GENETIC DIFFERENTIATION WITHIN AND BETWEEN POPULATIONS UNDER SELECTION – STUDIES ON DIVERSE CHICKEN POPULATIONS AND THE GÖTTINGEN MINIPIG

Dissertation

for the Doctoral Degree

at the Faculty of Agricultural Sciences,

Georg-August-University Göttingen

presented by

Christian Gärke

born in Thuine

Göttingen, May 2012

D 7

1. Supervisor: Prof. Dr. Henner Simianer

2. Co- supervisor: Prof. Dr. Georg Thaller

Date of disputation: 21 May 2012

for you

“A friend may well be reckoned
the masterpiece of nature.“

Ralph Waldo Emerson

Table of contents

Genetische Differenzierung innerhalb und zwischen Populationen unter Selektion - Studien zu unterschiedlichen Hühnerrassen und dem Göttinger Minischwein

Ziel dieser Arbeit war es, verschiedene Aspekte der Verwendung von genetischen Markern in der genomischen Charakterisierung und Analyse von verschiedenen Tierpopulationen zu untersuchen. Zu Beginn werden die verschiedenen Anwendungsbereiche von genetischen Markern in der Tierzucht beschrieben.

In der ersten Analyse wurden acht Hühnerrassen für 9‘216 Single Nucleotide Polymorphism (SNPs) und 29 Microsatelliten (Single Sequence Repeats, SSRs) typisiert. Um die Rassen zu differenzieren wurden zwei unterschiedliche Methoden herangezogen: (i) die Bayesian Model-based Clustering Analyse, die im Programm STRUCTURE (Version 2.3) implementiert ist und (ii) eine Hauptkomponenten-Analyse (Principal Component Analysis, PCA), bei der eine Differenzierung der Rassen aufgrund ihres Euklidischen Abstandes zueinander durchgeführt wurde.

Die Ergebnisse der STRUCTURE Analyse zeigten, dass die Wiederholbarkeit bei SNPs, unabhängig von Ihrer Anzahl, höher war als bei den SSR. Bei der Zuordnung eines Individuums zu einer der Rassen wurden die höchsten Werte für 29 SSRs und 100 SNPs errechnet. Die PCA-basierte Methode ergab, dass 2.4 SNPs je SSR benötigt werden, um eine vergleichbare Differenzierung zu erreichen. Dieses Ergebnis ist vergleichbar mit Untersuchungen, die an Menschen oder Rindern durchgeführt wurden. Unter Verwendung aller SNPs konnte im Vergleich zu allen vorhandenen SSRs eine genetisch inhomogene Rasse detektiert werden. Aufgrund dieser Ergebnisse und der deutlich größeren Anzahl an verfügbaren SNPs verglichen mit SSRs kann davon ausgegangen werden, dass SNP-basierte Ansätze weiter an Bedeutung gewinnen werden.

Des Weiteren wurde eine umfassende Kartierung von Selektionssignaturen auf den Autosomen des Göttinger Minischweins durchgeführt. Für die Suche nach Selektions-signaturen wurden die mittels des Illumina Porcine BeadChip 60K (Illumina, San Diego, USA) gewonnenen SNP Daten anhand von zwei Methoden untersucht: Zum einen wurde der Long Range Haplotype Test (LRH) angewendet, der im Softwarepacket SWEEP integriert ist. Zum anderen wurden die genomischen Anteile der drei Ausgangsrassen für jeden SNP aufgrund einer Bayes‘schen Methode geschätzt.

Es konnte eine signifikante Veränderung der Anteile der Ausgangsrassen verglichen mit den berechneten pedigree-basierten Erwartungswert festgestellt werden. Hierbei fiel auf, dass die Zuteilung der Allele zu einer der drei Ausgangsrassen sowohl zwischen als auch innerhalb der Chromosomen hoch variabel ist. Es wurde angenommen, dass eine lokale Abweichung der Zusammensetzung als Hinweis darauf interpretiert werden kann, dass diese Region unter gerichteter Selektion stand. Mit Hilfe dieses Indikators und den Ergebnissen des LRH-Testes konnte eine Vielzahl von Regionen identifiziert werden, die unter Selektion standen. Einige dieser Regionen beherbergen Kandidatengene, die funktionell mit den Zuchtzielen der Göttinger Minischweine im Zusammenhang stehen, z.B. *SOCS2*, *TXN*, *DDR2* und *GRB10*, die mit der Körpergröße in Verbindung gebracht werden, oder das *PRLR* Gen, das die Wurfgröße beeinflusst. Die Ergebnisse dieser Untersuchungen lassen Rückschlüsse darauf zu, dass die Beziehung der Gene *SOCS2* und *GRB10* zu dem *IGF-1* Gen der mögliche Grund für den Zwergwuchs im Göttinger Minischwein sein könnte.

In einem weiteren Schritt wurden die Ergebnisse der Kartierung der Selektionssignaturen auf den Autosomen des Göttinger Minischweins validiert. Hierfür wurde die Cross Population Extended Haplotype Homozygosity (XPEHH) berechnet. Diese Methode basiert, genau wie der LRH-Test, auf dem Auffinden von Regionen ausgedehnter Haplotypenhomozygotie. Die beiden Ansätze unterscheiden sich dadurch, dass beim LRH-Test innerhalb einer Rasse und beim XPEHH im Vergleich zweier Rassen Selektionssignaturen aufgedeckt werden. Als Vergleichsrassen wurden Tiere der Rassen Göttinger Minischwein, Deutsche Landrasse und Large White mit dem Illumina Porcine BeadChip 60K (Illumina, San Diego, USA) genotypisiert.

Mit Hilfe des XPEHH-Tests konnten weitere Regionen identifiziert werden, die unter Selektion standen. Aufgrund des Vergleichs von Großschweinerassen mit dem Göttinger Minischwein und dem erneuten Auffinden des *SOCS2* Gens wird die Vermutung bestärkt, dass es sich hierbei um eines der wichtigsten Gene für den Zwergwuchs beim Göttinger Minischwein handelt.

Zusammenfassend ergibt sich, dass SNP-basierte Ansätze einen deutlich besseren Einblick in die genomische Architektur von Populationen ermöglichen. Dadurch wird ein besseres Verständnis von Selektion und Differenzierung von Rasse auf genomischer Ebene erreicht.

Genetic differentiation within and between populations under selection – studies on diverse chicken populations and the Göttingen Minipig

In this thesis different aspects of the use of high density markers in the genetic characterisation and analysis of farm and experimental animal populations are addressed. First, a general introduction and an overview over the different fields of marker applications in animal breeding and farm animal genetics is given.

The first analysis deals with the marker-based differentiation of chicken populations. Eight chicken breeds were genotyped for 9'216 single nucleotide polymorphisms (SNPs) and 29 microsatellites (single sequence repeats, SSRs). Two different methods were applied to differentiate the breeds: (i) a Bayesian model-based clustering approach, which is implemented in the software STRUCTURE version 2.3 and (ii) a partitioning of the Euclidean distance matrix based on a principal component analysis (PCA).

In the model-based clustering, the similarity coefficient obtained with SNPs compared to SSRs showed significantly higher values between repeated runs. The membership coefficients, reflecting the proportion in which a fraction segment of the genome belongs to a particular cluster, showed the highest values for 29 SSRs and 100 SNPs, respectively. The PCA-based partitioning showed that 2.4 SNPs per SSR were required to achieve equivalent differentiation ability. This result is comparable with studies conducted on humans or farm animals reported in the literature. With the use of high SNP numbers it was possible to detect genetic heterogeneity of one breed which was completely missed when all available SSRs were used. The results of our study and the availability and cost-efficiency of larger numbers of SNPs compared to SSRs suggest that SNP-based approaches will probably become the technology of choice in farm animal genetic studies.

In a second analysis, a genome-wide mapping of selection signatures on the autosomes of the Göttingen Minipig (GMP) was carried out. To search for signatures of recent positive selection, genotypes obtained with the Illumina Porcine BeadChip 60K (Illumina, San Diego, USA) were analysed by two methods: (i) the Long Range Haplotype (LRH) test, which is integrated in the software SWEEP and (ii) the estimation of the membership coefficient of the three founder breeds for each SNP using a Bayesian method.

The breed composition of the Göttingen Minipig (i.e. the estimated proportion of the three founder breeds) was found to deviate significantly from to the proportions expected from pedigree information. The probability of alleles to originate from one of the three founder breeds of the GMP is highly variable between the chromosomes and even within each chromosome. It was assumed that selection is a genetic mechanism having a locus-specific impact on the composition of the genome and considerable local deviations from the genome-wide average can be interpreted as regions being under directional selection. Combining the membership coefficient and the results of the LRH test, several regions under selection were identified. Some regions of recent selection overlapped with candidate genes which are related to breeding goals of the Göttingen Minipig, e.g. *SOCS2*, *TXN*, *DDR2* and *GRB10*, which are connected with the body size, and the *PRLR* gene, which affects the litter size. The results suggest that the connection between the *SOCS2* and *GRB10* gene with the *IGF-1* gene might be one reason for the small body size of the Göttingen Minipigs.

In a further step the results of the LRH test and the membership coefficient were validated with the Cross Population Extended Haplotype Homozygosity (XPEHH) approach. This method is based on the detection of long haplotype homozygosity between different breeds. The XPEHH test was used to detect selective sweeps between the Göttingen Minipig and German Landrace and Large White pigs based on genotypes obtained with the Illumina Porcine 60K BeadChip.

The XPEHH test revealed additional regions that might have been under divergent selection in the GMP compared to the two normal-sized breeds. Again the region containing the *SOCS2* gene produced one of the most prominent signals, so that the conjecture is confirmed that this might be one of the most important genes involved in the dwarfism in the Göttingen Minipig.

In conclusion the applications of high density genotyping data reported in this study suggest that SNP-based approaches allow a much better insight in the genomic architecture of populations and, by this, lead to a better understanding of the mechanisms underlying selection and breed differentiation on the genomic level.

1st CHAPTER

General introduction

General Introduction

The first step of classical genetics is to identify genes concerned in inheritance and to locate them on linkage maps (Thoday 1961). The birth date of the genomic era was reached with the complete sequencing of the human genome in April 2003 (Guttmacher & Collins 2003). The advent of sequencing has significantly accelerated biological research. Genome sequencing is used for determining the exact order of the nucleotide bases in a molecule of DNA. Knowledge of genome sequences has become indispensable for biological research in human, plant and animal populations.

Genetic markers and their use in animal breeding represent one of the most powerful tools for the analysis of genomes. They are used to detect variation among individuals or between alleles in a particular segment of DNA. Genetic markers are variations in the DNA sequence which can be inherited from an ancestor or rarely arise as the result of a novel mutation. Many different kinds of genetic markers have been used over the last decades:

- Random amplified Polymorphic DNA (RAPD)
- Restriction Fragment Length Polymorphisms (RFLP)
- Amplified Fragment Length Polymorphisms (AFLP)
- Microsatellites (Simple Sequence Repeat, SSR)
- Diversity Array Technology (DArT)
- Single Nucleotide Polymorphisms (SNP)
- Copy Number Variations (CNV)

Vignal *et al*. (2002) mentioned two points before using markers for genetic studies: (i) from the molecular biologist's point of view it is necessary to produce them at low cost and as simple as possible in order to generate as much genotypes as possible. (ii) From the statistician's point of view the dominance relationships, information content, neutrality, map positions or genetic independence of markers is important. Genetic markers have developed rapidly over the last years. The two main markers at the moment are microsatellites (Simple Sequence Repeats, SSRs) and Single Nucleotide Polymorphisms (SNPs), now used in applications in genetic analysis (Duran *et al*. 2009).

SSRs are short sequences of DNA that can be highly polymorphic in terms of their length and number of repetitions (Tautz 1989; Weber & May 1989). An allele of an SSR is defined by the number of the same base pair-sequences, an example of a SSR is …CGCGCGCGCGCGCGCG… where the dinucleotide motif "CG" is repeated eight times. SSRs are quite common in the genome and have been found in higher prokaryotic and eukaryotic organisms to date (Toth *et al*. 2000; Katti *et al*. 2001). Microsatellites are a powerful genetic marker system due to their high genomic abundance, random distribution across the genome, genetic co-dominance, high polymorphism, multi-allelic variation and high reproducibility (Duran *et al*. 2009; Teneva 2009). In the past, SSRs were used predominantly to find loci that have a significant impact on a phenotypic trait (quantitative trait loci, QTL) and to uncover relationships between markers and QTLs. Using them is a powerful way of mapping genes controlling economic traits (Beuzen *et al*. 2000). The high variability made them invaluable for human genetic linkage studies (Weir *et al.* 2006) for a long time. They have been widely used in genetic studies of humans and livestock populations (Ball *et al*. 2010). The overall number of SSRs in the genome depends mainly on their complexity and size. More than 10'000 microsatellite sequences are identified for pigs (Karlskov-Mortensen *et al*. 2007) and in chicken about 7'300 polymorphic microsatellites (Brandström & Ellegren 2008) have been reported so far.

SNPs are single base changes in the DNA sequence with an alternative of two possible nucleotides at a given position. For example, at a certain position in the DNA there may be a C (cytosine) present in some gametes but a G (guanine) in others. SNPs are usually biallelic, which means that they are less informative compared to SSRs, but this is compensated by their high abundance (Schaid *et al.* 2004; Liu *et al*. 2005; Ball *et al.* 2010). The big number of SNPs can provide a high density of markers around a locus of interest (Duran *et al*. 2009). SNP genotyping is low in error rate, easily automated in high-throughput technologies and cost effective (Fries & Durstewitz 2001; Xing *et al*. 2004). In pigs one SNP appears about every 500 base pairs (Wiedmann *et al.* 2008) and in chicken about every 300 base pairs (Vignal *et al*. 2002). About 5.4 Million SNP in pigs and about 4 Million SNP in chicken are potentially exist. Today SNP arrays cover up to 50'000 SNPs in chicken, 60'000 SNPs in pigs, 700'000 SNPs in cattle and more than 1'000'000 SNPs in humans.

During the past decades several genetic marker technologies have been developed and applied for animal breeding. These technologies are predominantly for assessing differences between animals of one species (within and between different populations) and to find a relationship of genomic regions with important phonotypical traits. SNPs and SSR have many uses in genetic research, such as:

- Assessing genetic variation (Nei & Li 1979, Bennett et al. 2005)
- Linkage mapping (Geldermann 1975)
- Association studies (Botstein & Risch 2003; Klein et al. 2005)
- Marker assisted selection (Lande & Thompson 1990)
- Genomic selection (Meuwissen et al. 2001; Schaeffer 2006)

In chapter 2 genetic markers will be used to differentiate eight chicken populations. The aim is to determine the number of SNPs needed to achieve an equivalent differentiation power as with a given standard set of microsatellites. For this, eight different chicken populations, comprising both commercial and fancy breeds, were available.

Two classical statistical approaches are used to differentiate SNPs and SSRs: principal component analysis (PCA) -based partitioning of the distance matrix and a model-based clustering implemented in the software STRUCTURE.

The PCA is the most common data reduction method to differentiate populations based on allele frequency data (Morrison 1976; Laloë *et al*. 2007). It is a non-parametric linear dimension technique (Lee *et al*. 2009) to classify individuals on a reduced number of significant components (Dunteman 1989). Further the PCA is suitable for genetic markers with two alleles (SNPs) and also with more than two alleles (SSRs) (Patterson *et al*. 2006). This method is well qualified to uncover the population structure, even in the case of admixed populations without information about the origin of individuals (Paschou *et al*. 2007). The Euclidean distance based on the first two coordinates of the individuals was calculated to compare the differentiation ability for different types of genetic marker (SNP and microsatellites).

The second method used to determine the structure of populations is a model-based clustering implemented in the software STRUCTURE (Pritchard *et al*. 2000). STRUCTURE was used to cluster individuals to a defined number of assumed populations based on different genetic markers (SNPs and microsatellites). It is possible

to identify distinct genetic populations, assigning individuals to populations, and to identify admixed individuals (Pritchard *et al*. 2000). This software has been used in several studies for assessing the genetic structure of populations (Rosenberg *et al*. 2002; Bodzsar *et al*. 2009).

Another possibility to use genetic markers is to detect signatures of recent positive selection. In chapter 3 and 4 the detection of recent selection is done within and between pig breeds. This study focuses on the Göttingen Minipig (GMP) and its founder breeds to find genomic regions that may have undergone selection since their creation. The GMP is an animal model developed to meet the special demand for non-rodent animal models. Compared to normal pig breeds, the body weight of adult miniature breeds is much lower. In toxicological tests this can reduce the costs for experiments when the often extremely expensive test compounds are dosed per kg body weight. During the last years, the demand for minipigs grew more and more because of the high physiological and anatomical similarities to humans (Brandt *et al*. 1997). To evaluate the potential of minipigs as an animal model in medical research, the EU-project RETHINK was realised. Different authors summarised that minipigs will be useful for the testing of biotechnology products based on the close sequence homology between pigs and humans. Minipigs are the only non-rodent model where transgenic animals can be easily generated (Forster *et al*. 2010). The immune system of the pig is better characterized than that of the dog or primates (McAnulty 1999; Bode *et al*. 2010). The costs of minipigs as a medical model are not significantly higher than the costs for a study in dogs (Van der Laan *et al*. 2010). The most important minipig breeds used as non-rodent animal models in medical research are the Göttingen Minipig, the Minnesota Minipig, the Yucatan Minipig and the Hanford Minipig (Köhn 2007).

The GMP was developed in the 1960's at the University of Göttingen, Germany, by crossing Minnesota Minipigs (MMP), Vietnamese Potbelly Pigs (VPP) and German Landrace (GL). The first generations were obtained by crossing the MMP (low body weight) and the VPP (high fertility) which led to a small and coloured (black or spotted) pig breed. Because pigs with a white skin are more desirable for animal experiments in dermatology, the GL was introduced between 1965 and 1969 by artificial insemination. Since 1970 a distinct white line of the GMP was established (Glodek & Oldigs 1981). Due to the strong market demand for white GMPs, these animals were produced with a heavily expanding production and the production and maintenance of the coloured line

was stopped. Thus, the GMP is a closed breed since the beginning of the 1970s, making it a relatively young breed.

During the last 30 years the breeding goals of the GMP were aligned to the market demands (e.g. small body size, sufficient fertility, moderate inbreeding coefficient, unpigmented skin, calm temperament). At the beginning of the development, the main focus was to achieve a moderate inbreeding coefficient by a high exchange of breeding animals. After establishing this, the GMP was phenotypically selected for low body weight on the basis of birth and weaning weight (Glodek & Oldigs 1981). Since the 1970's the breeding goal focused more on litter size because of the high demand for the animals. This resulted in a positively correlated selection response on body weight reflecting the genetic and physiological antagonism between litter size and body weight in multiparous species (Ferguson *et al*. 1985, Simianer & Köhn 2010). Thus, since the mid 1970's the breeding goal combined low body weight and increased litter size. Another recent breeding objective is a calm temperament, especially in interaction with humans (Köhn *et al*. 2009). Besides a high uniformity in the pigs, e.g. concerning body weight, skin and eye colour, they should be as small as possible, show reduced hair coat, a calm temperament and no abnormalities. Glodek & Oldigs (1981) calculated the proportion of the original breeds based on pedigree information. In the white GMP line, the proportion was found to be 60 % VPP, 33 % MMP and 7 % GL. The current GMP is a white dwarf pig breed where all body parts are reduced in size. This type of dwarfism is often caused by growth hormone deficits, mainly of the insulin-like growth factor 1 (*IGF-1*) (Simianer & Köhn 2010).

The entire breeding population is located in three locations spread across the globe. In 1992 an exclusive licence contract was made with Ellegaard ApS in Denmark. Since 2002, the production and marketing for GMP in the USA is managed by Marshall Farms Inc. Since 2011, Ellegaard ApS is in negotiation with a Japanese company about a breeding herd in Japan. Besides the base population in Germany, the two Danish and one American population in full-barrier breeding facilities provide animals with the highest hygienic standard. The University of Göttingen is still in charge of the genetic management of all populations of the GMP.

Genomic regions controlling traits of interest are expected to exhibit footprints of selective breeding. This can give a better insight into the breeding history of the GMP

and will help to identify genomic areas which are functionally and selectively relevant for the GMP.

Searching for selection signatures started in the human genome which is assumed to be homogeneous, i.e. it is not a mixture of different ancestral races. Admixed populations may mask signals of recent selection but these populations can also be used to search for selection signatures (Akey *et al*. 2004; Lohmüller *et al*. 2011). Selection signatures can be classified into hard and soft sweeps (Hermisson & Pennings 2005; Pritchard *et al*. 2010). Hard sweeps reflect the classical model in which a new advantageous mutation arises and quickly expands to fixation. For soft sweeps two different scenarios are possible: (i) several variants with different surrounding haplotypes are selectively favoured due to multiple independent mutations at a single locus, or (ii) an existing allele becomes selectively favourable (e.g. due to changes in environment or a change of breeding goal) so that selection starts from 'standing variation', i.e. the surrounding haplotype is already heterogeneous. In both cases it is to be expected that the resulting statistical signal is heterogeneous and more difficult to detect.

To detect signatures of recent selection different methods were used. The Extended Haplotype Homozygosity (EHH) test is a method to identify genomic regions which have been under recent positive selection. It is defined as 'the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by decent for the entire interval' (Sabeti *et al*. 2002). The so called 'selective sweeps' reflect a fast increase in allele frequency of a core region and the surrounding haplotype (Maynard Smith & Haigh 1974; Nielsen 2005). The identified regions show alleles that have a frequency which has increased faster than it is possible only due to drift and natural selection (Sabeti *et al*. 2002). The 'speed' at which the allele frequency increases is indirectly measured by the length of the surrounding conserved haplotype: if the increase of an originally rare allele is due to genetic drift, the surrounding haplotype is reduced by recombination in each generation. This consequently leads to a low frequency of the haplotype when the frequency of the allele is high. If a high frequency of the allele is reached quickly by directional selection, the number of generations and thus the number of meioses is lower. Because of the lower numbers of possible recombinations, the surrounding haplotype will be longer. To correct the EHH test statistic for the local variability in recombination rates the 'Relative Extended

Haplotype Homozygosity' (REHH) was developed. The REHH test compares the EHH of a tested core haplotype to the EHH of other core haplotypes present at a locus.

Another method used to detect recent positive selection is the Cross Population Extended Haplotype Homozygosity (XPEHH) test. If two populations were separated and selected for different breeding goals, it is possible that the allele frequency, according to the breeding goal, changed only in one population. Large differences in allele frequency for defined genomic regions between populations are assumed to reflect selection (Rothammer 2011). The XPEHH test is based on the method of the EHH test but additionally compares two populations with each other. The XPEHH test will be used to detect genomic regions under selection in which the selected allele may have (almost) achieved fixation in one population but remains polymorphic among in both population together (Sabeti *et al*. 2007).

The third method used to detect selection signatures is the calculation of a Membership coefficient (MC). According to the quantitative genetics theory (Falconer & Mackay 1996), the composition of the genome of the GMP under absence of selection and genetic drift should have been maintained on average. If only short segments are considered, the variability of breed composition of the GMP may be changed relevantly due to genetic drift. The active management of the GMP is avoiding a high inbreeding rate so that genetic drift should not have a large impact of variability of the composition across the genome of the GMP.

Directional selection is expected to be a genetic mechanism having an impact on the composition of the genome of the GMP. In the established synthetic breed a small body size, high fertility, and a white coat colour was the breeding objective. For this it is assumed that alleles being responsible for breed specific characteristics should have been increased in frequency in the GMP. This shift in the allele frequency should not only be observed for a respective candidate gene, but also for the adjacent chromosome region being in linkage disequilibrium with the gene under selection. Hence, in a genomic region carrying a relevant mutation (allele) for a founder breed specific trait we expect the founder breed to be represented at a higher proportion than on average in the genome.

Scope of the Thesis

The aim of this thesis was to study the use of high density markers in the field of genetic characterisation and analysis of animal populations. In a first step, a marker based differentiation of chicken populations with different types of markers was performed. Genotypes of single nucleotide polymorphisms and microsatellites were analysed by two different methods and the required number of SNPs to reach the same differentiation as one microsatellite was determined. Chapter 2 reports the results of this study.

In a second step, a genome wide mapping of selection signatures on the autosomes of the Göttingen Minipig (GMP) was carried out. To search for signatures of recent positive selection, genotypes obtained with the Illumina Porcine BeadChip 60K (Illumina, San Diego, USA) were analysed with three different methods: (i) a Bayesian method to estimate the membership coefficient of the three founder breeds for each SNP was used, (ii) the 'Extended Haplotype Homozygosity' (EHH) was calculated to find signatures of recent positive selection within the GMP; (iii) the 'Cross Population Extended Haplotype Homozygosity' (XPEHH) was used to detect genomic regions of recent selection between the GMP and 'normal sized' pig breeds. These results are reported in chapter 3 and 4 of this thesis, respectively. Chapter 5 presents a general discussion of all results.

References

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. & Kruglyak, L. (2004): Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol., 2, 10, 1591-1599.

Ball, A.D., Stapley, J., Dawson, D.A., Birkhead, T.R., Burke, T. & Slate, J. (2010): A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (Taeniopygia guttata). BMC Geno., 11, 218-232.

Bennett, S.T., Barnes, C., Cox, A., Davies, L. & Brown, C. (2005): Toward the $1000 human genome. Pharmacogenomics, 6, 373-382.

Beuzen, N.D., Stear, M.J. & Chang, K.C. (2000): Molecular markers and their use in animal breeding. Vet. J., 160, 42-52.

Brandström, M. & Ellegren, H. (2008): Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Res., 18, 881-887.

Brandt, H., Möllers, B. & Glodek, P. (1997): Prospects for a genetically very small minipig. Scand. J. Anim. Sci. Suppl., 25, 93-96.

Bode, G., Clausing, P., Gervais, F., Loegsted, J., Luft, J., Nogues, V. Sims, J. & under the auspices of the steering group of the RETHINK Project (2010): The utility of the minipig as an animal model in regulatory toxicology. J. of Pharmacol. A. Toxicol. Meth., 62, 196–220.

Bodzsar, N., Eding, H., Revay, T., Hidas, A. & Weigend, S. (2009): Genetic diversity of Hungarian indigenous chicken breeds based on microsatellite markers. Anim. Genet., 40, 516-523.

Botstein, D. & Risch, N. (2003): Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet., 33, 228-237.

Dunteman, G.H. (1989): Principal Component Analysis. Sage Publications. Newbury Park, CA, USA.

Duran, C., Appleby, N., Edwards, D. & Batley, J. (2009): Molecular genetic markers: discovery, applications, data storage and visualization. Current Bioinform., 4, 16-27.

Falconer, D. S. & Mackay, T.F.C. (1996): Introduction to quantitative genetics. Longmans Green, 4, Harlow, Essex, UK.

Ferguson, P. W., Harvey, W.R. & Irvin, K.M. (1985): Genetic, phenotypic and environmental relationships between sow body weight and sow productivity traits. J. Anim. Sci., 60, 375-384.

Forster, R., Ancian, P., Fredholm, M., Simianer, H., Whitelaw, B. & under the auspices of the steering group of the RETHINK Project (2010): The minipig as a platform for new technologies in toxicology. J. of Pharmacol. A. Toxicol. Meth., 62, 227–235.

Fries, R. & Durstewitz, G. (2001): Digital DNA signatures for Animal Tagging. Nat. Biotech., 19, 508.

Geldermann, H. (1975): Investigations of quantitative characters in animals by gene markers. Theor. Appl. Genet., 46, 319-330.

Glodek, P. & Oldigs, B. (1981): Das Göttinger Miniaturschwein. Schriftenreihe Versuchstierkunde 7, Paul Parey Verlag, Berlin, Germany.

Guttmacher, A.E. & Collins, F.S. (2003): Welcome to the genomic Era. N. Engl. J. Med., 349, 996-998.

Hermisson, J. & Pennings, P.S. (2005): Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics, 169, 2335–2352.

Karlskov-Mortensen, P., Hu, Z.L., Gorodkin, J., Reecy, J.M. & Fredholm, M. (2007): Identification of 10 882 porcine microsatellite sequences and virtual mapping of 4528 of these sequences. Anim. Genet., 38, 401–405.

