaging met name m.b.t. het filteren van de data. In hoofdstuk 3 worden verschillende data filteringstrategieën met elkaar vergeleken om een maximale output aan SNPs over het genoom te genereren. Duizenden SNPs werden geïdentificeerd, maar het uitfilteren van SNPs met een lage kwaliteit heeft consequenties. SNPs met een lage kwaliteit veroorzaken ruis en verminderen de identificatie van varianten als ze toch gemapped worden op het referentie genoom. Het genoom is het totaal aan baseparen, ook wel nucleotiden genaamd. De genetische diverstiteit kan worden beschreven als het aantal SNPs in het genoom, en word dan ook wel nucleotide diversiteit genoemd. Het blijkt dat bij (sub)metacentrische chromosomen de diversiteit toeneemt naar het eind van de chromosomen (telomeren) en afneemt naar het midden (centromeer). De nucleotide diversiteit op het geslachtschromosoom (X) is lager in vergelijking met de autosomale chromosomen.

De beschreven methode blijkt kosteneffectief te zijn voor het bestuderen van genoomwijde variatie en kan worden toegepast bij alle diersoorten waarbij een (gedeeltelijk) referentie genoom bekend is. Daarom is in hoofdstuk 4 dezelfde methode gebruikt om de genetische variatie in meer rassen te beschrijven. Het betrof de volgende internationale rassen: Pietrain, Large White, Landras en Duroc en daarnaast het Europese wilde zwijn. Pietrain bleek de laagste nucleotide diversiteit te hebben gevolgd door het wilde zwijn. Dit bevestigde eerdere studies waaruit ook naar voren kwam dat het wilde zwijn weinig bijdraagt aan de genetische diversiteit in vergelijk met locale rassen. De lagere genetische variatie in Pietrain ten opzichte van andere internationale rassen was ook al eerder gevonden. Het patroon van diversiteit, wat gevonden was in hoofdstuk 3, bleek goed overeen te komen met de resultaten in hoofdstuk 4 met specifieke verschillen voor sommige rassen en het wilde zwijn.

Veredeling ten tijde van het domesticatieproces, alsook nu nog in de commerciele fokkerij, is hoofdzakelijk gebaseerd op het fenotype. Fokken op basis van het genotype is tot op heden lastig wegens gebrek aan kennis over hoe het fenotype, gecodeerd is in het genotype, het DNA. Het fenotype, ofwel het geheel aan kenmerken van het dier, wordt vaak door veel genen met elk een klein effect beïnvloed, hoewel sommige genen een groter effect kunnen hebben. De identificatie van genen die van invloed zijn op een fenotype is daarom niet eenvoudig. Veel studies naar deze genen (Quantitative Trait Loci: QTL) hebben enkel genen met een groot effect opgespoord.

Een andere benadering is het bestuderen van genomische variatie in populaties om daarmee de genetische basis van complexe kenmerken te identificeren. De aanname hierbij is dat genen die variatie vertonen die afwijkt van wat verwacht kan worden in afwezigheid van selectie, kandidaat genen zijn die het fenotype beïnvloeden. Het is de verwachting dat genen die biologisch interacties aangaan dezelfde evolutionaire historie zullen vertonen. Deze populatiegenomische benadering vereist dat de genetische variatie van veel loci gelijktijdig bestudeerd wordt. Tot nu toe ontbrak het aan voldoende genomische informatie om deze benadering toe te passen om de genetische basis van complexe kenmerken te ontrafelen. In Hoofdstuk 4 worden regio's/genen die vermoedelijk onder invloed zijn geweest van selectie beschreven. Het patroon van deze geselecteerde regio's over het genoom geeft de historie van selectie van de diverse rassen weer. Bij Large White en Landras zijn overeenkomstige regio's gevonden met een lage diversiteit. Duroc vertoonde het meest afwijkende patroon van regio's met een lage diversiteit. Hoewel de meeste signalen populatie specifiek waren zijn er ook signalen gevonden waarbij genen uit een zelfde biologisch netwerk of 'pathway' betrokken waren. Regio's met een lage diversiteit bleken genen te bevatten uit pathways die vachtkleur, hersenontwikkeling, spiervorming, groei en metabolisme beïnvloeden. Bij het Europese wilde zwijn vertoonden vooral genen met betrekking tot metabolisme en immuniteit minder diversiteit. Genen die van invloed zijn op het reukvermogen bleken vooral in gebieden met meer dan verwachte diversiteit voor te komen en in regio's waarin rassen van elkaar verschilden. Net als bij ander zoogdieren bleek een deel van deze 'reuk'-genen een 'balancing selection model' te volgen terwijl andere genen een 'positive selection model' volgden. Hoewel hoofdstuk 4 slechts een voorlopige kaart van geselecteerde regio's bevat zijn de resultaten hoopgevend omdat deze regio's overlap vertonen met eerder geïdentificeerde signalen van positieve selectie, zoals het *KIT* gen.