Katti, M.V., Ranjekar, P.K. & Gupta, V.S. (2001): Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evolut., 18, 1161-1167.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., Hoh, J. (2005): Complement factor H polymorphism in age-related macular degeneration. Science, 308, 385–393.

Köhn, F. (2007): Growth curve and body weight in Göttingen Minipigs - a phenotypic and genetic study. Diss. agr. GAU Göttingen, Germany.

Köhn, F., Sharifi, A.R. & Simianer, H. (2009): Genetic analysis of reactivity to humans in Göttingen Minipigs. Appl. Anim. Beh. Sci., 120, 68–75.

Laloë, D., Jombart, T., Dufour, A.B. & Moazami-Goudarzi, K. (2007): Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. Genet. Selec. Evol., 39, 545–567.

Lande, R. & Thompson, R. (1990): Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics, 124, 743-756.

Lee, C., Abdool, A. & Huang, C.-H. (2009): PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics, 10 Suppl 1, S73.

Liu, N., Chen, L., Wang, S., Oh, C. & Zhao, H. (2005): Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet., 6 Suppl 1, S26.

Lohmüller, K.E., Bustamante, C.D. & Clark, A.G. (2011): Detecting directional selection in the presence of recent admixture in African-Americans. Genetics, 187, 823–835.

Maynard-Smith, J. & Haigh, J. (1974): The hitch-hiking effect from a favourable gene. Genet. Res., 23, 23–35.

McAnulty, P.A. (1999): The value of the minipig in toxicity and other studies supporting the development of new pharmaceuticals. European Pharmaceutical Contractor, 82-86.

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

Morrison, D. F. (1976): Multivariate Statistical Methods. McGraw-Hill, New York, USA.

Nei, M. & Li, W.H. (1979): Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci., 76, 5269-5273.

Nielsen, R. (2005): Molecular Signatures of Natural Selection. Annual Rev. Genet., 39, 197-218.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W. & Drineas, P. (2007): PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet., 3, 1672-1686.

Patterson, N., Price, A.L. & Reich, D. (2006): Population structure and eigenanalysis. PLoS Genet., 2, 12, 2074-2093.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000): Inference of population structure using multicocus genotype data. Genetics, 155, 945-959.

Pritchard, J.K., Pickrell, J.K. & Coop, G. (2010): The genetics of human adaptation: hard sweeps, soft sweeps and polygenic adaptation. Current Biol., 20, 208–215.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. & Feldman, M.W. (2002): Genetic structure of human populations. Science, 298, 2381-2385.

Rothammer, S. (2011): Genomweite Detektion von Selektionssignaturen in divergent selektierten Rinderpopulationen mit anschließender Identifikation eines möglichen kausalen Gens. Diss. med. vet. LMU München, Germany.

Sabeti, P.C., Reich, D.E., Higgins, J.M. Levine, H., Richter, D.J., Schaffner, S., Gabriel, S.B., Platko, J., Patterson, N.J., McDonald, G.J., Ackerman, H., Campbell, S.J., Altshuler, D., Cooperk, R., Kwiatkowski, D., Ward, R. & Lander, E.S. (2002): Detecting recent positive selection in the human genome from haplotype structure. Nature, 419, 832–837.

Sabeti, P.C., Varilly, P., Fry, B., Lohmüller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A. Gaudet, R., Schaffner, S.F., Lander, E.S. & The International HapMap Consortium (2007): Genome wide detection and characterization of positive selection in human populations. Nature, 449, 913-918.

Schaeffer, L. R. (2006): Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet., 123, 218-223.

Schaid, D.J., Guenther, J.C., Christensen, G.B., Hebbring, S., Rosenow, C., Hilker, C.A., McDonnell, S.K., Cunningham, J.M., Slager, S.L., Blute, M.L. & Thibodeau, S.N. (2004): Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer–susceptibility loci. The Am. J. Human Genet., 75, 948–965.

Simianer, H. & Köhn, F. (2010): Genetic management of the Göttingen Minipig population. J. Pharmacol. Toxicol. Methods, 62, 3, 221-226.

Tautz, D. (1989): Hypervariability of simple sequences as a gerneral source of polymorphic DNA markers. Nucl. Acids Res., 17, 6643–6471.

Teneva, A. (2009): Molecular markers in animal genome analysis. Biotech. in Anim. Husb., 25, 1267-1284.

Thoday, J. M. (1961): Location of polygenes. Nature, 191, 368-370.

Toth, G., Gaspari, Z. & Jurka, J. (2000): Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res., 10, 967-981.

Van der Laan, J.W., Brightwell, J., McAnulty, P., Ratky, J., Stark, C. & under the auspices of the steering group of the RETHINK Project (2010): Regulatory acceptability of the minipig in the development of pharmaceuticals, chemicals and other products. J. Pharmacol. Toxicol. Methods, 62, 184–195.

Vignal, A., Milan, D., Sancristobal, M. & Eggen, A. (2002): A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol., 34, 275-305.

Weber, J.L. & May, P.E. (1989): Abundant class of Human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet., 44, 388-396.

Weir, B.S., Anderson, A.D. & Hepler, A.B. (2006): Genetic relatedness analysis: modern data and new challenges. Nature Reviews, Genetics, 7, 771–780.

Wiedmann, R.T., Smith, T.P.L. & Nonneman, D.J. (2008): SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet., 9, 81-87.

Xing, C., Schumacher, F.R., Xing, G., Lu, Q., Wang, T. & Elston, R.C. (2005): Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. BMC Genet., 6 Suppl 1, S29.

2nd CHAPTER

Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations

C. Gärke[*], F. Ytournel[*], B. Bed´hom[†], I. Gut[‡], M. Lathrop[‡], S. Weigend[§] and H. Simianer[*]

[*] Institute of Animal Breeding and Genetics, University of Göttingen, 37075 Göttingen

[†] INRA, AgroParisTech, UMR1313 Animal Genetics and Integrative Biology, Jouy-en-Josas, France.

[‡] Centre National de Génotypage, 91057 Evry, France

[§] Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, 31535 Neustadt, Germany

Published in

Animal Genetics (2011) 85: 84-92

Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations

C. Gärke[*], F. Ytournel[*], B. Bed´hom[†], I. Gut[‡], M. Lathrop[‡],

S. Weigend[§] and H. Simianer[*]

[*] Institute of Animal Breeding and Genetics, University of Göttingen, 37075 Göttingen

[†] INRA, AgroParisTech, UMR1313 Animal Genetics and Integrative Biology, Jouy-en-Josas, France.

[‡] Centre National de Génotypage, 91057 Evry, France.

[§] Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, 31535 Neustadt, Germany

Abstract

Many studies in human genetics compare informativeness of single-nucleotide polymorphisms (SNPs) and microsatellites (single sequence repeats; SSR) in genome scans, but it is difficult to transfer the results directly to livestock because of different population structures. The aim of this study was to determine the number of SNPs needed to obtain the same differentiation power as with a given standard set of microsatellites. Eight chicken breeds were genotyped for 29 SSRs and 9216 SNPs. After filtering, only 2931 SNPs remained. The differentiation power was evaluated using two methods: partitioning of the Euclidean distance matrix based on a principal component analysis (PCA) and a Bayesian model-based clustering approach. Generally, with PCA-based partitioning, 70 SNPs provide a comparable resolution to 29 SSRs. In model-based clustering, the similarity coefficient showed significantly higher values between repeated runs for SNPs compared to SSRs. For the membership coefficients, reflecting the proportion to which a fraction segment of the genome belongs to the *i*th

cluster, the highest values were obtained for 29 SSRs and 100 SNPs respectively. With a low number of loci (29 SSRs or ≤100 SNPs), neither marker types could detect the admixture in the Gödöllo¨ Nhx population. Using more than 250 SNPs allowed a more detailed insight into the genetic architecture. Thus, the admixed population could be detected. It is concluded that breed differentiation studies will substantially gain power even with moderate numbers of SNPs.

Keywords

chicken, microsatellites (SSR), population structure, SNP

Introduction

The main advantages of single nucleotide polymorphisms (SNPs) compared to microsatellites (single sequence repeats, SSRs) are a low mutation rate, a very low false genotyping rate and the abundance in the genome which makes them suitable for automation and standardisation in high throughput technologies (Fries & Durstewitz 2001; Martínez-Arias *et al.* 2001; Xing *et al.* 2005). The high number of SNPs may compensate the fact that they are only biallelic and thus less informative (Schaid *et al.* 2004) than SSRs. The latter are highly polymorphic and thus provide higher information content per locus (often more than 6 alleles as compared with strictly 2 alleles for an SNP) (Bahram & Inoko 2007).

SNP arrays covering up to one million SNPs in humans and many experimental and farm animal species are widely available. Studies in human genetics showed that, due to their reduced informativeness, more SNPs are required to achieve the same information content as obtained with microsatellites (Schaid *et al.* 2004; Xing *et al.* 2005). The number of SNPs needed to replace one SSR varied between 1.7 and 5.56 (Chakraborty *et al.* 1999; Glaubitz *et al.* 2003; Goddard & Wijsman 2002; Krawczak 1999; Kruglyak 1997; Thalamuthu *et al.* 2004). Many studies compared SSRs and SNPs applied in whole genome scans in humans, while this is a relatively novel research area. Because of the differences in structure, size and demography of human and livestock populations (Hayes *et al.* 2003), it is difficult to transfer results pertaining to the phylogenetic analysis regarding the use of different types of markers from human to livestock populations.

In poultry, Schopen *et al.* (2008) showed that the number of SNPs needed to compensate one SSR locus depended on the size of the marker set. The number of SNPs required providing the same information as one SSR increased with an increasing total number of SSRs. For 6 SSRs, about 1.3 SNPs and for 12 SSRs on average 2.3 SNPs per SSR were required to achieve equivalent information content. For Galloway cattle Herráez *et al.* (2005) found that the information content of 2.65 SNPs corresponded to that of one SSR.

Two widely used methods to assess genetic differentiation between populations are Principal Component Analysis (PCA) and model-based clustering, as for example implemented in the software package STRUCTURE (Pritchard *et al.* 2000, version 2.3). A PCA is a nonparametric linear dimension reduction technique (Lee *et al.* 2009). It is the most common data reduction method using allele frequency data to differentiate between populations (Laloë *et al.* 2007; Morrison 1976). In the multivariate setting, principal components (PCs) are linear combinations of the original variables (genetic marker) reflecting patterns of covariation in the data (Kirkpatrick & Meyer 2004). PCA is well suited to uncover the population structure for hundreds of individuals and thousands of loci without any modelling of the dataset. The differentiation power of a PCA was demonstrated by Paschou *et al.* (2007) who showed that the algorithm can be effectively used for the analysis of admixed populations, even without having the information about the origin of individuals.

To compare the differentiation power of microsatellites and SNPs we also used the model-based clustering algorithm implemented in the software STRUCTURE which allows to cluster individuals to K assumed populations. This software has been used in many studies for assessing the genetic structure and relatedness within and among populations (e.g. Rosenberg *et al.* 2002; Liu *et al.* 2005; Twito *et al.* 2007; Bodzsar *et al.* 2009).

The Food and Agriculture Organisation of the United Nations (FAO) has defined standardised species-specific sets of around 30 selected microsatellites for the assessment of genetic diversity between farm animal populations (FAO 2004). With the beginning of high throughput SNP genotyping, the implementation of an SNP-based alternative becomes an issue, both regarding information content and thus phylogenetic resolution, as well as genotyping cost and comparability of genotypes. The present study aimed at assessing the number of SNPs needed to reach the same differentiation

power as 29 SSRs to classify animals into eight chicken populations by using PCA and STRUCTURE.

Material & Methods

Chicken populations and markers:

Sixty-four individuals originating from eight chicken populations were used in this study (Table 1), comprising both commercial and fancy breeds. All animals were genotyped for 29 SSRs and 9,216 SNPs.

All but one (MCW0080) of the SSRs were from the FAO panel recommended for biodiversity studies in chicken (FAO 2004). The microsatellite loci were distributed across 15 chromosomes and between one and five loci were located on a single chromosome. The number of alleles ranged from two to fourteen for the SSR markers.

The SNPs were randomly distributed across the whole chicken genome. Genotyping was done using the Illumina GoldenGate array. Since we used an early SNP array available for chicken, we selected only markers with call rates of 100% for further analysis. During filtering, monomorphic markers and SNPs with unknown positions were deleted. After this, 2,931 SNPs were left to be used in this study. The data used in the paper can be downloaded here: ftp://ftp.tzv.fal.de/download/Chicken_SNPdata.zip.

Table 1: Population information.

Population name	sampling country	Abbreviation	Management *
Padova	Italy	PAD	STAND
Green legged Partidge	Poland	GLP	STAND
Orlov	Russia	ORL	STAND
Gödöllö Nhx	Hungary	GOD	CONSERV
White egg layer A	commercial	WL	SEL
Broiler dam line D	commercial	BDL	SEL
Brown egg layer C	commercial	BL_C	SEL
Brown egg layer D	commercial	BL_D	SEL

* STAND, fancy breeds which were selected for a given standard; SEL, selected for quantitative traits; CONSERV, conservation flocks. Adapted from Granevitze *et al*. (2007).

Statistical analyses

F statistics

To check the relatedness between the breeds we estimated F_{ST} values across the eight breeds for both marker types with the software GENEPOP on the web (Raymond & Rousset 1995, Rousset 2008), using the Weir and Cockerham (1984) approach. Values can range from 0 to 1, with high F_{ST} values indicating a higher degree of differentiation between populations.

Principal components analysis

PCA was used to classify individuals based on a reduced number of significant orthogonal principal components (PC) (Dunteman, 1989). Each PC is related to an *eigenvalue* describing the amount of total inertia covered in the component. That is to say the eigenvalue indicates the part of the total genetic variability that is represented by the associated PCs. The first PCs which are related to a high amount of inertia produce a structuring of the genetic data (Jombart *et al*. 2009). This method is applicable for both kinds of markers: SNPs with two alleles and also SSRs with two or more alleles (Patterson *et al.* 2006). We used the software R (version 2.9.1) and the package adegenet (Jombart *et al.* 2010) to conduct PCAs with different marker subsets: all microsatellite markers, all SNPs and various subsets of SNPs. For the SNP subsets the allele frequencies were scaled to compensate differences among alleles due to their underlying binominal nature (Jombart *et al*. 2009). Microsatellite allele frequencies were not scaled as this was considered to be unnecessary by several authors (e.g. Jombart *et al*. 2009; Patterson *et al*. 2006).

The different subsets for the SNPs were obtained by choosing random samples of 29, 100, 150, 200, 300, 400, 500, 1,000, 1,500, 2,000, and 2,500 of the 2,931 SNPs. Random selections of loci were repeated 100 times for each number of SNPs. To assess whether the results obtained with SSRs were affected by the chromosomal region the SSRs are positioned in, analyses were also carried out with a particular subset of SNPs containing 50 SNPs directly flanking for 25 of our 29 SSRs with known position.

PCA-based partitioning of the distance matrix

The first measure of differentiation reflects the separation of populations relative to the total variability in a space spanned by a defined number of principal components (abbreviated as *nc*). This approach is scale independent and therefore results obtained with SSRs and SNPs are directly comparable. We used the Euclidean distances calculated with the first two principal components (*nc* = 2).

The Euclidian distance between two animals' *j* and *j*' was:

$$d_{j,j'} = \sqrt{\sum_{k=1}^{nc}(x_{j,k}\, x_{j',k})^2}$$

Where $x_{j,k}$ is the value of individual *j* on the *k*-th principal component. Then, the accumulated distance of all animals within a breed *i* was:

$$d_{i.} = \sum_{j=1}^{7}\sum_{j'>j} d_{ij,ij'}$$

The accumulated distance between the animals of two breeds *i* and *i'* was:

$$d_{ii'} = \sum_{j=1}^{8}\sum_{j'=1}^{8} d_{ij,i'j'}$$

Finally, the sum of all distances in the sample can be partitioned in the proportion within breeds and the proportion between breeds, and the relative proportion of the within breed distances can be expressed as:

$$DA = \frac{\sum_{i=1}^{8} d_{i.}}{\sum_{i=1}^{8} d_{i.} + \sum_{i=1}^{7}\sum_{i'>i} d_{ii'}}$$

The parameter DA (differentiation ability) reflects the level of differentiation: the smaller it is the clearer is the differentiation.

Permutation test

In order to investigate the presence of structure in the fixed data sets (all microsatellites, 50 flanking SNPs and all SNPs), we used a permutation test (Mukherjee *et al.* 2003) with 10,000 replicates. In each replicate, we assigned each of the 64 animals randomly to one of the eight populations. To differentiate the populations we used the null hypothesis of no structure between the breeds, i.e. that all animals are sampled from the same population. The alternative hypothesis is that there is structure between the breeds.

We calculated the 10,000 DA values, the corresponding means and variances, and derived the empirical critical values corresponding to a one-sided type I error rate of 5%, 1% and 0.1%, respectively, for each fixed data set.

Curve fitting

To assess the number of SNPs required to the first two PCs of the PCA, we modelled the average DA according to the number of SNPs in each of the subsets and fitted a logarithmic form curve to the data.

Model-based clustering

Population structure was determined using a model-based clustering as implemented in the software package STRUCTURE (Pritchard *et al*. 2000). We applied an admixture model with correlated allele frequencies. The model was used with 20,000 iterations of burn-in and 50,000 iterations of MCMC. In a preliminary test with K = 8, the STRUCTURE algorithm could not differentiate two closely related breeds, BL_C and BL_D, while one cluster remained almost empty (data not shown). Furthermore the lowest F_{ST} values were achieved between the breeds BL_C and BL_D (Table 2). We therefore used K=7 as number of clusters for further analysis. Based on a random selection of loci, we created 20 subsets of each 15, 20, 30, 75, 100, 250, 500, 750 and 1,000 SNPs and 5, 15, 20, 25 SSRs, respectively. For the SSRs we also analysed the complete set of 29 markers. Analysis of each subset was repeated 100 times. Due to the high computing time demand of the algorithm in STRUCTURE, the maximum number of loci was restricted to 1,000 SNPs.

The admixture model produced a membership coefficient vector Q containing 7 values (one for each possible cluster) to denote the admixture proportions for each individual with values ranging between 0 and 1. These values describe the affiliation of an individual to each single cluster (K). The highest value (max Q) for each individual within each replicate was retained. All max Q values were then averaged over all individuals and all replicates (average maximum Q).

We also estimated pairwise similarity coefficients (C) among all 100 repeated solutions within each subset as described by Rosenberg *et al.* (2002). The C value attempts to maximise the measure of similarity between Q-matrices across all replicates over all possible alignments of the replicates. Based on a total of 4950 comparisons we

calculated the average C value for each of the 20 subsets for each marker type and number of loci.

The 100 STRUCTURE solutions of the subsets were averaged for each number of loci using the CLUMPP software (Jakobsson & Rosenberg 2007). CLUMPP permutes the cluster output of independent runs of clustering programs such as STRUCTURE, so that they match as closely as possible. We applied the Large Greedy option for aligning replicates. Mean membership coefficients Q for each subset and population were calculated. The maximum Q-values given for one of the seven clusters were compared to the varying numbers of loci and marker types, respectively. Graphical display of mean membership coefficients (Q) of each population for the seven clusters was performed using the DISTRUCT software (Rosenberg 2004).

Test of significance

In order to test for significant differences between means of the subsets (Q values and C values), a two-tailed non-parametric Mann-Whitney-U test (Mann & Whitney 1947) for two independent samples was applied. For the SNP subset we calculated the mean of all replicates. We used the set of 29 SSRs as reference and compared all sets of SNPs pairwise with this set.

Results & Discussion

In this study we investigated the differentiation power of genetic markers when varying their type (SNPs vs. SSRs) and number (from 29 to 2,961) on the basis of two classical statistical methods (PCA-based partitioning of the distance matrix and model-based clustering).

To get a general overview over the eight populations the F_{ST} value between breeds for both marker types was calculated (Table 2). The lowest F_{ST} (0.06) between the breeds BL_C and BL_D was achieved for the SSR data (mean = 0.28, SD = 0.09). The F_{ST} for the SNP data between these two breeds was 0.16 (mean = 0.32, SD = 0.1). The correlation between the F_{ST} results for both marker types was 0.87.

Table 2: F_{ST} values across the eight breeds based on SSRs (above the diagonal) and SNPs (below the diagonal).

Populations	PAD	GLP	ORL	GOD	WL	BDL	BL_C	BL_D
PAD		0.32	0.28	0.28	0.32	0.22	0.37	0.4
GLP	0.34		0.27	0.24	0.41	0.24	0.38	0.39
ORL	0.32	0.26		0.13	0.28	0.19	0.26	0.24
GOD	0.35	0.25	0.19		0.26	0.16	0.21	0.21
WL	0.43	0.43	0.4	0.43		0.28	0.39	0.56
BDL	0.33	0.23	0.19	0.13	0.41		0.25	0.28
BL_C	0.46	0.37	0.31	0.18	0.53	0.26		0.06
BL_D	0.48	0.36	0.33	0.22	0.42	0.28	0.16	

PCA-based partitioning of the distance matrix

Using PCA for classification purposes is not the first choice from a theoretical point of view. Linear discriminant analyses are more common for genetic data. However, in practice the PCA is a widely used method and in many cases one of the first analyses done with genetic data. Therefore, to implement a quick method for supervised classification purposes, the PCA-based partitioning of the distance matrix was used.

In literature different numbers of principal components were used. We decided to use the first two principal components because they describe almost the same percentage of the total variances, for 29 microsatellites 24.6% (13.8% and 10.8%) and for the complete set of 2,931 SNPs 23.6% (15.4% and 8.2%). We also tried different criteria, i.e. those suggested by Jolliffe (1972) or Kaiser (1960) to define a larger subset of components, but with these approaches, the number of components and the amount of variance explained varied too much between the different replicates to allow a sensible comparison.

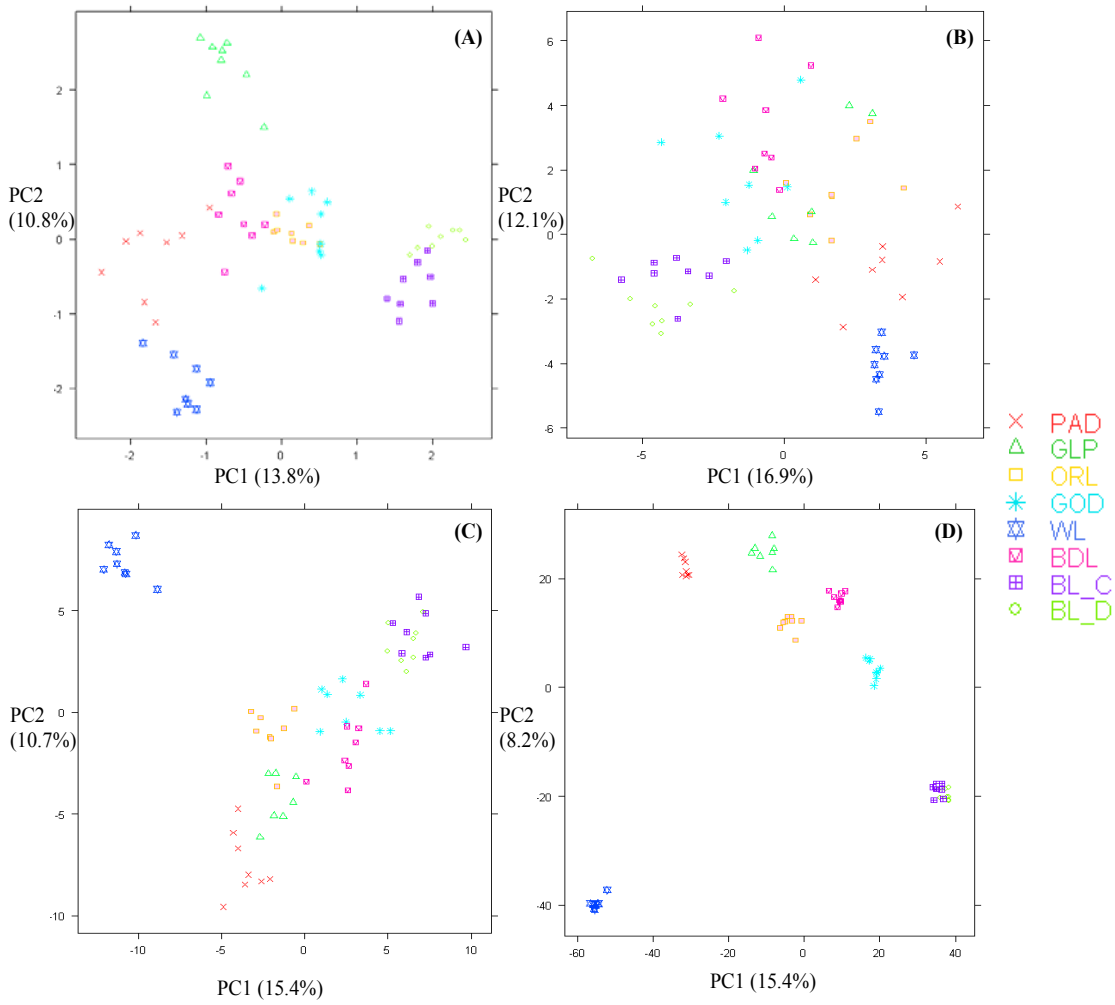

Figure 1: Plot of the two main principal components (PC) and their part of the total Variation in % for all 29 microsatellites (A), 29 random SNPs (B), 100 random SNPs (C) and all 2,931 SNPs (D).

Figure 1 illustrates the results of the first two principal components for the complete sets of SSRs and SNPs (A and D) as well as for one of the replicates with 29 SNPs (B) and one of the replicates with 100 SNPs (C). The results of the PCA partitioning of the distance matrix showed that 29 SSRs provided a more stringent differentiation of the eight breeds than 29 or 50 SNPs (50 SNPs not shown). A comparable result for the SSRs was reached with 100 SNPs (Figure 1 C). With more SNPs in the analyses the resulting DA values decrease (e.g. Figure 1 D). Increasing numbers of SNPs reduced the distances between animals within breeds, while the distances between breeds grew. Using a small set of SSRs, that in many cases were chosen to be informative in a standard panel of breeds, may cause some ascertainment bias (Ellegren *et al.* 1995). This may result in the fact, that breeds not being comprised in the standard panel seem more distant from the standard set than they would be with a non-preselected marker

set. This problem likely is less significant with large sets of SNP markers, although an ascertainment bias may still sustain if SNPs were also detected in standard breed panels.

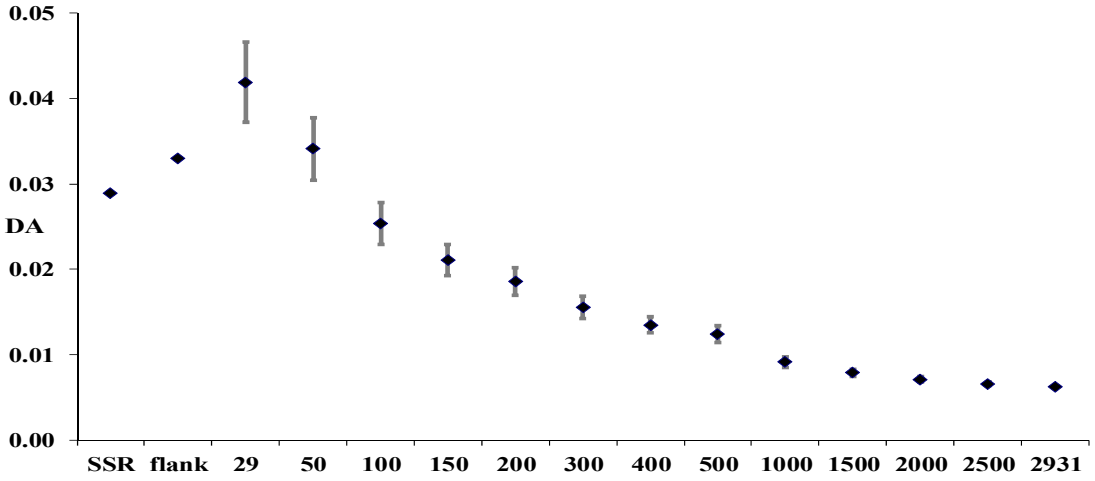

Figure 2: Mean and SD for the 29 SSRs (SSR), the 50 flanking SNPs (flank) and the different SNP subsets of the DA value for the first two principal components.

The results of the PCA partitioning were confirmed by the DA values reflecting the degree of differentiation with a lower value representing a more stringent differentiation. The mean DA values and the standard deviation for the 100 replicates of the different subsets based on the first two principal components are plotted in. The highest (worst differentiation) DA value was achieved with 29 SNPs. The DA values obtained with the subsets containing between 29 and 100 SNPs decreased quickly. The observed DA values for the subsets with ≥ 100 SNP were significantly lower ($p < 10^{-3}$) than those obtained with 29 SSRs. For the same number of loci as SSRs and for the subset with 50 SNPs the DA values were significantly higher ($p < 10^{-3}$). The DA value for the flanking SNPs is in the range of the results obtained with 50 randomly chosen SNPs. Thus a systematic effect of the position of the SNPs cannot be confirmed. One explanation for this might be the distance between the SSRs and the flanking SNPs. In some cases the next SNP was several Mb away from the adjacent SSR. In general: the more SNPs were used, the lower were the achieved DA values and thus the clearer was the observed differentiation. The standard deviation of the DA values decreased with increasing numbers of SNPs.

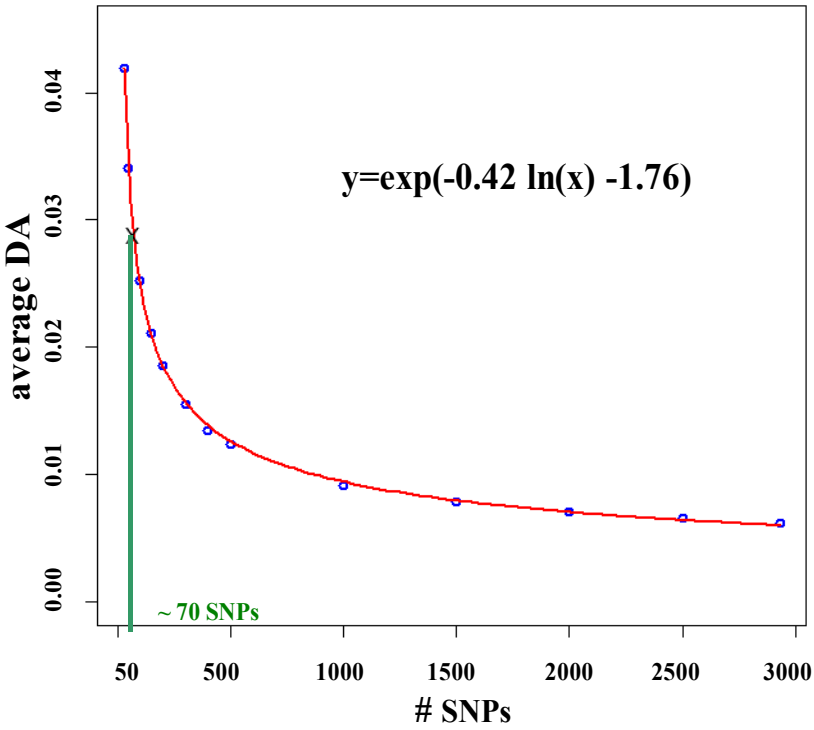

Figure 3: Curve fitting for the empirical data with a regression of the DA values on the number of SNPs.

To calculate the number of needed SNPs based on the differentiation ability, we fitted a function through the empirical results (Figure 3). From the obtained non-linear function, we found that the number of SNPs needed to reach the same differentiation ability as 29 SSR is for this investigation 70. The resulting number of 2.4 SNPs per one SSR is in the same range as reported by Schopen *et al*. (2008). Furthermore, the needed number of SNPs per SSR was found to be in the same range in studies with Galloway cattle (Herráez *et al.* 2005) or Humans (Chakraborty *et al.* 1999; Glaubitz *et al.* 2003; Goddard & Wijsman 2002; Krawczak 1999; Kruglyak 1997; Thalamuthu *et al.* 2004).

Table 3: Observed DA values and mean and empirical critical DA thresholds for three different test levels (one sided test), derived from 10'000 permutations of the three fixed marker sets.

Data set	Observed DA*	p = 0.05	p = 0.01	p = 0.001	Mean DA
29 SSRs	0.0289	0.1056	0.1025	0.0998	0.1112
50 SNPs	0.033	0.1054	0.102	0.0978	0.1111
2,931 SNPs	0.0062	0.1043	0.1006	0.0954	0.1111

*** Differentiation ability (DA) value calculated for two principal components. SSR, single-sequence repeats.**

The results of the permutation tests for the subsets with a fixed composition of markers (2,931 SNPs, 29 microsatellites and the flanking SNPs) are presented in Table 3. For all

three complete subsets, the observed DA values significantly deviated from the expectation under the null hypothesis with a one-sided type I error rate of 1%, showing that in all cases the existing population substructure was detected. This was not completely obvious from the PCA plots, especially with few markers (Figure 1A for instance).

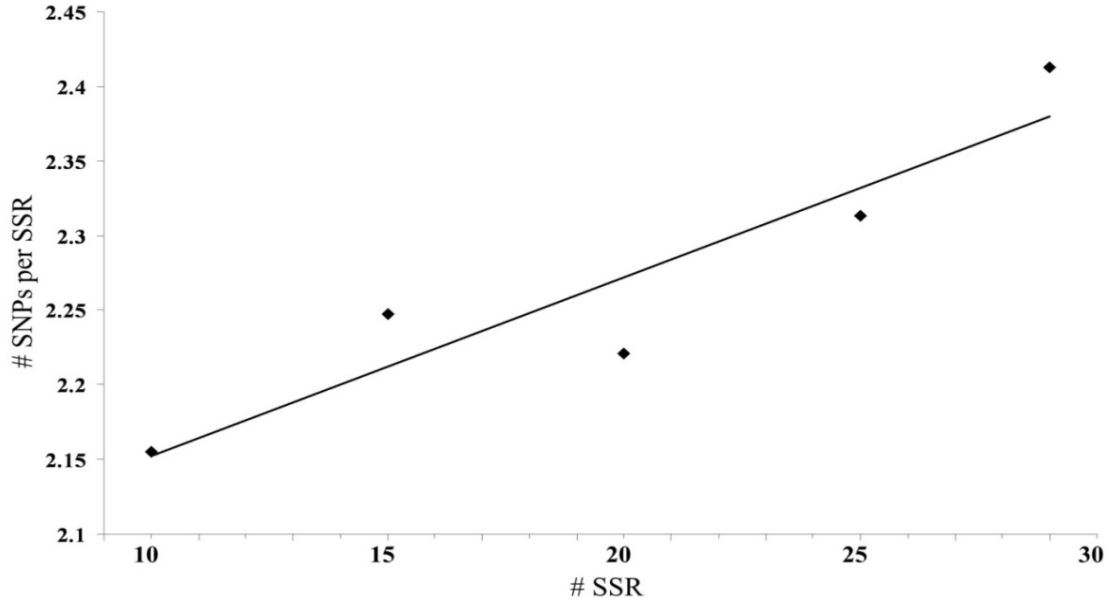

Figure 4: Required number of SNPs per SSR (dots) to obtain the same differentiation ability with different numbers of SSRs (Linear regression: y = 0.012x + 2.033; R2 = 0.863).

Figure 4 displays the linear regression of the number of SNPs needed to reach the same differentiation ability as with one SSR for different sizes of SSR subsets. As already described by Schopen *et al.* (2008) the ratio of SNPs per SSR increases with an increasing number of SSRs in the analysis. In our case this regression is linear indicating that for each added SSR the ratio of SNPs per SSR increases by 0.012.

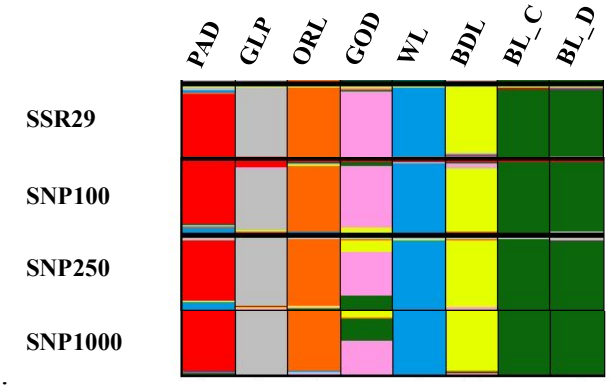

Figure 5: Cluster patterns of eight populations obtained by DISTRUCT for a fixed number of clusters (K=7) and different subsets of markers.

Model-based clustering

Figure 5 shows the DISTRUCT results for 29 SSRs, 100, 250 and 1,000 SNPs. While all breeds appeared to be well separated (with the exception of the two brown egg layer populations) with 29 SSRs, the Gödöllö Nhx (GOD) clearly turned out to be an admixed population with substantial admixture with broiler dam lines (BDL) and brown egg layer populations (BL_C and BL_D).

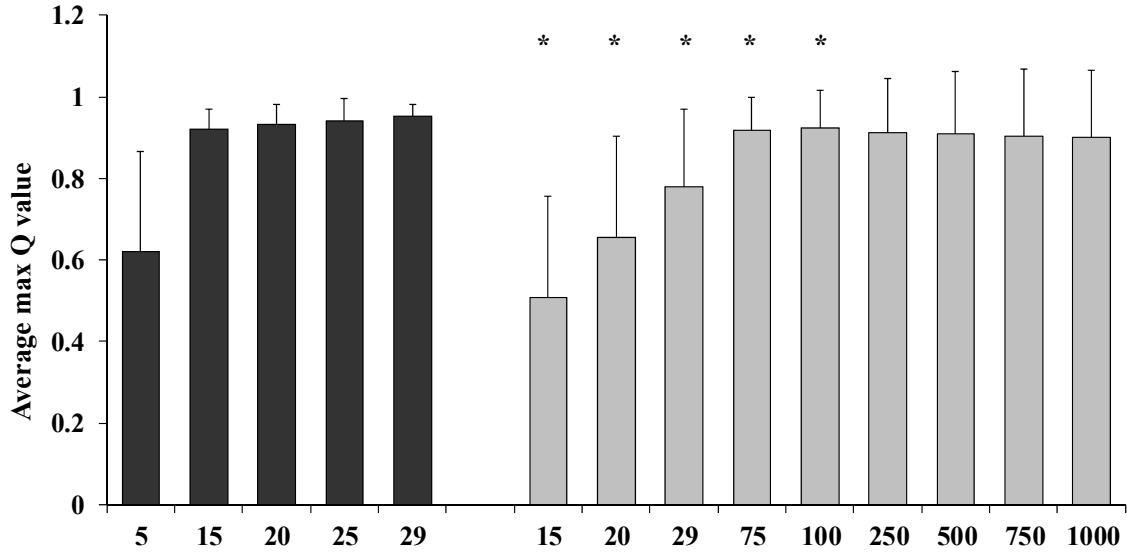

Figure 6: The calculated average for the max Q value and its standard deviations of 100 STRUCTURE runs within different marker subsets for a predefined number of clusters (K=7). On the left are the different SSR subsets and on the right the SNP subsets. The significant differences (* $p < 10^{-3}$) were calculated between 29 SSRs and the SNP subsets.

This admixture was only detectable when at least 250 SNPs were used. Similar results have also been reported by Rosenberg *et al.* (2001). This confirmed by the history of the breeds, as it is a synthetic population resulting from New Hampshire and Rhode Island White breeds, respectively. These two breeds have also been used to create Brown egg layer lines.

The level of differentiation for the different subsets is expressed in STRUCTURE by the membership coefficient Q. Figure 6 shows the results for the max Q values of the different subsets. The highest values were achieved for 29 SSRs and 100 SNPs, respectively. For the SNP subsets with fewer markers (15 to 100 SNPs) we achieved significant differences in comparison to the full set of 29 SSRs ($p < 10^{-3}$). The average max Q value for the SSRs increased with an increasing number of markers. At the same time the standard deviation decreased. For SNPs the average max Q value decreased when using more than 100 SNPs, while the standard deviation increased at the same

time. In Figure 7 the average max Q values are plotted for each population and for both marker types (SSRs and SNPs). For the SSRs, similar results were obtained for all subsets containing at least 15 markers. The highest breed specific max Q values were achieved with the complete set of microsatellites (29 SSRs) indicating that the largest proportion of the genome (above 0.8 for all breeds) originates from one of the seven clusters. A similar pattern was observed with SNPs, where a plateau was reached with around 75 to 100 SNPs for most chicken populations.

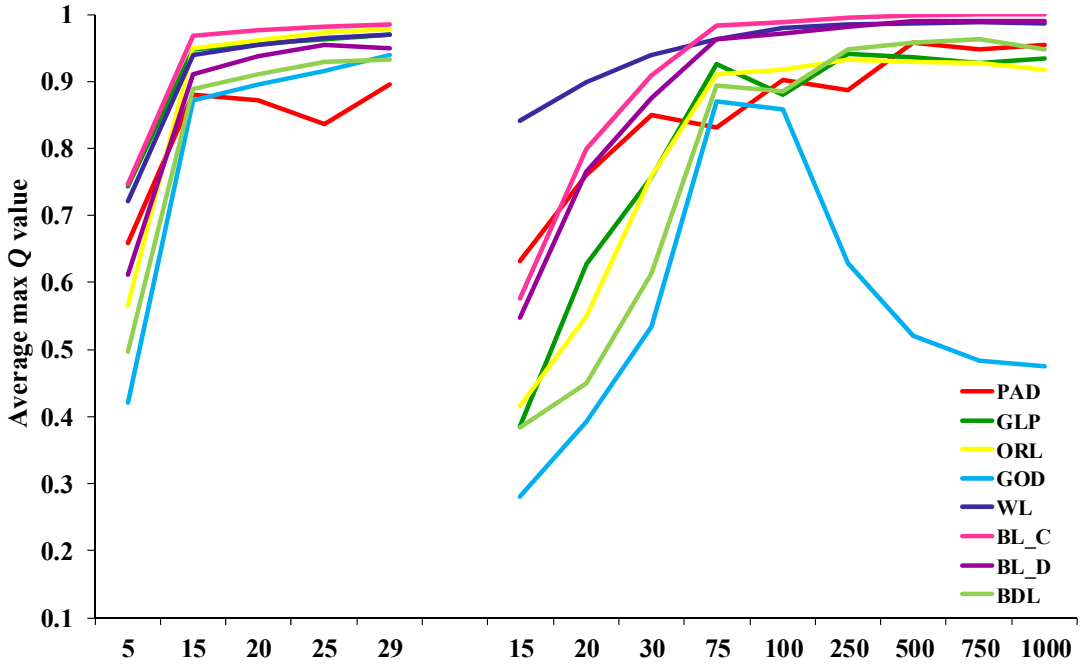

Figure 7: The calculated average for the max Q value for K=7 cluster and eight chicken populations. On the left are the different SSR subsets and on the right the SNP subsets.

This, however, is not true for the breed Gödöllö Nhx (GOD). Here, the average max Q value shows a sharp decrease when more than 100 SNPs are used, and with 1,000 SNPs an average max Q value of around 0.5 was reached. A possible interpretation of this result is that with 29 SSRs or 100 SNPs the power is not sufficient to reveal the admixed status of this breed, and both 29 SSRs and ≤ 100 SNPs provide a comparable insufficient resolution. Only with more SNPs (and probably with more SSRs, too) the method detects the admixture. This also explains the decline of the average max Q value and the increase in the corresponding standard deviation in Figure 6 with subset sizes of more than 100 SNPs, reflecting the mixture of 7 breeds with high average max Q values and the Gödöllö Nhx breed with declining average max Q values. The average max Q values for pure breeds (WL, BL_C, BL_D) increased with the use of more SNPs. For the Gödöllö Nhx the Q value decreased with more SNPs because of the more detailed

insight in their genetic structure, and hence possible admixture. For the other breeds the Q values fluctuated, but they persisted on the same level. The significant differences of averaged max Q values between 29 SSRs and 250 SNPs are due to the detection of the admixed population.

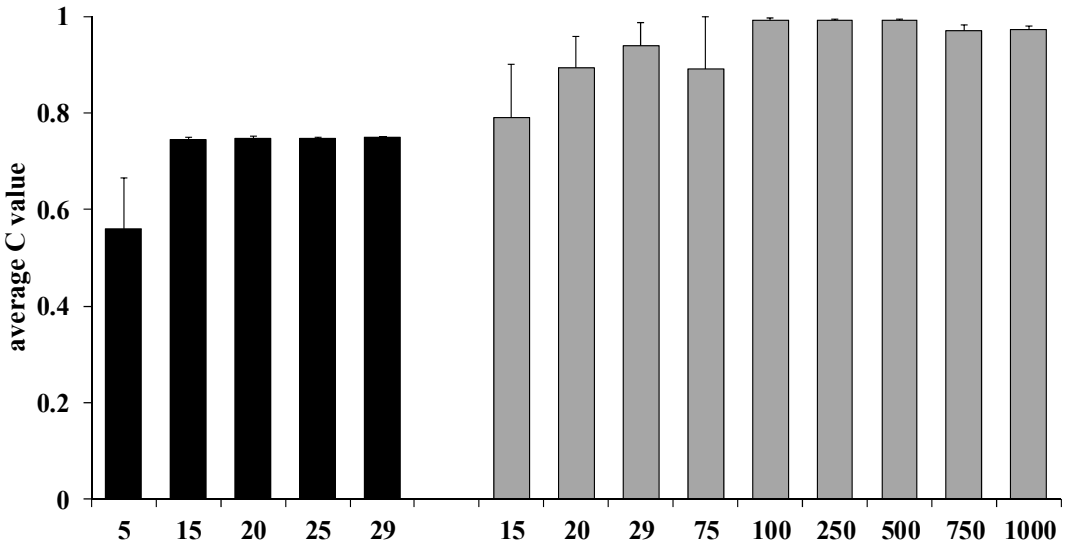

Figure 8: Average similarity values (C) and the standard deviations between 100 STRUCTURE runs for different marker subsets with a predefined number of clusters (K=7). On the left are the different SSR subsets and on the right the SNP subsets. The differences between 29 SSRs and the SNP subsets are significantly different ($p < 10^{-3}$) from the set with 29 SSRs.

The estimated pairwise similarity coefficients (C) for the different scenarios are displayed in Figure 8. For the SSRs the C value was low and highly variable with 5 SSRs only and invariantly showed a constant value with 15 to 29 SSRs. Similarity coefficients obtained with SNPs were higher for all scenarios and reached a plateau with $C > 0.95$ with 100 SNPs or more. According to the Mann-Whitney-U test, all SNP subsets had significantly higher C values ($p < 10^{-3}$) compared to the complete set of 29 microsatellites. Twito *et al.* (2007) found comparable results for neighbour-joining cladograms. With an increasing number of SNPs, they detected a significant increase in the average repeatability. The repeatability of SNPs is significantly higher than with 29 SSRs even with the same number of SNPs (29).

Conclusions

In the current study 70 SNPs are required to reach the same differentiation ability as 29 SSRs for the PCA-based partitioning. Similarity for SNPs is significantly better than for microsatellites between STRUCTURE runs within subsets and at a given level of clustering. To detect admixed populations, a minimum number of SNPs (in this study approximately ≥100-250 SNPs) is needed, with less SNPs or the standard set of 29 SSRs this important result cannot be obtained. The differentiation between breeds improved massively with an increasing number of SNPs when using the PCA. In model-based clustering we achieved a better insight into the architecture of the breeds by increasing the number of SNPs (which arguably might also be obtained with more SSRs).

Papachristou and Lin (2006) stated that SNPs already have become the genetic marker of choice. The results of our study suggest that analyses based on high throughput SNP genotyping will substantially improve the ability to detect and assess the breed differentiation even with a moderate number of SNPs.

Acknowledgements

We are grateful to Michèle Tixier-Boichard and Leif Andersson for providing SNP data. DNA samples were taken from the chicken DNA bank established during the EC AVIANDIV project. We also would like to thank Helmut Lichtenberg for his great support in developing an efficient infrastructure for the computational part of the data analysis, and Denis Laloë for his help on methodological aspects.

References

Bahram, S. & Inoko, H. (2007): Microsatellite markers for genome-wide association studies. Nat. Rev. Genet., 8, doi:10.1038/nrg1962-c1.

Bodzsar, N., Eding, H., Revay, T., Hidas, A. & Weigend, S. (2009): Genetic diversity of Hungarian indigenous chicken breeds based on microsatellite markers. Anim. Genet., 40, 516-523.

Chakraborty, R., Stivers, D.N., Su, B., Zhong, Y., Zhong, Y. & Budowle, B. (1999): The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis, 20, 1682-1696.

Dunteman, G.H. (1989): Principal Component Analysis. Sage Publications. Newbury Park, CA, USA.

Ellegren, H., Primmer, C.R. & Sheldon, B.C. (1995): Microsatellite evolution: Directionality or bias? Nat. Genet, 11, 36–362.

FAO (2004): Secondary Guidelines: Measurement of Domestic Animal Diversity (MoDAD): New Recommended Microsatellite Markers. (Available at: http://dad.fao.org/).

Fries R. & Durstewitz, G. (2001): Digital DNA signatures for animal tagging. Nat. Biotech., 19, 508.

Glaubitz, J.C., Rhodes, E. & Dewoody, J.A. (2003): Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Molec. Ecol., 12, 1039-1047.

Goddard, K.A.B. & Wijsman, E.M. (2002): Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. Genet. Epidemiol., 22, 205-220.

Granevitze, Z., Hillel, J., Chen, G.H., Cuc, N.T.K., Feldman, M., Eding, H. & Weigend, S. (2007): Genetic diversity within chicken populations from different continents and management histories. Anim. Genet., 38, 576-583.

Hayes, B.J., Visscher, P.M., McPartlan, H.C. & Goddard, M.E. (2003): Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res., 13, 635–643.

Herráez, D.L., Schäfer, H., Mosner, J., Fries, H.R. & Wink, M. (2005): Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. Z. Naturfo., 60c, 637–643.

Jakobsson, M. & Rosenberg, N.A. (2007): CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics, 23, 1801-1806.

Jolliffe, I.T. (1972): Discarding variables in a principal component analysis, I: Artificial data. Applied Stat., 21, 160-173.

Jombart, T., Pontier, D. & Dufour, A.B. (2009): Genetic markers in the playground of multivariate analysis. Heredity, 102, 330-341.

Jombart, T., Devillard, S. & Balloux, F. (2010): Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet., 11, 94.

Kaiser, H. F. (1960): The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Kirkpatrick, M. & Meyer, K. (2004): Direct estimation of genetic principal components: simplified analysis of complex phenotypes. Genetics., 168, 2295–2306.

Krawczak, M. (1999): Informatively assessment for biallelic single nucleotide polymorphisms. Electroph., 20, 1676-1681.

Kruglyak, L. (1997): The use of a genetic map of biallelic markers in linkage studies. Nat. Genet., 17, 21-24.

Laloë, D., Jombart, T., Dufour, A.B. & Moazami-Goudarzi, K. (2007): Consensus genetic structuring and typological value of markers using multiple co-inertia analysis. Genet. Sel. Evol., 39, 545–567.

Lee, C., Abdool, A. & Huang, C.-H. (2009): PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics, 10 Suppl 1, S73.

Liu, N., Chen, L., Wang, S., Oh, C. & Zhao, H. (2005): Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet., 6 Suppl 1, S26.

Mann, H. B., & Whitney, D. R. (1947): On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50–60.

Martínez-Arias R, Calafell, F., Mateu, E., Comas, D., Andrés, A. & Bertranpetit, J. (2001): Sequence variability of a human pseudogene. Genome Res., 11, 1071-1085.

Morrison, D. F. (1976): Multivariate Statistical Methods. McGraw-Hill, New York. USA

Mukherjee, S., Golland, P.& Panchenko, D. (2003): Permutation tests for classification. AI Memo 2003-019. Vol. Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory.

Papachristou, C. & Lin, S. (2006): Microsatellites versus Single-Nucleotide Polymorphisms in confidence interval estimation of disease loci. Genet. Epidemiol., 30, 3–17.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W. & Drineas, P. (2007): PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genetics, 3, 1672-1686.

Patterson, N., Price, A.L. & Reich, D. (2006): Population structure and eigenanalysis. PLoS Genet., 2, 12, 2074-2093.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000): Inference of population structure using multicocus genotype data. Genetics, 155, 945-959.

Raymond, M. & Rousset, F. (1995): GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered., 86, 248-249.

Rousset, F. (2008): Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. Mol. Ecol. Resources, 8, 103-106.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A.M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K. & Weigend, S. (2001): Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics, 159, 699–713.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. & Feldman, M.W. (2002): Genetic structure of human populations. Science, 298, 2381-2385.

Rosenberg, N.A. (2004): Distruct: a program for the graphical display of population structure. Mol. Ecol. Notes, 4, 137-8.

Schaid, D.J., Guenther, J.C., Christensen, G.B., Hebbring, S., Rosenow, C., Hilker, C.A., McDonnell, S.K., Cunningham, J.M., Slager, S.L., Blute, M.L. & Thibodeau, S.N. (2004): Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer–susceptibility loci. Am. J. Hum. Genet., 75, 948–965.

Schopen, G.C.B., Bovenhuis, H., Visker, M.H.P.W. & van Arendonk, J.A.M. (2008): Comparison of information content for microsatellites and SNPs in poultry and cattle. Anim. Genet., 39, 451-453.

Thalamuthu, A., Mukhopadhyay, I., Ray, A. & Weeks, D.E. (2004): A comparison between microsatellite and single-nucleotide polymorphism markers with respect to two measures of information content. BMC Genet., 6 Suppl 1, S27.

Twito, T., Weigend, S., Blum, S., Granevitze, Z., Feldman, M.W., Perl-Treves, R., Lavi, U. & Hillel, J. (2007): Biodiversity of 20 chicken breeds assessed by SNPs located in gene regions. Cytogenet. Genomoe Res., 117, 319-326.

Weir, B.S. & Cockerham, C.C. (1984):. Estimating F-statistics for the analysis of population structure. Evolution, 38, 1358-1370.

Yeung, K.Y. & Ruzzo, W.L. (2001): Principal component analysis for clustering gene expression data. Bioinformatics 17, 9, 763-774.

Xing, C., Schumacher, F.R., Xing, G., Lu, Q., Wang, T. & Elston, R.C. (2005): Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. BMC Genet., 6 Suppl 1, 29-33.

3rd CHAPTER

Footprints of recent selection and variability in breed composition in the Göttingen Minipig genome

C. Gärke[*], F. Ytournel[*], A. R. Sharifi[*], E. C. G. Pimentel[§], A. Ludwig[†] and H. Simianer[*]

[*] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

[§] Department of Animal Breeding, University of Kassel, 37213 Witzenhausen, Germany

[†] Leibniz Institute for Zoo and Wildlife Research, 10252 Berlin, Germany

Submitted in

Animal Genetics (2012)

Manuscript ID: AnGen-12-02-0033

Footprints of recent selection and variability in breed composition in the Göttingen Minipig genome

C. Gärke*, F. Ytournel*, A. R. Sharifi*, E. C. G. Pimentel§, A. Ludwig† and H. Simianer*

* Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

§ Department of Animal Breeding, University of Kassel, 37213 Witzenhausen, Germany

† Leibniz Institute for Zoo and Wildlife Research, 10252 Berlin, Germany

Abstract

The Göttingen Minipig (GMP) developed at the University of Göttingen is a synthetic breed which is widely used in medical research and toxicology. It combines the high fertility of the Vietnamese potbellied pig, the low body weight of the Minnesota Minipig and the white coat colour of the German Landrace. The aim of this study was to find genomic regions that may have undergone selection since the creation of the breed in the 1960s. Therefore the whole genome was screened for footprints of recent selection based on single nucleotide polymorphism (SNP) genotypes from the Illumina porcine 60k SNP chip with two methods: the extended haplotype homozygosity (EHH) test and the estimation of the genomic proportion of the three original breeds at each SNP using a Bayesian approach. Local deviations from the average genome-wide breed composition are tested with a permutation-based empirical test. Results for a comprehensive whole genome scan for both methods are presented. Several regions showing the highest p-values in the EHH test are related to breeding goals relevant in the Göttingen Minipig, such as growth (SOCS2, TXN, DDR2 and GRB10 gene) and white colour (PRLR gene). Additionally the calculated proportion of the founder breeds diverged in many regions significantly from the pedigree-based expectations and the genome average. The results provide a genome wide map of selection signatures in the

GMPs, which leads to a better understanding of selection that took place over the last decades in the GMP breed development.