In hoofdstuk 5 is dezelfde benadering als in hoofdstuk 4 toegepast op de genomische variatie van de kip. Twee vleeskuikenlijnen en twee leglijnen (een witte en een bruine ei-leglijn) werden vergeleken. Het bleek dat de genetische variatie bij kippen afhangt van de chromosoomgrootte, op kleinere (micro)chromosomen bleek de diversiteit hoger te zijn. Daarnaast bleek er een verband tussen de mate van recombinatie en de nucleotide diversiteit. Hoewel vrijwel het gehele genome gevolgd kon worden was als gevolg van de lage diversiteit in de witte leg lijn minder power voor analyses. Ondanks dit probleem hebben we toch regio's gevonden die overlapten met regio's waarin eerder 'selected sweeps' zijn beschreven, bijvoorbeeld de *IGF1*-regio. De resultaten bleken goed overeen te stemmen met de selectie historie van de commerciële lijnen. Zoals verwacht kon worden is er op verschillende pathways geselecteerd in de vlees- en leglijnen waar dan ook verminderde diversiteit gevonden is. Bij de leglijnen betrof het genen die samenhangen met folaat synthese dat een rol speelt bij de vruchtbaarheid. Bij de vleeskuikens bleken vooral regio's met genen die eiwitsynthese en suikermetabolisme bevatten een verhoogde diversiteit te vertonen.

In hoofdstuk 6 worden de eerder beschreven resultaten van dit proefschrift bediscussieerd. Er wordt betoogd dat de meeste variatie in varkens rassen voorkomt en niet in hun wilde soortgenoten. Hoewel in Europese rassen Chinese haplotypen voorkomen is het niet duidelijk waar deze hogere variatie in locale Europese rassen door veroorzaakt wordt, evenals de verschillen tussen de Europese en Chinese rassen. Toekomstige studies zouden zich dan ook meer moeten richten op het ontrafelen van de genetische variatie in Europese en Chinese wilde zwijnen, naast de bestaande rassen, om beter te kunnen begrijpen hoe genetische variatie is veranderd door domesticatie tot stand gekomen. Wellicht kan het toevoegen van ander soorten binnen het geslacht *Sus* helpen om het aandeel missende allelen vast te stellen.

Verder wordt er betoogd hoe de populatie genomische benadering gebruikt kan worden bij de studie van complexe kenmerken. Hoewel de methode, beschreven in hoofdstuk 4 en 5, verbeterd kan worden liet het zien dat geselecteerde regio's in het genoom ontdekt konden worden.

Het hoofdstuk wordt afgesloten met een discussie over de toepassing van 'massive parallel sequencing' voor het bestuderen van complexe kenmerken in gedomesticeerde dieren.

Aknowledgments

Aknowledgments

This thesis is the result of many hours at my desk, huge consumption of paper and many air miles and I thank everybody who gave me the impetus to face this challenge and helped me along the way.

First of all I would like to thank my family and friends for all their support, patience and for believing in me. A special thanks to my husband, Luis Filipe Fonseca to whom I dedicate this thesis and to my little sister Daniela. Thank you for being my harbor, where I could always get more supplies of love, strength and courage.

I am deeply grateful to my promoter Martien Groenen and my co-promotors Henri Heuven and Hendrik-Jan Megens for believing in me and for giving me this opportunity. Thank you for your guidance and support. During this time, Henri was the supervisor who always called me back to earth: “make a plan”,” before anything else make an outline of your paper”. Hendrik-Jan was the supervisor who at his own way challenged me and deviated me from the plan and was always available Thank you both.

I thank all my colleagues from the ABG group for their help and support. Albart thank you for all the R tips. Hinri thanks for all the PERL and bash tips. Haisheng thank you for all the nice discussions we had! Thank you Pieter and Martin for being such a nice roommates. Thank you Ada and Monique, for all your help.