Keywords

dwarfism, extended haplotype homozygosity, selection signature, SNP

Introduction

The Göttingen Minipig (GMP) is a synthetic breed of laboratory animals developed in the 1960's at the Georg-August-University Göttingen, Germany (Simianer & Köhn 2010). The GMP is an excellent model for medical research and toxicology, because it has many physiologic, anatomical, and metabolic similarities to humans and thus is a widely used non-rodent animal model in pharmaceutical safety testing (Brandt *et al*. 1997; Forster *et al*. 2010).

The first generations on the way to the GMP were obtained by crossing Minnesota Minipigs (MMP), Vietnamese Potbelly Pig (VPP), which led to a small and coloured (black or spotted) pig breed. Because pigs with a white skin are more desirable for animal experiments in dermatology, the German Landrace (GL) was introduced by artificial insemination (between 1965 and 1969), so that along with the coloured lines a distinct white line of the GMP (Glodek & Oldigs 1981) was established. Due to the strong market demand for white GMPs, the production of coloured lines was stopped in 1992. Since then, only white GMPs were produced with a strongly expanding production. Thus, the GMP is a closed breed since the beginning of the 1970s, making the GMP a relatively young breed. The present Göttingen Minipig is a white, dwarf animal where all body parts are reduced in size. This type of dwarfism is often caused by growth hormone deficit, especially of the insulin-like growth factor 1 (*IGF-1*) (Simianer & Köhn 2010).

During the last 30 years of breeding GMPs, the breeding goals were aligned to the situations and the market demands (e.g. small body size, sufficient fertility, moderate inbreeding coefficient, unpigmented skin, modest temperament). At the beginning of the development of the GMP, the main focus was to achieve the desired breed composition while avoiding too much inbreeding by a high exchange of breeding animals. After establishing a moderate inbreeding coefficient, the GMPs were selected for low body

weight first on the basis of their 154-day weight, later on weaning weight (Köhn *et al.* 2007, Simianer & Köhn 2010). The initial selection on low body weight alone resulted in a correlated selection response of reduced litter size (Simianer & Köhn 2010), due to the numerically positive genetic and phenotypic correlation between litter size and body weight in multiparous species. Since the mid 1970's the breeding goal is an index combining low body weight and increased litter size. Other more recent breeding objectives are for instance a moderate temperament, especially in the interaction with humans (Köhn *et al*. 2009).

Based on pedigree information, Glodek & Oldigs (1981) calculated the proportion of the three original breeds in the white Göttingen Minipig line to be 60% Vietnamese potbelly pigs, 33% Minnesota Minipigs and 7% German Landrace. According to quantitative genetics theory (Falconer & Mackay 1996), this composition of the genome on average should be stable under absence of selection and genetic drift. Average effective population size of the GMP population between 1975 and 2007 calculated with the software poprep (Groeneveld *et al.* 2009) was 63, which suggests that some deviations from the pedigree-based expected breed composition may have been caused by drift. Since genetic drift is a random and non-directional process, realised breed composition may differ from the expectation for small segments or single chromosomes, but these deviations should largely cancel out when averaged across the whole genome.

The second genetic mechanism expected to have an impact on the composition of the genome is selection. After formation of the synthetic breed, artificial selection favoured a small body size (mostly from the MMP and partly from the VPP), a high fertility (from the VPP), and a white skin (from the GL). It is thus expected, that alleles being responsible for these founder breed-specific characteristics should have been increased in their frequency in the GMP population. This shift in allelic frequency should not only be observed for a respective candidate gene, but also for the adjacent chromosome region being in linkage disequilibrium with the gene under selection. Hence, in a genomic region carrying, say, a relevant mutation (allele) for the trait 'white skin' we expect the GL being represented with a higher proportion than on average in the genome.

Another concept to identify genomic regions which have been under recent positive selection is the detection of 'selective sweeps' reflecting a fast increase in allele

frequency of a core region and a surrounding long conserved haplotype (Maynard Smith & Haigh 1974; Nielsen 2005). Sabeti *et al*. (2002) have suggested the extended haplotype homozygosity (EHH) statistic for this purpose. The EHH parameter is intended to identify regions that have allelic frequencies that increased faster than it is possible only due to drift and natural selection. In this concept, the 'speed' of the increase in allele frequency is indirectly measured by the length of the surrounding conserved haplotype: if an originally rare allele increases slowly due to drift alone, the originally surrounding haplotype is shortened in each generation by recombination and thus is expected to be small when a high allele frequency is reached. If, however, allele frequency is increased quickly by directional selection, it takes fewer generations and thus fewer meioses with possible recombination to reach a high frequency and thus the surrounding conserved haplotype will be longer.

The 'Relative Extended Haplotype Homozygosity' (REHH) statistic corrects EHH for the local variability in recombination rates. Detecting signals of recent selection with this approach was first performed for the human genome (Sabeti *et al*. 2002) based in ethnically defined samples, thus reflecting a genetically homogeneous population.

Admixed populations can also be used to search for signals of recent positive selection (Akey *et al.* 2004; Lohmüller *et al.* 2011). However, admixture may mask signals (Akey *et al.* 2004) so that less significant results might be found (Parra *et al*. 1998). Hermisson & Pennings (2005) and Pritchard *et al.* (2010) classify the selection signatures into hard sweeps, which is the classical model in which a new advantageous mutation arises and quickly expands to fixation, and soft sweeps. For the latter two different scenarios are possible: (i) an already existing allele becomes selectively favourable (due to changes in environment or, in an animal breeding context, a change of breeding goal) so that selection starts from 'standing variation', i.e. the surrounding haplotype is already heterogeneous; (ii) due to multiple independent mutations at a single locus several variants with different surrounding haplotypes are selectively favoured. In both cases it is expected that the resulting statistical signal is heterogeneous and more difficult to detect.

The GMP is a synthetic breed originating from three phylogenetic distantly related breeds (Thuy *et al*. 2006). During the cross-breeding of the founder breeds it is possible that the same mutation was transferred from different breeds with different surrounding haplotypes. The increase of the allele frequency for the mutation and the increase of

different haplotypes are comparable with 'standing variation' in the soft sweep definition. If different mutations in the different founder breeds appear, the increase of the frequency for one of the mutations is comparable to the second soft sweep scenario of Hermisson & Pennings (2005) and Pritchard *et al*. (2010).

The aim of the study was to identify genomic regions that may have undergone selection since the creation of the GMP breed by combining two approaches: the calculation of the allele-based proportion of the founder breeds of the GMP and the EHH statistic to detect selection signatures by the approach of Sabeti *et al.* (2002). We will argue that, given the heterogeneous genetic makeup of the GMP, regions showing both extreme (R)EHH values and a shifted estimated breed composition are the most likely candidate regions to reflect intensive selection since the establishment of the breed. Identifying regions of the genome that have been under recent selection will provide important insights into the breeding history of the Göttingen Minipig and will help identifying genomic areas which are functionally and selectively relevant for the important trait complexes.

Material & Methods

DNA samples and data preparation

Blood or tissue samples were obtained from 195 individuals originating from the following four populations:

- 159 Göttingen Minipigs (GMP) from three origins: the university owned stock (Versuchsgut Relliehausen, Germany), the population of Ellegaard Göttingen Minipigs ApS (Denmark) and the population of Marshall BioResources (USA).
- 18 Minnesota Minipigs (MMP): Sinclair Research Center (Columbia), USA
- 14 German Landrace (GL): University owned stock (Versuchsgut Relliehausen, Germany)
- 4 Vietnamese Potbelly Pigs (VPP): Tierpark Berlin-Friedrichsfelde, Germany.

Genotyping was carried out using the Illumina Porcine SNP60 BeadChip containing a total of 62’163 Single Nucleotide Polymorphisms (SNP). Four individuals were excluded from analyses (2 GMP and 2 GL), because of low call-rates (< 97%). SNPs with unknown chromosome or position, call-rates < 95% or monomorphic were deleted

from the dataset. A total of 191 animals and 50'279 markers passed the filtering and, excluding chromosome X, 49'077 SNPs were used in the further analyses.

Reconstruction of haplotypes and Linkage Disequilibrium analysis

For the analyses in this study, fully phased haplotype data were required. After the filtering process described above the haplotypes for every chromosome were reconstructed using fastPHASE (Scheet & Stephens 2006) applying the standard parameter settings. Linkage disequilibrium (LD) was estimated from these reconstructed haplotypes using the parameter r^2 (Hill & Robertson 1968). The LD was calculated only for the polymorphic markers of the GMP (31'536 SNPs and 157 individuals).

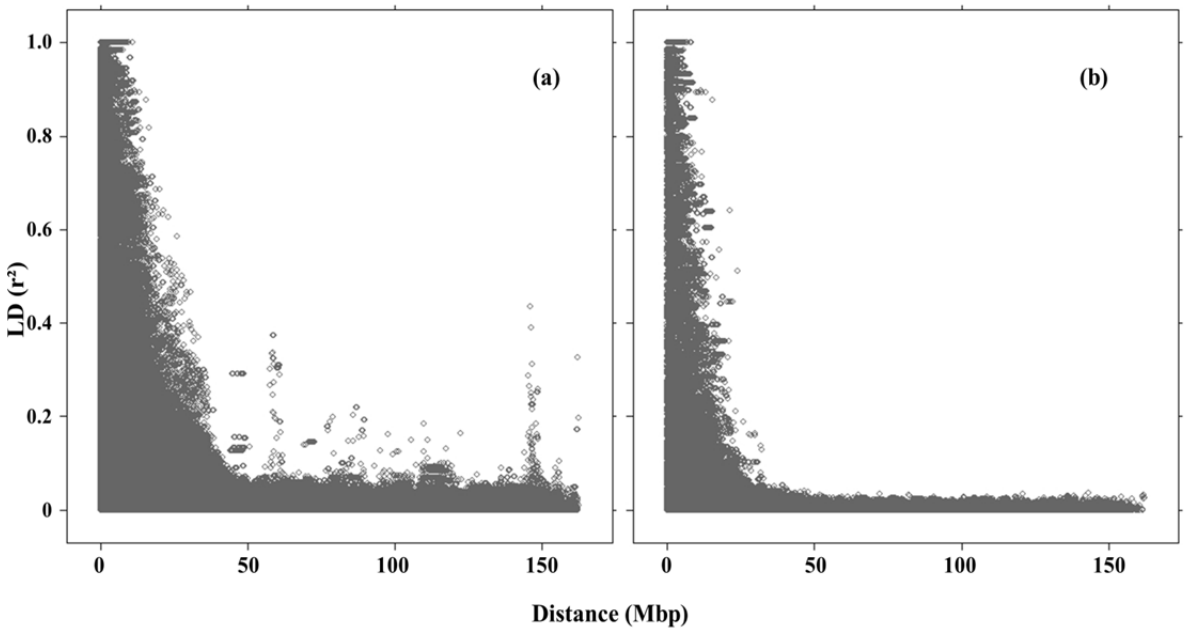

Figure 1: Plot of r^2 against the physical distance between (a) 3'157 SNPs before cleaning the data and (b) 2'979 SNPs after cleaning the data (exemplarily for chromosome 2).

To check for correct SNP positions and to identify LD outliers, the r^2-values for the GMP were plotted against the physical distance for each chromosome (Figure, exemplarily for chromosome 2). It was evident that in the area where the r^2-values reach an asymptotic value some massive outliers are present, which most likely can be attributed to incorrect mapping positions. To remove these outliers, the mean and standard deviation of the r^2-values in the asymptotic region were calculated and outliers were determined as being the points with an r^2-value exceeding the mean plus ten standard deviations. This very high threshold was chosen to remove technical artefacts,

but to keep values which may reflect biological variation. If a SNP was involved in two or more such extreme events, the SNP was removed from the dataset. Figureb shows the adjusted LD distribution for chromosome 2 after filtering. Applying this procedure, a total of 3'300 SNPs were removed, so that 28'236 SNPs segregating in the GMP breed and 45'777 SNPs for all breeds were used for further analyses.

Estimation of the breed composition

The probability of membership (membership coefficient) for every single allele of the GMPs to one of the three founder breeds (VPP, MMP and GL) was calculated. Denote A_{zj} the allele j ($j \in \{1,2\}$) of interest of individual z at a given SNP and x the founder population (x = 1, 2, 3). According to the Bayes theorem the membership coefficient for this allele, i.e. the conditional probability that the observed allele originates from population x given the allelic state, is:

$$P(z \in x \mid A_{zj}) = \frac{P(A_{zj} | z \in x) P(x)}{\sum_{y=1}^{n_f} P(A_{zj} | z \in y) P(y)}$$

where n_f is the number of founder breeds, in this study 3.

The *a priori* probability *P(y)* for all three founder populations (VPP, MMP and GL) was assumed to be 60%, 33% and 7%, based on pedigree calculation of Glodek & Oldigs (1981).

The membership coefficients were averaged across the genome, across each chromosome, and in core haplotypes or in regions containing a candidate gene. To ensure a normal distribution of the proportions, an arcsine-transformation for each chromosome was performed. A two-sided t-test (Sokal & Rohlf 1981) was applied to the transformed data to test for significant deviations of the observed average proportions of each breed for each chromosome from the pedigree-based priors of Glodek & Oldigs (1981).

To remove the huge SNP to SNP variability we averaged membership coefficients in sliding windows of eight subsequent SNPs. With an average marker distance of 65kb, a window of eight SNPs represents about 500 kb. In order to investigate the variability of the average membership coefficient and to identify regions with an abnormal representation of one or more founder breeds compared to the genome average, the empirical 95% confidence interval of the proportion of each founder breed was

calculated using a permutation test (Mukherjee *et al.* 2003) with 1'000 replicates. In each replicate, we shuffled the physical positions of all SNPs randomly, and then we calculated sliding windows consisting of eight subsequent SNPs. With the lowest and the highest window average of each replicate we calculated the of 2.5% and 97.5% quantile thresholds delimiting the genome-wide 95% confidence interval.

χ^2-test for the deviation of the observed from the expected breed composition

In addition to the membership coefficients, reflecting the representation of each founder breed, a general test was conducted whether the observed breed composition at a locus deviates from the expected composition. For this the following test statistic was calculated at each position:

$$X^2 = \sum_{x=1}^{3} \frac{[(Obs_x * na) - (Exp_x * na)]^2}{(Exp_x * na)}$$

where x is one of the three founder population x (x = 1, 2, 3), Obs_x the observed and Exp_x the expected frequency of the membership coefficient of the founder population and na is the number of the alleles {1,2} observed, which in this case is na = 314 alleles (157 GMP individuals). The test statistic was compared with the tabulated χ^2-values with 2 degrees of freedom.

Application of EHH and REHH

To identify core regions characterized by a strong LD among SNPs we applied the algorithm suggested by Gabriel *et al.* (2002) as implemented in SWEEP v.1.1 (Sabeti *et al*. 2002). This algorithm defines a pair of SNPs to be in strong LD, if the upper 95% confidence bound of the LD bound of D' is between 0.7 and 0.98. At least 3 SNPs were necessary to define a core region in this study.

The EHH statistic (Sabeti *et al*. 2002) was used to evaluate the decay of LD around core regions. This test is based on the contrast of one core haplotype showing a combination of high frequency and extended homozygosity with other core haplotypes at the same locus (Qanbari *et al*. 2010). The EHH is specified as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are homozygous for the entire interval from the core region to a given locus.

The EHH of a tested core haplotype *t* is calculated as:

$$EHH_t = \frac{\sum_{i=1}^{s}\binom{e_{ti}}{2}}{\binom{c_t}{2}}$$

where *s* is the number of unique extended haplotypes, e_{ti} is the number of samples of a particular extended haplotype i and c_t is the number of samples of a specific core haplotype *t*.

To correct the EHH for the variability in recombination rates, the 'Relative Extended Haplotype Homozygosity' (REHH) was used (Sabeti *et al*. 2002). REHH is computed as $EHH_t/\overline{EHH}$, where $\overline{EHH}$ is defined as the decay of EHH on all other core haplotypes. For this we used the following equation:

$$\overline{EHH} = \frac{\sum_{j=1,j\neq t}^{n}\left[\sum_{i=1}^{s}\binom{e_i}{2}\right]}{\sum_{i=1,i\neq t}^{ns}\binom{c_i}{2}}$$

where n is the number of different core haplotypes.

To evaluate the significance of the REHH values, the haplotypes were assigned into bins based on their frequency (ranges of 0-5%, 5-10%, etc.). The REHH for each common haplotype in a candidate region was compared to all equally frequent haplotypes. P-values were obtained by log-transforming the REHH-values within the bins to achieve normality, and the mean and the standard deviation (sd) were used to estimate p-values for the REHH values observed. As suggested by Sabeti *et al*. (2002), core haplotypes with extreme REHH in the distribution were considered to indicate a signature of recent selection.

Gene annotation

To identify genes close to the regions of interest, the map viewer option of the porcine genome sequence assembly (*Sus scrofa* 10; available online at http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9823) was used. For this, the regions around the detected signals of selection were expanded 1 Mb up and downstream to find candidate genes. In humans Sabeti *et al.* (2002) assume a distance around the detected signals on each side about 250 Kb. Because of the longer extent of LD in livestock compared to humans (Qanbari *et al*. 2010) we chose the length of 1 Mb.

Results & Discussion

Breed composition

Figure 2 depicts the averaged estimated membership coefficients, i.e. the average probability of an allele to come from one of the three original breeds, for all autosomes (SSC1 – SSC18) and for all autosomes together (all) calculated from genomic data. For the total genome (46'777 SNPs), the contribution of the founder breeds (GL = 0.085; MMP = 0.371; VPP = 0.544) was significantly different from the pedigree-based expectation, with a relative over-representation of the GL (+0.015) and the MMP (+0.041) while the VPP (-0.056) were under-represented.

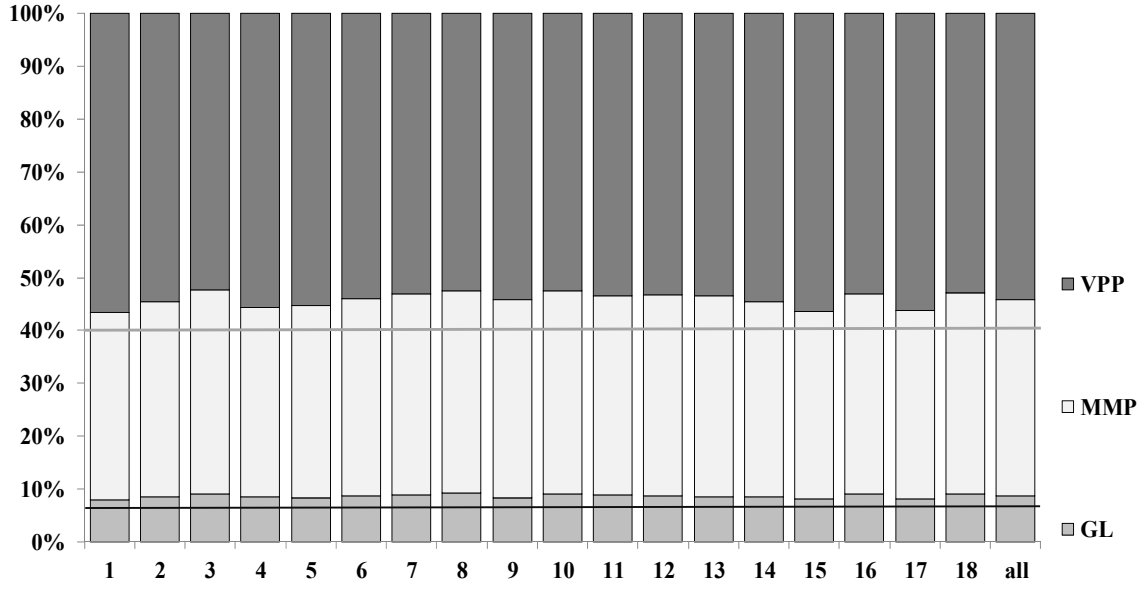

Figure 2: Allelic proportion of the three founder breeds of the GMP for each autosomal chromosome and all autosomes together (all). Horizontal lines represent the pedigree-based expected proportion of the founder breeds (black line = GL, grey line = VPP). All displayed subsets are significantly ($p < 0.05$) different from the pedigree based proportions.

The discrepancy between pedigree-based expectations and marker based estimates of the breed composition may be due to several reasons:

- We used current samples of the three original breeds, while the actual contribution to the GMP was made by breeds in the genetic composition of the 1960s. It must be assumed, that allele frequencies changed substantially over the last 50 years, where this change is mainly due to genetic drift in the small breeds (MMP and VPP), while it likely is due to drift and selection in GL.

- The sample size for the original breeds was small, especially so for the VPP, so that the estimated allele frequencies have a substantial error variance. While this is expected to affect results for single SNPs, it will have little impact on the genome-wide estimates, since even with a small sample the allele frequency estimates are still unbiased.

Composition of single chromosomes was variable (Figure 3), with GL ranging from 0.079 (SSC1) to 0.091 (SSC18), MMP ranging from 0.354 (SSC1) to 0.385 (SSC3), and VPP ranging from 0.523 (SSC3) to 0.565 (SSC1).

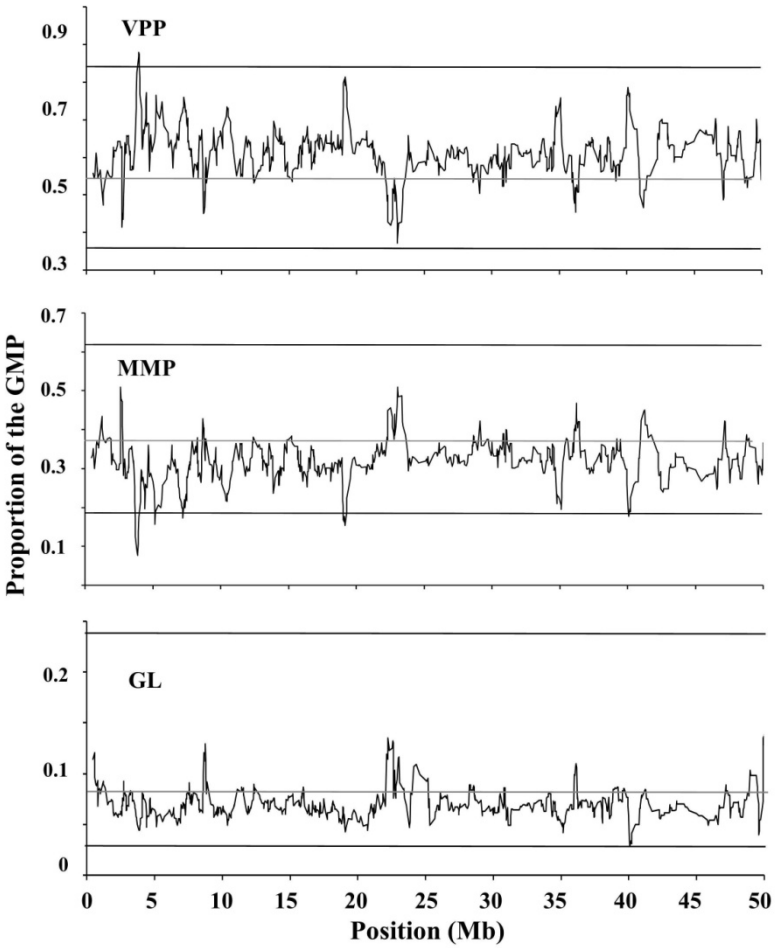

Figure 3: Average membership coefficients for the three breeds for sliding windows of eight adjacent SNPs exemplarily for the first 50 Mb of chromosome 15. The three proportions at each position add up to one. The black horizontal lines are the empirical 2.5% and 97.5% quantile thresholds; the grey line shows the pedigree-based expectation of the founder breed proportion (Glodek & Oldigs 1981).

The probability of membership of the GMP is variable between the chromosomes (Figure 2) and even within each chromosome. To illustrate the fluctuation of proportions within chromosomes, the composition for a region on chromosome 15

(between 0 Mb and 50 Mb) is exemplarily shown as moving average of 8 SNPs (Figure 3). We used a permutation test to calculate the limits of the empirical 95% confidence interval for the breed proportions (black lines). The SNP-based contribution of each founder breed is displayed as a grey line. The proportion of the three founder breeds sum up to one (GL + MMP + VVP = 1). The proportion for single alleles is highly variable, so that it can be assumed that different regions of the genome differ in proportion even within the chromosomes. The fluctuations within each chromosome may be due to several factors: a stochastic error, partly due to the low number of representatives of the original breeds, genetic drift due to the small effective population size of the GMP, but also directional effects due to selection and hitchhiking effects.

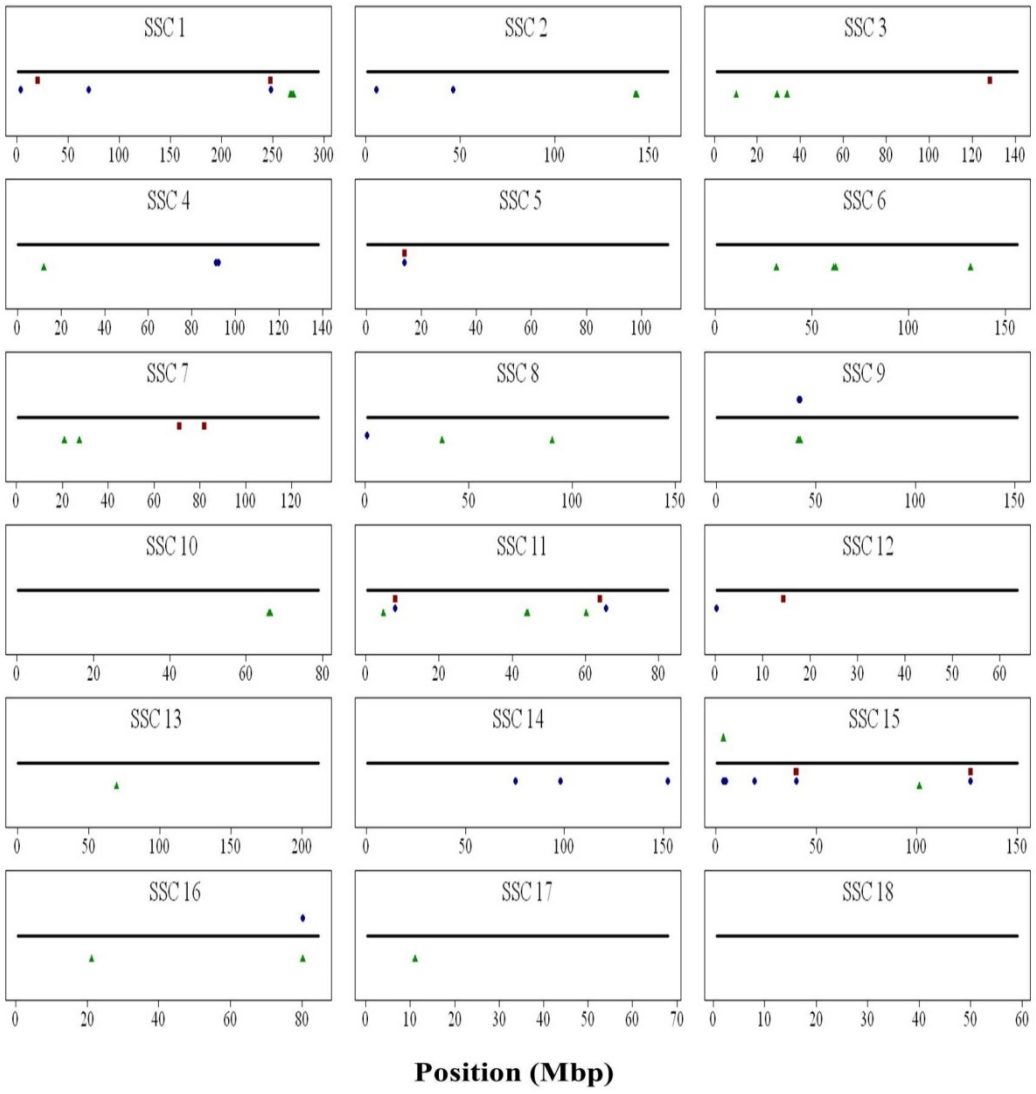

Figure 4: Positions of regions for which the average membership coefficients for sliding windows of eight adjacent SNPs in the lower (symbol below the horizontal line) or in the upper (symbol above the horizontal line) empirical 2.5 % quantile. Green triangles, blue circles, and red squares pertain to the VPP, MMP, and GL, respectively.

We used a permutation test to calculate the limits of the empirical 95% confidence interval for the breed proportions (black lines). The pedigree proportion (Glodek & Oldigs 1981) of each founder breed is displayed as a grey line. The proportion of the three founder breeds sum up to one (GL + MMP + VVP = 1).

The proportion for single alleles is highly variable, so that it can be assumed that different regions of the genome differ in proportion even within the chromosomes. The fluctuations within each chromosome may be due to several factors: a stochastic error, partly due to the low number of representatives of the original breeds, genetic drift due to the small effective population size of the GMP, but also directional effects due to selection and hitchhiking effects.

One aim was to identify regions of systematic deviations in the composition of the GMP compared to the overall mean. To do so, we averaged the membership coefficient of eight adjacent SNPs, respectively, and their positions in non-overlapping sliding windows. Figure 4 shows for each chromosome the position of those sliding windows outside the empirical 95% confidence interval for one or several of the founder breeds (coloured dots). Signals above or below the horizontal line indicate a significant over- or underrepresentation of the respective breed. In total, 60 signals were significant on the genome-wide five per cent level. Most signals were found on chromosomes 1, 11 and 15. Around the detected sliding windows we expanded the region 1 Mb up and downstream to annotate genes, which might be relevant in the formation of the GMP (Table 1).

Table 1: Average membership coefficients for sliding windows of eight adjacent SNPs outlying the 95% confidence interval and annotated genes at the respective position ± 1 Mb. Bold letters for the membership coefficient show the breed outlying the 95% confidence interval.

Chr	Position*	Genes	Membership coefficient		
			GL	MMP	VPP
1	3’871’346	PDE10A	0.056	**0.176**	0.768
1	20’391’092	RAB32,GRM1,SHPRH,FBXO30	**0.030**	0.229	0.741
1	70’071’353	POU3F2,FBXL4,USP45,MCHR2	0.073	**0.189**	0.738
1	248’036’914	RECK,CLTA,GNE,MELK,POLR1E, GRHPR,RG9MTD3,SHB	**0.037**	**0.190**	0.773
1	267’277’840	ZNF618, AMBP,KIFI2	0.119	0.522	**0.359**
1	267’935’513	COL27A1,ORM1	0.099	0.532	**0.369**
1	270’192’459	TNFRSF8,TNC	0.126	0.551	**0.323**
2	5’634’338	CTSF,MUS81,SNX32,DPF2,SLC22A,SYVN1,HD1,MEN1, MAP4K2,PYGM	0.037	**0.192**	0.771
2	46’346’264	PIK3C2A, PLEKHA7,SOX-6,INSC	0.120	**0.181**	0.699
2	142’956’408	PHF15,SAR1B,DDX46,H2AFY,CXCL14	0.126	0.569	**0.305**
2	143’605’553	H2AFY,CXCL14,IL9,TRPM2	0.107	0.521	**0.372**
3	10’164’941	NSUN5,LRWD1,RASA4,ZP3,HIP1,MDH2,HSP27,TRIM50,BAZ1B	0.102	0.547	**0.351**
3	29’209’865	ABCC6	0.132	0.510	**0.358**
3	33’645’239	ERCC4,PRM1,SOCS1	0.135	0.553	**0.312**
3	128’326’039	NBAS	**0.036**	0.225	0.739
4	12’033’793	MYC	0.108	0.519	**0.373**
4	90’950’150	PBX1,RGS5,HSD17B7,**DDR2**,UAP1,ATF6	0.143	**0.167**	0.690
4	92’348’607	FCRLA,FCGR2B	0.049	**0.187**	0.763
5	13’976’894	CRY1,MTERFD3,RIC8B,RFX4	**0.031**	**0.160**	0.808
6	31’591’830	GTP2	0.117	0.514	**0.369**
6	61’209’801	CHD5,ESPN,PLEKHG5,NOL9,ZBTB48,SLC2A7,CA6	0.111	0.523	**0.366**
6	62’369’168	PIK3CD,CLSTN1,RBP7,KIF1B,PGD,PEX14	0.140	0.498	**0.363**
6	131’982’981	PTGER3,CTH,ANKRD13C,SRSF11,LRRC7	0.109	0.567	**0.324**

Chr	Position*	Genes	Membership coefficient		
			GL	MMP	VPP
7	20’855’239	MRS2,SLC34A1,TRIM38	0.101	0.569	**0.330**
7	27’507’394	NRM,TUBB2A,DDR1,GTF2H4,VARS	0.094	0.540	**0.366**
7	71’086’207	NAPS3,AKAp6	**0.037**	0.292	0.671
7	81’950’302	CHD8, KLHL33,PNP,PARP2	**0.035**	0.276	0.689
8	827’027	MFSD7,WHSC2,POLN,TNIP2,ADD1,GRK4, NOP14,ADRA2C	0.052	**0.190**	0.758
8	37’127’435	GUF1	0.121	0.522	**0.358**
8	90’341’849	IL15.TBC1D9	0.112	0.582	**0.306**
9	41’075’675	CASP1.GRIA4,KBTBD3.CWF19L2	0.084	**0.551**	0.364
9	41’552’373	CWF19L2.ACAT1,NPAT	0.088	0.623	**0.289**
9	42’179’815	NPAT,KDELC2.EXPH5	0.085	0.624	**0.291**
10	65’933’765	CUL2	0.139	0.489	**0.372**
10	66’343’844	CUL2	0.146	0.498	**0.357**
11	4’802’967	USP12.LNX2.PDX1.CDX2.FLT1	0.116	0.520	**0.364**
11	8’068’949	HSPH1,FRY	**0.034**	**0.164**	0.801
11	43’915’684	KLHL1	0.114	0.563	**0.323**
11	44’261’222	KLHL1	0.112	0.536	**0.352**
11	60’278’412	SLITRK5,MIR20	0.107	0.522	**0.371**
11	63’983’108	MIR20	**0.034**	0.204	0.762
12	229’794	NARF.WDR45L,FASN	0.050	**0.150**	0.800
12	14’397’875	PRKCA,HELZ,PITPNC1,BPTF	**0.035**	0.236	0.728
13	69’662’006	LMCD1,OXTR	0.136	0.493	**0.371**
14	75’487’154	/	0.084	**0.125**	0.791
14	98’203’141	CHAT,TIMM23,MSMB,MARCH8	0.094	**0.178**	0.728
14	152’469’751	PWWP2B,INPP5A,KNDC1,MIR202.CYP2E1	0.046	**0.184**	0.771

Chr	Position*	Genes	Membership coefficient		
			GL	MMP	VPP
15	3’653’485	EPC2,ACVR2A	0.046	**0.086**	**0.869**
15	4’282’198	/	0.048	**0.180**	0.772
15	4’913’304	/	0.048	**0.187**	0.766
15	19’078’779	RAB3GAP1,MGAT5	0.039	**0.188**	0.772
15	39’990’037	/	**0.028**	**0.185**	0.787
15	101’228’790	ANKAR,SLC40A1.ORMDL1,MSTN,MFSD6,GLS	0.132	0.584	**0.284**
15	126’812’464	PECR,XRCC5,SMARCAL1,IGFBP5	**0.027**	**0.174**	0.798
16	21’195’188	**PRLR**,DNAJC21,RAI14	0.148	0.528	**0.324**
16	80’223’692	MTRR.ADCY2,PAPD7,NSUN2	**0.126**	0.668	**0.206**
17	11’045’115	IDO1,ADAM18	0.113	0.527	**0.360**

* averaged position in bp

Two of the genes found (*DDR2* and *PRLR*) show an obvious relation to the breeding goals of the GMP. The discoidin domain receptor 2 (*DDR2*) on Chromosome 4 is a member of a subfamily of receptor tyrosine kinases. Labrador *et al.* (2001) showed that the absence of the *DDR2* gene in mice leads to a smaller body size. Adult mice were up to 40% reduced in weight compared to wild type mice. Kano *et al*. (2008) indicated that the absence of *DDR2* in mice leads to growth retardation. This gene could possibly be responsible for the small body size in the GMP. The second gene of interest is the prolactin receptor gene (*PRLR*) on chromosome 16. This gene was suggested as a candidate gene influencing the number of piglets born alive and the number of teats (Drögemüller *et al.* 2001; Putnová *et al*. 2002). Interesting for the region harbouring this gene is the decrease of the proportion of VPP in comparison to the SNP-based contribution (-40.44%) and the increase in both other founder breeds (GL: +74.12% and MMP: + 42.32%). Higher proportion of VPP was expected for a region with an effect on piglets born alive. The reason for this decrease of the proportion remains unknown. The other genes found are less obviously linked to the breeding goals of the Göttingen Minipig.

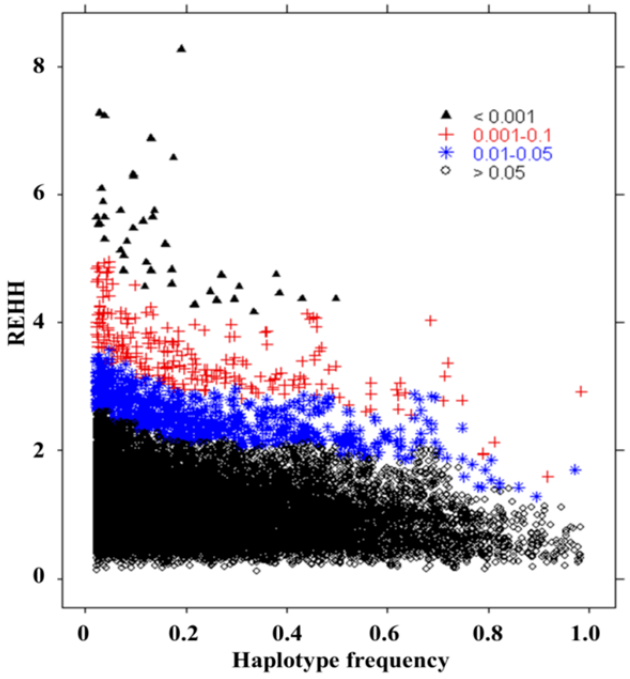

Figure 5: Distribution of the REHH against the haplotype frequencies.

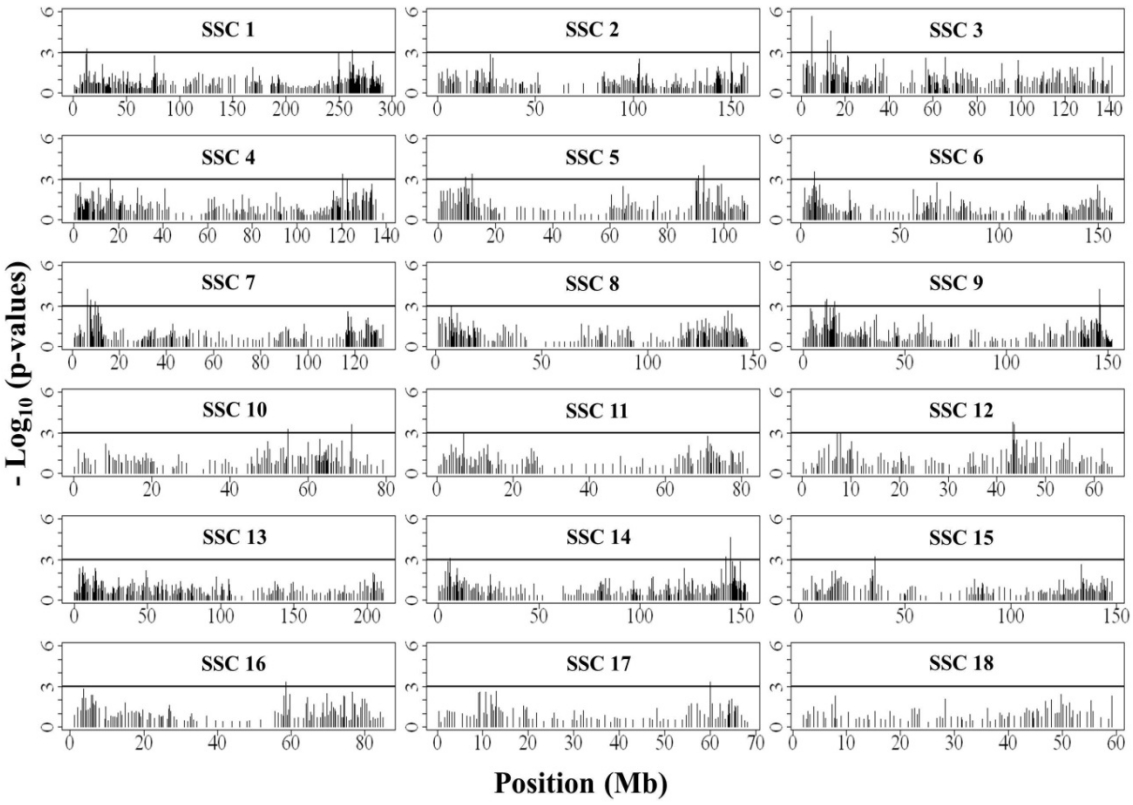

Figure 6: Distribution of the negative logarithm (-log10) of the REHH p-values for core haplotypes on each chromosome against their physical position. The horizontal line lines represent a threshold of 0.001.

Whole genome scans for signatures of recent selection

The REHH test for 1 Mb intervals was calculated on both the upstream and downstream side of each core region to find outlying core haplotypes. The distribution of REHH values vs. haplotype frequencies is presented in Figure 5. The graph shows that the core regions with the lowest p-values ($p < 0.001$) have low to medium haplotype frequency. The negative logarithm ($-\log_{10}$) of the p-values for core haplotypes with a threshold of 0.001 % was plotted against the SNP position on the chromosomes in Figure 6 to visualize the distribution of selection signatures in the genome. The highest signals were found on chromosomes 3, 5, 7, 9 and 14, while on these and several other chromosomes clusters of signals were observed.

Table 2: Summary of the core haplotypes showing the lowest p-values (p < 0.001) after REHH test and the mean membership coefficient of the involved SNPs for the three founder breeds. All displayed membership coefficients are significantly different from the mean autosomal proportion (t-test; p < 0.05).

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
1	12’857’284	13’221'128	5.2E-04	1.9E-04	CLDN20,TFB1M,TIAM2,SCAF8, OPRM1,MOR	3	0.070	0.187	0.744
1	262’501’682	262'660’359	6.5E-04	1.0E-01	IKBKAP,CTNNAL1,PTPN3,**TXN**	5	0.062	0.260	0.678
2	149’908’135	149’928’787	8.9E-04	5.1E-01	TAF7,FCHSD1,ARAP3,PCDH1,PCDH12, RNF14	3	0.059	0.282	0.659
3	4’652’086	4’783’662	2.0E-06	3.1E-04	TMEM130,TRRAP,LMTK2,PMS2	5	0.059	0.284	0.656
3	11’719’243	11’817’679	1.4E-04	1.2E-06	CLDN3,CLDN4,CLIP2	6	0.076	0.349	0.575
3	11’911’657	12’011’995	1.2E-04	8.1E-02	GTF2I,NCF1	4	0.072	0.345	0.583
3	13’122’665	13’313’373	1.5E-04	1.3E-01	/	6	0.061	0.301	0.638
3	13’431’447	13’516’282	2.4E-05	4.5E-04	/	4	0.064	0.272	0.664
4	16’030’238	16’329’425	9.0E-04	8.2E-01	ANXA13,NDUF8B9,FBXO32,ATAD2, ZHX2	7	0.067	0.288	0.645
4	120’313’113	120’759’203	4.1E-04	4.6E-02	/	5	0.104	0.420	0.476
4	122’393’290	122’672’263	9.1E-04	8.9E-03	AMY2,COLL11A1,OLFM3	4	0.190	0.356	0.454
5	9’428’689	9’671’319	7.2E-04	2.2E-05	APOL3	4	0.060	0.271	0.669
5	11’810’479	11’947’521	4.2E-04	4.0E-01	TIMP3,FBOX7,BPIL2,PRDM4	3	0.066	0.349	0.585
5	90’839’778	91’009’130	5.2E-04	2.9E-12	ELK3,AMDHD1,VEZT	4	0.107	0.453	0.441
5	92’813’352	92’997’549	1.0E-04	4.1E-07	VEZT,TMCC3,PLXNC1,**SOCS2**,EEA1	3	0.052	0.454	0.494
6	6’743’879	6’819’136	2.7E-04	1.3E-01	CDH13,PLCG2,GAN	3	0.056	0.285	0.660
7	5’852’455	7’250’790	5.3E-05	7.6E-06	SSR1,**BMP6**,MU,TFAP2A,GCNT2, NEDD9	19	0.079	0.401	0.520
7	7’690’627	7’747’189	9.2E-04	1.6E-01	TFAP2A,GCNT2,NEDD9	3	0.097	0.358	0.545
7	7’932’328	8’004’461	3.4E-04	1.4E-01	TFAP2A,GCNT2,NEDD9	4	0.057	0.284	0.659
7	9’569’676	9’935’631	4.9E-04	7.2E-05	EDN1,PHATR1,TBC1D7,RANBP9,SIRT5	7	0.120	0.278	0.603

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
9	10’885’003	11’028’079	4.5E-04	5.0E-03	XRRA1,ARRB1,RPS3,GDPD5,MOGAT2, DGAT2,WNT11,PRKRIR	5	0.050	0.192	0.758
9	11’454’080	11’634’383	3.2E-04	2.0E-01	RPS3,GDPD5,MOGAT2,DGAT2,WNT11, PRKRIR	4	0.057	0.300	0.643
9	15’469’094	15’691’961	4.5E-04	4.7E-02	**GAB2**,NARS2,ODZ4	7	0.087	0.407	0.506
9	145’783’544	145’872’227	5.6E-05	7.7E-02	IRF6,PLXNA2,CD34	3	0.047	0.221	0.732
9	146’086’379	146’322’354	7.8E-04	3.8E-01	CD34,CD46	5	0.072	0.340	0.588
10	54’849’188	55’090’849	5.6E-04	7.2E-01	ARMC4;ANKRD26,MASTL,ACBD5	4	0.082	0.373	0.545
10	71’033’588	71’298’088	2.4E-04	4.3E-01	ITIH2	4	0.114	0.258	0.628
12	43’240’520	43’258’838	1.7E-04	3.0E-02	TMEM132E,CCL2,CCL1,CCL8	5	0.069	0.339	0.592
12	43’461’543	43’702’270	2.4E-04	3.7E-01	TMEM132E,CCL2,CCL1,CCL8,SPACA3	3	0.073	0.360	0.568
14	5’702’121	5’768’481	7.6E-04	1.4E-01	/	4	0.091	0.470	0.439
14	142’198’242	143’069’982	6.4E-04	6.2E-04	ATE1,NSMCE4A	13	0.055	0.286	0.659
14	144’745’784	144’838’386	2.2E-05	2.9E-03	IKZF5,HMX3,BUB3,GPR26,CPXM2	3	0.072	0.213	0.716
15	35’572’310	36’136’999	6.3E-04	8.3E-01	PTPN18,PTPN4	7	0.109	0.425	0.466
16	58’597’955	58’727’224	4.5E-04	3.4E-10	LCP2,FOXI1,DOCK2,CCDC99,SLIT3, ODZ2	4	0.046	0.257	0.697
17	59’872’938	60’053’503	4.5E-04	1.3E-01	/	3	0.048	0.258	0.694

*position in bp

Most of the clusters with significant results appear to be near the end of the chromosomes, which may be due either to a mechanism increasing the probability of a selective sweeps towards the telomeres, or a suppression of selective sweeps close to the centromere. Obvious explanations like the variability of the recombination rate between genomic regions can be excluded, since the REHH approach already accounts for this. Since similar patterns also have been reported in other studies (e.g. Qanbari *et al*. 2010), further analyses are required to understand the underlying causes. A summary of the core haplotypes reflecting the lowest p-values ($p < 0.001$) of the REHH test and the respective estimated membership coefficients are listed in Table 2. The calculated mean proportion of the founder breeds over all SNPs of the particular region is reported because of the variability allocation of each allele mentioned before even within the selection signatures. In order to examine the composition of the proportion of the founder breeds, the membership coefficient for the core haplotypes were transformed using an arcsine-transformation and compared using a t-test against the pedigree information. All values besides one for the VPP on chromosome 7 (start: 9'569'676 bp) show a significant difference to the pedigree proportion of Glodek & Oldigs (1981). Some regions overlapped with genes of potential biological relevance for the GMPs. One of the strongest signals is adjacent to the suppressor of cytokine signaling-2 (*SOCS2*) gene on chromosome 5 (start: 92'813'352 bp). The *SOCS2* gene negatively regulates growth hormone and insulin-like growth factor-1 (*IGF-1*) and might play a negative regulatory role in the growth hormone *IGF-1* pathway (Metcalf *et al.* 2000). In mice it displays an excessive growth phenotype characterized by a 30–50% increase in mature body size (Greenhalgh *et al.* 2005). Piper *et al*. (2005) mapped the *SOCS2* gene on porcine chromosome 5. This gene might be one important cause for the reduced body size of the Göttingen Minipigs. The contributions of the founder breeds in this region confirm the assumption of implication of *SOCS2* in the small body size of the GMPs. The contribution of the MMP is significantly higher while it is significantly lower for GL compared to the pedigree values. Simianer and Köhn (2010) mentioned a possible influence of the insulin-like growth factor1 (*IGF-1)* gene for the small body size in the Göttingen Minipigs. The involvement of *SOCS2* in the *IGF-1* pathway supports this suggestion.

Another gene of interest found on chromosome 1 (start: 262'501'682 bp) is the thioredoxin gene (*TXN*, also known as *TRX*), which has a possible effect on growth-

related traits in pigs. In an association analysis in a Berkshire and Yorkshire F_2 population Yu *et al.* (2007) reported a significant effect of the *TXN* gene on growth and carcass traits. But further research is needed to elucidate the association between *TXN* and growth related traits.

The other genes found are less obviously linked to the breeding goals of the Göttingen Minipig. One of the core regions on chromosome 7 (start: 5'852'455 bp) harbours the *BMP6* gene. Bone morphogenetic proteins (*BMP*s) are a family of secreted signalling molecules that can generate bone growth (Jane *et al.* 2002). *BMPs* have a clear function in regulation of bone formation (Linkhart *et al.* 1996). Additionally members of the BMP family are involved in ovarian function of pigs and the follicular development (Brankin *et al*. 2005; Paradis *et al*. 2009).

More striking is that this signature has the most extended core region with 19 SNPs stretching over 1'398'335 bps. Another signal observed on chromosome 9 (start: 15'469'094 bp) corresponds to the *GAB2* gene. This gene is a member of the GRB2-associated binding protein (*GAB*) gene family. Lock *et al*. (2002) described a correlation for the *GAB* family (*GAB1* and *GAB2*) to various cytokines and growth factors, so that it could play an important role for the small size of the GMP. Surprisingly we did not find any strong signatures in the vincinity of prominent genes affecting coat colour, like KIT or MC1R, which may reflect either the limited power of the study or may reflect that genes involved in the inheritance of complex phenotypes may fail to lead to sufficiently strong signatures of selection.

Table 3: Summary of the core haplotypes showing the lowest p-values (p < 0.001 after Bonferroni correction) for the χ^2-test and the mean membership coefficient of the involved SNPs for the three founder breeds. All displayed membership coefficients are significantly different from the mean autosomal proportion (t-test; p < 0.05).

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
1	77'686'940	77'830'044	3.7E-02	1.5E-14	PDSS2,SCML4,SEC63	3	0.099	0.155	0.746
1	110'462'294	110'599'775	2.3E-01	5.4E-09	WDR7,ST8SIA3,FECH,SUMO1,OAZ2	3	0.098	0.528	0.374
1	195'332'077	195'385'688	2.7E-01	6.9E-10	SLC35F4,ARIDA4A	3	0.052	0.222	0.726
1	247'635'217	248'128'615	2.2E-01	6.1E-14	CD72,CA9,RRGP1,RECK,MELK,POLR1E, GRHPR,RG9MTD3	5	0.035	0.203	0.762
1	262'688'266	262'750'389	1.7E-02	9.4E-09	PTPN3,**TXN**	3	0.091	0.529	0.380
1	262'810'119	262'881'548	8.6E-03	1.2E-15	PTPN3,**TXN**	3	0.138	0.551	0.311
1	267'850'898	268'032'353	5.2E-02	4.0E-08	ORM1,TNC,PAPPA,ASTN2	6	0.098	0.519	0.383
2	26'265'758	26'465'138	1.6E-01	5.6E-16	RAG2,RAG1,TRAF6	4	0.103	0.584	0.313
2	106'575'056	106'648'293	3.5E-02	9.1E-09	CAST,LNPEP,RIOK2,CHD1	3	0.068	0.219	0.713
2	119'728'423	119'801'803	2.5E-01	3.1E-10	WDR36,CAMK4,STARD4,DCP2,APC	4	0.123	0.192	0.686
2	142'949'391	143'126'995	3.3E-02	1.2E-08	PPP2CA,PHF15,SAR1B,DDX46,H2AFFFFFY, CXCL14,IL9,TRPM2	12	0.100	0.523	0.377
2	152'596'156	152'735'587	3.6E-01	2.4E-12	/	3	0.183	0.434	0.383
3	1'494'573	1'844'210	2.7E-01	5.4E-12	AMZ1,IQCE,GRIFIN	5	0.140	0.516	0.344
3	6'125'394	6'157'593	5.9E-03	1.4E-15	TMEM130,TRRAP,SMURF1,ZKSCAN5, GPC2	3	0.117	0.571	0.312
3	14'387'825	14'579'974	5.8E-03	3.5E-13	/	5	0.177	0.461	0.361
3	16'136'745	16'150'048	8.8E-03	1.0E-18	TPST1,GUSB	3	0.228	0.334	0.438
3	21'104'387	21'236'245	1.7E-03	8.9E-09	HS3ST4,ZKSCAN2	4	0.049	0.237	0.715
3	22'836'583	22'951'302	1.5E-01	2.3E-08	ZKSCAN2,RBBP6,PRKCB,HS3ST2,OTOA	3	0.055	0.234	0.711
3	33'714'844	34'186'421	6.9E-03	2.5E-09	USP7,ABAT4	9	0.116	0.517	0.367
3	35'618'387	35'737'712	3.3E-02	3.3E-13	RBFOX1	4	0.180	0.456	0.365

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
3	65’647’565	65’733’807	3.7E-01	1.7E-15	CTNNA2,LRRTM1	5	0.074	0.593	0.333
3	100’306’964	100’631’347	5.5E-02	2.1E-08	SRBD1,SLC3A1	4	0. 059	0.230	0.711
4	2’166’585	2’539’776	1.4E-01	1.3E-08	TOP1MT,TRAPPC9	7	0.078	0.215	0.708
4	5’287’565	5’329’178	3.8E-02	5.8E-09	/	3	0.097	0.528	0.375
4	7’975’417	8’073’943	7.4E-03	9.9E-09	ST3GL1,TG,PHF20L1	3	0.061	0.225	0.714
4	11’776’710	11’992’812	1.3E-02	2.8E-09	MYC	6	0.108	0.523	0.368
4	82’326’151	82’590’402	1.0E-01	1.2E-15	PCMTD1,SNAI2	5	0.207	0.403	0.390
4	86’255’688	87’197’352	6.3E-02	1.7E-11	BLZF1,ATP1B1,XCL1,GPR161,BRP44, MPZL1,RCSD1,CD247,POU2F1,DUSP27	4	0.039	0.220	0.741
4	90’526’641	90’784’788	9.5E-03	1.5E-10	RXRG,PBX1,RGS5	5	0.188	0.285	0.527
4	92’296’719	92’317’795	2.2E-02	4.9E-13	HSD17B7,**DDR2**,UAP1,ATF6,FCRLA, FCGR2B	3	0.054	0.191	0.755
4	116’343’160	116’445’192	7.5E-03	1.1E-16	CHIA,CHI3L2,LRIF1,RBM15,SLC6A17, EPS8L3,ATXN7L2,SORT1,PSRC1,	3	0.103	0.589	0.308
4	122’844’747	122’890’243	5.5E-02	1.2E-13	AMY2,COLL11A1,OLFM3	3	0.171	0.482	0.348
4	126’181’098	126’280’184	9.8E-03	4.6E-12	SLC35A3,AGL,FRRS1,PALMD,DPYD	4	0.107	0.550	0.343
4	138’462’131	138’497’488	3.2E-01	1.3E-08	LMO4,HS2ST1,SEP15,DDAH1,BCL10, SYDE2	3	0.172	0.401	0.427
5	12’407’245	12’552’126	2.6E-02	4.7E-19	TIMP3,FBXO7,BPIL2,PRDM4,CRY1,RIC8B	4	0.226	0.373	0.401
5	14’023’315	14’154’240	2.5E-02	5.6E-09	CRY1,RIC8B,RFX4	3	0.049	0.234	0.717
5	64’698’606	64’787’530	3.4E-03	6.8E-09	RIMKLB,LPCAT3,LEPREL2,NCAPD2, NOP2,ACRBP,VWF,ANO2	4	0.165	0.429	0.405
6	39’821’723	39’987’390	2.2E-01	1.0E-12	USF2,HAMP,FFAR2,HAUS5,COX7A1	4	0.102	0.559	0.339
6	62’159’543	62’671’136	2.3E-01	9.5E-14	PARK7,CA6,TMEM201,PIK3CD,RBP7, CLSTN1,KIF1B,PGD,PEX14,**MTOR**	5	0.160	0.502	0.337

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
6	63’825’178	64’218’121	8.3E-02	2.3E-08	KIF1B,PGD,PEX14,**MTOR**,PTCHD2, MTHFR,BNP,VPS13D	3	0.177	0.369	0.455
6	77’549’698	78’711’487	2.6E-01	3.7E-11	PIGV,SFN,SLC9A1,SYTL1,ATIPIF1, PHACTR4,OPRD1	3	0.104	0.544	0.352
6	108’898’313	109’071’487	8.3E-02	1.1E-10	DSG1,DSG2,TTR,MEP18,KLHL14	5	0.192	0.343	0.465
6	147’516’035	147’610’825	3.6E-02	3.5E-17	SCP2,CC2D1B,CLIC1,NRD1,OSBPL9, EPS15,FAF1	3	0.089	0.600	0.311
7	11’837’643	12’116’601	6.1E-03	5.1E-15	JARID2,DTNBP1,MYLIP,GMPR,ATXN1	5	0.116	0.567	0.317
7	13’176’173	13’241’394	2.1E-01	2.2E-09	MYLIP,GMPR,ATXN1,CAP2,NUP153, KDM1B,DEK	3	0.178	0.392	0.429
7	13’261’225	13’449’875	1.4E-01	2.4E-10	MYLIP,GMPR,ATXN1,CAP2,NUP153, KDM1B,DEK	4	0.092	0.544	0.364
7	55’461’746	56’525’630	1.5E-01	4.6E-15	ARNT2,IL16,TMC3,EFTUD1,ADAMTSL3	20	0.213	0.335	0.452
7	56’611’064	56’695’601	7.2E-02	2.9E-18	ADAMTSL3	3	0.196	0.177	0.627
7	116’930’058	117’018’272	1.4E-02	3.5E-08	KCNK10,SPATA7,PTPN21,EML5,TTC8, FOXN3,TDP1,PSMC1	3	0.169	0.272	0.558
7	117’061’194	117’133’386	7.1E-02	4.6E-09	SPATA7,PTPN21,EML5,TTC8,FOXN3, TDP1,PSMC1	4	0.146	0.475	0.379
7	119’164’800	119’472’121	6.7E-02	4.5E-08	SMEK1,TC2N,TRIP11,ATXN3,CPSF2, LGMN,CHGA,UBR7,BTBD7	6	0.109	0.510	0.381
7	127’332’964	127’479’348	1.4E-01	4.3E-09	BCL11B,SETD3,CCNK	5	0.158	0.451	0.391
7	130’229’259	130’476’299	2.5E-01	3.1E-08	RAGE,WDR20,PLD4,UGPP	7	0.153	0.451	0.397
8	6’362’061	6’400’017	6.4E-03	2.7E-08	OTOP1,ZBTB49,PNAS5	3	0.064	0.227	0.709
8	8’105’894	8’179’460	2.0E-02	1.5E-17	RAB28,BOD1L	3	0.081	0.606	0.313
8	12’089’245	12’251’523	5.5E-02	3.4E-09	NCAPG,LCORL	3	0.119	0.513	0.368
8	92’050’153	92’180’550	3.3E-01	2.4E-09	MAML3,OSAP,NAA15,NDUFC1,CCRN4L	4	0.155	0.461	0.384

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
8	143’129’541	143’334’866	1.2E-01	1.3E-09	AGPAT9,HELQ,HPSE,COPS4,THAP9,LIN54, SCD5	4	0.044	0.233	0.723
9	10’885’003	11’028’079	4.5E-04	3.4E-10	XRRA1,ARRB1,RPS3,GDPD5,MOGAT2, DGAT2,PRKRIR	5	0.105	0.535	0.360
9	145’081’582	145’199’822	8.3E-02	1.2E-09	HHAT,SERTAD4,SYT14,IRF6	4	0.056	0.220	0.724
9	145’783’544	145’872’227	5.6E-05	2.9E-12	IRF6,PLXNA2,CD34	3	0.115	0.545	0.340
9	149’232’822	149’247’405	2.3E-01	6.3E-10	DDC,**GRB10**	3	0.188	0.325	0.487
10	4’074’194	4’285’151	8.4E-01	3.4E-10	/	3	0.189	0.331	0.479
10	17’228’854	17’397’772	2.2E-01	6.1E-13	EPHX1,LEFTY2,H3F3A,PARP1,PSEN1, ADCK3,EXO1	4	0.130	0.536	0.333
10	22’204’332	23’081’858	2.7E-01	6.1E-09	TLR5,KIF26B,SMYD3	5	0.178	0.383	0.439
10	54’545’722	54’619’033	5.9E-02	2.2E-08	ARMC4	3	0.046	0.243	0.710
10	56’026’229	56’157’952	4.8E-02	8.0E-16	ARMC4,ANKRD26,MASTL,ACBD5,GAD2, MYO3A	3	0.174	0.499	0.328
10	58’062’142	58’135’020	5.0E-02	2.0E-10	ARHGAP21	3	0.101	0.540	0.360
10	62’327’563	62’455’328	1.1E-01	3.8E-11	NEBL,PLXDC2	4	0.132	0.518	0.350
10	64’118’377	64’220’978	2.9E-02	1.2E-10	ITGB1,NRP1	3	0.098	0.543	0.358
10	65’779’329	65’851’836	1.7E-01	2.7E-11	CUL2	3	0.195	0.337	0.468
10	67’318’742	67’453’442	1.2E-02	7.1E-09	/	3	0.177	0.282	0.542
10	68’661’410	68’930’387	4.0E-03	2.3E-09	/	5	0.117	0.516	0.367
10	74’564’460	74’777’316	2.5E-01	2.2E-10	NET1,AKR1C4,AKR1C1,KLF6,PFKP	5	0.102	0.538	0.359
11	25’420’865	25’553’629	2.5E-02	1.9E-11	DGKH	4	0.117	0.537	0.347
12	74’980	321’972	1.6E-01	4.8E-12	NARF,FASN,ACTB,BAHCC1	7	0.055	0.199	0.746
12	8’154’483	8’366’664	6.9E-02	2.2E-13	SDK2,COG1,SSTR2,SOX9	3	0.205	0.344	0.451
12	18’361’958	18’663’673	3.9E-01	6.4E-10	ITGB3,PLCD3,GFAP,GJD3	11	0.145	0.486	0.368

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
12	20’197’154	20’313’915	2.7E-02	6.6E-09	PLCD3,GFAP,GJD3,UBTF,PLEKHH3,CCR10, CNTNAP1,RAMP2	5	0.136	0.488	0.375
12	34’376’265	34’621’947	3.8E-02	7.4E-09	SCPEP1,DGKE,SRSF1	5	0.166	0.428	0.406
12	44’053’529	44’192’011	3.5E-03	4.9E-09	SPACA3,CDK5R1,RHOT1,ADAP2,CRLF3, SUZ12	3	0.175	0.397	0.427
13	3’629’541	3’715’522	2.2E-02	3.4E-10	SH3BP5,RFTN1	3	0.111	0.530	0.359
13	30’380’898	30’542’260	2.2E-02	2.2E-12	ABHD5,KIF15,CDCP1,CCR9,SLC6A20	4	0.198	0.282	0.520
13	46’262’390	46’400’431	8.1E-02	8.9E-10	PTPRG,FEZF2,CADPS	4	0.152	0.473	0.375
13	70’091’801	70’253’957	3.7E-01	2.8E-09	XTR,SETD5,MTMR14,CIDEC,CPNE9, ARPC4,BRPF1,FANCD2	4	0.072	0.541	0.387
13	85’176’731	85’305’173	1.1E-01	2.7E-18	EXYT3,CEP70,FAIM,PIK3CB,FOXL2, COPB2,RBP2,RBP1	3	0.063	0.615	0.322
13	174’687’073	174’921’751	2.2E-01	4.3E-11	HTR1F,POU1F1,VGLL3	4	0.154	0.485	0.361
14	96’089’462	96’239’411	3.8E-01	1.6E-10	MMRN2,SNCG,BMPR1A,GLUD1,GPRIN2, PPYR1,GDF2,FRMPD2,MAPK8	6	0.077	0.552	0.372
14	96’670’353	96’730’391	3.4E-02	1.3E-14	PPYR1,GDF2,FRMPD2,MAPK8,ERCC6	3	0.043	0.588	0.369
14	119’174’868	119’557’100	2.9E-02	2.9E-13	PI4K2A,HPSE2,GOT1	4	0.162	0.494	0.344
14	122’417’459	122’581’041	6.9E-02	1.6E-11	SEMA4G,BTRC,FBXW4,MGEA5,PPRC1, NOLC1,HPS6,PSD,GBF1,SUFU,TRIM8	3	0.046	0.212	0.742
14	124’082’863	124’108’919	4.4E-02	1.5E-15	PSD,GBF1,SUFU,TRIM8,C14H10,CNNM2, INA,COL17A1,GSTO1	3	0.109	0.577	0.314
14	125’558’653	125’627’557	2.4E-02	6.1E-09	COL17A1,GSTO1,GSTO2,SORCS3	3	0.110	0.519	0.372
14	143’387’596	143’447’344	8.2E-02	1.4E-12	ATE1,PAF1,SAMD4B,IL28B,IKZF5,HMX3, BUB3	4	0.048	0.201	0.751
14	146’848’808	147’079’198	3.1E-03	2.3E-08	CCIP,DHX32	6	0.117	0.506	0.377
14	149’273’491	149’500’630	8.3E-03	4.5E-08	PTPRE	7	0.089	0.523	0.388

Chr	start *	end *	REHH p-value	χ2-test p-value	Genes	# SNPs	Membership coefficient GL	MMP	VPP
14	151’578’879	151’879’008	6.1E-02	1.8E-08	MGMT,PWWP2B,INPP5A	8	0.155	0.448	0.397
15	18’923’721	19’031’226	2.4E-02	3.1E-11	PAB3GAP1,MGAT5	4	0.033	0.229	0.737
15	34’832’370	35’108’036	1.7E-02	3.6E-10	GLI2,PTPN18,PTPN4	5	0.046	0.225	0.729
15	49’922’169	49’986’405	3.2E-01	3.5E-15	WWC2,CLDN22,CASP3,ACSL1	3	0.212	0.354	0.433
16	18’600’403	18’628’286	7.6E-02	1.9E-08	PDZD2	3	0.082	0.214	0.704
16	21’215’092	21’619’935	9.0E-02	1.6E-11	SLC45A2,AMACR,C1QTNF3, RAI14,DNAJC21,**PRLR**	6	0.128	0.526	0.346
17	12’301’328	12’415’934	5.8E-02	1.9E-11	IDO1,ADAM18,ADAM3A,CHRNA6	4	0.196	0.310	0.495
17	16’469’827	16’802’243	4.0E-02	2.2E-13	PRNP,RASSF2,PCNA,PROKR2,CDS2, GPCPD1,CRLS1,CHGB,MCM8	4	0.046	0.195	0.758
17	55’852’424	55’945’787	1.9E-02	7.4E-15	NCOA3,SULF2	3	0.124	0.559	0.317
18	7’865’943	7’936’927	5.0E-03	1.0E-14	ZYX,KEL,TRPV6,CLEC5A	3	0.135	0.546	0.319
18	32’766’555	33’037’455	1.1E-01	8.4E-11	CAV1,CAV2,TES,MDFIC,FOXP2	5	0.049	0.216	0.735
18	43’261’642	43’456’159	4.3E-02	1.1E-09	BMPER,BBS9,KBTBD2,AVL9	3	0.040	0.237	0.723
18	43’684’258	44’038’246	6.0E-02	2.1E-08	BMPER,BBS9,KBTBD2,AVL9,PDE1C	6	0.100	0.209	0.691
18	45’226’045	45’311’991	4.5E-02	6.2E-13	PDE1C,GHRHR,CRHR2,NOD1	4	0.170	0.476	0.355
18	51’216’800	51’375’627	1.7E-02	2.7E-08	NPVF,CYCS,OSBPL3,DFNA5,NPY	4	0.107	0.514	0.379
18	55’219’126	55’365’034	1.1E-02	2.1E-08	PPIA,MYO1G,NPC1L1,GCK,AEBP1,POLM, HECW	6	0.110	0.513	0.377

* position in bp

χ^2-test for the deviation of the breed composition

The core haplotypes with the lowest p-values of χ^2-test ($p < 0.001$ after Bonferroni correction) are presented in Table 3. One of the lowest p-values of the χ^2-test on chromosome 9 (start: 149'232'822 bp) is adjacent to the growth factor receptor-bound protein 10 (*GRB10*). Charalambous *et al.* (2003) suggested that *GRB10* belongs to a major foetal growth pathway. They showed that a disruption in the *GRB10* gene causes overgrowth such that the mutant mice are 30% larger than normal mice at birth. The functional role of *GRB10* in insulin signalling is controversial. Shiura *et al.* (2005) and Wang *et al.* (2007) demonstrated that *GRB10* negatively regulates the *IGF-1* gene whereas other authors found that *GRB10* is a positive regulator of *IGF-1* action (Smith *et al.* 2007; Deng *et al*. 2003). It must be noted that *GRB10* strongly influences animal growth and might be a reason for the small body size in the GMP.

Another gene found on chromosome 6 (start: 62'159'543 bp) is the mechanistic target of rapamycin gene (*MTOR*). The *MTOR* pathway is regulating growth factor signalling, and its inhibition leads to changes in the abundance of *GRB10* (Hsu *et al*. 2011; Zoncu *et al*. 2011). The relationship between *MTOR* and *GRB10* can increase the importance of *GRB10* for the small body size in the GMP.

Combining REHH and membership coefficient

In an admixed population chromosome segments under selection may be characterised by a selective sweep, by an allele frequency spectrum deviating from the genome average, or by a combination of both indicators. We therefore identified for each of the 2'807 core regions the haplotype with the highest REHH value and plotted this value against the p-value obtained from the χ^2-test for this core region (Figure 7).

The core haplotypes reflecting the p-values of the REHH-test below 0.001 are listed in Table 2, while the core haplotypes with the lowest p-values of χ^2-test ($p < 0.001$ after Bonferroni correction) are presented in Table 3.

Figure 7 shows that only two core regions that achieved both an extreme REHH value and a highly significant signal in the χ^2-test. This is in agreement with the hypothesis, that the two tests focus on different aspects of selection, namely extended homozygosity for the REHH test and deviation in breed composition for the χ^2-test. Of the 2458 intervals of 1 Mb length across the genome, 31 (1.26 %) intervals show a significant

REHH and 108 (4.4%) of all intervals are significant for the -test. If the two tests are complementary, we would expect both tests to be significant in 0.055% (0.0126 x 0.044 = 0.00055) or 1.36 of the 2458 intervals, and the observed number of 2 intervals matches this expectation well. The genes found in these two regions (Table 2): one on chromosome 5 (start: 90'839'778 bp) and the other one on chromosome 16 (start: 58'597'955 bp) are known to affect lung, kidney function, ear, cancer, etc. but no obvious relationship to the breeding goals of the GMP can be stated. Further studies are necessary to clarify the effect of these regions.

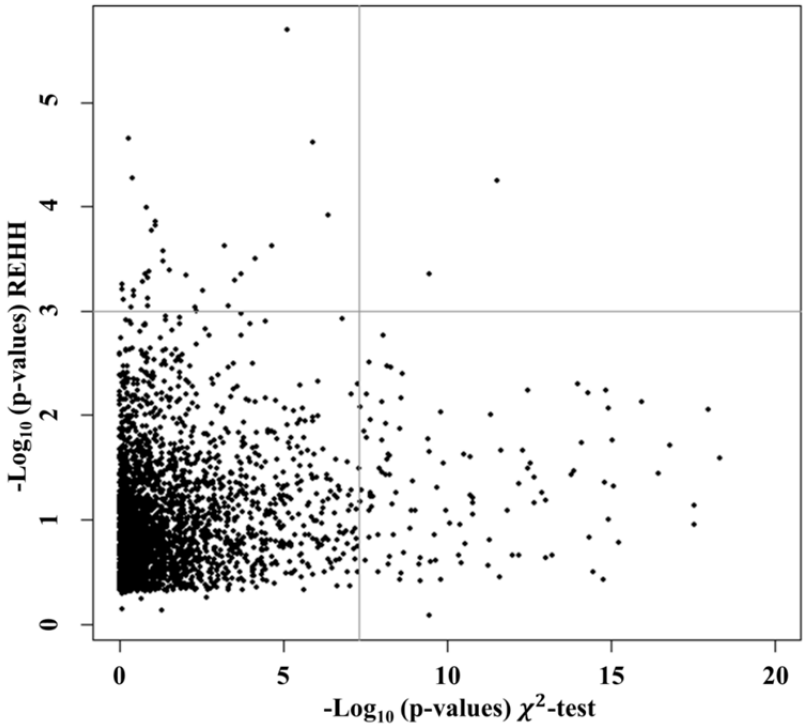

Figure 7: Plot of the negative logarithms of the p-values of the REHH versus the -test. Grey lines represent thresholds for the REHH (p < 0.001) and - test (0.1% Bonferroni level).

Conclusions

This study investigated the proportion of the three founder breeds based on SNP data. Since the calculation of Glodek und Oldigs (1981) based on pedigree data, the proportion of the founder breeds in the GMP populations apparently changed (GL = 0.085; MMP = 0.371; VPP = 0.544). To confirm the results studies with a larger

number of VPP animals are necessary. The assignment of a SNP allele to one of the founder breeds is highly variable with a fluctuation of the genetic composition for the different chromosomes and even within chromosomes. A massive local deviation of the breed composition from the genome average is interpreted as a potential indication that the region was under directional selection.

The assignment of a SNP allele to one of the founder breeds together with REHH values identified numerous regions harbouring candidate genes which appear to be functionally related to breeding goals of the Göttingen Minipig, c.f. *SOCS2, TXN, DDR2* and *GRB10* linked to body size, or *PRLR* being related to piglets born alive. These candidate genes can now serve as starting points for further studies. Other candidate regions do not show signatures of recent selection as expected, which may be either due to statistical or to biological reasons. The results suggest that the pathway connecting *SOCS2* and *GRB10* with the *IGF-1* might be causal for the small body size of the Göttingen Minipigs.

Acknowledgements

We are grateful to Ellegaard Göttingen Minipigs ApS (Denmark), Marshall BioResources (USA), Sinclair Research Center (USA) and the Tierpark Berlin-Friedrichsfelde (Germany) for providing DNA samples.

References

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. & Kruglyak, L. (2004): Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol., 2, 10, 1591-1599.

Brandt, H., Möllers, B. & Glodek, P. (1997): Prospects for a genetically very small minipig. Scand. J. Anim. Sci. Suppl., 25, 93-96.

Brankin, V., Quinn, R.L., Webb, R. & Hunter, M.G. (2005): Evidence for a functional bone morphogenetic protein (BMP) system in the porcine ovary. Dom. Anim. Endocrinol., 28, 367–379.

Charalambous, M., Smith, F.M., Bennett, W.R., Crew, T.E., Mackenzie, F. & Ward, A. (2003): Disruption of the imprinted Grb10 gene leads to disproportionate overgrowth by an Igf2-independent mechanism. Proc. Natl. Acad. Sci., 100, 8292–8297.

Deng, Y., Bhattacharya, S., Swamy, O.R., Tandon, R., Wang, Y., Janda, R. & Riedel, R. (2003): Growth factor receptor-binding protein 10 (Grb10) as a partner of phosphatidylinositol 3-kinase in metabolic insulin action. J. Biol. Chem., 278, 39311–39322.

Drögemüller, C., Hamman, H. & Distl, O. (2001): Candidate gene markers for litter size in different German pig lines. J. of Anim. Sci., 79, 2565–70.

Falconer, D. S. & Mackay, T.F.C. (1996): Introduction to quantitative genetics. Longmans Green, Edition 4, Harlow, Essex, UK.

Forster, R., Bode, G., Ellegaard, L. & van der Laan, J.W. (2010): The RETHINK project on minipigs in the toxicity testing of new medicines and chemicals: Conclusions and recommendations. J. of Pharmacological and Toxicological Methods, 62, 3, 236-242.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. & Altshuler, D. (2002): The structure of haplotype blocks in the human genome. Science, 296, 2225–2229.

Glodek, P. & Oldigs, B. (1981): Das Göttinger Miniaturschwein. Schriftenreihe Versuchstierkunde 7, Paul Parey Verlag, Berlin.

Greenhalgh, C.J., Rico-Bautista, E., Lorentzon, M., Thaus, A.L., Morgan, P.O., Willson, T.A., Zervoudakis, P., Metcalf, D., Street, I., Nicola, N.A., Nash, A.D., Fabri, L.J., Norstedt, G., Ohlsson, C., Flores-Morales, A., Alexander, W.S. & Hilton, D.J. (2005): SOCS2 negatively regulates growth hormone action in vitro and in vivo. J. Clin. Invest, 115, 397-406.

Groeneveld, E., Westhuizen, B.V.D., Maiwashe, A.,Voordewind, F. & Ferraz, J.B.S. (2009): POPREP: A Generic Report for Population Management. Gen. Mol. Res., 8, 1158–1178.

Hermisson, J. & Pennings, P.S. (2005): Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics, 169, 2335–2352.

Hill, W.G. & Robertson, A. (1968): Linkage Disequilibrium in finite populations. Theoretical and Applied Genetics, 38, 226-231.

Hsu, P.P., Kang, S.A., Rameseder, J., Zhang, Y., Ottina, K.A., Lim, D., Peterson, T.R., Choi, Y., Gray, N.S., Yaffe, M.B., Marto, J.A. & Sabatini, D.M. (2011): The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signalling. Science, 332, 1317-1322.

Jane, J.A., Dunford, B.A., Kron, A., Pittman, D.D., Sasaki, T., Li, J.Z., Li, H., Alden, T.D., Dayoub, H., Hankins, G.R., Kallmes, D.F. & Helm, G.A. (2002): Ectopic osteogenesis using adenoviral bone morphogenetic protein (BMP)-4 and BMP-6 gene transfer. Mol. Therapy, 6, 4, 464-470.

Kano, K., Marin de Evsikova, C., Young, J., Wnek, C., Maddatu, T.P., Nishina, P.M. & Naggert, J.K. (2008): A novel dwarfism with gonadal dysfunction due to loss-of-function allele of the collagen receptor gene, DDR2, in the mouse. Mol. Endocri., 22, 1866–1880.

Köhn, F., Sharifi, A.R. & Simianer, H. (2007): Modelling the growth of the Göttingen Minipig. J Anim. Sci., 85, 84-92.

Köhn, F., Sharifi, A.R. & Simianer, H. (2009): Genetic analysis of reactivity to humans in Göttingen Minipigs. Applied Anim. Beh. Sci., 120, 68–75.

Labrador, J.P., Azcoitia, V., Tuckermann, J., Lin, C., Olaso, E., Manes, S., Bruckner, K., Goergen, J.L., Lemke, G., Yancopoulos, G. Angel, P., Martínez-A, C. & Klein, R. (2001): The collagen receptor DDR2 regulates proliferation and its elimination leads to dwarfism. Embo. Rep., 2, 446–452.

Linkhart, T.A., Mohan, S. & Baylink, D.J. (1996): Growth factors for bone growth and repair, IGF, TGFβ and BMP. Bone, 19, 1-12.

Lock, L.S., Maroun, C.R., Naujokas, M.A. & Park, M. (2002): Distinct Recruitment and function of Gab1 and Gab2 in Met Receptor-mediated Epithelial Morphogenesis. Mol. Biology of the Cell, 13, 2132–2146.

Lohmüller, K.E., Bustamante, C.D. & Clark, A.G. (2011): Detecting directional selection in the presence of recent admixture in African-Americans. Genetics, 187, 823–835.

Maynard-Smith, J. & Haigh, J. (1974): The hitch-hiking effect from a favourable gene. Genetical Research, 23, 23–35.

Metcalf, D., Greenhalgh, C.J., Viney, E., Wilson, T., Starr, R., Nicola, N., Hilton, D., Alexander, W.S. (2000): Gigantism in mice lacking suppressor of cytokine signalling-2. Nature, 405, 1069–1073.

Mukherjee, S., Golland, P.& Panchenko, D.(2003): Permutation tests for classification. AI Memo 2003-019. Vol. Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory.

Nielsen, R. (2005): Molecular Signatures of Natural Selection. Annual Review Genetics, 39, 197-218.Paradis, F., Novak, S., Murdoch, G.K., Dyck, M.K., Dixon, W.T. & Foxcroft, G.R. (2009): Temporal regulation of BMP2, BMP6, BMP15, GDF9, BMPR1A, BMPR1B, BMPR2 and TGFBR1 mRNA expression in the oocyte, granulosa and theca cells of developing preovulatory follicles in the pig. Reproduction, 138, 115–129.

Paradis, F., Novak, S., Murdoch, G.K., Dyck, M.K., Dixon, W.T. & Foxcroft, G.R. (2009): Temporal regulation of BMP2, BMP6, BMP15, GDF9, BMPR1A, BMPR1B, BMPR2 and TGFBR1 mRNA expression in the oocyte, granulosa and theca cells of developing preovulatory follicles in the pig. Reproduction, 138, 115–129.

Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., & Shriver, M.D. (1998): Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet., 63, 1839-1851.

Piper, E., Chen, Y. & Moran, C. (2005): Assignment of suppressor of cytokine signalling-2 (SOCS2) to porcine chromosome 5 with radiation hybrids. Cytogenetic Genome Res., 111, 96B.

Pritchard, J.K., Pickrell, J.K. & Coop, G. (2010): The genetics of human adaptation: hard sweeps, soft sweeps and polygenic adaptation. Current Biology, 20, 208–215.

Putnová , L., Knoll, A., Dvorák, J. & Cepica, S. (2002): A new HpaII PCR-RFLP within the porcine prolactin receptor (PRLR) gene and study of its effect on litter size and number of teats. J. Anim. Breed. Genet. 119, 57–63.

Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R. & Simianer, H. (2010): A genome-wide scan for signatures of recent selection in Holstein cattle. Anim. Genetics, 41, 377-389.

Sabeti, P.C., Reich, D.E., Higgins, J.M. Levine, H., Richter, D.J., Schaffner, S., Gabriel, S.B., Platko, J., Patterson, N.J., McDonald, G.J., Ackerman, H., Campbell, S.J., Altshuler, D., Cooperk, R., Kwiatkowski, D., Ward, R. & Lander, E.S. (2002): Detecting recent positive selection in the human genome from haplotype structure. Nature, 419, 832–837.

Scheet, P. & Stephens, M. (2006): A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet., 78, 629-644.

Shiura, H., Miyoshi, N., Konishi, A., Wakisaka-Saito, N., Suzukif R., Muguruma, K., Kohda, T., Wakana, D., Yokoyama, M., Ishino, F. & Kaneko-Ishino, T. (2005): Meg1/Grb10 overexpression causes postnatal growth retardation and insulin resistance via negative modulation of the IGF1R and IR cascades. Biochem. a. Biophy., 329, 3, 909-916.