I thank Andy Law and Alan Archibald, who supervised me in my external period at Roslin Institute. Andy, thank you so much for transmitting me important foundations in bioinformatics, which I still apply in my daily working routine. Thank you for being always available for teaching and helping me. I also thank to my colleagues at Andy’s group for being so helpful and supportive while I was there.

I thank the girl power from the Southern European cluster at South wing in Roslin. Valentina, Ciupi, Stefania, Silvia and Fausta for being so supportive. Thank you for making this time so pleasant. For three months I didn’t visit home and it would have been much more difficult without you.

I thank Edinburg!!!!! The city where something is always happening and where is so fun to live.

I thank Miguel Peréz-Enciso, Luca Ferretti and Sebas Ramos, for all of their support in Barcelona. Thank you for all your patience and for being such good buddies.

I thank Cecile and Ane for being my out of office mates in Wageningen. I really enjoyed the time we've spent together.

I thank Pieter!!!!! Thank you for listening me, for cooking pizza for me and for all the great time we spent together.

I thank Ana Sousa for being so understanding and giving some days from work in order to finish my thesis.

Finally, thank you Wageningen and the Netherlands. I will miss my bike rides along the Rhine, which relaxed me. I will miss the smell of flowers in the spring, the apple pie and all the lovely places where I and Filipe had such a nice time.

Curriculum vitae

Short biography

Andreia de Jesus Amaral Gomes Barbosa Fonseca was born on May 28 1976 in Cascais, Portugal. In 1994, she graduated from high school *Escola Secundária de Alvide* and started her study in Agronomic Engineering at *Instituto Superior de Agronomia,* Technical University of Lisbon. She specialized in Animal Production and she performed a research thesis at the *Centro de Botânica Aplicada à Agricultura* in evaluation of "GxE interations in vineyard clones". After graduation she worked in consultancy and as an officer for the management of European Funds of the Common Agricultural Policy at "Instituto Nacional de Garantia à Agricultura". In 2004 she started her Master study in Conservation Biology in *Faculdade de Ciências*, University of Lisbon. She developed a research thesis at *Centro de Biologia Aplicada* entitled **"**Study of genetic integrity of *Coturnix coturnix* Linnaeus, 1758 in Portugal and detection of hybridization with an exotic species *Coturnix japonica* Temminck and Schlegel, 1849". In 2006 she obtained her MSc degree and in the same year, she started her PhD study at the Animal Breeding and Genomics Centre of Wageningen University. The results of the research on the characterization of nucleotide variation and the effect of selection in the porcine and chicken genomes are described in this thesis. Since January 2010, she is part of the research team headed by Prof. Dr. Ana E. Sousa, at the Clinical Immunology Unit, *Instituto de Medicina Molecular, Faculdade de Medicina,* University of Lisbon. She is studying the modulation of expression of miRNAs in T-cells during HIV infection.

List of publications

Papers in refereed journals

Amaral A. J.*, Megens H-J., Kerstens, H.D., Heuven H. C. M., Crooijmans R. P. M. A., den Dunnen, J.T., Groenen M. A. M. (2009) **Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome.** *BMC Genomics*, 10:374, *corresponding author.

Ramos A. M., Crooijmans R.P.M.A., Affara N., **Amaral A.J.,** Archibald A. , Beever J., Bendixen C., Dehais P., Hansen M., Hedegaard J., Hu Z., Kerstens H.D., Law A., Megens H-J, Milan D., Nonneman D., Rohrer G., Rothschild M., Smith T., Schnabel R.D., Van Tassell C., Clark, R. Churcher C., Taylor J., Wiedmann R. , Schook L.B., Groenen M.A.M. (2009) **Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology.** *PlosONE,* 4(8):e6524.

Amaral A. J., Megens H-J, Heuven H. C. M., Crooijmans R. P. M. A., Groenen M. A. M.(2008) **Linkage Disequilibrium Decay and Haplotype Block Structure in the Pig.** *GENETICS,*179:569-579.

Amaral A. J., Grosso A.R., Chicki L., Bastos-Silveira C., Dias D. (2007) **Detection of hybridization and species identification in domesticated and wild quails using genetic markers.** *FOLIA ZOOLOGICA* 56 (3): 285-300.