Simianer, H. & Köhn, F. (2010): Genetic management of the Göttingen Minipig population. J. of Pharmacological and Toxicological Methods, 62, 3, 221-226.

Smith, F.M., Holt, L.J., Garfield, A.S., Charalambous, M., Koumanov, F., Perry, M., Bazzani, R., Sheardown, S.A., Hegarty, B.D., Lyons, R.J., Cooney, G.J., Daly, R.J. & Ward, A. (2007): Mice with a disruption of the imprinted GRB10 gene exhibit altered body composition, glucose homeostasis, and insulin signalling during postnatal life. Mol. A. Cell. Biol., 27, 16, 5871–5886.

Sokal, R.R. & Rohlf, F.J. (1981): Biometry – The principles and practice of statistics in biological research; second edition. W.H. Freeman and Company, San Francisco.

Thuy, N.T.D., Melchinger-Wild, E., Kuss, A.W., Cuong, N.V., Bartenschlager, H. & Geldermann, H. (2006): Comparison of Vietnamese and European pig breeds using microsatellites. J. Anim. Sci., 84, 2601-2608.

Wang, L., Balas, B., Christ-Roberts, C.Y., Kim, R.Y., Ramos, F.J., Kikani, C.K., Li, C., Deng, C., Reyna, S., Musi, N., Dong, L.Q., DeFronzo, R.A. & Liu, F. (2007): Peripheral disruption of the Grb10 gene enhances insulin signaling and sensitivity in vivo. Mol Cell Biol 27, 6497–6505.

Yu, M., Geiger, B., Deeb, N. & Rothschild, M.F. (2007): Investigation of TXNIP (thioredoxin-interacting protein) and TRX (thioredoxin) genes for growth-related traits in pigs. Mamm Genome 18, 197-209.

Zoncu, R., Efeyan, A. & Sabatini, D.M. (2011): mTOR: from growth signal integration to cancer, diabetes and ageing. Nat. Rev. Mol. Cell. Biol., 12, 21-36.

4th CHAPTER

The cross population extended haplotype homozygosity test reveals differences between the Göttingen Minipig and two normal sized conventional breeds

C. Gärke*, F. Ytournel*, N. Kemper# and H. Simianer*

* Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

Institute of Agricultural and Nutritional Sciences, Martin-Luther- University Halle-Wittenberg, Halle, Germany

Submitted in

Animal Genetics (2012)

Manuscript ID: AnGen-12-02-0054

The cross population extended haplotype homozygosity test reveals differences between the Göttingen Minipig and two normal sized conventional breeds

C. Gärke*, F. Ytournel*, N. Kemper# and H. Simianer*

* Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

Institute of Agricultural and Nutritional Sciences, Martin-Luther- University Halle-Wittenberg, Halle, Germany

Abstract

The Göttingen Minipig (GMP) developed at the University of Göttingen is a synthetic breed that combines the high fertility of the Vietnamese Potbelly Pig, the low body weight of the Minnesota Minipig and the white coat colour of the German Landrace. The aim of this study was to detect genomic regions of recent selection between the GMP and two normal sized breeds. The whole genome was screened for footprints of recent selection based on single nucleotide polymorphism genotypes from the Illumina porcine 60k SNP chip with the 'Cross Population Extended Haplotype Homozygosity' (XPEHH) test. Regions showing the most extreme p-values harbour genes related to breeding goals relevant in one of the tested breeds, such as growth (*PLG* and *SOCS2*), skin and coat colour *(MPLH, SNAI2* and *HPS3*), or carcass traits (*PLIN2*).

Keywords

Pigs, SNP, XPEHH

Introduction

The Göttingen Minipig (GMP) is a non-rodent animal model for medical research and toxicology. It was developed in the 1960's by crossing Minnesota Minipigs, Vietnamese Potbelly Pig and German Landrace (GL) (Glodek & Oldigs 1981). The GMP is a white dwarf pig breed where all body parts are reduced in size. This type of dwarfism is often caused by growth hormone deficits, mainly of the insulin-like growth factor 1 (*IGF-1*) (Simianer & Köhn 2010).

Material & Methods

For the present study, 157 GMP animals, 50 GL animals and 55 Large White (LW) animals were genotyped with the Illumina Porcine BeadChip 60k (Illumina, San Diego, USA). Single nucleotide polymorphisms (SNPs) with unknown chromosome or position, call-rates < 95% or being monomorphic were deleted from the dataset. The haplotypes for every chromosome were reconstructed using fastPHASE (Scheet & Stephens 2006) applying the standard parameter settings. Linkage disequilibrium (LD) was estimated from these reconstructed haplotypes within each breed using the parameter r^2 (Hill & Robertson 1968). To remove SNPs with incorrect positions, an LD correction was used to remove identified LD outliers. For this, r^2-values were plotted against the physical distance for each chromosome and LD between a pair of SNPs was identified as outlier if its r^2 value was 10 standard deviations above the mean of all intervals with a similar physical distance. A SNP was deleted from the dataset if it was identified as an outlier in at least two breeds. A total of 47'578 SNPs passed the filtering and, after excluding chromosome X, 46'390 SNPs were used in the further analyses.

Table 1: Summary of the core regions showing the lowest XPEHH values (GMP).

Chr	start *	end *	max XPEHH LW (rank)	max XPEHH GL (rank)	sum (rank)	Gene
1	5'052'876	7'459'340	3.43 (3)	3.35 (1)	4	PDE10A,QKI,PARK2,MAP3K4,AGPAT4,**PLG**, SLC22A3,OCT,IGF2R,MRPL18,WTAP
8	17'499'845	19'904'302	3.31 (6)	2.74 (20)	26	GPR125,DHX15,SEL1L3,RBPJ
8	19'904'302	22'308'758	3.54 (1)	2.99 (14)	15	DHX15,SEL1L3,RBPJ,STIM2
8	34'331'042	36'735'499	3.33 (4)	3.14 (10)	14	NSUN7,APBB2,UCHL1,LIMCH1,SLC30A9, BEND4,GUF1
8	137'722'680	140'127'137	2.63 (18)	2.88 (16)	34	MMRN1,SNCA,NAP1L5,HERC3,PKD2,IBSP, SPARCL1
12	59'520'239	61'898'054	2.89 (10)	3.10 (12)	22	ZNF18,DNAH9,MYOCD,HS3ST3a1,TRPV
13	195'643'657	198'058'606	3.19 (7)	3.27 (4)	11	BACH1,GRIK1,CLDN17,CLDN8
13	198'058'606	200'473'554	2.86 (12)	3.25 (6)	18	MIR155,APP,ADAMTS1
15	145'383'595	147'765'726	3.47 (2)	3.12 (11)	13	GBX2,COL6A3,SCLY,ILKAP,KLHL30,EPSPNL, PER2,**MLPH**,RAMP1
18	12'056'681	14'442'627	2.82 (14)	3.17 (9)	23	SVOPL,CREB3L2,PTF-BETA

* position in bp

The Cross Population Extended Haplotype Homozygosity (XPEHH) test in the implementation of Pickrell *et al*. (2009) was used to detect selective sweeps in a two-population dataset where the selected allele may have (almost) achieved fixation in one population, but is still polymorphic in both populations together. The XPEHH approach is an indirect test to infer selection between two breeds where low values suggest selection in population A and high values in population B (Sabeti *et al*. 2007).

To limit the number of false positive signals the genome was subdivided into 1'000 windows of equal physical length (2.41 Mb). The results of both XPEHH calculations (GMP vs. GL, GMP vs. LW) were then arranged in ascending order (rank) according to the accumulated XPEHH statistic of the respective SNPs (Rothammer 2011). The lowest and highest 1%, respectively, of the accumulated ranked regions was used for the following annotation.

Results & Discussion

Table 1 shows the regions with accumulated lowest window ranks, indicating regions harbouring genes that were selected for in the GMP relative to the large breeds. Annotation revealed two genes that are clearly related to the breeding goals of the GMP. The melanophilin gene (*MLPH*) has an effect on the coat colour. Matesic *et al.* (2001) showed that pigmentation of mammalian hair and skin is a multistep process in which the *MLPH* gene is involved. In dogs a mutation within this gene gives rise to lighter skin or coat colour (Drögemüller *et al*. 2007). The plasminogen gene (*PLG*) plays a role in many biological processes influencing health and growth in mice. PLG-deficient mice show normal viability but retarded growth (Ploplis *et al.* 1995). Four of the ten top regions are on SSC8, suggesting additional studies for this chromosome.

Table 2 shows the regions with accumulated highest window ranks, indicating regions harbouring genes that were selected for in the large breeds relative to the GMP. Several genes found here are related to diverging breeding goals between GL and LW on one side versus GMP. The *PLIN2* gene (SSC1) is linked to carcass traits (Gandolfi *et al*. 2011), *GPR149* (SSC13) to fertility traits (Edson et al. 2010), and *SNAI2* (SSC4) and *HPS3* (SSC13) are linked to skin and coat colour characteristics (Sánchez-Martín *et al*. 2003; Santiago-Borrero *et al*. 2006). Interestingly, the suppressor of cytokine signaling-2 (*SOCS2*) gene on chromosome 5 was found. The *SOCS2* gene plays a negative regulatory role in the *IGF-1* pathway (Metcalf *et al*. 2000). A deficiency in *SOCS2*

causes an increase in mature body size including body weight, bone lengths and the weight of internal organs up to 50% (Greenhalgh *et al*. 2005). Locating the region of the *SOCS2* gene as a signature of recent selection supports the major effect of this gene on the small body size of the GMPs. Six out of the top regions for selection regions in the 'normal' sized breeds are on SSC13, suggesting additional studies for this chromosome.

Table 2: Summary of the core regions showing the highest XPEHH values (GL & LW).

Chr	start *	end *	min XPEHH LW (rank)	min XPEHH GL (rank)	sum (rank)	Gene
1	224'041'135	226'447'600	-2.20 (989)	-3.09 (999)	1988	IFNA1,MIR491,DENND4C,**PLIN2**,IFNB1, MLLT3,RPS6,HAUS6
4	85'974'159	88'361'579	-2.10 (956)	-3.13 (997)	1953	**SNAI2**,MCM4,METTL11B,KIFAP3,SCLY3, BLZF1
5	93'750'031	96'148'909	-3.33 (986)	-3.38 (995)	1981	TMCC3,PLNCX1,**SOCS2**,EEA1,BRG1,DCN
13	89'385'935	91'800'883	-1.74 (955)	-2.36 (984)	1939	CLSTN2,TRIM42,ACPL2,RASA2,RNF7,GRK7, XRN1
13	91'800'883	94'215'831	-1.94 (982)	-1.96 (976)	1958	XRN1,TRPC1,U2SURP,PLOD2
13	94'215'831	96'630'780	-2.21 (988)	-1.70 (963)	1951	PLOD2,ZIC,CPB1,HLTF,HPS3,CP
13	96'630'780	99'045'728	-2.41 (987)	-2.03 (965)	1952	CPB1,HLTF,**HPS3**,CP,WWTR1,ANKUB1,EIF2A, SELT,SIAH2,P2Y12R,MED12L,GPR87,IGSF10
13	99'045'728	101'460'676	-2.80 (994)	-2.33 (974)	1968	SERP1,EIF2A,SELT,SIAH2,P2Y12R,MED12L, GPR87,IGSF10,MBNL1,P2RY1,DHX36
13	101'460'676	103'875'624	-3.34 (996)	-2.83 (992)	1988	MBNL1,P2RY1,DHX36,**GPR149**,MME,PLCH1, GMPS,KCNAB1
15	73'919'690	76'301'820	-3.18 (1000)	-2.32 (985)	1985	DAPL1,TANC1,BAZ2B,MARCH7,ITGB6,TANK,P SMD14,TBR1,DPP4,GCG,FAP,IFIH1

* position in bp

Conclusion

The cross-breed approach revealed numerous genes related to divergent breeding goals of the GMP versus large breeds (GL and LW) like *MLPH, PLG, PLIN2, GPR149*, and *SOCS2*. The XPEHH approach is shown to be an efficient tool to reveal genes under divergent selection across lines.

Acknowledgements

We are grateful to Sophie Rothammer for helpful discussions on our analysis and results. Ellegaard Göttingen Minipigs ApS (Denmark), Marshall BioResources (USA) for providing DNA samples. German Landrace and Large White samples were genotyped in the research project "geMMA – structural and functional analysis of the genetic variation of the MMA-syndrome" (FKZ0315138), funded by the German Federal Ministry of Education and Research (BMBF).

References

Drögemüller, C., Philipp, U., Haase, B. et al. (2007): A noncoding melanophilin gene (MLPH) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat colour dilution in dogs. Journal of Heredity, 98, 468–473.

Edson, M.A., Lin, Y. & Matzuk, M.M. (2010): Deletion of the novel oocyte-enriched gene, GPR149, leads to increased fertility in mice. Endocrinology, 151, 358–368.

Gandolfi, G., Mazzoni, M., Zambonelli, P. et al. (2011): Perilipin 1 and perilipin 2 protein localization and gene expression study in skeletal muscles of European cross-breed pigs with different intramuscular fat contents. Meat Science, 88, 631–637.

Glodek, P. & Oldigs, B. (1981): Das Göttinger Miniaturschwein. Schriftenreihe Versuchstierkunde 7, Paul Parey Verlag, Berlin.

Greenhalgh, C.J., Rico-Bautista, E., Lorentzon, M. et al. (2005): SOCS2 negatively regulates growth hormone action in vitro and in vivo. J. Clin. Invest, 115, 397-406.

Hill, W.G. & Robertson, A. (1968): Linkage Disequilibrium in finite populations. Theoretical and Applied Genetics, 38, 226-231.

Matesic, L.E., Yip, R., Reuss, A.E. et al. (2001): Mutations in Mlph, encoding a member of the Rab effector family, cause the melanosome transport defects observed in leaden mice. PNAS, 98, 10238–10243.

Metcalf, D., Greenhalgh, C.J., Viney, E. et al. (2000): Gigantism in mice lacking suppressor of cytokine signalling-2. Nature, 405, 1069–1073.

Pickrell, J.K., Coop, G., Novembre, J. et al. (2009): Signals of recent positive selection in a worldwide sample of human populations. Genome Res., 19, 826-837.

Ploplis, V.A., Carmeliet, P., Vazirzadeh, S. et al. (1995): Effects of disruption of the plasminogen gene on thrombosis, growth, and health in mice. Circulation, 92, 2585-2593.

Rothammer, S. (2011): Genomweite Detektion von Selektionssignaturen in divergent selektierten Rinderpopulationen mit anschließender Identifikation eines möglichen kausalen Gens. Dissertation, München, Germany.

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

Sánchez-Martín, M., Pérez-Losada, J., Rodríguez-García, A. et al. (2003): Deletion of the SLUG (SNAI2) gene results in human piebaldism. Am. J. Med. Genet., 122A, 125–132.

Santiago Borrero, P.J., Rodríguez-Pérez, Y., Renta, Y.J. et al. (2006): Genetic testing for Oculocutaneous Albinism type 1and 2 and Hermansky–Pudlak syndrome type 1 and 3 mutations in Puerto Rico. J. Investigative Dermatology, 126, 85–90.

Scheet, P. & Stephens, M. (2006): A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Gene., 78, 629-644.

Simianer, H. & Köhn, R. (2010): Genetic management of the Göttingen Minipig population. J. Pharmacological and Toxicological Methods, 62, 221-226.

5th CHAPTER

General discussion

General Discussion

This thesis has focused on the use of high density markers in genetic characterisation and analysis of farm and experimental animal populations. For this, different genetic markers were used to differentiate populations and to find signatures of recent positive selection. Because many different genetic markers are available, one of our intentions was to compare two of the most commonly used genetic markers. For this, the number of needed SNPs (Single Nucleotide Polymorphisms) to achieve the same differentiation ability of one microsatellite (Single Sequence Repeats, SSR) was calculated with different approaches for eight chicken populations. Based on these results we selected the genetic marker for all prospective studies. In a second step, a whole genome scan for signatures of recent positive selection within and between pig populations was done.

Today, several different genetic markers are available: Restriction Fragment Length Polymorphic markers (RFLP), microsatellites (Single Sequence Repeats, SSR), Amplified Fragment Length Polymorphisms (AFLP), Single Nucleotide Polymorphism (SNP), etc. Due to the effectiveness of genetic markers, the demand of using genomic information is increasing in animal breeding. Two of the today's most important genetic markers in animal breeding are SNPs and SSRs. SNPs are only biallelic and thus less informative (Schaid *et al*. 2004), they have a low mutation rate, a very low false genotyping rate and the high number of SNPs in the genome may compensate the low number of alleles compared to SSRs. Further, an automatic genotyping with high-throughput technologies is easy to carry out (Fries & Durstewitz 2001; Martínez-Arias *et al*. 2001; Xing *et al*. 2005). On the other hand SSRs are highly polymorphic and for a single locus they often show large number of alleles (Bahram & Inoko 2007), while the abundance in the genome is marginal compared to SNPs. SNP arrays based on high-throughput technologies cover up to 60'000 SNPs in pigs, 700'000 SNPs in cattle and more than one million SNPs in humans.

To compare SNPs and SSRs two classical statistical approaches were used in this study: PCA-based partitioning of the distance matrix and a model-based clustering implemented in the software STRUCTURE.

Using PCA for classification purposes is not the first choice, but in practice in many cases, one of the first analyses performed with genetic data. It is well adapted to uncover the population structure even of admixed populations (Paschou *et al*. 2007). For the calculation

of the PCA-based partitioning different criteria were tested, e.g. those suggested by Jolliffe (1972) or Kaiser (1960). For the different criteria the amount of variance explained varied too much between the different marker subsets thus a judicious comparison between different replicates was not possible. The number of principal components used was set to two, because in both marker types the amount of variance explained by the first two principal components was almost the same. The second method used is a model based clustering implemented in STRUCTURE (version 2.3; Pritchard *et al*. 2000). Several studies have used this software for assessing the genetic structure and relatedness within and among populations (e.g. Rosenberg *et al*. 2002; Liu *et al*. 2005; Twito *et al*. 2007; Bodzsar *et al*. 2009). STRUCTURE allows clustering individuals to a defined number of assumed populations.

Because of highly genetic relatedness of two breeds (BL_C and BL_D), the used number of clusters for the STRUCTURE analysis was K=7. Both marker types were able to differentiate the chicken breeds into these seven clusters. The estimated pairwise similarity between repeated STRUCTURE runs for SNPs was significantly better at a given level of clustering compared to SSRs. Both methods improved massively with an increasing number of SNPs and thus we achieved a better insight into the architecture of the breeds.

According to the PCA-based partitioning, about 2.4 SNPs were found to achieve equivalent information content as one SSR. The number of SNPs having the equivalent information content as one SSR varied between different studies and even more between species (Table 1).

Table 1: The number of SNPs per SSR providing equivalent information content for different species.

Species	# of SNPs per microsatellite	Literature
Cattle	2.7	Herráez *et al.* 2005
	3.0-3.4	Schopen *et al.* 2008
Human	4.0-5.0	Lindholm *et al.* 2004
	2.2-2.5	Kruglyak 1997
	5.6	Glaubitz *et al* . 2003
	4.3	Krawczak 1999
	3.8	Thalamuthu *et al.* 2004
	1.9-3.5	Chakraborty *et al.* 1999
Poultry	1.0-2.3	Schopen *et al.* 2008

The highest number of studies comparing SNPs and SSRs were done in humans. Here the number of SNPs needed to replace one SSR varied between 1.9 and 5.56. Hayes *et al*. (2003) indicated that transferring results from human to livestock populations is difficult because of the differences in structure, size, and demography. Even the investigations in livestock about the required number of SNPs do not provide a uniform result. Between 1 and 3.4 SNPs per SSR are necessary in livestock populations.

Regardless of the species it was observed that the number of needed SNPs per SSR depends on the specific genetic properties of the SNPs. The information content of both marker types depends on the genomic position and/or the allele frequency distribution (Chakraborty *et al*. 1999; Xiong & Jin 1999). In our study, the SNPs flanking the SSRs produce a comparable result to randomly chosen sets of SNPs of the same size. Furthermore, the number of SNPs needed to compensate one SSR locus depends on the size of the marker set. Figure 1 shows the number of needed SNP per SSRs for a varying number of SSRs (Schopen *et al*. 2008).

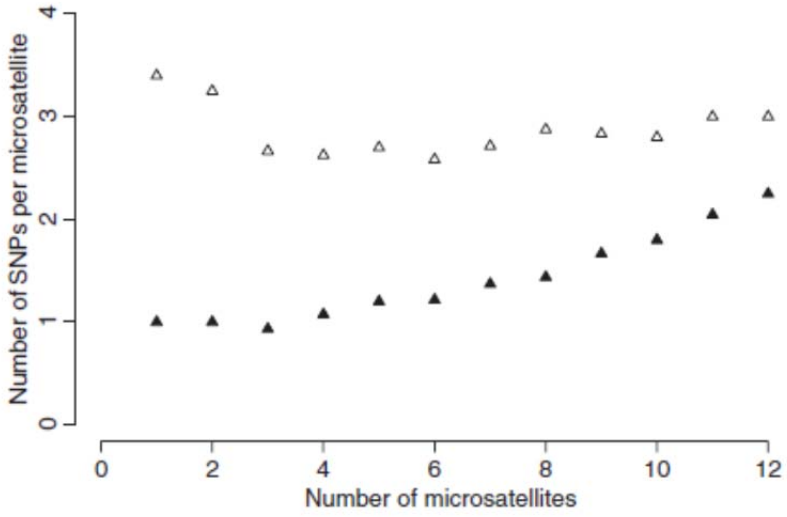

Figure 1: Number of SNPs per SSR needed to obtain the same information content with an increased number of microsatellites for poultry (▲) and cattle (Δ) (Schopen *et al*. 2008).

In poultry the number of SNPs per SSR increased with the total number of SSRs. For one SSR, about one SNP, for six SSRs, about 1.3 SNPs, and for 12 SSRs, on average 2.3 SNPs per SSR were required to achieve equivalent information content. In cattle the number of SNPs was stable or slightly decreases with an increasing number of SSRs. Our results in chapter 2 confirm the finding of Schopen *et al*. (2008). In our study the number of needed SNPs for 10 SSRs was about 2.16 SNPs and for 25 SSRs about 2.3 SNPs per SSR.

Due to specific details of the genetic properties of SNPs and SSRs and the methods used to compare these marker types, it is difficult to provide a uniform result which applies to all species. Our estimate that about 2.4 SNPs are necessary to replace one SSR is comparable to those in the literature. Furthermore, the results suggest that the ability to detect and assess breed differentiation will substantially improve for analyses based on high-throughput SNP genotyping, even with a moderate number of SNPs. When large numbers of SNPs are available they are today's genetic marker of choice for investigations of genetic architecture (Liu *et al*. 2005; Papachristou & Lin 2006; Qanbari *et al*. 2009).

As a result of the comparison of SNPs and SSRs (chapter 2) and the increasing number of SNPs available we had the opportunity to analyse the genome in a dense way with genotypes of the Illumina Porcine SNP60 BeadChip. A whole genome scan for signatures of recent positive selection was done. Genomic regions controlling traits of economic importance, e.g. body size, fertility, or coat colour are expected to exhibit footprints of selective breeding. Their detection is an important tool to identify genes which will improve phenotypes of interest (Hayes *et al*. 2008).

We used the Göttingen Minipig (GMP) for the detection of recent positive selection. The GMP is a synthetic breed derived from three founder breeds and combines the high fertility of the Vietnamese Potbellied Pig (VPP), the low body weight of the Minnesota Minipig (MMP) and the white coat colour of the German Landrace (GL).

Since the creation of the GMPs in the 1960s the breeding goals were aligned to the market demand. After establishing a moderate inbreeding coefficient and stopping the production of coloured GMP in 1992, the breeding goal changed to low body weight. The GMP is a miniature breed and is characterized by an adult body weight of 35-45 kg (Bollen *et al.* 1998). The body size is an important trait in the GMP. It can reduce the costs for experiments when the test compounds are dosed per kg of body weight of the recipient. A further reduction of body size in the future and the finding of genes involved in this trait could therefore be of considerable economical advantage for the GMP. Selection on low body weight resulted in a negatively correlated selection response on litter size (Simianer & Köhn 2010). Similarly, after a 10-year period of selection on low body weight (140-d weight) in MMP a decrease of one piglet per litter was observed (Dettmers *et al*. 1971). In general there is a genetic and physiological antagonism between litter size and body weight in multiparous species. Another more recent breeding objective is for instance a calm temperament, especially in the interaction with humans (Köhn *et al*. 2009).

The present GMP is a white, dwarf animal where all body parts are reduced in size. This type of dwarfism is often caused by growth hormone deficit. Simianer & Köhn (2010) suggested that the insulin-like growth factor 1 (*IGF-1*) may have a major role in small body size of GMPs. Different types of dwarfism are known, most of them being due to mutations in the genome. The most common types of dwarfism in humans are achondroplasia and pituitary dwarfism. Achondroplasia is caused by a mutation of the fibroblast growth factor receptor 3 gene (*FGFR3*). This kind of dwarfism is caused by an abnormal bone growth and a disproportional body (Shiang *et al.* 1994). The pituitary dwarfism, where the body is proportional but the stature is short, is caused by a growth hormone deficiency (Burns 1990). The dwarf gene (DW) in chicken is linked to the sex chromosome and known for many years. The DW genotype results in significantly smaller adult body weight and bone length (Hutt 1959). Sutter *et al.* (2007) found a strong association between the *IGF-1* gene and small body size in dogs. A single *IGF-1* SNP haplotype could be found in almost all small dog breeds and this haplotype is nearly absent in all giant dog breeds.

The GMP has been intensively selected during the last decades, and it has thus achieved tremendous phenotypic changes over the past 30 years of breeding. The identification of selection signatures in the GMP, which are associated with phenotypic changes based on the breeding goals, could improve the breeding of GMPs and thus allowed to align them even better to the market demand.

The search for selection signatures was first carried out in human populations. The human population is assumed to be homogeneous, i.e. it is not a mixture of different ancestral races. In admixed populations signals of recent selection may be masked but the classical methods to detect selection signatures can also be applied (Akey *et al.* 2004; Lohmüller *et al.* 2011). To avoid a possible admixture effect we used different methods for searching signatures of recent positive selection:

- Membership Coefficient (MC)
- Extended Haplotype Homozygosity (EHH)
- Cross Population Extended Haplotype Homozygosity (XPEHH)

The relative composition of the genome arising from the three original breeds should have been maintained on average in a closed population like the GMP under the assumption of the absence of selection and genetic drift (Falconer & Mackay 1996). The idea was to calculate the probability of membership of every allele of the GMPs in one of the three founder breeds. A membership coefficient (MC) for each SNP was calculated. We expected that a genomic region carrying a relevant allele would harbour a higher proportion of one founder breed the respective allele originates from, than on average in the whole genome. When a genomic region carries for instance a relevant allele for the trait 'white skin', we expect that the proportion of the GL which is responsible for this characteristic will be overrepresented compared to its average proportion in the whole genome. Genetic drift could cause variability of breed composition if only small chromosome segments are considered. It should not have a large impact across the whole genome in the GMP, due to the active management avoiding a high inbreeding rate. So it is assumed that selection is the genetic mechanism having a long range impact on the composition of the genome. This shift in allelic frequency should be observed for a chromosome region in linkage disequilibrium with a respective candidate gene under selection.

Sabeti *et al*. (2002) have suggested the extended haplotype homozygosity (EHH) statistic for the detection of 'selective sweeps'. This method reflects a fast increase in allele frequency of a core region and a surrounding long conserved haplotype. The EHH test is intended to identify regions that show allelic frequencies which have increased faster than expected only due to drift and selection. To correct the EHH test for the local variability in recombination rates the 'Relative Extended Haplotype Homozygosity' (REHH) was developed. Several authors used this approach in humans to find disease genes or to detect

population genetic structure (e.g. Sabeti *et al.* 2002; Oleksyk *et al*. 2010). Hayes *et al.* (2008) and Qanbari *et al*. (2009) used this approach to detect genes that might reflect selection on important economic breeding traits in cattle.

The third method used to detect signatures of recent selection is the Cross Population Extended Haplotype Homozygosity (XPEHH) test (Sabeti *et al*. 2007). This test is based on the EHH test with the enhancement of detecting positive selection by comparing two populations. If a selected allele achieved fixation in one population, the EHH test is not able to detect selection for this region. If a selected allele has almost achieved fixation in one population, but is still polymorphic in the second population (and thus in both populations together), the XPEHH can detect recent positive selection. The XPEHH test was used in different human populations as well as in cattle populations to detect breed (regional) specific signatures of recent selection (c.f. Pickrell *et al*. 2009; Bray *et al*. 2010; Noyes *et al*. 2011; Rothammer 2011). This method was used to verify the results of the two methods previously explained. Thus, the GMPs were compared to two 'normal sized' breeds German Landrace (GL) and Large White (LW) to check for signatures of recent selection, especially for regions related to growth or coat colour traits.

In chapter 3 and 4 genotypes of the Illumina Porcine SNP60 BeadChip were used to find signatures of recent positive selection. After the required filtering criteria (SNPs with unknown chromosome or position, call-rates < 95% or monomorphic markers were excluded) for the respective test, an additional test of the correct positions of the SNPs was carried out. The position of many SNPs changes between different genome builds. Therefore, a test for correct SNP position based on linkage disequilibrium (LD) was performed. To this end, haplotypes were reconstructed using fastPHASE (Scheet & Stephens 2006) and the LD was estimated using the parameter r^2 (Hill & Robertson 1968) within the breeds. To identify LD outliers, the LD was plotted against the physical distance for each chromosome. When plotting the r^2 against the position, an asymptotic value was expected to be achieved for great distances. All pairs of SNPs with r^2-values exceeding the mean plus ten standard deviations in the asymptotic region were considered as outliers. If a SNP was involved in two or more such outliers, it was removed from the dataset. For the calculation of selection signatures between breeds, the LD filtering has been done in each breed and SNPs involved in outliers of two or more breeds were deleted. After applying this procedure, a total of 3'300 SNPs were removed to search for selection signatures within the GMPs and 2'745 SNP to search for selection signatures between breeds. Of

course not all SNPs with wrong positions could be detected, but this is a recommendable pragmatic approach to decrease the number of SNPs with a presumable wrong position.

To annotate genes close to the regions of recent selection, the map viewer option of the porcine genome sequence assembly was used. Starting from the selection signature, the region was expanded 1 Mb up and downstream to find candidate genes. A distance of 1 Mb around the detected signal was assumed because of the large extent of LD in livestock populations compared to the human population (Qanbari *et al*. 2009).

Table 2: Most interesting genes found for the selection signatures of the Göttingen Minipig.

Chr	Gene	Function	Method for detection	Reports in literature
1	*PLG*	growth	XPEHH	Ploplis *et al.* 1995
1	*TXN / TRN*	growth	EHH; EHH/MC	Heppell-Parton *et al.* 1995 Yu *et al.* 2007
4	*DDR2*	body size	MC; EHH/MC	Labrador *et al.* 2001 Kano *et al.* 2008
4	*SNAI2 / SLUG*	coat colour	XPEHH	Sánchez-Martín *et al* . 2003
5	*SOCS2*	body size	EHH; XPEHH	Favre *et al.* 1999 Metcalf *et al.* 2000 Greenhalgh *et al.* 2002 Alexander & Hilton 2004 Greenhalgh *et al.* 2005 Piper *et al.* 2005
6	*MTOR*	growth	EHH/MC	Hsu *et al.* 2011 Zoncu *et al.* 2011
7	*BMP6*	bone growth	EHH	Linkhart *et al.* 1996 Jane *et al.* 2002
9	*GRB10*	growth	EHH/MC	Charalambous *et al* . 2003 Deng *et al.* 2003 Shiura *et al.* 2005 Wang *et al.* 2007 Smith *et al.* 2007
9	*GAB2*	growth	EHH	Lock *et al.* 2002
13	*GPR149*	fertility	XPEHH	Edson *et al.* 2010
13	*HPS3*	albinism	XPEHH	Santiago-Borrero *et al.* 2006
15	*MLPH*	coat colour	XPEHH	Matesic *et al.* 2001 Drögemüller *et al.* 2007
16	*PRLR*	fertility	MC; EHH/MC	Drögemüller *et al.* 2001 Putnová *et al.* 2002

Some regions of recent selection overlapped with genes of potential biological relevance for the GMPs. Table 2 displays the most interesting genes found by at least one of the tests (MC, EHH and XPEHH). For searching selection signatures within the GMP, the MC and the EHH test were combined to avoid a possible cross-breeding effect. For this the results

of both tests were plotted against each other. Genes overlapping with signatures of this combination are marked with EHH/MC. Several regions harbouring candidate genes which appear to be functionally related to breeding goals of the Göttingen Minipig could be identified, e.g. *SOCS2*, *GRB10* and *DDR2* linked to body size, *PRLR* and *GPR149* related to fertility traits, or *SNAI2*, *HPS3* and *MLPH* for coat colour traits and some others with putative regions suggested being under selection in the GMP.

On chromosome 9 the growth factor receptor-bound protein 10 (*GRB10*) exhibited a signal of positive selection. A disruption in the *GRB10* gene causes an overgrowth in mice up to 30% compared to normal mice (Charalambous *et al*. 2003). The functional role of the *GRB10* gene is controversially discussed in several studies: It is still unclear if the *GRB10* gene regulates the insulin-like growth factor-1 (*IGF-1*) negatively or positively (c.f. Shiura *et al*. 2005; Wang *et al*. 2007; Smith *et al*. 2007; Deng *et al*. 2003).

The *GRB10* gene was detected with the combination of MC and EHH tests, but neither of the tests on its own could detect the region. Only the combination of these two tests facilitated avoiding a possible admixture effect and made detection possible. *GRB10* strongly influences animal growth and might be a reason for the small body size of the GMP. Furthermore, the mechanistic target of the rapamycin gene (*MTOR*) on chromosome 6, which influences the growth factor signalling and *GRB10* (Hsu *et al*. 2011; Zoncu *et al*. 2011) could be found. The connection confirmed our assumption about the influence of *GRB10* for the small body size in the GMP.

One of the most interesting discoveries was the signal for of the suppressor of cytokine signaling-2 (*SOCS2*) gene on chromosome 5 (Piper *et al*. 2005). Metcalf *et al.* (2000) and Greenhalgh *et al*. (2002) detected a negative regulation of the *SOCS2* gene with the growth hormone (*GH*) and *IGF-1 gene*. In mice an excessive growth phenotype characterized by a 30–50% increase in mature body size (Greenhalgh *et al.* 2005) was reported, when other authors suggested that the *SOCS2* gene can both positively and negatively regulate the body size (Favre *et al*. 1999; Alexander & Hilton 2004). The methods used in this study detected the region of the *SOCS2* gene as a strong signature of recent selection for the EHH test within the GMP as well as XPEHH test between the GMP and two 'normal sized' breeds (GL and LW). Though, with the XPEHH test, the signal of the *SOCS2* gene was found for the 'normal sized' breeds. Our findings confirm the results of Favre *et al*. (1999) and Alexander & Hilton (2004) that the *SOCS2* gene can both positively and negatively regulate the body size. The *SOCS2* gene also has important effects in the

regulation of other processes such as metabolism, cancer or the response to infection (Rico-Bautista *et al*. 2006), but the regulation of the *GH* and *IGF-1* genes by the *SOCS2* gene might be one important cause for the reduced body size of the Göttingen Minipigs. Simianer and Köhn (2010) mentioned a possible influence of the *IGF-1* gene for the small body size in the GMP and the finding of the *SOCS2* gene supports this suggestion. However, to confirm the previous findings and their roles directly or indirectly in the GMP more reference data (larger dataset) will be required.

General conclusion

The number of SNPs needed to achieve comparable differentiation ability as one microsatellite is about 2.4, but it is difficult to provide a uniform result for all species and number of markers. However, the PCA-based partitioning of the distance matrix is a good technique to detect and measure differentiation between breeds even with low numbers of SNPs. Because large numbers of SNPs are available in combination with easy automatic genotyping using high-throughput technologies, SNPs are presently the genetic markers of choice for investigation of genetic architecture (c.f. Liu *et al*. 2005; Papachristou & Lin 2006; Qanbari *et al.* 2009). Due to the availability of the complete genome sequence assembly the next step could be a genotyping by sequencing.

Based on the results comparing SNPs and SSRs we decided to use dense SNP genotypes to identify signatures of recent positive selection that potentially contain genes contributing to within and inter-breed phenotypic variation. To avoid a possible cross-breeding effect where admixture may mask selection signatures, a combination of the membership coefficient and the Extended Haplotype Homozygosity was used. All three methods of detecting selection signatures identified several regions harbouring candidate genes which appear to be functionally related to breeding goals of the GMP, e.g. *PRLR* and *GPR149* related to fertility traits, *SOCS2*, *GRB10* and *DDR2* linked to body size, or *SNAI2*, *HPS3* and *MLPH* for coat colour traits. The finding of the *SOCS2* gene within the GMPs as well as between the GMP and the 'normal sized' breeds increase the relevance of this gene for the small body size of the GMP. The pathway connecting between *SOCS2* and *GRB10* with the *IGF-1* gene might be causal for the small body size of the Göttingen Minipigs.

Many of the regions showing extreme values of the used statistics seem to play important roles in economically important traits of the GMPs. These regions can now serve as starting points to improve the breeding of GMPs and thus to align them even better to the

market demand. It could also be interesting to proceed to fine mapping studies to see if they confirm our results.

It can be concluded that SNP-based approaches allow a much better insight in the genomic architecture of populations. Further, the uses of SNP-based approaches improves the understanding of the genetic mechanisms underlying selection and breed differentiation.

References

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. & Kruglyak, L. (2004): Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol., 2, 10, 1591-1599.

Alexander, W.S. & Hilton, D.J. (2004): The role of suppressors of cytokine signalling (SOCS) proteins in regulation of the immune response. Annu. Rev. Immunol., 22, 503–529.

Bahram, S. & Inoko, H. (2007): Microsatellite markers for genome-wide association studies. Nat. Rev. Genet., 8, doi:10.1038/nrg1962-c1.

Bodzsar, N., Eding, H., Revay, T., Hidas, A. & Weigend, S. (2009): Genetic diversity of Hungarian indigenous chicken breeds based on microsatellite markers. Anim. Genet., 40, 516-523.

Bollen, P., Andersen, A. & Ellegaard, L. (1998): The behaviour and housing requirements of minipigs. Scand. J. Lab. Anim. Sci. Suppl., 25, 23-26.

Bray, S.M., Mulle, J.G., Dodd, A.F., Pulver, A.E., Wooding, S. & Warren, S.T. (2010): Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. PNAS, 37, 16222-16227.

Burns, H. (1990): Growth promoters in humans. Proc. Nutr. Soc., 49, 467-472.

Chakraborty, R., Stivers, D.N., Su, B., Zhong, Y., Zhong, Y. & Budowle, B. (1999): The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis, 20, 1682-1696.

Deng, Y., Bhattacharya, S., Swamy, O.R., Tandon, R., Wang, Y., Janda, R. & Riedel, R. (2003): Growth factor receptor-binding protein 10 (Grb10) as a partner of phosphatidylinositol 3-kinase in metabolic insulin action. J. Biol. Chem., 278, 39311–39322.

Dettmers, A.E., W.E. Rempel & Hacker, D.E. (1971): Response to current mass selection for small size in swine. J. Anim Sci., 33, 212-215.

Drögemüller, C., Hamann, H. & Distl, O. (2001): Candidate gene markers for litter size in different German pig lines. J. of Anim. Sci., 79, 2565–70.

Drögemüller, C., Philipp, U., Haase, B., Günzel-Apel, A.-R., & Leeb, T. (2007): A noncoding melanophilin gene (MLPH) SNP at the splice donor of exon 1 represents a candidate causal mutation for coat colour dilution in dogs. J. Hered., 98, 468–473.

Edson, M.A., Lin, Y. & Matzuk, M.M. (2010): Deletion of the novel oocyte-enriched gene, GPR149, leads to increased fertility in mice. Endocrinology, 151, 358–368.

Favre, H., Benhamou, A., Finidori, J., Kelly, P.A. & Edery, M. (1999): Dual e¡ects of suppressor of cytokine signaling (SOCS-2) on growth hormone signal transduction. FEBS Letters, 453, 63-66.

Falconer, D. S. & Mackay, T.F.C. (1996): Introduction to quantitative genetics. Longmans Green, Edition 4, Harlow, Essex, UK.

Fries R. & Durstewitz, G. (2001): Digital DNA signatures for Animal Tagging. Nat. Biotech., 19, 508.

Glaubitz, J.C., Rhodes, E. & Dewoody, J.A. (2003): Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Mol. Ecol., 12, 1039-1047.

Greenhalgh, C.J., Bertolino, P., Asa, S.L., Metcalf, D.,Corbin, J.E., Adams, T.E., Davey, H.E., Nicola, N.A., Hilton. D.J. & Alexander, W.S. (2002): Growth enhancement in suppressor of cytokine signaling 2 (SOCS-2)-deficient mice is dependent on signal transducer and activator of transcription5b (STAT5b). J. Mol. Endocrinol., 16, 1394–1406.

Greenhalgh, C.J., Rico-Bautista, E., Lorentzon, M., Thaus, A.L., Morgan, P.O., Willson, T.A., Zervoudakis, P., Metcalf, D., Street, I., Nicola, N.A., Nash, A.D., Fabri, L.J., Norstedt, G., Ohlsson, C., Flores-Morales, A., Alexander, W.S. & Hilton, D.J. (2005): SOCS2 negatively regulates growth hormone action in vitro and in vivo. J. Clin. Invest, 115, 397-406.

Hayes, B.J., Visscher, P.M., McPartlan, H.C. & Goddard, M.E. (2003): Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res., 13, 635–643.

Hayes, B. J., Lien, S., Nilsen, H., Olsen, H. G., Berg, P., Maceachern, S., Potter, S. & Meuwissen, T. H. (2008): The origin of selection signatures on bovine chromosome 6. Anim. Genet., 39, 105-111.

Heppell-Parton, A., Cahn, A., Bench, A., Lowe, N., Lehrach, H., Zehetner, G. & Rabbitts, P. (1995): Thioredoxin, a Mediator of Growth Inhibition, Maps to 9q31. Genomics, 26, 379-381.

Herráez, D.L., Schäfer, H., Mosner, J., Fries, H.R. & Wink, M. (2005): Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. Z. Naturfo., 60c, 637–643.

Hill, W.G. & Robertson, A. (1968): Linkage Disequilibrium in finite populations. Theor. Appl. Genet., 38, 226-231.

Hsu, P.P., Kang, S.A., Rameseder, J., Zhang, Y., Ottina, K.A., Lim, D., Peterson, T.R., Choi, Y., Gray, N.S., Yaffe, M.B., Marto, J.A. & Sabatini, D.M. (2011): The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signalling. Science, 332, 1317-1322.

Hutt, F.B. (1959): Sex-linked dwarfism in the fowl. J. Hered., 50, 209-221.

Jane, J.A., Dunford, B.A., Kron, A., Pittman, D.D., Sasaki, T., Li, J.Z., Li, H., Alden, T.D., Dayoub, H., Hankins, G.R., Kallmes, D.F. & Helm, G.A. (2002): Ectopic osteogenesis using adenoviral bone morphogenetic protein (BMP)-4 and BMP-6 gene transfer. Mol. Therapy, 6, 4, 464-470.

Jolliffe, I.T. (1972): Discarding variables in a principal component analysis, I: Artificial data. Appl. Statist., 21, 160-173.

Kaiser, H. F. (1960): The application of electronic computers to factor analysis. Educ. Psychol. Meas., 20, 141-151.

Kano, K., Marin de Evsikova, C., Young, J., Wnek, C., Maddatu, T.P., Nishina, P.M. & Naggert, J.K. (2008): A novel dwarfism with gonadal dysfunction due to loss-of-function allele of the collagen receptor gene, DDR2, in the mouse. Mol. Endocri., 22, 1866–1880.

Köhn, F., Sharifi, A.R. & Simianer, H. (2009): Genetic analysis of reactivity to humans in Göttingen Minipigs. Appl. Anim. Beh. Sci., 120, 68–75.

Krawczak, M. (1999): Informatively assessment for biallelic single nucleotide polymorphisms. Electrophoresis, 20, 1676-1681.

Kruglyak, L. (1997): The use of a genetic map of biallelic markers in linkage studies. Nat. Genet., 17, 21-24.

Labrador, J.P., Azcoitia, V., Tuckermann, J., Lin, C., Olaso, E., Manes, S., Bruckner, K., Goergen, J.L., Lemke, G., Yancopoulos, G. Angel, P., Martínez-A, C. & Klein, R. (2001): The collagen receptor DDR2 regulates proliferation and its elimination leads to dwarfism. Embo. Rep., 2, 446–452.

Lindholm, E., Hodge, S.E. & Greenberg, D.A. (2004): Comparative informativeness for linkage of multiple SNPs and single microsatellites. Hum. Hered., 58, 164–170.

Linkhart, T.A., Mohan, S. & Baylink, D.J. (1996): Growth factors for bone growth and repair, IGF, TGFβ and BMP. Bone, 19, 1-12.

Liu, N., Chen, L., Wang, S., Oh, C. & Zhao, H. (2005): Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet., 6 Suppl 1, S26.

Lohmüller, K.E., Bustamante, C.D. & Clark, A.G. (2011): Detecting directional selection in the presence of recent admixture in African-Americans. Genetics, 187, 823–835.

Lock, L.S., Maroun, C.R., Naujokas, M.A. & Park, M. (2002): Distinct Recruitment and function of Gab1 and Gab2 in Met Receptor-mediated Epithelial Morphogenesis. Mol. Biol. Cell, 13, 2132–2146.

Matesic, L.E., Yip, R., Reuss, A.E., Swing, D.A., O'Sullivan, T.N., Fletcher, C.F., Copeland, N.G. & Jenkins, N.A. (2001): Mutations in Mlph, encoding a member of the Rab effector family, cause the melanosome transport defects observed in leaden mice. PNAS, 98, 10238–10243.

Martínez-Arias R, Calafell, F., Mateu, E., Comas, D., Andrés, A. & Bertranpetit, J. (2001): Sequence variability of a human pseudogene. Genome Res., 11, 1071-1085.

Metcalf, D., Greenhalgh, C.J., Viney, E., Wilson, T., Starr, R., Nicola, N., Hilton, D., Alexander, W.S. (2000): Gigantism in mice lacking suppressor of cytokine signalling-2. Nature, 405, 1069–1073.

Noyes, H., Brass, A., Obara, I., Anderson, S., Archibald, A.L., Bradley, D.G., Fisher, P., Freeman, A., Gibson, J., Gicheru, M., Hall, L., Hanotte, O., Hulme, H., McKeever, D., Murray, C., Jung Oh, S., Tate, C., Smith, K., Tapio, M., Wambugu, J., Williams, D.J., Agaba, M. & Kemp, S.J. (2011): Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. PNAS, 22, 9304-9309.

Oleksyk, T.K., Smith, M.W & O'Brien, S.J. (2010): Genome-wide scans for footprints of natural selection. Phil. Trans. R. Soc. B., 365, 185–205.

Papachristou, C. & Lin, S. (2006): Microsatellites versus Single-Nucleotide Polymorphisms in confidence interval estimation of disease loci. Genet. Epidem., 30, 3–17.

Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W. & Drineas, P. (2007): PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet., 3, 1672-1686.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. & Pritchard, J.K. (2009): Signals of recent positive selection in a worldwide sample of human populations. Genome Res., 19, 826-837.

Piper, E., Chen, Y. & Moran, C. (2005): Assignment of suppressor of cytokine signalling-2 (SOCS2) to porcine chromosome 5 with radiation hybrids. Cytogenet. Genome Res., 111, 96B.

Ploplis, V.A., Carmeliet, P., Vazirzadeh, S., Van Vlaenderen, I., Moons, L., Plow, E.F. & Collen, D. (1995): Effects of disruption of the plasminogen gene on thrombosis, growth, and health in mice. Circulation, 92, 2585-2593.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000): Inference of population structure using multicocus genotype data. Genetics, 155, 945-959.

Putnová , L., Knoll, A., Dvorák, J. & Cepica, S. (2002): A new HpaII PCR-RFLP within the porcine prolactin receptor (PRLR) gene and study of its effect on litter size and number of teats. J. Anim. Breed. Genet., 119, 57–63.

Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R. & Simianer, H. (2009): A genome-wide scan for signatures of recent selection in Holstein cattle. Anim. Genet., 41, 377-389.

Rico-Bautista, E., Flores-Morales, A. & Fernández-Pérez, L. (2006): Suppressor of cytokine signaling (SOCS) 2, a protein with multiple functions. Cytokine Growth F. Rev., 17, 431–439.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. & Feldman, M.W. (2002): Genetic structure of human populations. Science, 298, 2381-2385.

Rothammer, S. (2011): Genomweite Detektion von Selektionssignaturen in divergent selektierten Rinderpopulationen mit anschließender Identifikation eines möglichen kausalen Gens. Diss. med. vet. LMU München, Germany.

Sabeti, P.C., Reich, D.E., Higgins, J.M. Levine, H., Richter, D.J., Schaffner, S., Gabriel, S.B., Platko, J., Patterson, N.J., McDonald, G.J., Ackerman, H., Campbell, S.J., Altshuler, D., Cooperk, R., Kwiatkowski, D., Ward, R. & Lander, E.S. (2002): Detecting recent positive selection in the human genome from haplotype structure. Nature, 419, 832–837.

Sabeti, P.C., Varilly, P., Fry, B., Lohmüller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A. Gaudet, R., Schaffner, S.F., Lander, E.S. & The International HapMap Consortium (2007): Genome wide detection and characterization of positive selection in human populations. Nature, 449, 913-918.

Sánchez-Martín, M., Pérez-Losada, J., Rodríguez-García, A., González-Sánchez, G., Korf, B.R., Kuster, W., Moss, C., Spritz, R.A. & Sánchez-García, I. (2003): Deletion of the SLUG (SNAI2) gene results in human piebaldism. Am. J. Med. Genet., 122A, 125–132.

Santiago Borrero, P.J., Rodríguez-Pérez, Y., Renta, Y.J., Izquierdo, N.J., del Fierro, L., Muñoz, D., Molina, N.L., Ramírez, S., Pagán-Mercado, G. Ortíz1, I., Rivera-Caragol, E., Spritz, R.A. & Cadilla, C.L. (2006): Genetic testing for Oculocutaneous Albinism type 1and 2 and Hermansky–Pudlak syndrome type 1 and 3 mutations in Puerto Rico. J. Invest. Dermatol., 126, 85–90.

Schaid, D.J., Guenther, J.C., Christensen, G.B., Hebbring, S., Rosenow, C., Hilker, C.A., McDonnell, S.K., Cunningham, J.M., Slager, S.L., Blute, M.L. & Thibodeau, S.N. (2004): Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer–susceptibility loci. Am. J. Hum. Genet., 75, 948–965.

Scheet, P. & Stephens, M. (2006): A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet., 78, 629-644.

Schopen, G.C.B., Bovenhuis, H., Visker, M.H.P.W. & van Arendonk, J.A.M. (2008): Comparison of information content for microsatellites and SNPs in poultry and cattle. Anim. Genet., 39, 451-453.

Shiang, R., Thompson, L., Zhu, Y., Church, D., Fielder, T., Bocian, M., Winokur, S. & Wasmuth. J. (1994): Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. Cell, 78, 335-342.

Shiura, H., Miyoshi, N., Konishi, A., Wakisaka-Saito, N., Suzukif R., Muguruma, K., Kohda, T., Wakana, D., Yokoyama, M., Ishino, F. & Kaneko-Ishino, T. (2005): Meg1/Grb10 overexpression causes postnatal growth retardation and insulin resistance via negative modulation of the IGF1R and IR cascades. Biochem. a. Biophy., 329, 3, 909-916.

Simianer, H. & Köhn, F. (2010): Genetic management of the Göttingen Minipig population. J. Pharmacol. Toxicol. Methods, 62, 3, 221-226.

Smith, F.M., Holt, L.J., Garfield, A.S., Charalambous, M., Koumanov, F., Perry, M., Bazzani, R., Sheardown, S.A., Hegarty, B.D., Lyons, R.J., Cooney, G.J., Daly, R.J. & Ward, A. (2007): Mice with a disruption of the imprinted GRB10 gene exhibit altered body composition, glucose homeostasis, and insulin signalling during postnatal life. Mol. A. Cell. Biol., 27, 16, 5871–5886.

Sutter, N.B., Bustamante, C.D., Chase, K., Gray, M.M., Zhao, K., Zhu, L., Padhukasahasram, B. Karlins, E., Davis, S., Jones, P.G., Quignon, P., Johnson, G.S., Parker, H.G., Fretwell, N., Mosher, D.S., Lawler, D.F., Satyaraj, E., Nordborg, M., Lark, K.G., Wayne, R.K. & Ostrander, E.A. (2007): A single IGF1 allele is a major determinant of small size in dogs. Science, 316, 112-115.

Thalamuthu, A., Mukhopadhyay, I., Ray, A. & Weeks, D.E. (2004): A comparison between microsatellite and single-nucleotide polymorphism markers with respect to two measures of information content. BMC Genet., 6 Suppl 1, S27.

Twito, T., Weigend, S., Blum, S., Granevitze, Z., Feldman, M.W., Perl-Treves, R., Lavi, U. & Hillel, J. (2007): Biodiversity of 20 chicken breeds assessed by SNPs located in gene regions. Cytogenet. Genome Res., 117, 319-326.

Wang, L., Balas, B., Christ-Roberts, C.Y., Kim, R.Y., Ramos, F.J., Kikani, C.K., Li, C., Deng, C., Reyna, S., Musi, N., Dong, L.Q., DeFronzo, R.A. & Liu, F. (2007): Peripheral disruption of the Grb10 gene enhances insulin signaling and sensitivity in vivo. Mol. Cell. Biol., 27, 6497–6505.

Xing, C., Schumacher, F.R., Xing, G., Lu, Q., Wang, T. & Elston, R.C. (2005): Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. BMC Genet., 6 Suppl 1, S29.

Xiong, M. & Jin, L. (1999): Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. Am. J. Hum. Genet., 64, 629-640.

Yu, M., Geiger, B., Deeb, N. & Rothschild, M.F. (2007): Investigation of TXNIP (thioredoxin-interacting protein) and TRX (thioredoxin) genes for growth-related traits in pigs. Mamm Genome 18, 197-209.

Zoncu, R., Efeyan, A. & Sabatini, D.M. (2011): mTOR: from growth signal integration to cancer, diabetes and ageing. Nat. Rev. Mol. Cell. Biol., 12, 21-36.

I would like to thank:

Henner Simianer for the opportunity to work on this exciting topic, for acting as my main supervisor and for accepting this thesis.

Georg Thaller for acting as second supervisor and for accepting this thesis

Christoph Knorr for acting as third supervisor.

Ellegaard Göttingen Minipigs ApS, Marshall Farms Inc, Sinclair Research Center and the Tierpark Berlin-Friedrichsfelde for providing the minipig data. Special thanks to Jens Ellegaard and his family for a very interesting and motivating week in September 2009.

Friederike Köhn, Helge Täubert and Ralf Fischer for helping me with all problems concerning the minipig database.

Oskar Lippstreu and Knut Salzmann for the help with the minipigs in Relliehausen.

Reza Sharifi, Steffen Weigend and Michael Auwers for all help and advices.

All colleagues from the Dep. of Animal Sciences, especially the Animal Breeding and Genetics Group, for help, inspiration and funny time. Especially Ute Döring for helping me any time I needed it and for her never ending patience.

Florence Ytournel for sharing the office and for your help. You always answered all my questions, no matter how stupid they were. Thank you for three great years.

Christina, Claudia K., Nina, Patricia, Florian and Mahmood for motivation, funny lunches, office games and interesting talks. You made every boring working day much more colourful.

Thierry for accompany me all the scientific and non-scientific way. I could not imagine a better roommate.

Birthe for her patience, motivation and helpful advices in all situations. You are enriching my life and words can just hardly express my gratefulness.

Meiner Familie danke ich für ihre unglaubliche Unterstützung und die nicht endende Geduld. Ohne Euer Vertrauen und nicht zuletzt die finanzielle Unterstützung, wäre das alles nicht möglich gewesen. Ihr habt mich immer wissen lassen, dass alles gut wird.

For those people whose names are missing in this chapter, please accept my sincere gratitude.