Papers in preparation

Amaral A.J., Megens H-J, Zare Y., Nie H., Heuven H.C.M., Ferretti L., Ramos-Onsins S.E., Perez-Enciso M., Vereijken A., MacEachern S., Muir W.M., Cheng H.H.' Groenen M.A.M. (in preparation) **Localizing recent selection in the chicken genome from DNA pools.**

Amaral A.J., Ferretti L., Megens H-J, Crooijmans R.P.M.A., Nie H., Ramos-Onsins S.E., Perez-Enciso M., Schook L.B., Groenen M.A.M.(in preparation) **Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing.**

<u>Conference abstracts</u>

Amaral A.J., Nie H., Ferretti L., Megens H-J, Crooijmans R.P.M.A., Ramos-Onsins S.E., Perez-Enciso M., Schook L.B, Groenen M.A.M. (2009) **Massive parallel sequencing reveals footprints**

of selection in the porcine genome. *Oral presentation, Jornadas de Bioinformática*. Lisbon, Portugal.

Amaral A.J., Ferretti L., Megens H-J, Crooijmans R.P.M.A., Nie H., Ramos-Onsins S.E., Perez-Enciso M., Schook L., Groenen M.A.M. (2009) **Finding selection footprints in the swine genome using massive parallel sequencing.** *Poster, Conference on Next Generation Sequencing: Challenges and Opportunities. Barcelona, Spain.*

Amaral A.J., Archibald A.L., Groenen M.A.M., Law A. and the International Porcine SNP Chip Consortium (2008) **Mapping SNPs to un-sequenced regions of the pig genome using comparative genomics.** *Poster, 3rd SABRE Conference: Welfare and Quality Genomics, Foulum (University of Aarhus), Denmark.*

Amaral A. J., Kerstens H.H.D., Megens H-J, Dibits B., Crooijmans R. P. M. A, Dunnen J., Groenen M. A. M (2008) **Genome wide SNP discovery in pig using 1G Genome Analyzer: How to play around with sequence length, quality level and mapping qualities.** *Poster, Conference of the International Society for Animal Genetics, Amesterdam, The Netherlands.*

Amaral A. J., Megens H-J, Heuven H. C. M., Crooijmans R. P. M. A., Groenen M. A. M. (2008). LD In Pigs: Differences Between China And Europe**.** *Abstract Plant & Animal Genome XVI Conference, San Diego, USA.*

Amaral A. J., Megens H-J, Heuven H. C. M., Crooijmans R. P. M. A., Groenen M. A. M.(2007). **Comparative analysis and linkage disequilibrium in European and Chinese pig breeds.** *In: Book of Abstracts : NWO retraite genetica, Rolduc, Netherlands.*

Training and Supervision Plan

The Basic Package (3 credits[1])	
WIAS Introduction Course	2007
Course on philosophy of science and/or ethics	2007
Scientific Exposure (12 credits)	
International conferences *(type presentation)*	
16th ZonMw genetica retraite (oral)	2007
Plant & Animal Genomes XIV Conference (oral)	2008
ISAG meeting (poster)	2008
SABRE 2008 (poster)	2008
NGS meeting (poster)	2009
Jornadas de Bioinformática 2010 (oral)	2009
Seminars and workshops	
Ecological implications of adaptive behavior	2006
Darwinian agriculture: the evolutionary ecology of agricultural symbiosis	2007
Browsing genes and genomes with Ensembl	2007
In-Depth Studies (22 credits)	
Disciplinary and interdisciplinary courses	
Perl for Bioinformatics	2007
Molecular Evolution, Phylogenetics and Adaptation	2008
Coalescent theory	2009
Advanced statistics courses	
Linear Models in Animal Breeding	2007
Bayesian Statistics course	2007
Association Mapping	2009
Genomic selection	2009
PhD students' discussion groups	
Quantitative Genetics study group	2006/08
MSc level courses	
Modern Statistics in life sciences	2007

Professional Skills Support Courses (4 credits)	
Industrially relevant skills course	2007
Dutch language	2007
Course Techniques for Scientific Writing	2008
Career Orientation course	2008
Research Skills Training (8 credits)	
Preparing own PhD research proposal	2006
External training period	2008

Management Skills Training (5 credits)	
Organisation of seminars and courses	
WIAS science day	2008
Membership of boards and committees	
Education Committee Board	2007/2008

Total credits: 54

[1] One credit equals a study load of approximately 28 hours.

Colophon

This study was conducted as part of the SABRETRAIN Project, funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission.