

Institute of Animal Science  
Department of Animal Genetics and Breeding  
Prof. Dr. Jörn Bennewitz



UNIVERSITY OF  
HOHENHEIM

# **Using genome-wide association studies to map genes for complex traits in porcine F2 crosses**

Dissertation

submitted in fulfillment of the regulations to acquire the degree

Doktor der Agrarwissenschaften

(Dr. sc. agr. in Agricultural Science)

to the

Faculty of Agricultural Science

presented by

Markus Schmid

born in Böblingen

Baden-Württemberg

2018

The dissertation was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).



Institute of Animal Science  
Department of Animal Genetics and Breeding  
Prof. Dr. Jörn Bennewitz

# Using genome-wide association studies to map genes for complex traits in porcine F2 crosses

Dissertation

submitted in fulfillment of the regulations to acquire the degree

Doktor der Agrarwissenschaften

(Dr. sc. agr. in Agricultural Science)

to the

Faculty of Agricultural Science

presented by

Markus Schmid

born in Böblingen

Baden-Württemberg

2018

Day of the Oral Examination: July 6, 2018

Dean of the Faculty:

Prof. Dr. Ralf T. Vögele

Chairman of the Examining Board:

Prof. Dr. Markus Rodehutschord

Supervisor and Reviewer:

Prof. Dr. Jörn Bennewitz

Second Reviewer:

Prof. Dr. Georg Thaller

Additional Examiner:

Prof. Dr. Karl Schmid

The dissertation was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).



**TABLE OF CONTENTS**

GENERAL INTRODUCTION .....5

GENERAL SUMMARY (ENGLISH) .....9

GENERAL SUMMARY (GERMAN).....11

CHAPTER ONE .....15

CHAPTER TWO.....43

CHAPTER THREE.....65

CHAPTER FOUR.....97

GENERAL DISCUSSION.....109

ACKNOWLEDGMENTS.....123

LIST OF PUBLICATIONS .....125

CURRICULUM VITAE .....127



## GENERAL INTRODUCTION

Performance traits are environmentally influenced quantitative genetic traits and therefore do not solely depend on genes, resulting in a moderate heritability around 20 to 50 % in pigs (Clutter 2011). In order to improve the breeding success of such traits in economically relevant pig breeds (e.g. Piétrain), individuals are selected based on their genomic estimated breeding values (GEBVs) (Wellmann et al. 2013). A maximally high accuracy of the GEBVs is indispensable for making correct selection decisions. Accordingly, an increase of accuracy is still of great interest in animal breeding research. Results of genome-wide association studies (GWASs) provide knowledge about the genetic architecture of quantitative traits and may improve the accuracy of genomic selection (Goddard et al. 2016). In the past, besides mapping experiments with purebred populations, many porcine F2 cross designs were conducted to map genomic regions affecting complex traits. The founder breeds of such experimental designs were frequently chosen from divergent lineages, e. g. one European type breed and one Asian type breed, to create maximally informative F2 individuals for linkage analysis (Rothschild et al. 2007, Frantz et al. 2013, Rückert & Bennewitz 2010). Also F2 designs with two European type founder breeds were established (Boysen et al. 2010). Such family designs were required since linkage analysis relies on linkage information, and linkage phases are not consistent across families. Whereas large phenotypic data is available for many traits, genotype information was often limited in those studies due to the sparse resolution of microsatellite markers in the genome. Another issue, making precise mapping difficult in linkage analysis, is the lack of information coming from historical meiosis. Applied to F2 data, solely meiosis of two generations remain as information. In contrast, GWASs use all meiosis (also historical) and are based on linkage disequilibria (LD) between causative mutations and single nucleotide polymorphism (SNP) markers.

In the present research, the use of existing porcine F2 data (Rückert & Bennewitz 2010, Boysen et al. 2010) to map quantitative trait loci (QTL) in the genomic era was investigated. A special focus was on mapping genes that also segregate within the sire line Piétrain since genomic selection is implemented in this breed (Wellmann et al. 2013). The evaluations were based on data of four consisting F2 crosses, three crosses had Piétrain as one founder breed (Rückert & Bennewitz 2010, Boysen et al. 2010). The datasets contained phenotypic data of up to 230 traits, however they originally were sparsely genotyped using microsatellite markers as they were established for linkage mapping and SNP chips were not available back then. To update the

genotypic information into the genomic era, the genotypes were extended towards dense SNP chip information using the Porcine60k BeadChip (Ramos et al. 2009).

**Chapter 1** is a review article of statistical models and experimental populations applied in livestock GWASs. This chapter gives an overview of methods to conduct GWASs using single-marker models and multi-marker models. Further, approaches taking non-additive genetic effects or genotype-by-environment interactions into account are described. Finally, post-GWAS analysis possibilities and GWAS mapping populations are discussed.

In **chapter 2**, the power and precision of GWASs in F2 populations derived from closely and distantly related founder breeds, as well as a segregating population, was investigated using simulated whole-genome sequence data. Further, the effect of pooling data was determined.

Within and across LD structures of such F2 populations were examined in **chapter 3** by separately and jointly (pooled dataset) analyzing the existing F2 datasets mentioned above. The pooled dataset was also used to map QTL for economically important traits applying single-marker and Bayesian multi-marker regressions.

To infer the suitability of F2 data to map genes in a segregating breeding population, GWAS results of a pooled F2 cross were validated in two samples of the German Piétrain population (**chapter 4**).

The thesis ends with a **general discussion**. This section provides additional GWAS results for seldom investigated traits (e.g. organ weights), debates the use of F2 data for GWAS and the possibilities for the application in a purebred breed, and proposes future research directions.

**References**

- Boysen T.J., Tetens J. & Thaller G. (2010). Detection of a quantitative trait locus for ham weight with polar overdominance near the ortholog of the callipyge locus in an experimental pig F2 population. *J Anim Sci* 88: 3167–3172.
- Clutter A.C. (2011). Genetics of Performance Traits. In: Rothschild M., Ruvinsky A. (Eds.). *The Genetics of the Pig*. Cambridge, 325–354.
- Frantz L.A., Schraiber J.G., Madsen O., Megens H.J., Bosse M., Paudel Y., Semiadi G., Meijaard E., Li N., Crooijmans R.P., Archibald A.L., Slatkin M., Schook L.B., Larson G. & Groenen M.A. (2013). Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*. *Genome Biol* 14: R107.
- Goddard M.E., Kemper K.E., MacLeod I.M., Chamberlain A.J. & Hayes B.J. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc Biol Sci* 283: 20160569.
- Ramos A.M., Crooijmans R.P.M.A., Affara N.A., Amaral A.J., Archibald A.L., Beever J.E., Bendixen C., Churcher C., Clark R., Dehais P., Hansen M.S., Hedegaard J., Hu Z.L., Kerstens H.H., Law A.S., Megens H.J., Milan D., Nonneman D.J., Rohrer G. A., Rothschild M.F., Smith T.P.L., Schnabel R.D., van Tassell C.P., Taylor J.F., Wiedmann R.T., Schook L.B. & Groenen M.A.M. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* 4: e6524.
- Rothschild M.F., Hu Z.L. & Jiang Z. (2007). Advances in QTL mapping in pigs. *Int J Biol Sci* 3: 192–197.
- Rückert C. & Bennewitz J. (2010). Joint QTL analysis of three connected F2-crosses in pigs. *Genet Sel Evol* 42: 40.
- Wellmann R., Preuß S., Tholen E., Heinkel J., Wimmers K. & Bennewitz J. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. *Genet Sel Evol* 45: 28.



## GENERAL SUMMARY (ENGLISH)

In the era of genomics, genome-wide association studies (GWASs) have become the method of choice for gene mapping. This is still of great interest to infer the genetic architecture of quantitative traits and to improve genomic selection in animal breeding. Formerly, linkage analyses were conducted in order to map genes. Therefore, many F2 cross populations were generated by crossing genetically divergent lineages in order to create informative experimental populations. However, a small number of markers and the limited meiotic divisions led to imprecise mapping results. The main objective of the present study was to investigate the use of existing porcine F2 cross data, extended towards single nucleotide polymorphism (SNP) chip genotype information, for quantitative trait loci (QTL) mapping in the genomic era. A special focus was on mapping genes that also segregate within the Piétrain breed since this is an important sire line and genomic selection is applied in this breed.

**Chapter 1** is a review article of statistical models and experimental populations applied in GWASs. This chapter gives an overview of methods to conduct GWASs using single-marker models and multi-marker models. Further, approaches taking non-additive genetic effects or genotype-by-environment interactions into account are described. Finally, post-GWAS analysis possibilities and GWAS mapping populations are discussed.

In **chapter 2**, the power and precision of GWASs in different F2 populations and a segregating population was investigated using simulated whole-genome sequence data. Further, the effect of pooling data was determined. GWASs were conducted for simulated traits with a heritability of 0.5 in F2 populations derived from closely and distantly related simulated founder breeds, their pooled datasets, and a sample of the common maternal founder breed. The study showed that the mapping power was high (low) in F2 crosses derived from distantly (closely) related founder breeds and highest when several F2 datasets were pooled. By contrast, a low precision was observed in the cross with distantly related founder breeds and the pooling of data led to a precision that was between the two crosses. For genes that also segregated within the common founder breed, the precision was generally elevated and, at equal sample size, the power to map QTL was even higher in F2 crosses derived from closely related founder breeds compared with the founder breed itself.

Within and across linkage disequilibrium (LD) structures of such F2 populations were examined in **chapter 3** by separately and jointly (pooled dataset) analyzing four F2 datasets generated from different founder breeds. All individuals were genotyped with a 62k SNP chip. The LD decay was faster in crosses derived from closely related founder breeds compared with crosses from phylogenetically distantly related founder populations and fastest when the data of all crosses were pooled. The pooled dataset was also used to map QTL for the economically important traits dressing out and conductivity applying single-marker and Bayesian multi-marker regressions. For these traits, several genome-wide significant association signals were mapped.

To infer the suitability of F2 data to map genes in a segregating breeding population, GWAS results of a pooled F2 cross were validated in two samples of the German Piétrain population (**chapter 4**). All individuals were genotyped using standard 62k SNP chips. The pooled cross contained the data of two F2 crosses, both had Piétrain as one founder breed, and consisted of 595 individuals. Initially, GWASs were conducted in the pooled F2 cross for the production traits dressing yield, carcass length, daily gain and drip loss. Subsequently, QTL core regions around significant trait associated peaks were defined. Finally, SNPs within these core regions were tested for association in the two samples of the current Piétrain population (771 progeny tested boars and 210 sows) in order to validate them in this breed. In total, 15 QTL were mapped and 8 (5) of them were validated in the boar (sow) validation dataset. This approach takes advantage of the high mapping power in F2 data to detect QTL that may not be found in the segregating Piétrain population. The findings showed that many of the QTL mapped in F2 crosses derived from Piétrain still segregate in this breed, and thus, these F2 datasets provide a promising database to map QTL in the Piétrain breed.

The thesis ends with a general discussion.

## GENERAL SUMMARY (GERMAN)

Im Zeitalter der Genomik sind genomweite Assoziationsstudien (GWASs) zur Methode der Wahl für die Genkartierung geworden. Dies ist noch immer von großem Interesse, um die genetische Architektur von quantitativen Merkmalen abzuleiten und die genomische Selektion in der Tierzucht zu verbessern. In früheren Zeiten wurden Kopplungsanalysen durchgeführt, um Gene zu kartieren. Hierfür wurden viele F<sub>2</sub>-Kreuzpopulationen erzeugt, indem genetisch divergente Linien gekreuzt wurden, um informative experimentelle Populationen zu kreieren. Diese Studien waren jedoch durch eine geringe Anzahl von Markern und der begrenzten Anzahl von Meiosen limitiert, was zu ungenauen Kartierungsergebnissen führte. Das Hauptziel der vorliegenden Studie war es, die Verwendung von vorhandenen porcinen F<sub>2</sub>-Kreuzungsdaten, welche auf Einzelnukleotidpolymorphismus (engl.: single nucleotide polymorphism, SNP) Chip-Genotypinformation erweitert wurden, für die Kartierung von quantitativen Merkmalsgenorten (engl.: quantitative trait loci, QTL) im Zeitalter der Genomik zu untersuchen. Ein besonderer Schwerpunkt lag auf der Kartierung von Genen, die auch innerhalb der Rasse Piétrain segregieren, da diese eine wichtige Vaterlinie darstellt und genomische Selektion in dieser Rasse angewandt wird.

**Kapitel 1** ist ein Übersichtsartikel über statistische Modelle und experimentelle Populationen, die in GWASs Anwendung finden. Dieses Kapitel gibt einen Überblick über Methoden zur Durchführung von GWASs mit Einzelmarker- und Multimarkernmodellen. Darüber hinaus wurden Ansätze beschrieben, die nichtadditive genetische Effekte oder Genotyp-Umwelt-Interaktionen berücksichtigen. Final wurden post-GWAS Analysemöglichkeiten und Kartierungspopulationen für GWASs diskutiert.

In **Kapitel 2** wurden die Power und Präzision von GWASs in verschiedenen F<sub>2</sub>-Populationen und einer segregierenden Population unter Verwendung simulierter Gesamtgenom-Sequenzdaten untersucht. Außerdem wurde der Effekt des Poolens von Daten ermittelt. Es wurden GWASs für simulierte Merkmale mit einer Heritabilität von 0,5 in F<sub>2</sub>-Populationen, die von nah- und fernverwandten simulierten Ausgangsrassen abstammen, ihren gepoolten Datensätzen und einer Stichprobe der gemeinsamen maternalen Ausgangsrasse, durchgeführt. Die Studie zeigte, dass die Power zur Genkartierung in F<sub>2</sub>-Kreuzungen mit fernverwandten (nahverwandten) Ausgangsrassen hoch (niedrig) war, und am höchsten, wenn mehrere F<sub>2</sub>-Datensätze gepoolt wurden. Im Gegensatz dazu wurde eine geringe Kartierungspräzision in den Kreuzungen mit

fernverwandten Ausgangsrassen beobachtet. Das Poolen von Daten führte zu einer Genauigkeit, die zwischen den beiden Kreuzungen lag. Für Gene, die auch innerhalb der gemeinsamen Ausgangsrasse segregierten war die Präzision generell erhöht und bei gleichem Stichprobenumfang war die Fähigkeit QTL zu kartieren, in F<sub>2</sub>-Kreuzungen die von nahverwandten Ausgangslinien abstammten, im Vergleich zur Ausgangsrasse selbst, noch höher.

Kopplungsungleichgewichtsstrukturen innerhalb solcher F<sub>2</sub>-Populationen und über F<sub>2</sub>-Populationen hinweg wurden in **Kapitel 3** untersucht, indem vier F<sub>2</sub>-Datensätze mit verschiedenen Ausgangsrassen separat und gemeinsam (gepoolter Datensatz) analysiert wurden. Alle Individuen wurden mit einem 62k SNP-Chip genotypisiert. Der LD-Zerfall (Kopplungsungleichgewicht, engl.: linkage disequilibrium, LD) war bei Kreuzungen, die von nahverwandten Ausgangsrassen abstammten schneller als bei Kreuzungen aus phylogenetisch fernverwandten Ausgangspopulationen, und am schnellsten, wenn die Daten aller Kreuzungen gepoolt wurden. Der gepoolte Datensatz wurde auch genutzt, um QTL für die ökonomisch wichtigen Merkmale, Schlachtausbeute sowie Leitfähigkeit, unter Anwendung von Singlemarker- und Bayes-Multimarker-Regressionen, zu kartieren. Für diese Merkmale kartierten wir mehrere genomweit signifikante Assoziationssignale.

Um die Eignung von F<sub>2</sub>-Daten für die Kartierung von Genen in einer segregierenden Zuchtpopulation abzuleiten, wurden GWAS-Ergebnisse einer gepoolten F<sub>2</sub>-Kreuzung in zwei Stichproben der deutschen Piétrainpopulation validiert (**Kapitel 4**). Alle Individuen wurden unter Verwendung von 62k Standard-SNP-Chips genotypisiert. Die gepoolte Kreuzung enthielt die Daten zweier F<sub>2</sub>-Kreuzungen, welche beide Piétrain als eine Ausgangsrasse hatten, und bestand aus 595 Individuen. Zuerst wurden GWASs in der gepoolten F<sub>2</sub>-Kreuzung für die Produktionsmerkmale Schlachtausbeute, Schlachtkörperlänge, Tageszunahme und Tropfsaftverlust durchgeführt. Anschließend wurden QTL-Kernregionen um signifikante merkmalsassoziierte Peaks definiert. Schließlich wurden die SNPs innerhalb dieser Kernregionen in den zwei Stichproben der aktuellen Piétrainpopulation (771 Nachkommen-geprüfte Eber und 210 Sauen) auf Assoziation getestet, um sie in dieser Rasse zu bestätigen. Insgesamt wurden 15 QTL kartiert und 8 (5) davon wurden im Ebervalidierungsdatensatz (Sauenvalidierungsdatensatz) bestätigt. Dieser Ansatz nutzt die große Teststärke in F<sub>2</sub>-Daten um QTL zu detektieren, die möglicherweise in der segregierenden Piétrainpopulation nicht gefunden werden würden. Die Ergebnisse zeigten, dass viele der QTL die in von Piétrain abstammenden F<sub>2</sub>-Kreuzungen kartiert

wurden, noch immer in dieser Rasse segregieren. Somit bieten diese F2-Datensätze eine vielversprechende Datengrundlage zur Kartierung von QTL in der Rasse Piétrain.

Die Arbeit endet mit einer allgemeinen Diskussion.



## CHAPTER ONE

**Invited review: Genome-wide association analysis for quantitative traits in livestock – a selective review of statistical models and experimental designs**

**Markus Schmid and Jörn Bennewitz**

University Hohenheim, Institute of Animal Science, Garbenstrasse 17, 70599 Stuttgart, Germany

Corresponding author: [j.bennewitz@uni-hohenheim.de](mailto:j.bennewitz@uni-hohenheim.de)

Published in:

**Archives Animal Breeding** (2017) 60: 335-346.

doi: [10.5194/aab-60-335-2017](https://doi.org/10.5194/aab-60-335-2017).

## **Abstract**

Quantitative or complex traits are controlled by many genes and environmental factors. Most traits in livestock breeding are quantitative traits. Mapping genes and causative mutations generating the genetic variance of these traits is still a very active area of research in livestock genetics. Since genome-wide and dense SNP panels are available for most livestock species, genome-wide association studies (GWASs) have become the method of choice in mapping experiments. Different statistical models are used for GWASs. We will review the frequently used single-marker models and additionally describe Bayesian multi-marker models. The importance of non-additive genetic and genotype-by-environment effects along with GWAS methods to detect them will be briefly discussed. Different mapping populations are used and will also be reviewed. Whenever possible, our own real-data examples are included to illustrate the reviewed methods and designs. Future research directions including post-GWAS strategies are outlined.

## **1 Introduction**

Quantitative or complex traits are controlled by many genes and environmental factors. Most traits in livestock breeding are quantitative traits, and there is a tremendous interest in analyzing these traits, e.g., with the aim to estimate breeding values of selection candidates or to map the underlying genes or chromosomal regions (quantitative trait loci, QTL). In earlier QTL mapping studies sparse genetic marker maps and linkage analysis were used to map QTL in experimental populations like F2 crosses or half-sib designs (e.g., Weller et al., 1990). Although many QTL were mapped, the mapping precision was usually low and only in a few exceptional cases was the underlying gene identified.

With the advent of high-density SNP arrays for most of the livestock species, it became possible to apply genome-wide association studies (GWASs). The underlying principle of GWASs is to test the SNPs for trait associations. The interpretation of statistically significant SNP trait associations is that the SNP is in linkage disequilibrium (LD) with a causative gene and that gene and the SNP are tightly linked. The latter is the case because the level of LD is a function of the distance between two loci on the chromosome.

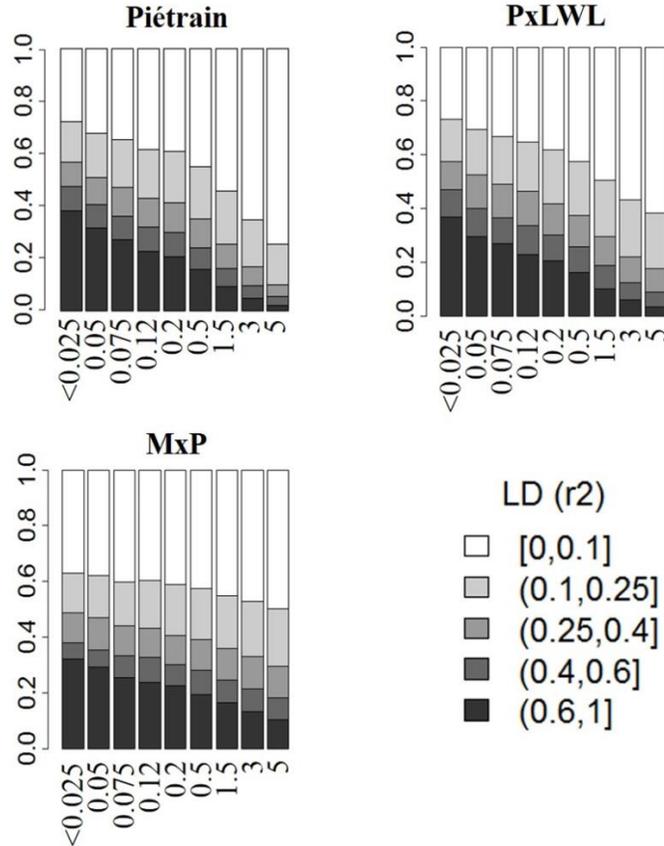
One of the main reasons for mapping QTL was to use mapped QTL for selection purposes in marker-assisted selection schemes (Dekkers, 2004). However, the success of these selection schemes was only very limited, mainly because the explained variance by the mapped QTL was very small. In order to overcome these limitations, Meuwissen et al. (2001) transferred marker-assisted selection on a genome-wide scale and developed statistical models to estimate genomic breeding values that rely on genome-wide and dense marker data but not on results from mapping experiments. The selection of breeding candidates based on genomic estimated breeding values became known as genomic selection, and it is implemented in many livestock genetic breeding programs, where it accelerates genetic gain substantially.

Despite the success of genomic selection, mapping genes for complex traits is still a burning issue in livestock genetics. Goddard et al. (2016) listed three main reasons for this. The first is to improve genomic selection. Second, GWAS results can increase biological knowledge about trait expression. The function of GWAS-identified genes can be used to derive and validate hypotheses about trait synthesis. This is of special interest for novel traits that eventually will be included in the selection goal or that might be controlled by tailored drugs or feeding strategies, like feather pecking in laying hens, greenhouse gas emission in ruminants or nutrient efficiency. Third, GWAS results can provide information on the genetic architecture of the quantitative trait; i.e., we may be interested in how many genes control the genetic variance, what the effect sizes are, how important non-additive genetic effects are and so on.

Different statistical methods and types of populations have been used in livestock GWAS experiments. In this study, we will review the most commonly used methods and mapping populations. First, single- and multi-marker GWAS models are presented. Next, we describe the importance of non-additive genetic and genotype-by-environment effects and show how these can be modeled in GWASs. Different mapping populations are used and these will be described in the following section. This review ends with a discussion, where future research directions including post-GWAS strategies are outlined. Whenever possible, our own real-data examples are included to illustrate the reviewed methods and designs. Because GWASs rely on LD between SNPs and causative genes, we start with a brief description of the most commonly used LD measure and its expectation.

## 2 Linkage disequilibrium measure $r^2$ and its expectation

Assume two loci,  $A$  and  $B$ , with two alleles each, i.e., alleles  $A$  and  $a$  and alleles  $B$  and  $b$ , with allele frequencies  $f_A$ ,  $f_a$ ,  $f_b$  and  $f_B$ . The haplotype frequencies of the haplotypes  $AB$ ,  $Ab$ ,  $aB$  and  $ab$  are denoted by  $f_{AB}$ ,  $f_{Ab}$ ,  $f_{aB}$  and  $f_{ab}$ , respectively. Following Hill and Weir (1994) the LD between these loci can be calculated as  $r^2 = \frac{(f_{AB}f_{ab} - f_{Ab}f_{aB})^2}{f_A f_a f_B f_b}$ . This measure has some convenient properties. It is bounded between 0 and 1, with 1 being a perfect LD. Assume locus  $A$  is a gene and  $B$  is an SNP used in a GWAS. In this case, the fraction of gene variance explained by the SNP is  $r^2$  (although the LD and the gene variance remain unknown until the causative gene itself is identified). The expectation of  $r^2$  can be expressed as  $E(r^2) = 1 / (1 + 4N_e c)$ , with  $N_e$  being the effective population size, and  $c$  denotes the recombination rate between the two loci (Sved, 2009; Tenesa et al., 2007). From this expression, it becomes obvious that the expected LD decays fast with increasing distances between loci, especially if the effective population size is large. The following example illustrates this. Stratz et al. (2014) investigated the LD structure of a segregating Piétrain pig population. They used SNP chip genotypes (porcine 60K BeadChip, Illumina Inc., San Diego, CA) of nearly 900 Piétrain boars for the LD  $r^2$  calculation for SNP pairs with a maximum distance of 5 megabases (Mb). The results are shown for *Sus scrofa* chromosome 1 (SSC 1) in Fig. 1 as a histogram of mean  $r^2$  for bins of SNP pair distances. The level of LD decreases strongly for larger distances. Compared to humans, long-range LD blocks are more common in livestock, especially in dairy cattle. This is due to the intensive use of relatively few sires for breeding the next generation, which results in a relatively small effective population size.



**Figure 1.** Linkage disequilibrium decay as a function of the marker distance in a purebred Piétrain population and in F2 crosses derived from closely (Piétrain x Landrace/Large White) and distantly (Piétrain x Meishan) related founder breeds (*Sus scrofa* chromosome 1).

### 3 Single-marker models

Single-marker GWASs fit one SNP at a time, usually in a mixed linear model (Yang et al., 2014). When assuming a single SNP  $j$  with genotypes coded as the number of copies of the allele with the minor frequency at the SNP for each individual  $i$  ( $x_{ij} = 0, 1$  or  $2$ ), the following model is frequently used:

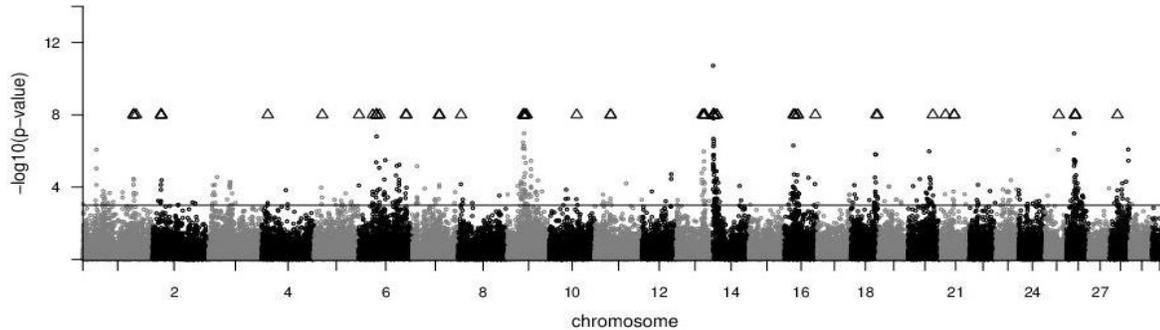
$$y_i = \mu + b_j x_{ij} + u_i + e_i. \quad (1)$$

Thereby,  $y_i$  is the trait record of individual  $i$ ,  $\mu$  denotes the fixed mean (assuming no other fixed effects exist) and  $b_j$  is the regression coefficient for SNP  $j$  to be estimated. In this parametrization, the SNP effect represents the gene substitution effect (Falconer and Mackay,

1996). The term  $e_i$  denotes the random residual and  $u_i$  the random polygenic effect of the individual. The distributional assumption of the polygenic effects is  $u \sim N(0, A\sigma_u^2)$ , with  $A$  being the relationship matrix either to be estimated from pedigree or from SNP data and  $\sigma_u^2$  being the polygenic variance component. The test for trait association is done by testing  $b_j$ , being different from 0, which results in an error probability or  $p$  value. In a GWAS, one SNP at a time is fitted to the model, resulting in multiple tests. In order to correct for these multiple tests, several approaches can be applied. The most common ones are the Bonferroni correction and the false discovery rate (FDR). Often, the Bonferroni correction is used to determine genome-wide significance thresholds and the FDR to assess how many of the associations reaching the significance level are false positives. The level of multiple testing can be enormous, especially if dense SNP chips or sequence data are used, and these SNPs are in LD and thus do not segregate independently. In these common situations, the Bonferroni correction is very stringent, and thus the results are conservative. More details about corrections for multiple testing in QTL mapping can be found in Fernando et al. (2004).

The polygenic component in Eq. (1) is important to capture population stratification effects and thus to prevent an inflation of type-I errors (e.g., MacLeod et al., 2010). Unlike in plant breeding, it is very convenient that for many livestock mapping populations the pedigree is known, and hence the relationship matrix needed to model this component adequately can be calculated using this information. If this is not possible, genetic markers can be used to set up a genomic relationship matrix (GRM) (VanRaden, 2008). If GRMs are used, the question is whether the SNP to be tested for association (or indeed the SNPs being in LD with this SNP) should also be used to set up the GRM or not. In the case of an inclusion, the SNP appears twice in the model and is treated once as a fixed and once as a random effect. Consequently, the SNP has to compete against itself, which seems somewhat counterintuitive. Indeed, Yang et al. (2014) showed that this results in a reduced mapping power. These authors recommended the exclusion of all SNPs that are located on the same chromosome as the SNP to be tested from the GRM. However, a recent article by Gianola et al. (2016) on GWASs with a GRM suggests that double-fitting the SNP effects (as fixed and random effects) is a less severe problem than previously thought. Another way of modeling population structure is to fit principal components (Patterson et al., 2006), but, as Hayes (2013) pointed out, it is not exactly traceable which variation source they

actually remove. It may be noted that removing population structure effects is not straightforward when generalized linear models (e.g., Poisson models) are applied (Lutz et al., 2017).



**Figure 2.** Test statistics of a single-marker GWAS for the trait milk protein yield in a sample of the German Holstein population. The solid line corresponds to a significance level of  $P = 0.001$ . Significant SNPs are indicated by triangles (taken from Streit et al., 2013a).

In a recent study we conducted a single-marker mixed model GWAS in Holstein dairy cattle data (Streit et al., 2013a). In brief, there were around 2300 progeny-tested bulls available, which were genotyped with the bovine 50K Bead-Chip (Illumina Inc., San Diego, CA). Qanbari et al. (2010) investigated the LD structure in this population. The trait considered was protein milk yield, and the relationship matrix of the bulls was established using pedigree information. The data set was split into a discovery data set (about 1800 bulls) for GWASs and a validation data set (500 bulls). The latter was exploited to confirm significant SNP associations identified in the discovery data set. FDR was applied to account for multiple testing. The results are shown in a so-called Manhattan plot in Fig. 2, with the negative decadic logarithm of the  $p$  value for each SNP on the y axis and the chromosomal position on the x axis (a common way of presenting GWAS results). Overall, 450 significant SNPs were identified with an FDR of maximally 7 %. Of these, 69 associations were also significant in the validation data set. Hence, these associations could be confirmed in the same population. Some of the identified trait-associated SNP clusters are located closely to well-known candidate genes segregating in the population (e.g., *DGAT1* on *Bos taurus* autosome (BTA) 14).

#### 4 Bayes-multi-marker models

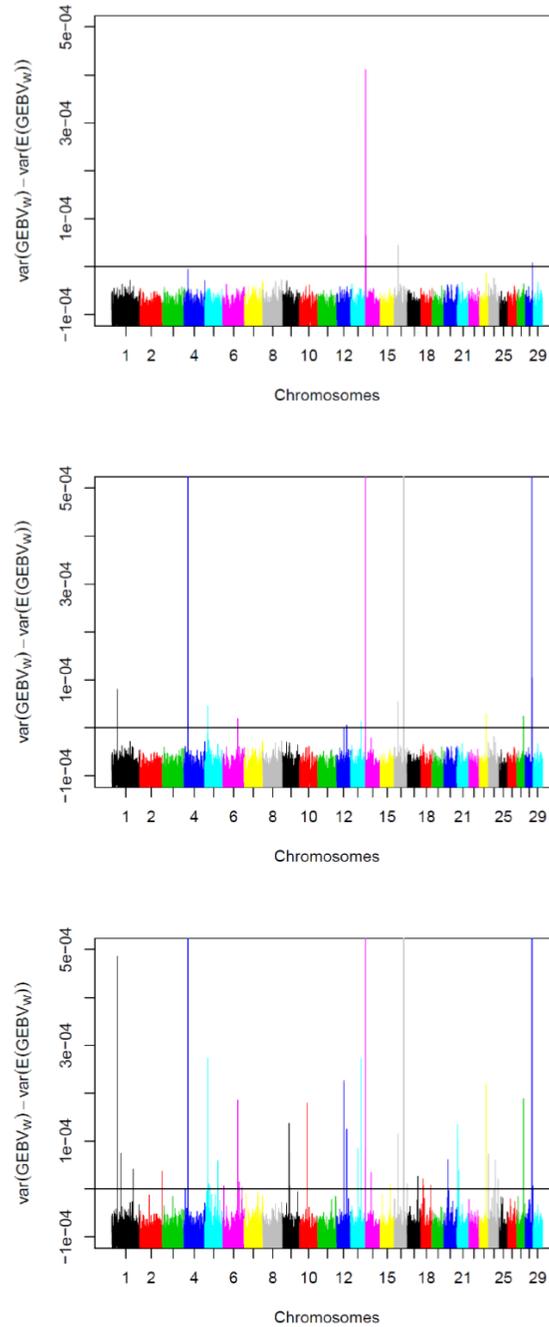
As stated above, the level of multiple testing can be enormous with dense SNP data, and stringent thresholds are needed in order to prevent an inflation of type-I errors. In addition, it is possible that the effect of a gene is only in part captured by a single marker due to imperfect LD but might be better explained jointly by the SNPs surrounding the gene. In order to overcome these limitations, multi-marker models that fit all SNPs simultaneously as random effects in the model were introduced for GWASs. Such models are able to deal with the problem that the number of SNPs often exceeds the number of observations. A general form of the model is as follows:

$$y_i = \mu + \sum_{j=1}^{n\_SNP} b_j x_{ij} + e_i. \quad (2)$$

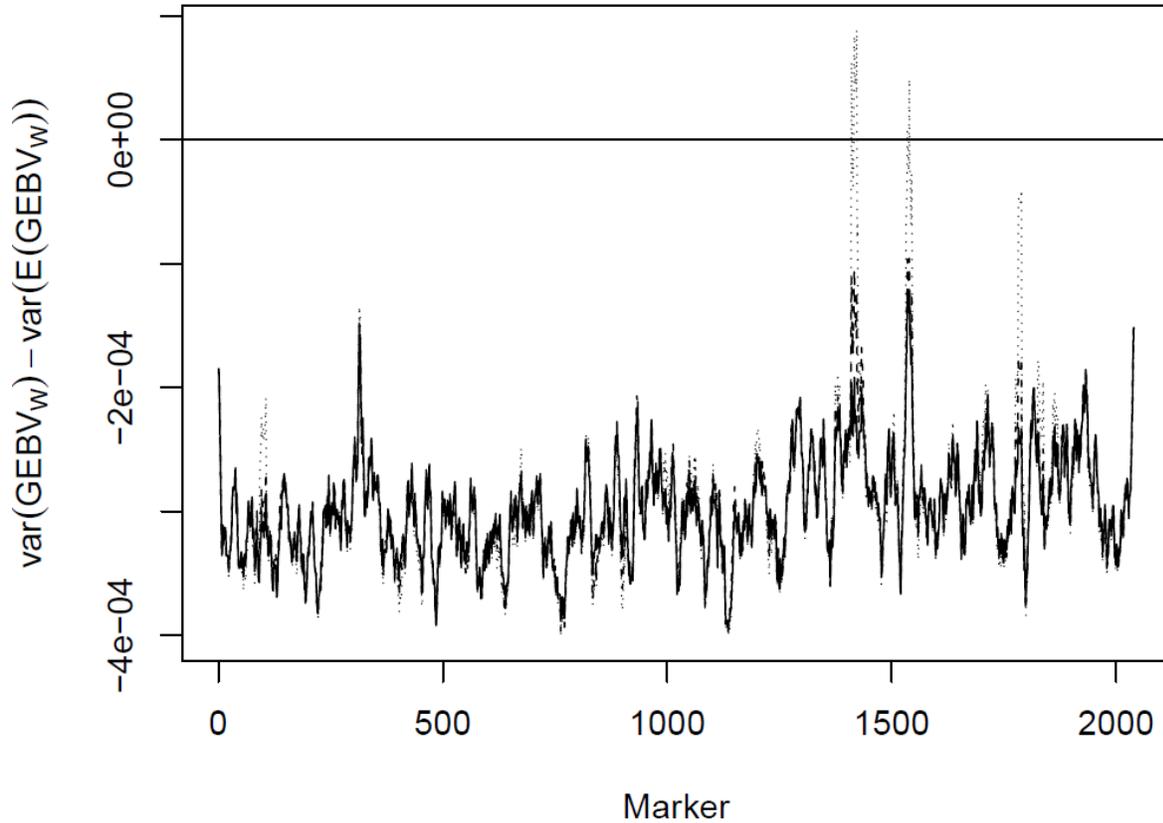
Compared to Eq. (1), the main difference is that all SNPs are fitted simultaneously as random effects. These models were originally developed for genomic selection purposes (Meuwissen et al., 2001) but have been shown to be very useful also for GWASs (Sahana et al., 2011; Goddard et al., 2016). The distributional assumptions of the SNP effects differ from model to model. The SNP-BLUP (Best Linear Unbiased Prediction) model assumes that all SNP effects come from one normal distribution with a small variance. This implies that the trait genetic variance is more or less equally distributed over the genome. This is a strong assumption and probably unrealistic for many quantitative traits. For this reason, Meuwissen et al. (2001) proposed two Bayesian models. The method called BayesA assumes a  $t$  distribution of the SNP effects, which is thicker-tailed compared to the normal distribution, depending on the degrees of freedom. BayesB models assume that only a fraction of the SNPs ( $\pi$ ) has an effect on the variance of a trait. For this fraction, a  $t$  distribution is assumed. Since the landmark paper by Meuwissen et al. (2001), further Bayes models were introduced (reviewed by Gianola, 2013). Verbyla et al. (2009) and Verbyla et al. (2010) proposed Bayesian stochastic search variable selection, which was also named BayesC by these authors. This model assumes two  $t$  distributions: one with a large variance for the  $\pi$  SNP fraction and one with a small variance for the  $1 - \pi$  fraction (e.g., 100 times smaller). SNPs belonging to the latter fraction hardly contribute to the genetic variance of a trait (or do not do so at all), and their effects are close to 0. The assumptions of BayesR, introduced by Erbe et al. (2012), are based on a mixture of normal distributions for the SNP effects.

Inference about a SNP trait association can either be drawn by the effect of a single SNP or by the posterior probability that the SNP effect comes from a distribution with large variance (in BayesB, C and R). The SNP effect is a random effect and a marginal effect, i.e., an effect corrected for all other SNP effects. This effect is sometimes also denoted as a conditional marker effect because the effects are drawn from conditional posterior distributions. The marginal marker effect is different from the effect obtained in Eq. (1) and, indeed, very sensitive to the SNP density. With increasing SNP density, the level of shrinkage towards 0 becomes stronger. Thus, it seems more straightforward to draw an inference by considering SNP effects within a window of defined size (e.g., 1 centimorgan (cM)) jointly and estimate the window genetic variance. Fernando et al. (2017) used the window genetic variance to calculate the window posterior probability of association (WPPA). This criterion has some nice properties. If a WPPA threshold of, e.g., 0.95 is used to declare an association as plausible, this results in a proportion of false positives of 0.05. This holds true if the data-generating model and the data-analysis models are similar. The WPPA criterion is convenient to compute, does not suffer from increasing marker density and produces an association criterion that is directly interpretable as the probability of window trait association.

For genomic predictions, the Bayesian methods often outperformed the SNP-BLUP model in computer simulations (e.g., Meuwissen et al., 2001), but this was often not the case in real data. This is probably due to the fact that many genes affect a trait and due to the long-range LD in livestock breeds, which results in many SNPs being in LD with a gene. However, this equal performance of the models does not hold for their use in GWASs. We used the Holstein dairy cattle data set mentioned above (Streit et al., 2013a) to compare the models SNP-BLUP, BayesA and BayesC in a GWAS for milk protein yield. In the BayesA and BayesC models,  $t$  distributions with 4 degrees of freedom (df) were assumed. The fraction of SNPs coming from the distribution with the large variance was  $\pi = 0.2$ . In BayesC, the variance of this distribution was 100 times larger than the variance of the a priori  $1 - \pi$  fraction of SNP effects. Gibbs sampling was used to draw samples from the posterior distributions using the program BayesDsamples (Wellmann and Bennewitz, 2012).



**Figure 3.** Results of a window-based multi-marker GWASs in a sample of the German Holstein population using the models SNP-BLUP (top panel), BayesA (middle panel) and BayesC (bottom panel). For each window, the deviation of the variance of the genomic estimated breeding value from its expected value is shown. The solid line corresponds to a deviation of 0.



**Figure 4.** Comparison of GWAS results generated with SNP-BLUP (solid line), BayesA (dashed line) and BayesC (dotted line) on BTA 6 in a sample of the German Holstein population. For each window, the deviation of the variance of the genomic estimated breeding value from its expected value is shown. The horizontal solid line corresponds to a deviation of 0.

The SNP effect estimates were used to calculate window genomic breeding values for windows of five consecutive SNPs ( $GEBV_W$ ) using standard notations (Falconer and Mackay 2007; Bennewitz et al., 2017). From these, the expected  $GEBV_W$  ( $E(GEBV_W)$ ) was subtracted in order to pinpoint trait-associated chromosomal regions. The  $E(GEBV_W)$  was calculated under the assumption of an equal distribution of the additive genetic variance across the genome; i.e., it was assumed that all genomic regions contribute equally to the additive genetic variance (for further details, see Bennewitz et al., 2017, Appendix). A putative QTL was assumed in those windows that showed a deviation greater than 0, i.e.,  $GEBV_W - E(GEBV_W) > 0$ . The plot of the  $GEBV_W$  deviations are shown in Fig. 3 for all three methods. When applying SNP-BLUP, only the window surrounding *DGATI* on BTA 14 showed evidence for trait association. BayesA produced around 10 additional and BayesC around 30 additional signals. The results are shown for BTA 6

in detail in Fig. 4, for which the single-marker GWAS (Eq. 1) revealed a confirmed trait-associated region (Fig. 2). BayesC clearly produced two signals on this chromosome, which were not detected by the two other methods. Following this, it seems that the Bayes methods, especially BayesC, are much more able to zoom into the genome and to pinpoint causative genomic regions. BayesR, which used a mixture of normal distributions with four components, was not investigated in this study but was propagated as a suitable method for the GWAS by Goddard et al. (2016).

Compared to single-marker GWASs, the application of these multi-marker methods is not straightforward and needs some carefully chosen parameters. For SNP effect estimation, the most important ones are the Markov chain Monte Carlo (MCMC) length,  $\pi$ , the variance scaling factor and the degrees of freedom. To our best knowledge, the length of the MCMC suitable for GWASs has not been sufficiently investigated until now. A small number of  $df$  results in a heavy-tailed  $t$  distribution and only large-effect SNPs will be identified (small effects will be regressed back to 0). Consequently, the number of false positives might be small but this compromises the power. The opposite holds true for larger numbers of  $df$ . The size of the windows for inference purposes, e.g., by the WPPA criterion, affects the power additionally. Larger window sizes result in an increased power but also in a reduced precision, i.e., the size of a trait-associated genomic region is larger (Bennewitz et al., 2017). There is a trade-off between power and precision. An obvious solution would be to start with larger window sizes, e.g., of 1 cM, to find significant trait-associated chromosomal regions and subsequently to reduce the windows size to fine-map the region.

We further tested non-parametric additive regression models originally adopted for genomic selection (Bennewitz et al., 2009) for GWASs using this data set. In contrast to Bayesian-methods, no prior information is needed. However, this method did not produce very clear GWAS signals, which were similar to the SNP-BLUP model (not shown).

## 5 Non-additive genetic and interaction effects

### 5.1 *Dominance and Imprinting*

The most important non-additive genetic effects are dominance and epistasis (Falconer and Mackay, 1996). It is well known that additive genetic variance is most important, and compared to this, dominance and epistatic variances are in general much smaller in size (Hill et al., 2008). However, this does not mean that there are no dominance effects of a detectable size (Wellmann and Bennewitz, 2011). Recent SNP-based investigations revealed that dominance variance can be substantial (e.g., Ertl et al., 2014; Su et al., 2012). Bolormaa et al. (2015) used a large-scale experiment with about 10 000 cattle, which were phenotyped for 16 quantitative traits and genotyped with dense SNP panels. They conducted a GWAS using single-marker regression analysis and found many trait-associated SNPs with a dominance effect. Moreover, the estimates of the dominance variance as a proportion of the phenotypic variance across the traits was between 0 and 42 % with a median of 5 %. Hence, it seems that dominance is an important source of genetic variation for some traits, and it seems appropriate to use this additional variation if the data structure permits it (i.e., genotypes and phenotypes are collected from the same individual). For example, the data set of Streit et al. (2013a) used in the previous section does not allow for dominance effect estimation because daughter yield deviations were used.

Single-marker association models (Eq. 1) can be extended straightforwardly towards modeling dominance. In addition to the regression on the SNP gene content, a regression on a heterozygous indicator variable is included, which represents the dominance deviation effect (Falconer and Mackay, 1996). Because dominance is modeled explicitly, the regression coefficient on the gene content no longer represents the gene substitution effect but the additive gene effect. This parametrization invokes one additional parameter to be estimated. Wellmann and Bennewitz (2011) showed that dominance and additive effects are dependent on each other in a complicated manner. Large dominance effects are usually observed for genes with large additive effects, which means that overdominance is a rare event. Therefore, in single-marker GWAS, a two-step procedure is often applied. In step one, only additive effects are fitted to the model. In the second step, dominance is included, and this extended model is applied only to SNPs with significant additive effects. This way of modeling dominance in single-marker GWAS models was chosen by e.g., Bolormaa et al. (2015).

The BayesC model was extended towards accounting for dominance, resulting in BayesD (Wellmann and Bennewitz, 2012). This model uses priors for the additive and dominance effects and the gene frequencies that resemble the complicated relationship between them. Roughly speaking, for small additive effects, the dominance deviations fluctuate around 0. With increasing additive effect sizes, the dominance deviation becomes larger and points in general to the homozygous genotype associated with the larger phenotypic value. The sign of additive and dominance effects depend on the gene frequency. Following this, it is unlikely that the contribution of the gene to the overall genetic variance is large. The latter is assumed because selection shifts the gene frequency towards a value where the variance contribution is small. Details can be found in Wellmann and Bennewitz (2011). In a recent study, we compared BayesC and BayesD for GWASs using simulated and real data sets (Bennewitz et al., 2017). We used the WPPA criterion for inference purposes and found a shift in power that was between  $-2$  and  $9$  %. Dominance is an interaction effect of the two alleles at a locus. Their effects are captured in the association analysis by matched haplotype pairs, i.e., diplotypes. Diplotypes show a faster decay around a focal point in the genome compared to haplotypes. Hence, it can be expected that BayesD improves the mapping precision as well, but this needs higher marker densities.

Imprinting seems to be a non-negligible source of variation for some quantitative traits in livestock. Trait-associated SNPs with imprinting effects can be detected by linkage analysis and GWASs. Models to do such analyses are presented in Mantey et al. (2005) and Hu et al. (2015).

## 5.2 *Epistasis*

The statistical interaction between SNPs is termed epistasis. The role of epistasis in the manifestation of quantitative traits has been subject to some debate during the last decades (e.g., Carlborg and Haley, 2004; Hill et al., 2008, among others). Detecting pairwise epistatic trait-associated SNPs can be done in principle by extending Eq. (1) by a second SNP and interaction terms between them. Even in this simple form of epistasis, i.e., pairwise epistasis, the model becomes much more complex because four interaction terms have to be fitted (additive-by-additive, additive-by-dominance, dominance-by-additive and dominance-by-dominance). In addition, the search for epistatic effects involves expanding from one dimension genome screenings (as for additive effects) towards two or even higher dimensions. This requires many

statistical tests and thus increases the problem of multiple testing enormously. Therefore, in addition to the need of dense SNP maps, a large sample size is needed in order to obtain a sufficient power to detect epistatic effects. It is sometimes argued that SNPs involved in epistasis also show additive effects. Based on this assumption, epistatic interactions are sometimes fitted only for SNPs that were significant in a previous GWAS run without fitting epistasis. This reduces the number of tests dramatically. Wei et al. (2014) reviewed statistical models to detect epistasis by GWASs.

### 5.3 *Genotype-by-environment interaction*

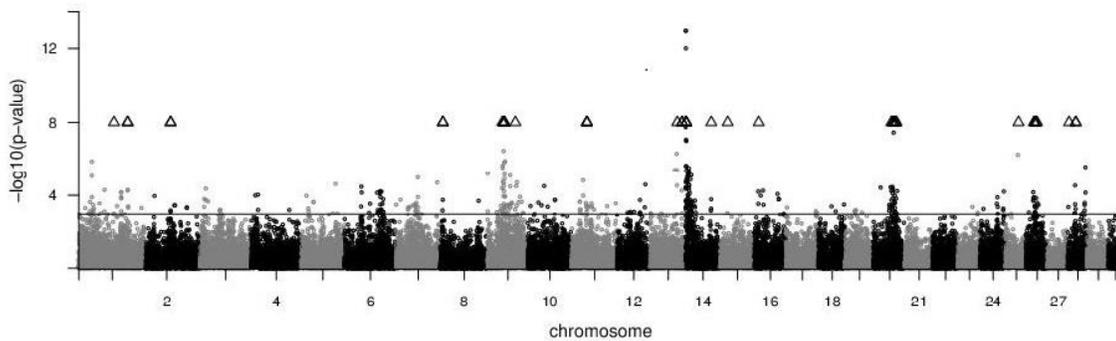
Genotype-by-environment interactions (G×E) are defined as the difference between genotype effects measured in different environments. A recent review of G×E in livestock can be found in Hayes et al. (2016). G×E can result in re-ranking effects; i.e., one genotype is superior in one environment, but inferior in the other environment. G×E scaling effects refer to the same ranking of genotypes, but the differences are larger in one environment compared to another environment. In general, two statistical methods are applied to test for G×E. Multiple-trait models treat the phenotypic records of a trait collected in different environments as different traits and calculate a genetic correlation between them. A deviation of this correlation from 1 (e.g., < 0.8) can be interpreted as evidence for G×E. In reaction norm models, the environment is described by a continuously distributed environmental descriptor and the phenotype is modeled as a function of the environment, where the phenotype is produced. Typical environmental descriptors are temperature-humidity indices (Hayes et al., 2003), average herd production levels as an indicator of the feeding level (Calus et al., 2002; Hayes et al., 2003) or herd disease levels (e.g., somatic cells score as an indicator of udder health and infection pressure on the farm; Streit et al., 2013b). Hayes et al. (2009) proposed a two-step reaction norm GWAS model to identify SNPs that showed G×E effects. In the first step, a random regression reaction norm model is applied to sires with sufficient progeny information in different environments as follows:

$$y_{ijk} = \mu + \sum_{m=0}^1 S_{jm} * E_k^m + e_{ijk} \quad (3)$$

Hereby,  $y_{ijk}$  is the observation of offspring  $i$  of sire  $j$  recorded in herd  $k$  with average level of the environmental descriptor  $E_k$ ,  $s_{jm}$  is the random sire effect of sire  $j$  of order  $m$  and  $e$  denotes the residual.

The covariance structure of the sire regression coefficients is  $var \begin{bmatrix} s_0 \\ s_1 \end{bmatrix} = A \otimes \begin{bmatrix} \sigma_{s_0}^2 & \sigma_{s_0 s_1}^2 \\ \sigma_{s_0 s_1}^2 & \sigma_{s_1}^2 \end{bmatrix}$ .

Note that this is a sire model. The residuals contain about three-quarters of the genetic variance. Thus, if  $G \times E$  is present, the residuals are heterogeneous, and this should be modeled as well. This model estimates two sire effects: one for the slope and one for the intercept of the reaction norm. If the mean of the environmental descriptor is set to 0, the intercept solutions of the sire regression coefficients are sire estimates for the general production level, i.e., the production level in the average environment. The sire's reaction norm slope effects represent the environmental sensitivity of the sire. In the second step, the sire's intercept and slope solutions are used as observations in a GWAS model, e.g., Eq. (1). GWAS hits for the slope identify environmentally sensitive trait-associated SNPs and thus SNPs involved in  $G \times E$ . Equation (3), shown above, is a random regression model



**Figure 5.** Test statistics of a single-marker GWAS for SNP environmental sensitivity for milk protein yield in a sample of the German Holstein population. The solid line corresponds to a significance level of  $P = 0.001$ . Significant SNPs are indicated by triangles (taken from Streit et al., 2013a).

As a result, the sire solutions for intercept and slope are regressed back to 0, which might compromise the power for a subsequent GWAS. Alternatively a fixed regression could be applied (known as the Finlay-Wilkinson regression in plant breeding), but the behavior of such a model for GWAS purposes needs to be investigated in detail.

In an earlier study, we used the two-step approach described above to map G×E SNPs in German Holsteins (Streit et al., 2013a). We used milk production test-day records of around 1.3 million daughters sired by 2300 sires with 12 million first lactation test records. We applied a two-step procedure to map SNPs associated with protein production G×E. Initially, a reaction norm random regression model (Eq. 3) was applied to the data, and subsequently we used the slope sire solutions as observations in a single-marker association model (Eq. 1). The results are shown in Fig. 5. We detected 351 significant trait-associated G×E SNPs, of which 44 could be confirmed in the same population. Generally, the results are very similar to those of the general milk protein production (see Fig. 2). Indeed, many trait-associated SNPs were also involved in G×E. This is discussed in detail in Streit et al. (2013a).

## **6 Mapping populations**

### **6.1 Segregating populations**

In contrast to QTL linkage mapping, for GWASs no experimental population (e.g., half-sib or F2 design) has to be established because genome-wide LD is assumed and used for mapping purposes (in linkage mapping the LD is generated within families by the mating design, and this allows for the use of low marker densities). Nevertheless, the study design affects the outcome of a GWAS. A few important aspects will be mentioned in the following. First, the sample size of the experiment affects the power. It is often stated that at least 1000 genotyped and phenotyped individuals have to be included, even for simple traits. This is the minimum number of individuals required for statistical analysis (except for mapping major genes, which are rare for quantitative traits). Larger numbers can be obtained by analyzing several mapping populations jointly, e.g., Holsteins, Jersey and Red cattle breeds, as done by Mao et al. (2016). This leads to a substantially larger mapping population, and the mapping resolution is much higher as well. The latter is because the genetic diversity within the mapping population is much larger (i.e., the hypothetical effective population size is larger, which in turn affects the LD pattern, as described

above). It was frequently shown that such across-breed analysis leads to clearer SNP association signals for genes that segregate in all breeds. At the same time, such an approach can be used to validate significant SNP trait associations across breeds. Another validation approach is to use a sample from the same breed, as done by Streit et al. (2013a) (Figs. 2 and 5). Across-breed analyses can be done either by pooling the data and analyzing them jointly or by a meta-analysis, where the results from the within-breed analysis (effect estimates and  $p$  values) are combined. The latter is more convenient to apply because each breed has its own fixed and random explanatory variables to be included in the GWAS models.

The density of the SNP panel is an additional important driver for the success of a GWAS experiment. From the expectation of the LD (shown above), it becomes obvious that higher densities are needed for populations with larger effective population sizes. For cattle, besides the standard chip (50k), there is a high-density (HD) SNP panel (777k) available. Especially for across-breed GWAS experiments, dense SNP data are beneficial due to the large hypothetical effective population size. For many breeds, influential sires were re-sequenced and these sequences can be used for imputation (Daetwyler et al., 2014). Hence, with the aid of HD-SNP chip data and the sequence information of some key ancestors, the whole-genome sequence variants can be inferred for all individuals within a mapping population. This, in turn, can be used for GWASs. A prerequisite for association mapping is a high LD between the marker and the causative mutation. A paradigm shift takes place when using genome sequence variants because all variants (i.e., SNPs and causative mutations) are included in the data set. Now the challenge is to identify the causative mutations among all polymorphisms and to separate them from SNPs that are solely in LD with the mutation. The success of a GWAS with genome sequence variants depends strongly on the quality of imputation of these variants in the study population. This is not always ensured but will not be reviewed here. Another problem is the level of multiple testing which increases towards several million correlated tests. A Bonferroni correction is too conservative. A possible solution for this problem is to map the QTL using SNP chip data in a first GWAS run applying standard multiple testing corrections (Bonferroni or FDR). In a second step, fine-mapping of the significant regions can be done using imputed genome sequence variants. Since it is assumed that the regions are significant, no additional stringent significance level has to be applied during fine-mapping.

## 6.2 *F2-Designs*

Many F2 crosses were established during the last decades, especially in pig breeding (Rothschild et al., 2007). Often, the F2 individuals were phenotyped under standardized conditions (e.g., on experimental farms) for traits that are interesting but very hard to measure, like efficiency traits or meat quality traits. Founder breeds were frequently chosen from Asian and from European breeds. Phylogenetic analysis of whole-genome sequence data revealed distinct lineages of these two types of breeds. In addition, F2 crosses within European types of breeds were established. In many cases one commercially used breed was one of the two founder breeds, e.g., the F2 crosses described in Boysen et al. (2010) and Rückert and Bennewitz (2010); both had Piétrain as one founder breed, which is an important sire line breed in Europe. We studied the LD pattern within an F2 cross derived from distantly related founders (i.e., a Meishan  $\times$  Piétrain cross) and within a cross derived from closely related founder breeds (i.e., Landrace/Large White  $\times$  Piétrain cross), using porcine SNP chip data. The results for SSC 1 are visualized for each of the two crosses in Fig. 1. As shown there, the LD is high and almost did not decrease with increasing marker distances up to 5 Mb in the Meishan  $\times$  Piétrain cross, which implies a poor mapping resolution. In contrast, the LD pattern of the Landrace/Large White  $\times$  Piétrain cross is similar to the pattern observed within the Piétrain breed (Fig. 1). Consequently, this results in a similar mapping resolution of such F2 designs and their founder breeds.

The question is whether is possible to map and fine-map genes in porcine F2 designs by SNP chip genotyping and GWASs. Ledur et al. (2009) studied the power of GWASs in F2 crosses by means of simulations and compared it to classical linkage analysis mapping. They found an increase in power and a smaller rate of false positive results in F2 crosses with large sample sizes and high marker densities. In order to continue these investigations, we simulated the two types of porcine F2 crosses described above (Schmid et al., 2017). Thereby, we created a situation where the genome sequence variants of all F2 individuals were available. The results showed that existing F2 crosses generated from closely related founder breeds with whole-genome sequence data available for all individuals could be used to map genes that segregate within a founder breed with a high precision. This is due to the high mapping resolution within this type of cross. Such genes are of interest for breeding purposes, e.g., in the genomic selection program established in the Piétrain breed. In contrast, the mapping precision was very poor in the cross derived from distantly related founder breeds, as expected. The results of the simulation study

showed that it might be a worthwhile effort to genotype existing F2 crosses derived from closely related founder breeds with dense SNP panels and conduct GWASs in order to make use of the existing information in the F2 crosses, especially with regard to the special traits that were collected in these individuals.

## **7 Post-GWAS analyses**

The final aims of a mapping experiment are to detect the underlying gene and the causative mutation within the gene. On the level of the DNA, the causality of a mutation can be identified (although not formally proved) by collecting pieces of evidence. The following facts strongly support the causality of a mutation (Mackay, 2001; Meuwissen, 2010). (1) If a mutation is included in the statistical model, no further polymorphism in strong LD with this mutation shows a significant effect. (2) The genotype effects can be validated and are similar in size in different populations and show the same algebraic sign. (3) The complete linkage disequilibrium test (CLD test) and (4) the concordance test have positive results. To verify point 2, one needs multiple populations. Due to the small number of individuals in experimental populations, this requirement is often difficult to fulfill. The CLD test (point 3; Uleberg and Meuwissen, 2011) is based on two analytic steps. First, all SNPs are tested one by one for association and the test statistics are noted. The second step consists of analyzing the difference in the test statistics. The hypothesis is based on the assumption that the causative mutation explains more variation than any SNP, which is in incomplete LD with the mutation. The concordance test (point 4; Ron and Weller, 2007; Weller and Ron, 2011) tests whether the same SNP allele identifies the same QTL allele (Q or q) in multiple families of QTL-heterozygous parents (which are identified by markers, for instance by using multiple marker regression; see Knott, 2005). Proving the causality of a mutation requires functional studies, but this is not the subject of this review.

## **8 Concluding remarks**

Mapping trait-associated SNPs and genes underlying the genetic variance of quantitative traits is still a burning issue in livestock genetic research. In future, two developments can be expected. On the one hand, we already observe that the data sets available for GWAS are increasing from day to day, and in the near future, we will be able to use several hundreds of thousands of

individuals. This holds true for traits that are widely used in animal breeding and for which large-scale phenotyping is thus implemented in routine data collection. Combined with improved annotated reference genomes and genome sequence databases, it will be possible to infer the whole-genome sequence variants of the individuals. Thus, it can be expected that the number of detected causative variants will increase for these mainstream traits, especially in across-breed analyses (within breeds the LD structure might prohibit the detection of many causative mutations even in large data sets). On the other hand, phenotypic records of genetically simpler traits can be collected in experimental populations by in-depth phenotyping (e.g., metabolic traits or gene expression traits). The detection of causative genes for these traits requires less large data sets, but a high precision in data recording and a well-defined experimental structure are needed.

**Data availability**

No data sets were used in this article.

**Competing interests**

The authors declare that they have no conflict of interest.

**Acknowledgements**

This study was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

**References**

- Bennewitz, J., Solberg, T., and Meuwissen, T. H. E.: Genomic breeding value estimation using nonparametric additive regression models, *Genet. Sel. Evol.*, 41:20, doi:10.1186/1297-9686-41-20, 2009.
- Bennewitz, J., Edel, C., Fries, R., Meuwissen, T. H. E., and Wellmann, R.: Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis, *Genet. Sel. Evol.*, 49:7, doi:10.1186/s12711-017-0284-7, 2017.
- Bolormaa, S., Pryce, J. E., Zhang, Y., Reverter, A., Barendse, W., Hayes, B. J., and Goddard, M. E.: Non-additive genetic variation in growth, carcass and fertility traits of beef cattle, *Genet. Sel. Evol.*, 47:26, doi:10.1186/s12711-015-0114-8, 2015.
- Boysen, T. J., Tetens, J., and Thaller, G.: Detection of a quantitative trait locus for ham weight with polar overdominance near the ortholog of the callipyge locus in an experimental pig F2 population, *J. Anim. Sci.*, 88, 3167–3172, doi:10.2527/jas.2009-2565, 2010.
- Calus, M. P. L., Groenen, A. F., and de Jong, G.: Genotype  $\times$  Environment Interaction for Protein Yield in Dutch Dairy Cattle as Quantified by Different Models, *J. Dairy Sci.*, 85, 3115–3123, doi:10.3168/jds.S0022-0302(02)74399-3, 2002.
- Carlborg, Ö. and Haley, C. S.: Opinion: Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.*, 5, 618–625, doi:10.1038/nrg1407, 2004.
- Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen R., Brøndum, R. F., Liao, X., Djari, A., Rodriguez, S. C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P. J., Coote, D., Chamberlain, A. J., Anderson, C., van Tassell, C. P., Hulsegge, I., Goddard, M. E., Gulbrandsen, B., Lund, M. S., Veerkamp, R. F., Boichard, D. A., Fries, R., and Hayes, B. J.: Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle, *Nat. Genet.*, 46, 858–865, doi:10.1038/ng.3034, 2014.
- Dekkers, J. C. M.: Commercial application of marker- and gene-assisted selection in livestock : Strategies and lessons, *J. Anim. Sci.*, 82, E313-E328, 2004.
- Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason,

- B. A., and Goddard M. E.: Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels, *J. Dairy Sci.*, 95, 4114–4129, doi:10.3168/jds.2011-5019, 2012.
- Ertl, J., Legarra, A., Vitezica, Z. G., Varona, L., Edel, C., Emmerling, R., and Götz, K.-U.: Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle, *Genet. Sel. Evol.*, 46:40, doi:10.1186/1297-9686-46-40, 2014.
- Falconer, D. S. and Mackay, T. F. C.: *Introduction to Quantitative Genetics*, 4th ed., Longman Group Ltd, London, 1996.
- Fernando, R., Toosi, A., Wolc, A., Garrick, D., and Dekkers, J. C. M.: Application of Whole-Genome Prediction Methods for Genome-Wide Association Studies: A Bayesian Approach, *J. Agric. Biol. Environ. Stat.*, 22, 172–193, doi:10.1007/s13253-017-0277-6, 2017.
- Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller, M.: Controlling the Proportion of False Positives in Multiple Dependent Tests, *Genetics*, 166, 611–619, doi:10.1534/genetics.166.1.611, 2004.
- Gianola, D.: Priors in whole-genome regression: The Bayesian alphabet returns, *Genetics*, 194, 573–596, doi:10.1534/genetics.113.151753, 2013.
- Gianola, D., Fariello, M. I., Naya, H., and Schön, C.-C.: Genome-Wide Association Studies with a Genomic Relationship Matrix: A Case Study with Wheat and Arabidopsis, *G3*, 6, 3241–3256, doi: 10.1534/g3.116.034256, 2016.
- Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., and Hayes, B. J.: Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture, *Proc. Biol. Sci.*, doi: 10.1098/rspb.2016.0569, 2016.
- Hayes, B. J.: Overview of Statistical Methods for Genome-Wide Association Studies (GWAS), in: *Genome-Wide Association Studies and Genomic Prediction*, Gondro, C., van der Werft, J., Hayes, B. J., Springer Protocols, New York, 149–169, 2013.
- Hayes, B. J., Carrick, M., Bowman, P. J., and Goddard, M. E.: Genotype×Environment Interaction for Milk Production of Daughters of Australian Dairy Sires from Test-Day

- Records, *J. Dairy Sci.*, 86, 3736–3744, doi:10.3168/jds.S0022-0302(03)73980-0, 2003.
- Hayes, B. J., Bowman, P. J., Chamberlain, A. J., Savin, K. W., van Tassell, C. P., Sonstegard, T. S., and Goddard, M. E.: A validated genome wide association study to breed cattle adapted to an environment altered by climate change, *PLoS One*, 4, 1–8. doi:10.1371/journal.pone.0006676, 2009.
- Hayes, B. J., Daetwyler, H. D., and Goddard, M. E.: Models for Genome x Environment interaction: Examples in livestock, *Crop Sci.*, 56, 2251–2259, doi:10.2135/cropsci2015.07.0451, 2016.
- Hill, W. G. and Weir, B. S.: Maximum-likelihood estimation of gene location by linkage disequilibrium, *Am. J. Hum. Genet.*, 54, 705–714, 1994.
- Hill, W. G., Goddard, M. E., and Visscher, P. M.: Data and theory point to mainly additive genetic variance for complex traits, *PLoS Genet.*, 4, doi:10.1371/journal.pgen.1000008, 2008.
- Hu, Y., Rosa, G. J. M., and Gianola, D.: A GWAS assessment of the contribution of genomic imprinting to the variation of body mass index in mice, *BMC Genomics*, 16:576, doi:10.1186/s12864-015-1721-z, 2015.
- Knott, S. A.: Regression-based quantitative trait loci mapping: robust, efficient and effective, *Philos. Trans. R. Soc. B., Biol. Sci.*, 360, 1435–1442, doi:10.1098/rstb.2005.1671, 2005.
- Ledur, M. C., Navarro, N., and Pérez-Enciso, M.: Large-scale SNP genotyping in crosses between outbred lines: how useful is it?, *Heredity*, 105, 173–182. doi:10.1038/hdy.2009.149, 2009.
- Lutz, V., Stratz, P., Preuß, S., Tetens, J., Grashorn, M. A., Bessei, W., and Bennewitz, J.: A genome-wide study in a large F2-cross of laying hens reveals novel genomic regions associated with feather pecking and aggressive behavior, *Genet. Sel. Evol.*, 49:18, doi:10.1186/s12711-017-0287-4, 2017.
- Mackay, T. F. C.: The Genetic Architecture of Quantitative Traits, *Annu. Rev. Genet.*, 35, 303–339. doi:10.1146/annurev.genet.35.102401.090633, 2001.
- MacLeod, I. M., Hayes, B. J., Savin, K. W., Chamberlain, A. J., McPartlan, H. C., and Goddard,

- M. E.: Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms, *J. Anim. Breed. Genet.*, 127, 133–142. doi:10.1111/j.1439-0388.2009.00831.x, 2010.
- Mantey, C., Brockmann, G. A., Kalm, E., and Reinsch, N.: Mapping and exclusion mapping of genomic imprinting effects in mouse F<sub>2</sub> families, *J. Hered.*, 96, 329–338, doi:10.1093/jhered/esi044, 2005.
- Mao, X., Sahana, G., De Koning, D.-J., and Guldbbrandtsen, B.: Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data, *J. Anim. Sci.*, 94, 1426–1437, doi:10.2527/jas.2015-9838, 2016.
- Meuwissen, T. H. E.: Use of whole genome sequence data for QTL mapping and genomic selection, in: *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany, 1-6 August 2010, 0018, 2010.
- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E.: Prediction of total genetic value using genome-wide dense marker maps, *Genetics*, 157, 1819–1829, 2001.
- Patterson, N., Price, A. L., and Reich, D.: Population structure and eigenanalysis, *PLoS Genet.*, 2, 2074–2093, doi:10.1371/journal.pgen.0020190, 2006.
- Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R., and Simianer, H.: The pattern of linkage disequilibrium in German Holstein cattle, *Anim. Genet.*, 41, 346–356, doi:10.1111/j.1365-2052.2009.02011.x, 2010.
- Ron, M., and Weller, J. I.: From QTL to QTN identification in livestock - Winning by points rather than knock-out: A review, *Anim. Genet.*, 38, 429–439, doi:10.1111/j.1365-2052.2007.01640.x, 2007.
- Rothschild, M. F., Hu, Z. L., and Jiang, Z.: Advances in QTL mapping in pigs, *Int. J. Biol. Sci.*, 3, 192–197, doi:10.7150/ijbs.3.192, 2007.
- Rückert, C. and Bennewitz, J.: Joint QTL analysis of three connected F<sub>2</sub>-crosses in pigs, *Genet. Sel. Evol.*, 42:40, doi:10.1186/1297-9686-42-40, 2010.
- Sahana, G., Guldbbrandtsen, B., and Lund, M. S.: Genome-wide association study for calving traits in Danish and Swedish Holstein cattle, *J. Dairy Sci.*, 94:479–486,

doi:10.3168/jds.2010-3381, 2011.

Schmid, M., Wellmann, R., and Bennewitz, J.: Power and precision of QTL mapping in simulated multiple F2 crosses using whole-genome sequence information, submitted, 2017.

Stratz, P., Wimmers, K., Meuwissen, T. H. E., and Bennewitz, J.: Investigations on the pattern of linkage disequilibrium and selection signatures in the genomes of German Piétrain pigs, *J. Anim. Breed. Genet.*, 131, 473–482, doi:10.1111/jbg.12107, 2014.

Streit, M., Wellmann, R., Reinhardt, F., Thaller, G., Piepho, H. P., and Bennewitz, J.: Using genome-wide association analysis to characterize environmental sensitivity of milk traits in dairy cattle, *G3*, 3, 1085–1093, doi:10.1534/g3.113.006536, 2013a.

Streit, M., Reinhardt, F., Thaller, G., and Bennewitz, J.: Genome-wide association analysis to identify genotype  $\times$  environment interaction for milk protein yield and level of somatic cell score as environmental descriptors in German Holsteins, *J. Dairy Sci.*, 96, 7318–7324, doi:10.3168/jds.2013-7133, 2013b.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., and Lund, M. S.: Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers, *PLoS One*, 7, 1–7, doi:10.1371/journal.pone.0045293, 2012.

Sved, J. A.: Linkage disequilibrium and its expectation in human populations, *Twin Res. Hum. Genet.*, 12, 35–43, doi:10.1375/twin.12.1.35, 2009.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M.: Recent human effective population size estimated from linkage disequilibrium, *Genome Res.*, 2, 520–526, doi:10.1101/gr.6023607.8, 2007.

Uleberg, E. and Meuwissen, T. H. E.: The complete linkage disequilibrium test: a test that points to causative mutations underlying quantitative traits, *Genet. Sel. Evol.*, 43:20, doi:10.1186/1297-9686-43-20, 2011.

VanRaden, P. M.: Efficient methods to compute genomic predictions, *J. Dairy Sci.*, 91, 4414–4423, doi:10.3168/jds.2007-0980, 2008.

- Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E.: Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle, *Genet. Res.*, 91, 307–311. doi: 10.1017/S0016672309990243, 2009.
- Verbyla, K. L., Bowman, P. J., Hayes, B. J., and Goddard, M. E.: Sensitivity of genomic selection to using different prior distributions, *BMC Proc.*, 4:S5, doi:10.1186/1753-6561-4-S1-S5, 2010.
- Wei, W.-H., Hemani, G., and Haley, C. S.: Detecting epistasis in human complex traits, *Nat. Rev. Genet.*, 15, 722–733, doi:10.1038/nrg3747, 2014.
- Weller, J. I. and Ron, M.: Invited review: Quantitative trait nucleotide determination in the era of genomic selection, *J. Dairy Sci.* 94, 1082–1090, doi:10.3168/jds.2010-3793, 2011.
- Weller, J. I., Kashi, Y., and Soller, M.: Power of Daughter and Granddaughter Designs for Determining Linkage Between Marker Loci and Quantitative Trait Loci in Dairy Cattle, *J. Dairy Sci.*, 73, 2525–2537, 1990.
- Wellmann, R. and Bennewitz, J.: The contribution of dominance to the understanding of quantitative genetic variation, *Genet. Res.*, 93, 139–154, doi:10.1017/S0016672310000649, 2011.
- Wellmann, R. and Bennewitz, J.: Bayesian models with dominance effects for genomic evaluation of quantitative traits, *Genet. Res.*, 94, 21–37, doi:10.1017/S0016672312000018, 2012.
- Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L.: Advantages and pitfalls in the application of mixed-model association methods, *Nat. Genet.*, 46, 100–106, doi:10.1038/ng.2876, 2014.



## CHAPTER TWO

### **Power and precision of QTL mapping in simulated multiple porcine F2 crosses using whole-genome sequence information**

**Markus Schmid\*, Robin Wellmann, Jörn Bennewitz**

Institute of Animal Science, University Hohenheim, Garbenstrasse 17, 70599 Stuttgart, Germany

\*Corresponding author: markus\_schmid@uni-hohenheim.de

Published in:

**BMC Genetics** (2018) 19: 22.

doi: 10.1186/s12863-018-0604-0.

## **Abstract**

### ***Background***

During the last two decades, many QTL (quantitative trait locus) mapping experiments in pigs have been conducted using F2 crosses established from two outbred founder breeds. The founder breeds were frequently chosen from the Asian and European type breeds. A combination of next-generation sequencing, SNP (single nucleotide polymorphism) genotyping technology using SNP-chips, and genotype imputation techniques, can be used to infer the sequence information of all F2 individuals in a cost-effective way. The aim of the present simulation study was to analyze the power and precision of genome-wide association studies (GWASs) with whole-genome sequence data in several types of F2 crosses, including pooled crosses.

### ***Methods***

Based on a common historical population, three breeds representing two European type breeds (EU1 and EU2) and one Asian type breed (AS) were simulated. Two F2 designs of 500 individuals each were simulated. The cross EU1xEU2 (ASxEU2) was simulated using the phylogenetically closely related breeds EU1 and EU2 (or distantly related breeds AS and EU2) as the founder breeds. The simulated genomes comprised ten chromosomes, each with a length of 1 Morgan, and whole-genome sequence information. A polygenic trait with a heritability of 0.5, which was affected by approximately 20 QTL per Morgan, was simulated. GWASs were conducted using single marker mixed linear models, either within the crosses or in their pooled datasets. Additionally, the studies were conducted in the breed EU2, which was a founder breed in both simulated crosses.

### ***Results***

The power to map QTL was high (low) in the ASxEU2 (EU1xEU2) cross and was highest when the data of both crosses were analyzed jointly. By contrast, the mapping precision was the highest in the EU1xEU2 cross. Pooling data led to a precision that was in between the precision of the

EU1xEU2 cross and the ASxEU2 cross. A higher mapping precision was observed for QTL segregating within a founder breed.

***Conclusions***

These results suggest that the existing F2 crosses are promising databases for QTL mapping when the founder breeds are closely related or several crosses can be pooled. This conclusion is particularly applicable for QTL that segregate in a founder breed.

***Keywords***

Genome-wide association studies, Mapping power, Mapping precision, Pooling data, Porcine F2 crosses, Simulation study, Whole-genome sequence data

## Background

QTL (quantitative trait locus) mapping and the identification of causative single nucleotide polymorphisms (QTNs, quantitative trait nucleotides) is still of high importance in animal breeding. The results of genome-wide association studies (GWASs) provide knowledge about the evolution and genetic architecture of traits and may improve the accuracy of genomic prediction [1], especially if the studies rely on genomic sequence data [2]. Before large-scale single nucleotide polymorphism (SNP) genotyping using next-generation sequencing technology was possible in pig breeding, QTL mapping was frequently performed by applying linkage analyses using sparse genetic maps, which were often built by microsatellite markers. The necessary linkage disequilibrium (LD) was established within families by generating experimental crosses. Many pig F<sub>2</sub> crosses have been generated during the last few decades, and numerous QTL for various traits have been reported [3, 4]. Often, the F<sub>2</sub> individuals were phenotyped under standardized conditions (e.g., on experimental farms) for interesting but hard-to-measure traits, such as efficiency traits or meat quality traits. Founder breeds were frequently chosen from Asian and European pig breeds. Phylogenetic analysis of whole-genome sequence data revealed distinct lineages of these two types of breeds [5]. However, F<sub>2</sub> crosses within European breeds were also established (e.g., [6]). In many cases, a commercially used breed was one of the two founder breeds. For example, the F<sub>2</sub> crosses described in [6, 7] both had Piétrain as one founder breed, which is an important sire line breed in Europe.

Since the availability of dense SNP maps and the possibility to conduct large-scale SNP genotyping with SNP-chips, QTL mapping is usually performed in genome-wide association studies (GWASs) within breeding populations or in admixed populations [1]. For example, QTL mapping was performed in the Piétrain breed (mentioned above) by Stratz et al. [8] using the Illumina PorcineSNP60 Beadchip [9]. Ledur et al. examined whether it is worthwhile to conduct large-scale SNP genotyping in F<sub>2</sub> crosses [10]. They studied the power of GWAS in F<sub>2</sub> crosses that were genotyped with large-scale SNP maps using simulations and compared the results with classical linkage analysis mapping. Their findings showed an increase in power and a smaller rate of false positive results in F<sub>2</sub> crosses with large sample sizes and high marker densities. A recent simulation study analyzed the mapping resolution and the linkage disequilibrium structures around causal genes of several simulated pig F<sub>2</sub> crosses at a maximized marker density (sequence information available for all individuals) [11]. It was shown that the mapping resolution is high

for genes that are also segregating in a founder breed, especially for F2 crosses established from two closely related founder breeds. In a few cases, the mapping resolution was even higher compared with a single outbred founder population due to the variation of LD between markers and QTNs among the founder breeds. Toosi et al. [12] reported similar results from a simulation of admixed cattle genomes. Thus the numerous past established F2 crosses might be underused experimental populations for mapping QTL and QTNs. This hypothesis might especially hold true for QTNs that segregate in the founder breeds. These QTNs are of interest for improving genomic predictions conducted within the founder breed. For example, mapping Piétrain segregating QTNs could improve the accuracy of genomic selection, which was implemented in this breed [13].

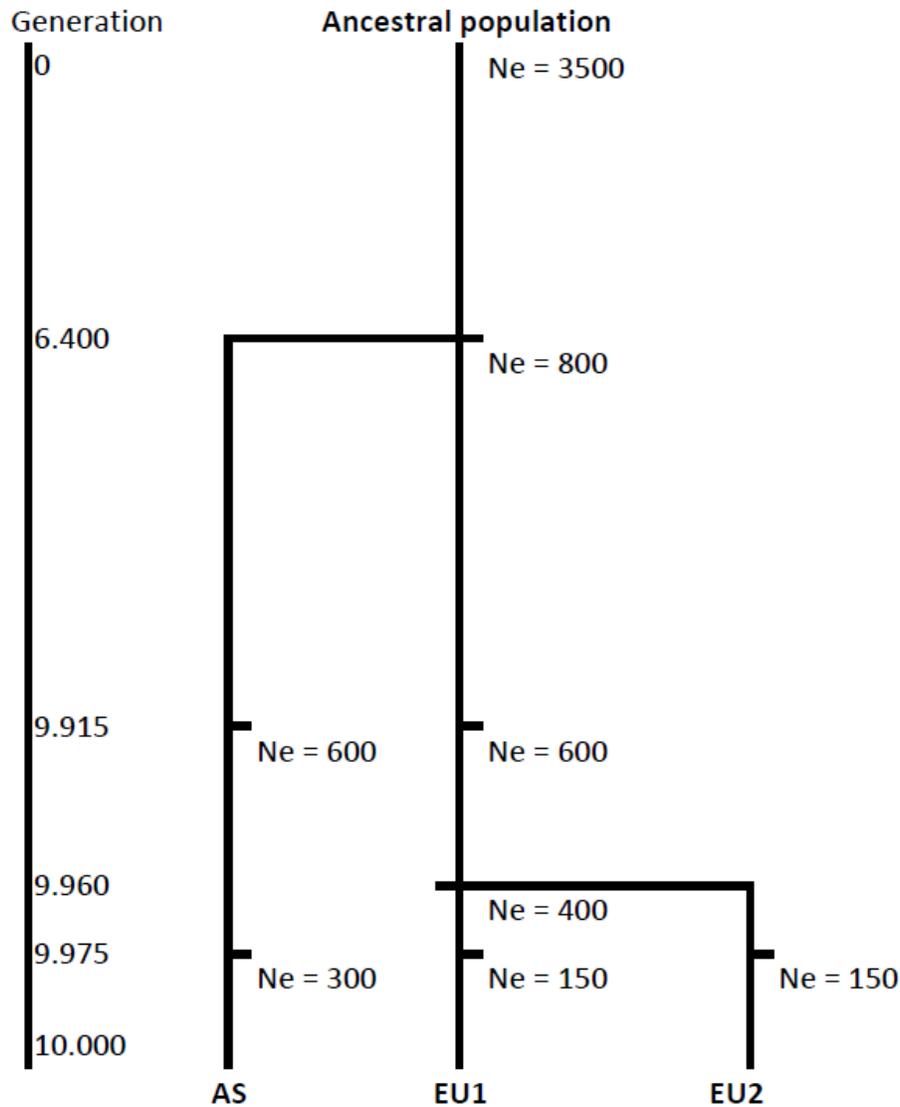
The aim of the present study was to analyze the power and precision of GWASs with whole-genome sequence data in several types of F2 crosses, including pooled crosses. Particular emphasis was paid to founder breed segregating QTNs because these are of interest for breeding purposes. The crosses were established using distantly or closely related founder breeds using stochastic simulations. The results were compared with those obtained from pooled F2 crosses, which increased the sample size and putatively reduced the LD. For comparison purposes, we also simulated one of the founder breeds and conducted GWASs within this breed.

## **Methods**

### ***Simulation of founder and F2 cross individuals***

Two porcine F2 crosses were simulated, one with closely related founder breeds and the other with distantly related founder breeds. One founder breed was the same in both crosses. A forward simulation approach was used to generate a Fisher-Wright diploid ancestral population, from which the founder breeds descended. The protocol to simulate the founder breeds is based on the knowledge of the phylogeny of pig breeds, especially the distinct lineages of European breeds and Asian breeds [5] and a sharp reduction of the effective population size over time due to intensified breeding schemes [14]. This protocol is described in detail in the following section and is also shown in Fig. 1. The ancestral population was simulated for 6400 generations with an

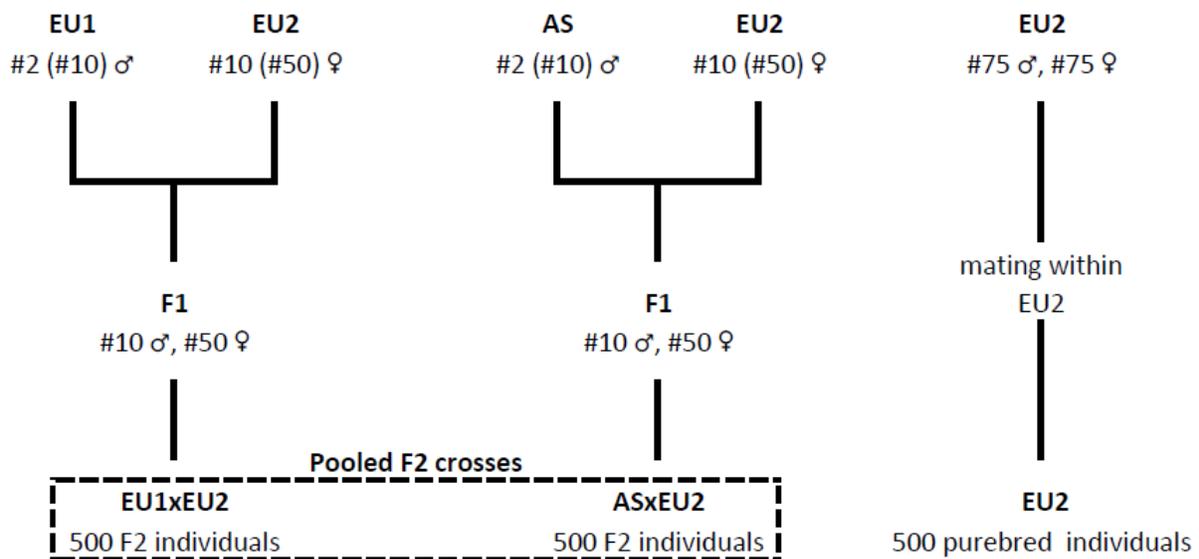
effective population size ( $N_e$ ) of 3500. In this generation, the ancestral population was split into two distinct lineages: the European and Asian lineages.



**Fig. 1** Phylogenetic history of the simulated founder breeds AS, EU1, and EU2

These lineages were simulated independently from each other from this generation onward. The Asian lineage was simulated until generation 9915 with a  $N_e$  of 800, from generation 9915 until generation 9960 with a  $N_e$  of 600 and 9975 until 10,000 generations with a  $N_e$  of 300. The last generation represented the Asian founder breed (AS). The European lineage was simulated from generation 6400 until generation 9915 with a  $N_e$  of 800 and from generation 9915 until

generation 9960 with a  $N_e$  of 600. In this generation, two breeds were generated from this lineage (breeds EU1 and EU2), which were simulated independently, until generation 9975 with a  $N_e$  of 400 and from generation 9975 until generation 10,000 with a  $N_e$  of 150. The level of genetic differentiation of the founder breeds was assessed by estimating the population differentiation index  $F_{ST}$ , using the formula (8) in Weir and Cockerham [15].



**Fig. 2** F2 schemes.

F2 schemes derived from phylogenetically closely (left) and distantly (middle) related founder breeds based on a small (large) number of F0 individuals as well as two generations of mating EU2 (right) as the purebred experimental population

From the three simulated founder breeds, two F2 crosses were generated, one with the closely related founder breeds EU1 and EU2 (EU1xEU2) and the other one with the distantly related founder breeds AS and EU2 (ASxEU2). The EU1xEU2 cross was established as follows. Founder animals were randomly selected from the founder breeds. The number of founder animals to establish an F2 crosses varied in real experiments, with usually a lower number of males compared to females (e.g. [6, 7]). In order to mimic this variable number of founder animals in our simulation, two different numbers were selected: two and ten males were selected from EU1 and ten and 50 females were selected from EU2. These animals were mated to create

ten male and 50 female F1 offspring. Each F1 male was mated to five F1 females with an assumed litter size of ten. Each female was allocated to only one male. This mating scheme resulted in 500 F2 EU1xEU2 individuals. Hence, we simulated two EU1xEU2 crosses, one with many and one with few founder animals. The same protocol was used to simulate the ASxEU2 cross; however, AS was the paternal founder breed. Both crosses shared one founder breed (EU2), but the founder animals from this breed were different. The datasets were pooled for the joint analyses of both crosses, also shown in Fig. 2.

### *Genomes and traits*

Ten chromosomes of one Morgan (M) length each were simulated. The pig genome consists of more than ten chromosomes, but we restricted this number for computational reasons. Recombination events were simulated according to the Haldane mapping function. The mutation rate was adjusted in a simplified manner so that two mutations per chromosome (20 per genome), on average, were expected to occur per meiosis. All SNPs were generated solely by the mutations within the evolution of the simulated populations. This protocol was repeated ten times. For each population, five traits were simulated, which resulted in 50 replicates in total. For each trait, 20 SNPs per chromosome were randomly selected to become a QTN, which resulted in 200 QTN to mimic the polygenic nature of quantitative traits [13]. Because the QTN were randomly selected, the traits were assumed to be unselected. This might be a simplification, because in reality some traits in F2 crosses are under selection in the founder breeds. However, considering this in a simulation is not straightforward and would result in additional assumptions. The minimum distance between QTNs was 2 Centimorgan (cM). The additive effects were sampled from a  $t$ -distribution with four degrees of freedom and were assumed to be the same for the two single crosses. This result roughly resembled the distribution of additive effects in porcine F2 crosses [16]. Breeding values were calculated for the individuals as follows [17]. For an individual with genotype  $x$  ( $x$  representing the number of copies of the mutant allele at QTN,  $x = 0, 1, \text{ or } 2$ ), the breeding value ( $BV$ ) is

$$BV(x) = \sum_{j=1}^q (x_j - 2p_j) a_j,$$

with  $a_j$  being the simulated additive effect,  $p_j$  the frequency of the mutant allele, and  $Q$  the number of simulated QTN. The additive genetic variance was calculated as the variances of the breeding values within the pooled dataset. Hence, the additive genetic variance differed slightly between the crosses due to different gene frequencies at the QTNs. The gene frequencies were more intermediate in the ASxEU2 cross compared to the EU1xEU2 cross, because in the former cross the two founder breeds were less related. However, in general the differences of the gene frequencies in both crosses were small. A random residual was added to the breeding values to complete the phenotypes of the individuals, assuming a heritability of 0.5 in the pooled dataset. In addition, 1000 EU2 individuals were simulated using the same procedures. Note that the LD structure of these types of simulated F2 crosses were investigated in detail in an earlier study [11], and hence it was not included in this study.

### *Association mapping*

All SNPs and QTNs with a minor allele frequency (MAF) below 0.05 in the individual crosses were removed from the following analyses. GWASs were conducted for each SNP (also for each QTN) separately by using the following regression model and the software GCTA (Genome-wide Complex Trait Analysis) [18]:

$$y_i = \mu_{k_i} + b_j x_{ij} + g_i + e_i.$$

Here,  $y_i$  denotes the phenotypic value of individual  $i$ ,  $\mu$  denotes the overall mean of the cross  $k$  to which individual  $i$  belongs,  $x_{ij}$  denotes the number of copies of a randomly chosen allele of SNP  $j$  ( $x_{ij} = 0, 1, \text{ or } 2$ ) and  $b_j$  is the regression coefficient for SNP  $j$ . The random polygenic effect of the individual ( $g_i$ ) was fit to capture population stratification effects. The covariance structure of the polygenic effects was modeled using a genomic relationship matrix (GRM) [18]. To avoid the pitfall of double fitting the SNP to be tested simultaneously as a fix and a random effect, a leave-one-chromosome-out approach was applied, as recommended [18]. This approach meant that when the SNP effects were tested for significance on a certain chromosome, the SNPs on this chromosome were excluded from the calculation of the GRM. The correction for multiple testing was conducted using the Bonferroni method. The two crosses were analyzed both separately and jointly (i.e., the pooled datasets). The slightly larger additive genetic variance of the cross

ASxEU2 was accommodated by the GRM, in which the off-diagonal elements were larger for the individuals in this cross compared with the corresponding elements of the EU1xEU2 cross.

### ***Scenarios***

Two GWAS scenarios were considered. In the *all segregating genes* (ASG) scenario, the aim was to map all available QTNs. Association mapping was performed on the full set of SNPs and QTNs. However, from a breeder's perspective, GWAS results are most important for QTNs segregating in the breed of interest. As economically relevant breeds (e.g., the Piétrain breed) were often used as founder breeds in F2 crosses, a second scenario, the so-called *founder segregating genes* (FSG) scenario was considered. The aim was to map QTNs that segregated in the common founder breed EU2. Consequently, all SNPs and QTNs that did not segregate in the common founder breed EU2 were removed from the simulated datasets because they could be excluded beforehand as putative QTNs. The association analyses were subsequently conducted using these reduced data sets. Consequently, the number of tests were much smaller and so were the levels of multiple testing corrections using Bonferroni. Both GWAS scenarios (ASG and FSG) were applied to all simulated F2 crosses.

For comparison purposes, the simulated purebred EU2 data set was also analyzed.

### ***Calculation of QTN and QTL mapping power and QTN mapping precision***

The set of SNPs was denoted as  $S$ . In the ASG scenario,  $S$  contained all segregating SNPs ( $MAF > 0.05$  in the respective cross), but in the FSG scenario, the set included only SNPs also segregating in the common founder breed ( $MAF > 0.05$  in EU2). Thus,  $Q_\alpha \subseteq Q \subseteq S$ , where  $Q$  is the set of QTNs, and  $Q_\alpha$  contains all simulated QTNs with a Bonferroni corrected  $p$ -value less than  $\alpha$ . We calculated the power to map QTNs as the proportion of QTNs with a Bonferroni corrected  $p$ -value smaller than  $\alpha$ , i.e.,  $QTN\ power = \frac{\#Q_\alpha}{\#Q}$ , where  $\#Q$  denotes the number of elements in set  $Q$ . This definition is in agreement with classical statistical test theory. QTL power was defined as the proportion of QTNs, which are either mapped per se or by a significant SNP in LD with the QTN. To determine whether a QTN  $i$  can be detected through a SNP in high LD, a window  $W_i$  was defined spanning 1 Centimorgan (cM) with the QTN in the center. This window

defined the QTL region. If the SNP with the smallest  $p$ -value was significant in such a QTL window, the QTN was indicated by this SNP and, therefore, was mapped. Hence, the QTL power was calculated as  $QTL\ power = \frac{\#W_\alpha}{\#Q}$ , with  $\#W_\alpha$  being the number of windows, which contained a QTN and at least one significant SNP within these windows.

The windows were also considered to specify the precision of mapping QTNs. For each window containing a significant QTN, the proportion of SNPs showing a higher significance than the QTN itself was computed. The QTN mapping precision was then calculated by subtracting this proportion from 1. This step ensured that the maximum achievable mapping precision was one, which implied that the QTN showed the highest significance among all SNPs in the window. By contrast, a precision close to 0.5 indicated that 50 % of the significant SNPs were more significant than the causal mutation.

The parameters QTN power, QTL power, and QTN precision were calculated for each analyzed data set and then averaged across the simulated replicates.

## Results and Discussion

### *Simulation structure*

Maximum marker density was simulated, which resembles a situation where the whole genome sequence variants are known from each F2 individual. In real porcine F2 crosses, sequencing all F2 individuals is still unaffordable, but the Illumina PorcineSNP60 BeadChip [9] with approximately 62 k SNPs can be used to impute sequence data from founder individuals in the F1 and subsequently in the F2 generation utilizing mainly pedigree information. Thus, the sequence data of F2 individuals can be generated by sequencing the founder individuals and SNP-chip genotyping the F1 and F2 generation, which is affordable in many situations. Although this strategy was not evaluated so far, it can reasonably be assumed that the imputation accuracy will be high.

The  $F_{ST}$  value calculated between the two simulated European breeds EU1 and EU2 was  $F_{ST} = 0.02$ , and between the Asian breed AS and the European breed EU2 it was  $F_{ST} = 0.36$ . These values implied a small (large) genetic differentiation between EU1 and EU2 (AS and EU2) [19].

Hence, although simplified assumptions during the establishment of the simulation protocol had to be made, it fits roughly the genetic differentiation of typical real pig founder breeds.

**Table 1** Number of SNPs (MAF > 0.05) within the respective datasets for the ASG and FSG scenarios

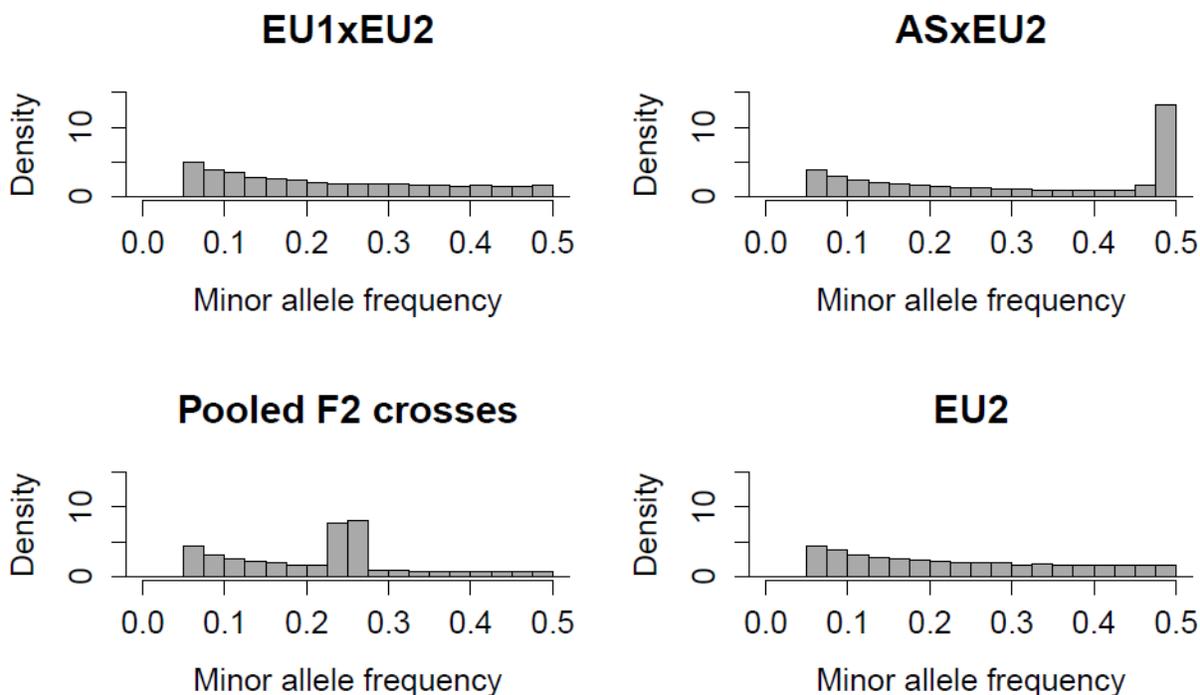
Scenario	ASG		FSG	
	mean	sd	mean	sd
EU2	100,783	683		
EU1xEU2 (small F0)	97,490	974	83,228	717
EU1xEU2 (large F0)	104,797	965	89,192	657
ASxEU2 (small F0)	237,574	698	79,106	619
ASxEU2 (large F0)	247,726	1,026	81,688	657
Pooled F2 crosses (small F0)	248,302	1,805	86,444	693
Pooled F2 crosses (large F0)	240,115	897	89,607	746

The means and standard deviations across all simulated datasets are shown

The average number of SNPs across all replicates in the ASG scenario with an MAF > 0.05 within the respective populations is given in Table 1. The MAF of SNPs with a MAF > 0.05 within the experimental populations are shown in Fig. 3 for a randomly chosen replicate. In most scenarios, the number of segregating SNPs in the F2 designs was higher compared with the founder population even though the numbers of founder individuals of the F2 crosses were limited. This increase was substantial, especially in the ASxEU2 cross. This was due to the numerous SNPs that were divergently fixed (or close to fixation) in the distantly related founder breeds but was segregating in the F2 cross, as shown in Fig. 3. Pooling data from both designs increased the number of SNPs only slightly (Table 1). The LD structure of the simulate crosses can be found in [11].

For the FSG scenario, the average numbers of SNPs with an MAF > 0.05 in the EU2 and the F2 crosses are given in Table 1. The numbers of SNPs were similar in all crosses and were lower than the number for the purebred population. The smallest number was observed in ASxEU2, which was derived from a small number of F0 individuals, because AS had many private alleles, and, therefore, shared fewer SNPs with the EU2 breed. A higher number of SNPs could be

observed if the F2 designs were based on a larger number of founder individuals. The number of SNPs was the highest in pooled data.



**Fig. 3** Minor allele frequencies for a randomly chosen trait.

Minor allele frequencies for all SNPs with a MAF > 0.05 within the two F2 crosses (top line), their pooled data and the purebred experimental population (bottom line)

### *Mapping power and precision*

The power to detect a QTN or at least one significant SNP within a 1 cM window around a QTN (i.e., a QTL) is given in Table 2 for the ASG scenario. This result showed that the mapping power was higher in ASxEU2 than in EU1xEU2. That was attributed to various mutations that were divergently fixed in the distantly related founder breeds and therefore were segregating with a high MAF in the F2 generation. By contrast, QTNs segregating in EU1xEU2 had more extreme allele frequencies. Because the QTL contributions to the total genetic variance strongly depended on allele frequencies, the power in ASxEU2 was substantially higher than in EU1xEU2. Additionally, the LD blocks are larger F2 crosses derived from distantly related founder breeds (like the ASxEU2) [11], and several QTNs may have been in LD with the QTN being tested, which may increase the effect explained by the QTN. Hence, the mapping power (especially QTL

mapping power) is higher in such designs where more SNPs are in LD with a QTN. The power was highest when the datasets were pooled and analyzed jointly, which resulted from the larger sample size. The mapping power depended only slightly on the number of founder individuals (Table 2).

A low mapping precision was observed in F2 crosses with phylogenetically strongly divergent founder breeds (Table 2). This is because the number of divergently fixed alleles was very high (Fig. 3) and, therefore, LD blocks large. The pooling of data resulted in a precision that was between the precision of both F2 crosses.

**Table 2** QTN and QTL mapping power and QTN mapping precision in the ASG scenario at a genome-wide significance level of  $\alpha = 0.05$

Parameter	QTN Power		QTL Power		QTN Precision	
	mean	sd	mean	sd	mean	sd
EU1xEU2 (small F0)	0.014	0.017	0.019	0.027	0.784	0.183
EU1xEU2 (large F0)	0.009	0.012	0.012	0.014	0.823	0.098
ASxEU2 (small F0)	0.064	0.044	0.127	0.081	0.672	0.105
ASxEU2 (large F0)	0.063	0.040	0.122	0.068	0.619	0.116
Pooled F2 crosses (small F0)	0.070	0.047	0.141	0.088	0.678	0.108
Pooled F2 crosses (large F0)	0.070	0.042	0.134	0.074	0.661	0.134

The means and standard deviations across all simulated replicates are shown

In the FSG scenario, the EU2 was the breed of interest, and the aim was to map QTNs segregating in this breed. The results for QTL and QTN power in this scenario are shown in Table 3. In ASxEU2, the QTN power was higher than in EU1xEU2. The reasons were the same as in the ASG scenario, such as more QTNs may have intermediate allele frequencies and several QTNs may be in LD with the QTN being tested. Pooling the data again led to an increase in power due to a larger sample size. However, it could not reach the mapping power in the purebred population at an equal sample size whose trait was simulated to have the same heritability. The reason for this result may be that the distribution of allele frequencies was U-shaped in the purebred population. Consequently, the distribution of the contributions of QTNs to the phenotypic variance was more heavy-tailed, as it can be seen in Fig. 3. The QTN mapping

power as a function of the QTN size is shown in Fig. 4. The QTL power was substantially larger than the QTN power in crosses with distantly related founder breeds and a small number of founders because more SNPs were in strong LD with the QTNs.

**Table 3** QTN and QTL mapping power and QTL mapping precision in the FSG scenario at a genome-wide significance level of  $\alpha = 0.05$

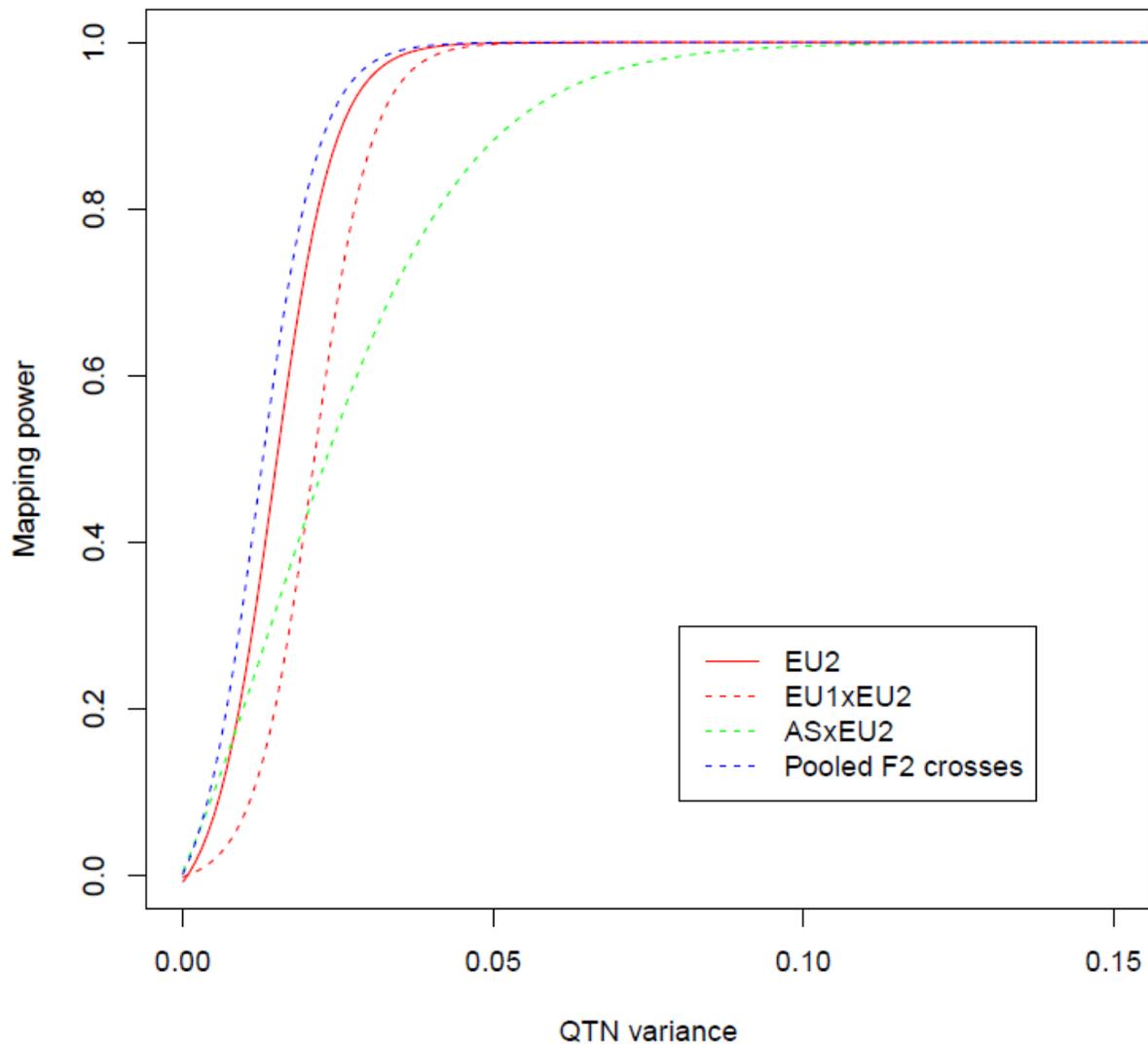
Parameter	QTN Power		QTL Power		QTN Precision	
	mean	sd	mean	sd	mean	sd
EU2 (500)	0.031	0.015	0.037	0.017	0.748	0.202
EU2 (1000)	0.076	0.026	0.090	0.028	0.874	0.080
EU1xEU2 (small F0)	0.015	0.019	0.020	0.026	0.758	0.191
EU1xEU2 (large F0)	0.011	0.014	0.013	0.016	0.834	0.093
ASxEU2 (small F0)	0.036	0.031	0.125	0.081	0.725	0.212
ASxEU2 (large F0)	0.033	0.024	0.114	0.068	0.626	0.195
Pooled F2 crosses (small F0)	0.050	0.036	0.138	0.080	0.788	0.181
Pooled F2 crosses (large F0)	0.038	0.024	0.113	0.065	0.757	0.182

The means and standard deviations across all simulated replicates are shown

As shown in Table 3, the precision in the FSG scenario was the highest for the EU1xEU2 cross with a large number of founder animals. The precision was even higher than in the purebred EU2 population with 500 individuals in the analysis. The high precision of the closely related cross resulted from the fact that LD blocks in crossbred populations may have been shorter than in the purebred populations [12]. The lowest precision was observed in the crosses of distantly related breeds.

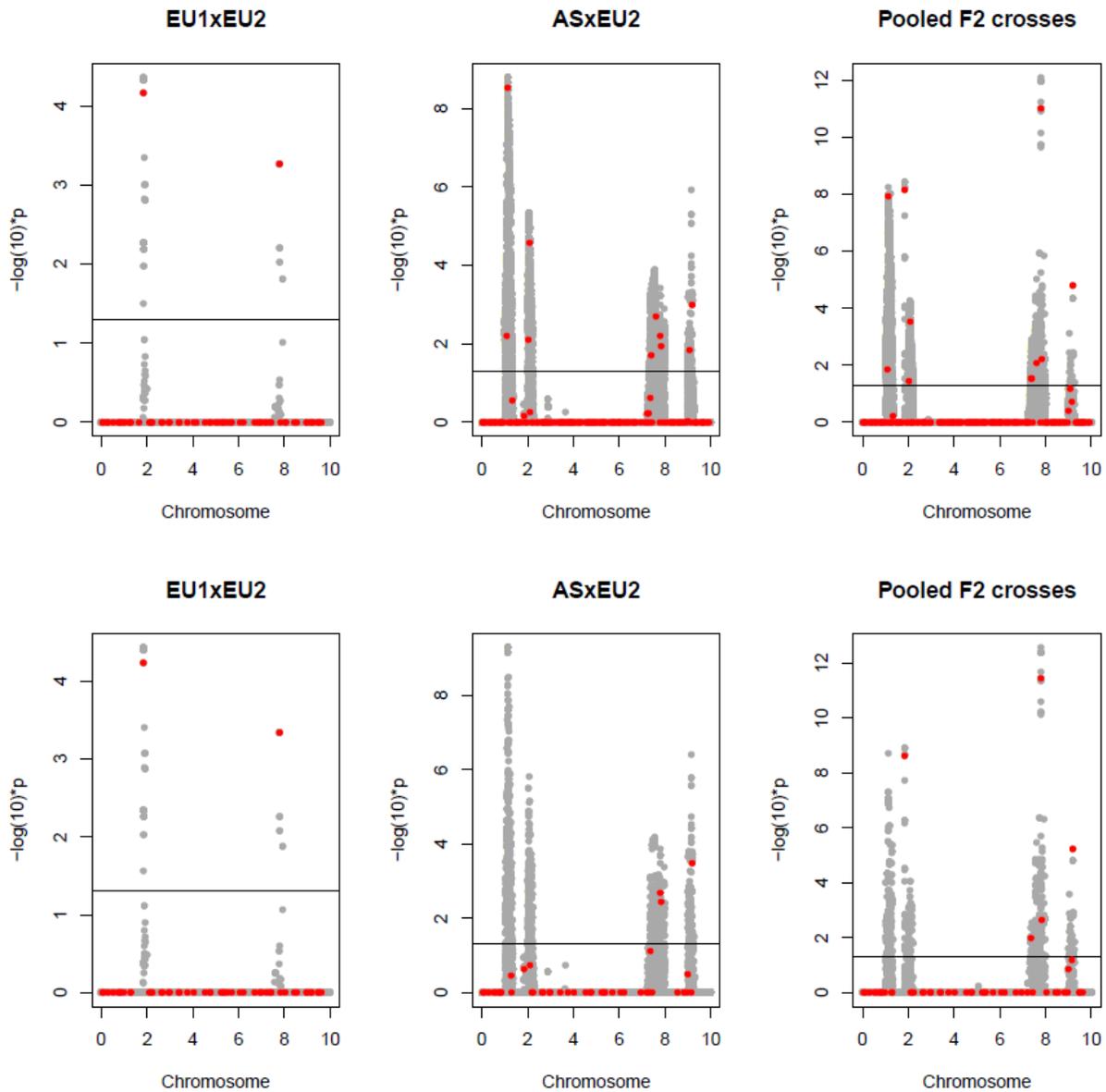
The precision in the FSG scenario was always above the precision in the ASG scenario. This result is in agreement with [11] for which the highest mapping resolution was observed in F2 populations for genes that also segregated in a founder breed.

The general pattern of the mapping power and precision results in the simulated populations and scenarios as described above is visualized by a comparison of the Manhattan plots for a randomly chosen replicate in Fig. 5.



**Fig. 4** QTN mapping power as a function of the QTN variance.

Mapping power as a function of the QTN variance (contribution to the phenotypic variance (VP)) averaged across all replicates



**Fig. 5** Manhattan plots of both scenarios for a randomly chosen replicate.

Test statistics ( $-\log(10)*p$ -value) and the position of SNPs (gray dots) and QTN (red dots) segregating in the F2 crosses (ASG scenario, top line) and also within the common founder breed EU2 (FSG scenario, bottom line) for both F2 crosses and their pooled data. The solid line corresponds to a genome-wide significance level of  $\alpha = 0.05$

## Conclusions

Based on the results of this simulation study, it can be concluded, that the existing F2 crosses are promising databases for gene mapping in the era of genomics when the founder breeds are closely related or when crosses can be pooled. For the fine-mapping of QTNs, F2 crosses from distantly related founder breeds should be pooled with data from additional populations in which the QTNs of interest are segregating. This step could substantially increase the precision. By contrast, the mapping precision could be even higher in F2 crosses from closely related founder breeds than in the purebred population; thus, no pooling would be required if the sample size and the numbers of founders in the F0 are sufficiently high. This conclusion is particularly true for QTNs that segregate in a founder breed.

## Abbreviations

AS: Asian founder breed; ASG: All segregating genes; ASxEU2: F2 cross derived from distantly related founder breeds; BV: Breeding value; cM: Centimorgan; EU1: European founder breed 1; EU1xEU2: F2 cross derived from closely related founder breeds; EU2: European founder breed 2; FSG: Founder segregating genes;  $F_{ST}$ : Population differentiation index; GCTA: Genome-wide complex trait analysis; GRM: Genomic relationship matrix; GWAS: Genome-wide association study; LD: Linkage disequilibrium; M: Morgan; MAF: Minor allele frequency;  $N_e$ : Effective population size; QTL: Quantitative trait locus; QTN: Quantitative trait nucleotide; SNP: Single nucleotide polymorphism

## Acknowledgements

This study was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

### **Funding**

This research was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). The funding bodies did not contribute to the design of the study or collection, analysis and interpretation of data and writing the manuscript.

### **Availability of data and materials**

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

### **Authors' contributions**

MS performed the statistical analyses. JB and RW initiated the study and oversaw the statistical analyses. RW designed the population simulation software. All authors wrote the paper and read and approved the final manuscript.

### **Ethics approval**

This is a simulation study. Animals have not been directly used in this study.

### **Consent for publication**

No applicable.

### **Competing interests**

The authors declare that there are no competing interests.

**References**

1. Goddard ME, Kemper KE, MacLeod IM, Chamberlain AJ, Hayes BJ. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc Biol Sci.* 2016;283:1-9.
2. Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol.* 2015;47:43.
3. Rothschild MF, Hu ZL, Jiang Z. Advances in QTL mapping in pigs. *Int J Biol Sci.* 2007;3:192-7.
4. Hu ZL, Park CA, Reecy JM. Developmental progress and current status of the animal QTLdb. *Nucl Acids Res.* 2016;44:827-33.
5. Frantz LA, Schraiber JG, Madsen O, Megens HJ, Bosse M, Paudel Y, et al. Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*. *Genome Biol.* 2013;14:R107.
6. Boysen TJ, Tetens J, Thaller G. Detection of a quantitative trait locus for ham weight with polar overdominance near the ortholog of the callipyge locus in an experimental pig F2 population. *J Anim Sci.* 2010;88:3167-72.
7. Rückert C, Bennewitz J. Joint QTL analysis of three connected F2-crosses in pigs. *Genet Sel Evol.* 2010;42:40.
8. Stratz P, Wellmann R, Preuss S, Wimmers K, Bennewitz J. Genome-wide association analysis for growth, muscularity and meat quality in Piétrain pigs. *Anim Genet.* 2014;45:350-6.
9. Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One.* 2009;4:e6524.
10. Ledur MC, Navarro N, Pérez-Enciso M. Large-scale SNP genotyping in crosses between outbred lines: how useful is it? *Heredity.* 2009;105:173-82.
11. Bennewitz J, Wellmann R. Mapping Resolution in Single and Multiple F2 Populations using Genome Sequence Marker Panels. In: *Proceedings of the 10th World Congress on*

- Genetics Applied to Livestock Production: 17-22 August 2014; Vancouver.  
[https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/184\\_paper\\_8399\\_manuscript\\_106\\_0.pdf?sfvrsn=2](https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/184_paper_8399_manuscript_106_0.pdf?sfvrsn=2).
12. Toosi A, Fernando RL, Dekkers JCM. Genomic selection in admixed and crossbred populations. *J Anim Sci*. 2010;88:32-46.
  13. Wellmann R, Bennewitz J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genet Res (Camb)*. 2012;94:21-37.
  14. Wang J, Zou H, Chen L, Long X, Lan J, Liu W, et al. Convergent and divergent genetic changes in the genome of Chinese and European pigs. *Sci Rep*. 2017;7:8662.
  15. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984;38:1358-70.
  16. Bennewitz J, Meuwissen TH. The distribution of QTL additive and dominance effects in porcine F2 crosses. *J Anim Breed Genet*. 2010;127:171-9.
  17. Falconer DS, Mackay TFC. *Introduction to quantitative genetics*. London: Longman; 1996.
  18. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88:76-82.
  19. Nielsen R, Slatkin M. *An Introduction to population genetics*. Sunderland: Sinauer Associates; 2013.



## CHAPTER THREE

### **Linkage disequilibrium pattern and genome-wide association mapping for meat traits in multiple porcine F<sub>2</sub> crosses**

**Patrick Stratz<sup>\*</sup>, Markus Schmid<sup>\*</sup>, Robin Wellmann<sup>\*</sup>, Siegfried Preuß<sup>\*</sup>, Iulia Blaj<sup>†</sup>, Jens Tetens<sup>‡</sup>, Georg Thaller<sup>†</sup>, and Jörn Bennewitz<sup>\*</sup>**

<sup>\*</sup>Institute of Animal Science, University Hohenheim, Garbenstraße 17, 70599 Stuttgart, Germany

<sup>†</sup>Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel, Hermann-Rodewald-Straße 6, 24118 Kiel, Germany

<sup>‡</sup>Functional Breeding Group, Department of Animal Science, Georg-August-University Göttingen, Burckhardtweg 2, 37077 Göttingen, Germany

Corresponding author: Patrick.Stratz@uni-hohenheim.de

Published in:

**Animal Genetics** (2018)

doi: 10.1111/age.12684.

**Summary**

In the present study, data from four F2 crosses were analysed and used to study the linkage disequilibrium (LD) structure within and across the crosses. Genome-wide association analyses (GWASes) for conductivity and dressing out meat traits were conducted using single-marker and Bayesian multi-marker models using the pooled data from all F2 crosses. Porcine F2 crosses generated from the distantly related founder breeds Wild Boar, Piétrain and Meishan, as well as from a porcine F2 cross from the closely related founder breed Piétrain and an F1 Large White x Landrace cross were pooled. A total of 2572 F2 animals were genotyped using a 62K SNP chip. The positions of the SNPs were based on genome assembly Sscrofa11.1. After post-alignment and genotype filtering, approximately 50K SNPs were usable for LD studies and GWASes. The main findings of the present study are that the breakdown of LD was faster in crosses from closely related founder breeds compared to crosses from distantly related founders. The fastest breakdown of LD was observed by pooling the data. Based on the single-marker results and LD structure, clusters and windows were built for 1-Mb intervals. For conductivity and dressing out, 183 and 191 nominal significant associations respectively and six and five clusters respectively were found. Dominance was important for conductivity, and considering dominance in GWASes improved the mapping signals. Most clear signals were found for conductivity on SSC6, 8 and 15 and for dressing out on SSC2 and 7. Considering dominance might contribute to the accuracy of genomic selection and serve as a guide for choosing mating pairs with good combining abilities. However, further research is needed to investigate if dominance is also important in crossbreed pig breeding schemes.

**Keywords**

association analysis, population structure, porcine F2-cross design, pork

## Introduction

In pig breeding, numerous F2 crosses have been previously established to map QTL (Rothschild *et al.* 2007). The individuals were typically genotyped for a relatively small number of microsatellite markers, and QTL mapping was performed using linkage analysis. Several QTL for a wide variety of trait complexes were identified (Hu *et al.* 2005), but the mapping precision was generally low due to the small number of individuals, low marker densities and the use of linkage analysis, which utilizes only meioses that occurred within the last generations. Analysing several F2 crosses jointly has proven to be a suitable tool for increasing the sample size and mapping resolution and thus the power and precision in QTL linkage mapping (Rückert & Bennewitz 2010). With the advent of the Porcine 62K SNP chip (Ramos *et al.* 2009), it is possible to conduct genome-wide association analyses (GWASes) in pig breeding. Schmid *et al.* (2016) used stochastic simulation studies to analyse the prospects of GWASes in porcine F2 crosses genotyped with dense SNP marker panels. The main results of Schmid *et al.* (2016) were that power and precision in GWASes were high when the founder breeds used to establish the cross were closely related (e.g. two European-type founder breeds) and when the sample size can be enlarged (e.g. by pooling data from several F2 crosses). For the latter, it is beneficial if at least one common founder breed is used in the crosses to be pooled.

Piétrain is an important sire line pig breed in Europe, selected for muscularity and lean meat content, but meat quality within this breed is an issue. The breed was used as a founder breed in three F2 crosses (Boysen *et al.* 2010; Rückert & Bennewitz 2010). As genomic selection is practised in this breed (Wellmann *et al.* 2013), it might be of interest to map SNPs associated with meat quality to improve genomic predictions.

In the present study, the data from the three F2 crosses mentioned above and one additional cross without Piétrain as a founder breed were pooled. Two crosses included Asian pigs, which had meat-quality-associated haplotypes introgressed into European pigs during the Industrial Revolution (Bosse *et al.* 2014). The aim of the present study was to examine the linkage disequilibrium (LD) structure, and thus mapping resolution, within and across the crosses. A second objective was to identify SNPs associated with meat traits, applying single-SNP and Bayesian GWASes. A further objective was to compare the results of both GWAS methodologies and stress their potential contribution to genetics, selection or breeding for meat traits.

## Methods

### *Animals*

Three porcine F2 crosses were generated from three distantly related founder breeds Meishan (five individuals), Piétrain (14 individuals) and Wild Boar (one individual), resulting in Wild Boar x Piétrain (WxP), Wild Boar x Meishan (WxM) and Meishan x Piétrain (MxP) crosses (Rückert & Bennewitz 2010). The MxP cross was obtained by mating one Meishan boar with eight Piétrain sows. The WxP cross was generated by mating one Wild Boar with nine Piétrain sows, some of which were the same as in the MxP cross. The WxM cross was obtained by mating the same Wild Boar with four Meishan sows. The numbers of F1 individuals in the MxP, WxP and WxM crosses were 22, 28 and 23 respectively. A total of 291 F2 individuals were in the WxP group, 304 F2 individuals were in the MxP group and 312 F2 individuals were in the WxM group. The experimental design of closely related founder breeds was a cross of the Piétrain breed and an F1 Large White 9 Landrace (PxLWL) population (Boysen *et al.* 2010). A total of 1665 F2 individuals were in the PxLWL group, which consisted of eight full-sib families (FS1-FS8) by mating five Piétrain boars to eight LWL crossbred sows. A total of 134 F2 individuals were in family FS1, 225 in family FS2, 205 in family FS3, 228 in family FS4, 234 in family FS5, 211 in family FS6, 165 in family FS7 and 263 in family FS8.

### *Genotypes*

All individuals were genotyped using the Porcine 62K SNP chip (Ramos *et al.* 2009) and for the *RYR1*:g.1843C>T mutation (subsequently denoted as *RYR1*). The number of *RYR1* genotypes is presented in Stratz *et al.* (2013). The locations of the markers on the Porcine 62K SNP chip were retrieved from genome assembly Sscrofa11.1 (Warr *et al.* 2016). The command line standalone BLAST+ (Camacho *et al.* 2009) was applied to align a nucleotide query to a nucleotide database. The queried sequences were obtained from the Porcine 62K SNP v2.0 Annotation File, available from Illumina, Inc. The FASTA format of the genomic sequences in the assembly (GCA\_000003025.5\_Sscrofa11\_genomic.fna) was used to set up the BLAST database. After alignment, post-alignment filtering criteria were applied, a step that led to the identification of 59 762 SNPs. Markers exceeding the stringent cut-offs are listed in Table S1, and their identifiers

and positions are provided. Markers on the sex chromosomes (represented by X) were excluded prior to genotype filtering.

Genotypes from F2 individuals were filtered with respect to call rate (removal of SNPs with a call rate less than 95 %), parent-progeny conflicts (removal of SNPs with parent-progeny conflicts greater than 0), MAF (exclusion of SNPs with a minor allele frequency smaller than 0.01 %), call frequency (exclusion of SNPs with a call frequency below 0.95), cluster separation (exclusion of SNPs with a cluster separation below 0.4) and heterozygosity excess (exclusion of SNPs with a heterozygosity excess greater or equal than 0.15). Genotype filtering was performed using GENOMESTUDIO software and the guidelines from Illumina ([https://www.illumina.com/Documents/products/technotes/technote\\_infinium\\_genotyping\\_data\\_analysis.pdf](https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf)).

After post-alignment and genotype filtering, 49 690 autosomal SNPs were retained for further analyses. The distribution of the autosomal SNPs in the porcine genomic sequence is shown in Fig. S1. The absolute number of SNPs ranged from 1272 SNPs on *Sus scrofa* chromosome (SSC) 18 to 5998 SNPs on SSC1. The SNP density in 1-Mb intervals was between 0 and 52 SNPs. The mean heterozygosity of the SNPs was 0.370.

### ***Phenotypes***

In the present study, one meat quality trait, cond45 (conductivity in mS/cm 45 min post-mortem) as well as dressing out (%; as a ratio of carcass weight to live weight at slaughter), were investigated. For cond45, measurements were made in the longissimus dorsi muscle between the 13th and 14th ribs. Descriptive statistics are provided in Table 1. Herein, the coefficient of variation, as a standardized measurement of the variance, was between 3.21 for dressing out and 71.37 for cond45. The traits were pre-corrected for environmental effects (Boysen *et al.* 2010; Rückert & Bennewitz 2010) and the effect of *RYR1* (Fujii 1991). To account for the population structure, the effect of the four crosses was pre-corrected. The traits were transformed to a mean of 0 and variance of 1. Phenotypes and genotypes are available on request from the corresponding author.

**Table 1** Mean, standard deviation (SD), minimum (Min) and maximum (Max) of the phenotypic observations and coefficient of variation (CV).

Trait	Mean	SD	Min	Max	CV
Cond45 (mS/cm)	5.17	3.69	2.34	24.90	71.37
Dressing out (%)	78.00	2.50	70.00	89.60	3.21

Pooled data from 2572 observations.

Cond45, conductivity 45 min post-mortem.

### *Estimation of the $F_{ST}$ index*

The population differentiation index  $F_{ST}$  was used to quantify the levels of differentiation between the founder breeds (Weir & Cockerham 1984; Holsinger & Weir 2009). A small  $F_{ST}$  (e.g.  $F_{ST} < 0.05$ ) indicates that allelic frequencies in both subpopulations are similar, whereas an  $F_{ST}$  above 0.05 indicates that allelic frequencies are different.  $F_{ST}$  values were estimated for each SNP between the founder breeds Piétrain and Meishan and the F1 LWL cross (Weir & Cockerham 1984). The  $F_{ST}$  estimates were combined across the SNPs by taking the average. For Wild Boar, no estimation was possible because there was only one founder individual available.

### *Extent of LD*

Haplotypes were reconstructed, and sporadically missing genotypes were imputed for the F2 individuals using default settings in BEAGLE 3.3.2 (Browning & Browning 2008). Reconstructed haplotypes were used to estimate  $r^2$  values (Hill & Robertson 1968) using PLINK software (Purcell et al. 2007) for marker pairs less than 5 Mb apart over the autosomes for the four crosses. The extent of LD was calculated separately for the WxP, MxP, WxM and PxLWL crosses as well as for the pooled crosses and was visualized using the R package SYNBREED (Wimmer *et al.* 2012). The fraction of marker pairs with different  $r^2$  levels was calculated for the autosomes for different inter-marker distances (in Mb) for the following bins: [0, 0.025), [0.025,0.05), [0.05, 0.075), [0.075, 0.12), [0.12, 0.2), [0.2, 0.5), [0.5, 1.5), [1.5, 3) and [3, 5). The extent of  $r^2$  was also visualized for 50-kb inter-marker distances over the autosomes.

### ***Population structure***

A segment-based realized relationship matrix was built on segments comprising at least 25 SNPs with a minimum length of 1 Mb and converted into a dissimilarity matrix. The contribution of the Piétrain founder breeds to the F2 individuals was calculated from the segments. For these calculations, the R package OPTISEL (Wellmann 2017) was used. The R-package SMACOF (de Leeuw & Mair 2009) was used to solve the stress target function for symmetric dissimilarities using the majorization approach and to report the normalised stress value. The genetic dissimilarity of two individuals was defined such that a low stress value was obtained. The structure of the population was visualized using multidimensional scaling (MDS) analysis.

### **Statistical analyses**

#### ***Variance component estimation and single-marker association analysis***

Variance component analysis and single-marker association analysis was performed using data from the pooled crosses. Variance components were estimated using a mixed model, where the pre-corrected trait of an F2-individual  $i$  ( $y_i$ ) is regressed on the fixed effect of the overall mean and the random effects of the genome-wide additive ( $g_{ai}$ ) and dominance genetic ( $g_{di}$ ) values, with a covariance structure  $g_{ai} \sim N(0, A\sigma_a^2)$  and  $g_{di} \sim N(0, D\sigma_d^2)$ , where  $A$  and  $D$  are the additive and dominance genomic relationship matrices (GRMs) including all SNPs respectively. In the AD Model,  $\sigma_a^2$  and  $\sigma_d^2$  were estimated by the restricted maximum likelihood approach, relying on the estimated GRMs. In the reduced model (A Model),  $g_{di} = 0$  and only  $\sigma_a^2$  was estimated. The errors  $e_i$  are assumed to be identical and independently distributed.

For GWAS, the A Model was applied to each SNP separately. The SNP effect was included as a fixed (gene substitution) effect in the model, and the pre-corrected phenotype was regressed on the number of copies of the allele with the lower frequency at the SNP  $m$  ( $x_{im} = 0, 1, 2$ ), i.e. the allelic content. The models were solved using GENOME-WIDE COMPLEX TRAIT ANALYSIS (GCTA; Yang *et al.* 2011). The GRM that was set up for the models included all autosomal SNPs except the autosome harbouring the SNP to be tested for association. The SNP was fitted as a fixed regression effect and a test statistic as well as point-wise error probability

( $p$ ) for the trait association, which was obtained in a frequentist manner, resulting in 49 690 nominal  $P$ -values. The genome-wide significance level was obtained using Bonferroni correction and was set at  $P_{genome-wide} \leq 0.05$ . Because Bonferroni correction is conservative, a nominal significance level (i.e.  $P < 5 \times 10^{-5}$ ) was also used. The number of false positives among the significant SNPs was calculated with the false discovery rate (FDR) using QVALUE software (Storey 2002; Storey & Tibshirani 2003). The FDR  $q$ -value of the significant SNP with the largest  $P$ -value provided an estimate of the proportion of false positives among the significant SNPs.

Clusters were built using the LD structure and nominally significant SNPs (Lutz *et al.* 2017). A cluster contained at least two genome-wide significant SNPs ( $P_{genome-wide} \leq 0.05$ ), with a maximum distance of 2 Mb between them. Starting from the midpoint of two genome-wide significant SNPs and moving in both directions up to 1 Mb on each side, we searched for nominally significant SNPs. The nominally significant SNPs at a maximum distance of 1 Mb from the cluster midpoint in both directions were used as the cluster bounds.

### ***Multi-marker association analysis***

In cases of imperfect LD, a single marker will only partly capture the effect of a causal mutation and a group of markers surrounding the causal variant may better explain the effect jointly. To account for this effect, multi-marker methods are proposed that fit all SNPs simultaneously as random effects in the model and account for the LD structure between the SNPs and the causal mutation (Fernando & Garrick 2013).

BayesC (Verbyla *et al.* 2009), which accounts for the distribution and the proportion of important SNPs in the prior assumption, and BayesD (Wellmann & Bennewitz 2012), which considers additionally both the additive and dominance effects, were applied. In the BayesD model, the prior assumption considers a small probability for a dominance effect to have a much larger magnitude than the additive effect; thus, overdominance is a rare but not negligible event. Only the hyperparameters, which were used to set up the prior distributions for BayesC and BayesD, will be described here. From the A Model,  $h_a^2$ , and from the AD Model,  $h_a^2$  and  $h_d^2$  were used as prior information for BayesC and BayesD respectively. The models assume that the distribution

of the effect of SNP  $m$  is a mixture of two  $t$ -distributions that differ by a scaling factor, which was set to 0.01. The marker effect is either allocated in the  $t$ -distribution with the larger variance with prior probability  $pLD = 0.02$  ( $pLD = 0.02 \approx \frac{1000}{49\ 690}$ ) or in the  $t$ -distribution with the smaller variance with prior probability  $1 - pLD$  ( $variance_{small} = 0.01 * variance_{large}$ ). The  $pLD$  was chosen based on the assumption that the number of SNPs associated with a QTL was 1000 and the number of SNPs for modelling the population structure was 48 690 (Stratz *et al.* 2014a).  $T$ -distributions were set up with 3 degrees of freedom for the additive effects for BayesC and BayesD. The Markov chain was generated by Gibbs sampling. For the joint posterior distribution of the additive and dominance effects see Wellmann & Bennewitz (2012). To assure that the SNP effects converged, 100 000 Gibbs sampling iterations were performed, the first 50 000 discarded as burn-in. Every 25th sample of the additive and dominance effects was stored for inference purpose. The models were solved using the R package BAYESDSAMPLES (Wellmann & Bennewitz 2012).

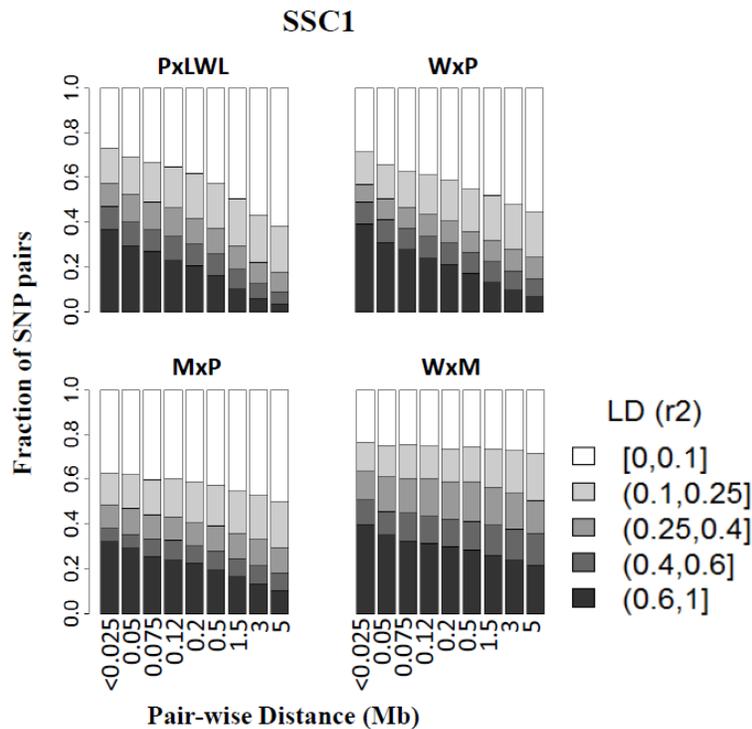
Bayesian posterior probabilities obtained from a single Markov Chain Monte Carlo analysis were used to make inferences on genomic windows. We calculated the window posterior probability of association (WPPA) criterion (Fernando & Garrick 2013). This method controls the proportion of false positives by calculating the posterior probability of a trait association for each SNP or each window of several consecutive SNPs. A sliding window approach was used to identify the most informative regions. WPPA values were calculated for 1-Mb sliding windows. A detailed description of the WPPA calculation for BayesC and BayesD can be found in Bennewitz *et al.* (2017). The 97.5% quantile of the WPPA was calculated, and windows that exceeded the threshold were declared as outliers and were stored.

Outlier WPPAs were compared to the results of the single-marker association analysis. The focus was on the allocation of significant SNPs to WPPAs. For SNPs flanking clusters built from single-marker association analysis results, the cluster interval was spanned up until the first (last) SNP in the sliding window. The overlapping results were written out thereafter.

## Results

### *F<sub>ST</sub> index and extent of LD*

The mean  $F_{ST}$  index over the SNPs was  $F_{ST} \approx 0.17$  for the Piétrain and Meishan founder breeds. Between the LWL and the Piétrain and Meishan founder breeds the index was  $F_{ST} \approx 0.07$  and  $F_{ST} \approx 0.25$  respectively.

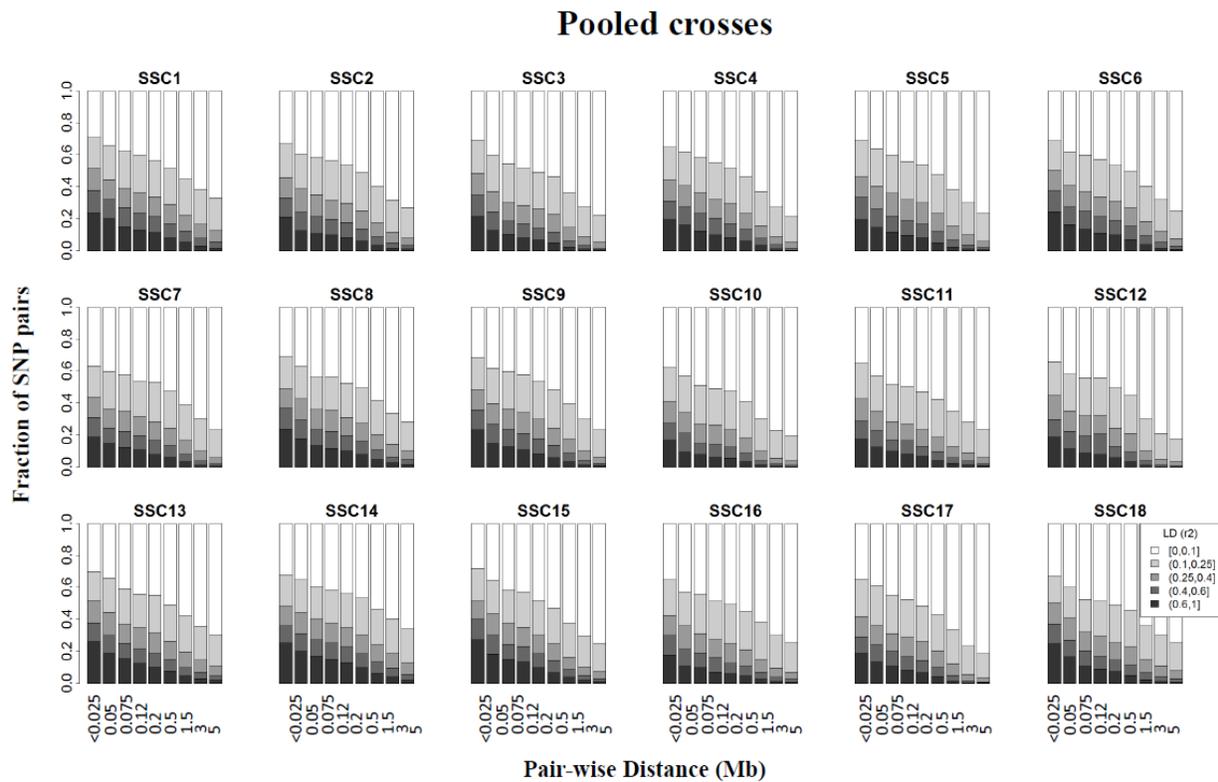


**Figure 1** Extent of linkage disequilibrium (LD) for SSC1 within the crosses.

Level of LD decay in each single cross [Piétrain x (F1 Large White x Landrace) (P x LWL), Wild Boar x Piétrain (W x P), Meishan x Piétrain (M x P) and Wild Boar x Meishan (W x M)] as a function of distance between pairs of SNPs up to 5 Mb for the large chromosome 1 (SSC1). The fraction of marker pairs with different  $r^2$  levels is shown for different distances between loci (in Mb) for the following bins: [0, 0.025), [0.025, 0.05), [0.05, 0.075), [0.075, 0.12), [0.12, 0.2), [0.2, 0.5), [0.5, 1.5), [1.5, 3) and [3, 5].

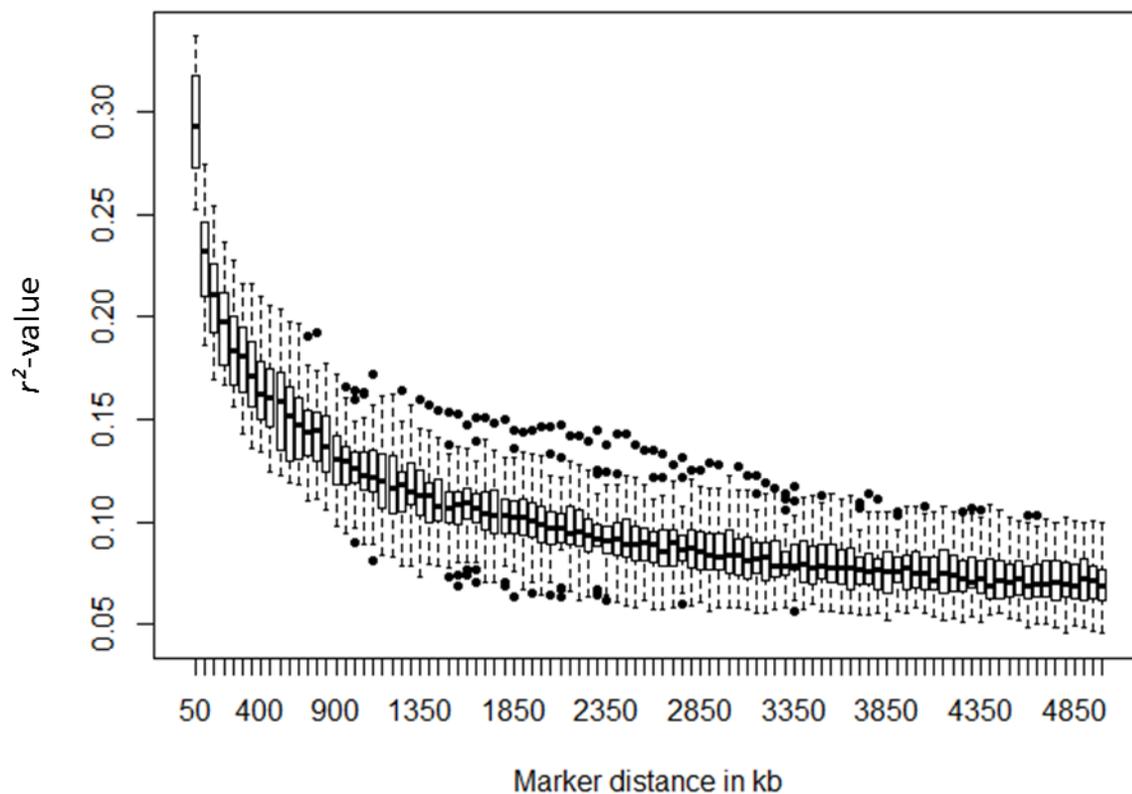
The extent of LD for the four crosses is shown exemplarily in Fig. 1 for the large SSC1. A complete visualization of the extent of LD in the crosses is shown in Fig. S2 for every autosome.

For the WxM cross, the LD is highest and decreases slowest compared to the other crosses. For the MxP cross, the decrease of LD is also slow, whereas the decrease of LD was fastest in the other crosses (PxLWL, WxP). Compared to the single crosses (Figs. 1 & S2), the LD in the pooled crosses is even lower and decreases faster (Fig. 2). For small distances, the level of LD is highest and decreases with an increase in distance, especially for distances greater than 0.5 Mb and greater than 1.5 Mb, depending on the autosomes (Figs. 2 & S2). It is obvious that the extent of LD for each inter-marker distance bin varies across the autosomes (Fig. 3).



**Figure 2** Extent of linkage disequilibrium (LD) for the pooled crosses over the chromosomes.

Level of LD decay in the pooled crosses [Piétrain x (F1 Large White x Landrace) (P x LWL), Wild Boar 9 Piétrain (W x P), Meishan x Piétrain (M x P) and Wild Boar x Meishan (W x M)] as a function of the distance between pairs of SNPs up to 5 Mb for the autosomes (SSC1–SSC18). The fraction of marker pairs with different  $r^2$  levels is shown for different distances between loci (in Mb) for the following bins: [0, 0.025), [0.025, 0.05), [0.05, 0.075), [0.075, 0.12), [0.12, 0.2), [0.2, 0.5), [0.5, 1.5), [1.5, 3), and [3, 5].



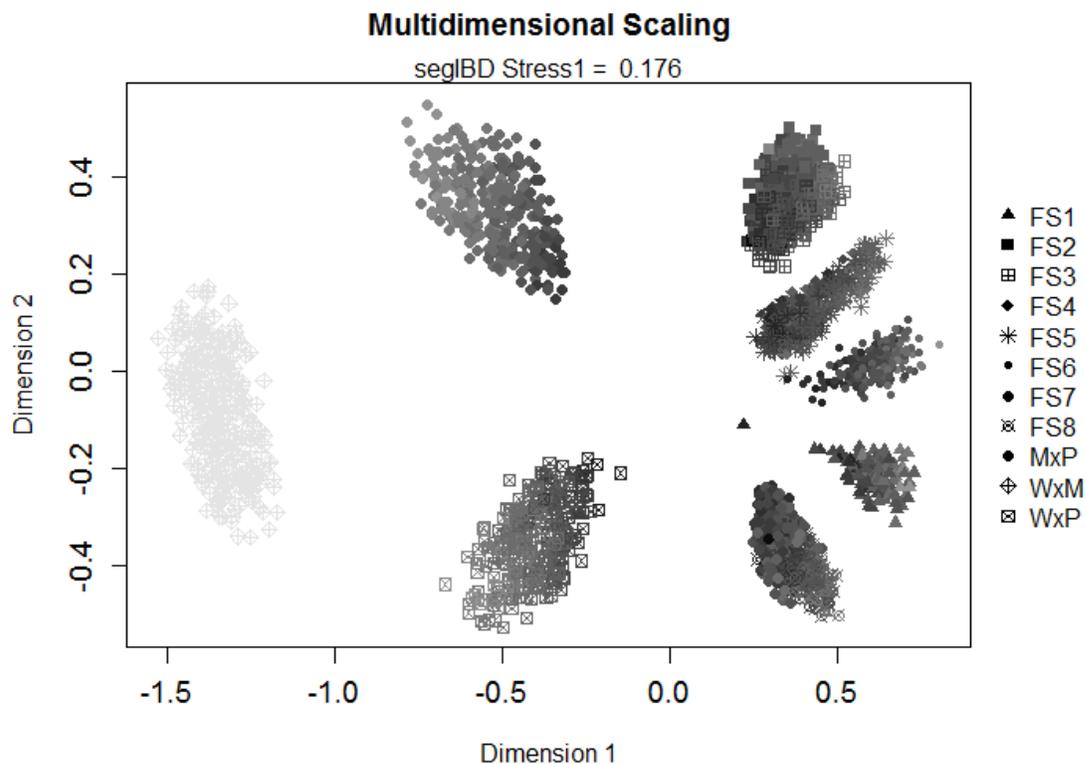
**Figure 3** Genome-wide decay of linkage disequilibrium (LD) over distance for the pooled crosses.

The x-axis denotes inter-marker distance in kb and the y-axis the LD for all potential marker pairs separated by less than 5 MB using  $r^2$ . Within each distance bin, the distribution of the mean LD over the autosomes is shown as a boxplot.

### *Population structure*

The results of the MDS analysis representing the underlying population structure are shown in Fig. 4. F2 individuals were arranged such that the distance between them corresponded to their genetic dissimilarity as much as possible. In MDS, the stress value was 0.176. A clear distinction of the eight full-sib families from the PxLWL cross according to their genetic dissimilarities was possible (right side). Individuals from the distantly related founder breeds Meishan, Piétrain and

Wild Boar clearly subdivided into the three MxP, WxM and WxP crosses (left side). The grey colour gradation indicates the proportion of the genome of the F2 individuals having segments shared with the Piétrain founder individuals. F2 individuals shown in the middle of Fig. 4 shared more segments with the Piétrain founder breed than did the F2 individuals in the peripheral locations.



**Figure 4** Multidimensional scaling analysis of the segment-based realized relationship matrix. Only segments comprising more than 25 SNPs and having a minimum length of 1 Mb were considered.

The x-axis is the first dimension (dimension 1), and the y-axis is the second dimension (dimension 2). The different symbols represent the eight Piétrain x (F1 Large White x Landrace) full-sib families (FS1–FS8) and the Meishan x Piétrain (MxP), Wild Boar x Meishan (WxM) and Wild Boar x Piétrain (WxP) crosses. The grey colour gradation indicates the proportion of the genome of the F2 individuals having segments shared with the Piétrain founder individuals. The darker the shade, the more segments the individuals had in common with Piétrain.

## Statistical analyses

### *Variance component estimation and marker association analysis*

The heritability estimates are listed in Table 2. It is obvious that, despite the correction for *RYR1*, there is still genetic variation in the traits (results from the A model). The broad-sense heritability is above the narrow-sense heritability, indicating that dominance is important, especially for cond45.

**Table 2** Heritability estimates.

Trait	A Model	AD Model		
	$h_a^2$	$h_a^2$	$h_d^2$	$h_g^2$
Cond45	0.202 (0.032)	0.174 (0.037)	0.157 (0.033)	0.331 (0.039)
Dressing out (%)	0.324 (0.033)	0.301 (0.035)	0.048 (0.020)	0.349 (0.034)

Estimated broad ( $h_g^2$ ) and narrow-sense heritability ( $h_a^2$ ) with standard error (in brackets);  $h_d^2$  is the proportion of dominance variance of the phenotypic variance; Cond45, conductivity 45 min post-mortem.

The results of the single marker association analysis are shown as Manhattan plots in Fig. 5. For cond45 and dressing out, 183 and 191 SNPs respectively were found with a significance of  $P < 5 * 10^{-5}$  and FDR q-values below 0.013 and 0.010, which indicates that approximately three and two SNPs respectively were false positive. Besides the nominally significant SNPs, 33 and 26 SNPs were also genome-wide significant for cond45 and dressing out respectively. Nominally significant SNPs are listed in Table S2. Genome-wide significant SNPs were located on SSC1, 2, 3, 6, 8, 11 and 15 for cond45 and on SSC1, 2, 7 and 8 for dressing out. Among the significant SNPs, six clusters were identified for cond45 on SSC1, 6, 8 and 15. For dressing out, five clusters on SSC1, 2 and 7 were detected. BayesC and BayesD WPPA values for the traits are shown in Figs S3 & 6 respectively. BayesC and BayesD led to similar mapping signals, which differed, however, in the level of the WPPA signals. BayesD led to higher WPPA signals than for BayesC. For cond45 and dressing out, 686 and 683 BayesD windows respectively exceeded the 97.5% quantile of the WPPA values (Table S3). In BayesD, WPPA signals above 0.8 were observed for cond45 on SSC1, 2, 6, 8, 11, 14 and 15. The highest signals for dressing out were observed on SSC2, 7, 10 and 12.

**Table 3** Overlap between clusters and windows. Cluster number, chromosomal position, number of significant SNPs, number of WPPA window outliers, the chromosomal position of the first and last SNP in the first (last) sliding window and the window length.

Trait	Cluster number <sup>1</sup>	SSC	Cluster Length in Mb	Number of significant SNPs <sup>2</sup>	Number of WPPA windows outliers <sup>3</sup>	Start/end position of the first/last SNP <sup>4</sup>	Window Length in Mb
Cond45	2	6	0.044	2 (2)	8	43,786,594–45,759,89	1.973
	5	8	1.046	12 (7)	4	69,048,896–70,741,34	1.692
	6	15	0.764	5 (3)	17	116,944,035–118,912,39	1.968
Dressing out (%)	3	2	2.465	18 (9)	16	70,14–4,058,654	3.989
	5	2	1.611	4 (2)	22	9,224,138–11,646,346	2.422
	6	7	1.416	12 (3)	47	42,708,898–45,746,958	3.038

<sup>1</sup>Cluster number for the clusters built from single marker association analysis results.

<sup>2</sup> Number of nominally significant SNPs ( $p < 5 \times 10^{-5}$ ) with the number of genome-wide SNPs ( $p_{genome-wide} < 0.05$ ) in brackets.

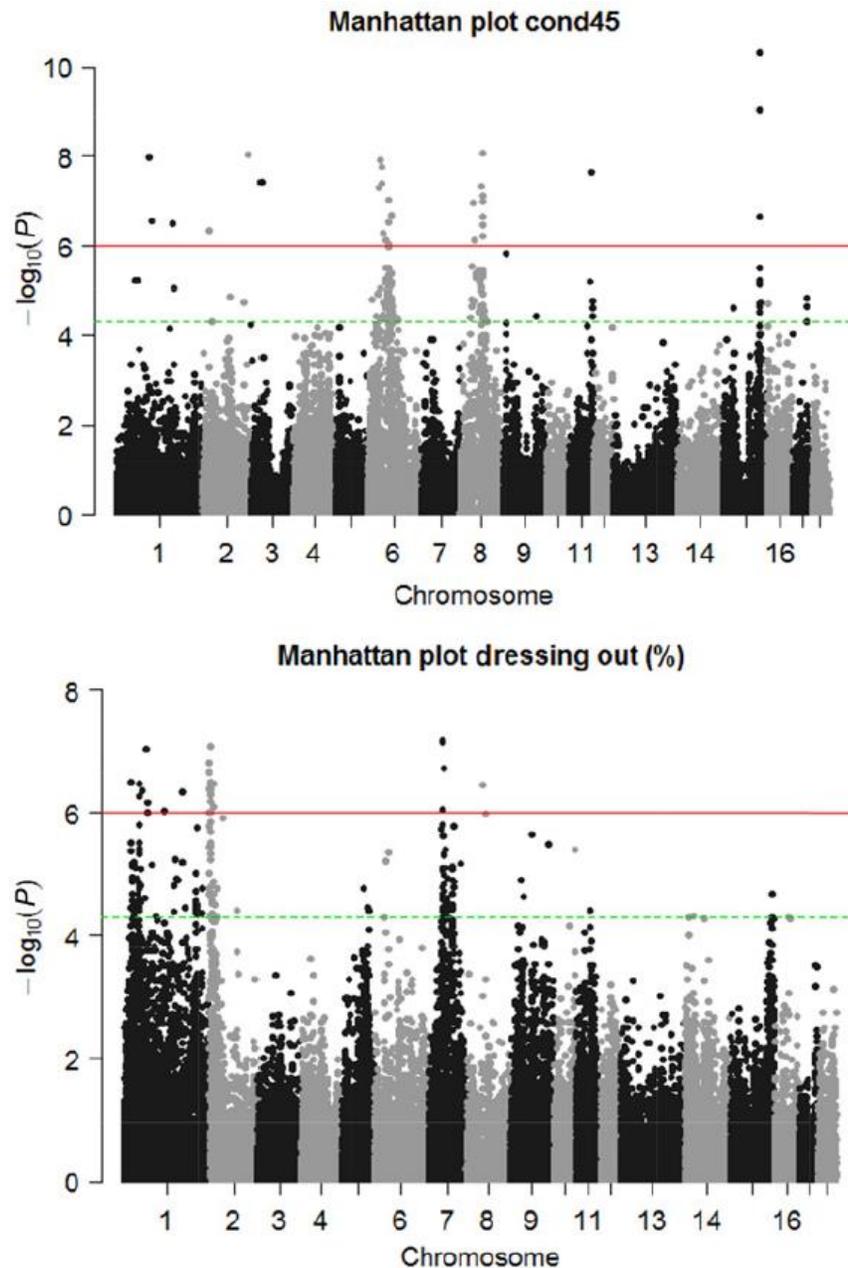
<sup>3</sup> Number of sliding windows, including nominally significant SNPs.

<sup>4</sup> Start and end position of the first and last marker in the first and last sliding window, which includes the first and last nominally significant SNP from the cluster.

Cond45, conductivity 45 min post-mortem; SSC, *Sus Scrofa* chromosome number.

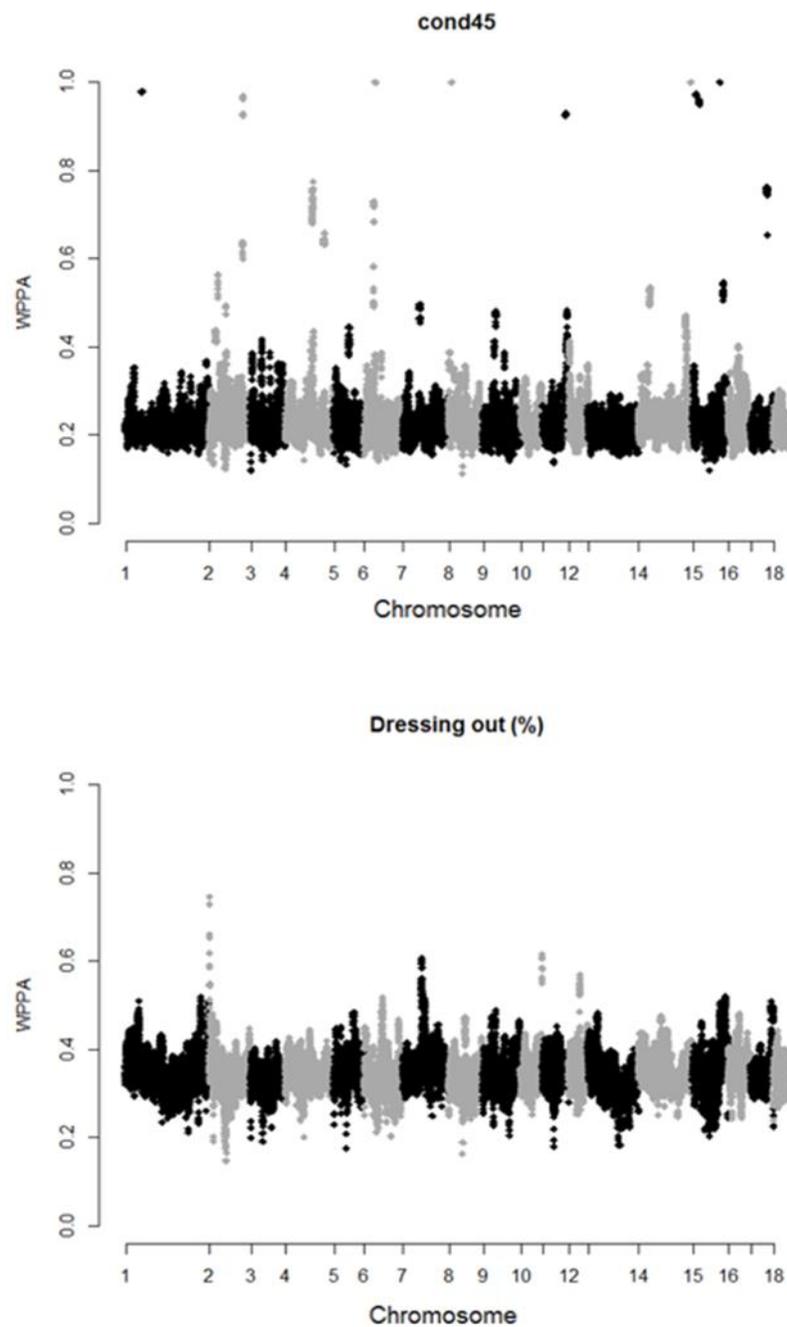
We found concordance between single-marker as well as the Bayesian WPPA results. For cond45 and dressing out, 34 and 74 nominally significant SNPs respectively were detected in outlier windows in BayesD. Considering a WPPA value greater than 0.8, 11 nominally significant SNPs were still found for cond45, whereas seven SNPs were also genome-wide significant in BayesD. From the nominally significant SNPs, two and five were found in clusters 2 and 6 respectively. SNPs with  $P < 5 \times 10^{-5}$  located in outlier windows were located on SSC2, 6, 8, 11, 15 and 17 for cond45 and on SSC1, 2, 5, 7, 10 and 15 for dressing out.

Overlaps between clusters and windows were found on SSC6, 8 and 15 for cond45 and on SSC2 and 7 for dressing out (Table 3). The SNP *MARC0039619*, located on SSC6, was significant for cond45 and dressing out and was located in cluster 2 and in an outlier window for cond45.



**Figure 5** Manhattan plots for the associations between 49 709 SNPs and the (top) cond45 (conductivity, 45 min post-mortem) and (bottom) dressing out traits with the single-marker model.

The x-axis shows the position in the genome, and the y-axis shows the observed  $-\log(P\text{-value})$ . The two lines correspond to an error probability of a nominal  $P \approx 5 \times 10^{-5}$  for the dashed line and  $P_{\text{genome-wide}} \approx 0.05$  for the solid line.



**Figure 6** BayesD WPPA calculated for 1-cM windows for the (top) cond45 (conductivity 45 min post-mortem) and (bottom) dressing out traits over the autosomes.

The x-axis denotes the 18 chromosomes, represented as alternating black and dark grey colours, and the y-axis denotes the WPPAs.

## Discussion

Previous F2 cross designs were set up to map QTL using linkage analysis. The power of the F2 cross design is highest if the founder breeds are alternatively homozygous for the QTL. Therefore, European-type and Asian-type breeds were typically used to set up porcine F2 crosses. In the present study, great differentiation of the breeds was found between the European- and Asian-type breeds, which is supported by their phylogeny (Groenen *et al.* 2012), although haplotypes of Asian breeds are found in European breeds (Bosse *et al.* 2014, 2015). Within the European-type breeds, the  $F_{ST}$  value was 0.07, indicating the smallest differentiation of the breeds.

The fastest breakdown of LD was observed in crosses set up with European-type breeds. In the crosses involving Piétrain as a founder (WxP, MxP and PxLWL), the LD decreased faster than for the WxM cross. This was also expected because of the relatively large effective population size of this breed ( $N_e > 200$ ; BLE 2010) and the heterogeneity in the Piétrain population (Stratz *et al.* 2014b). It seems that the number of sires used as founders might also impact the decay of LD, because in the PxLWL cross, for which five Piétrain boars were used, the decay of LD was fastest. By pooling the data the fastest LD decrease was observed, which positively impacts the mapping resolution in GWASes. The LD results of the present study are consistent with those of Bennewitz & Wellmann (2014) and Schmid *et al.* (2016) and support the benefit of pooling data from several F2 crosses.

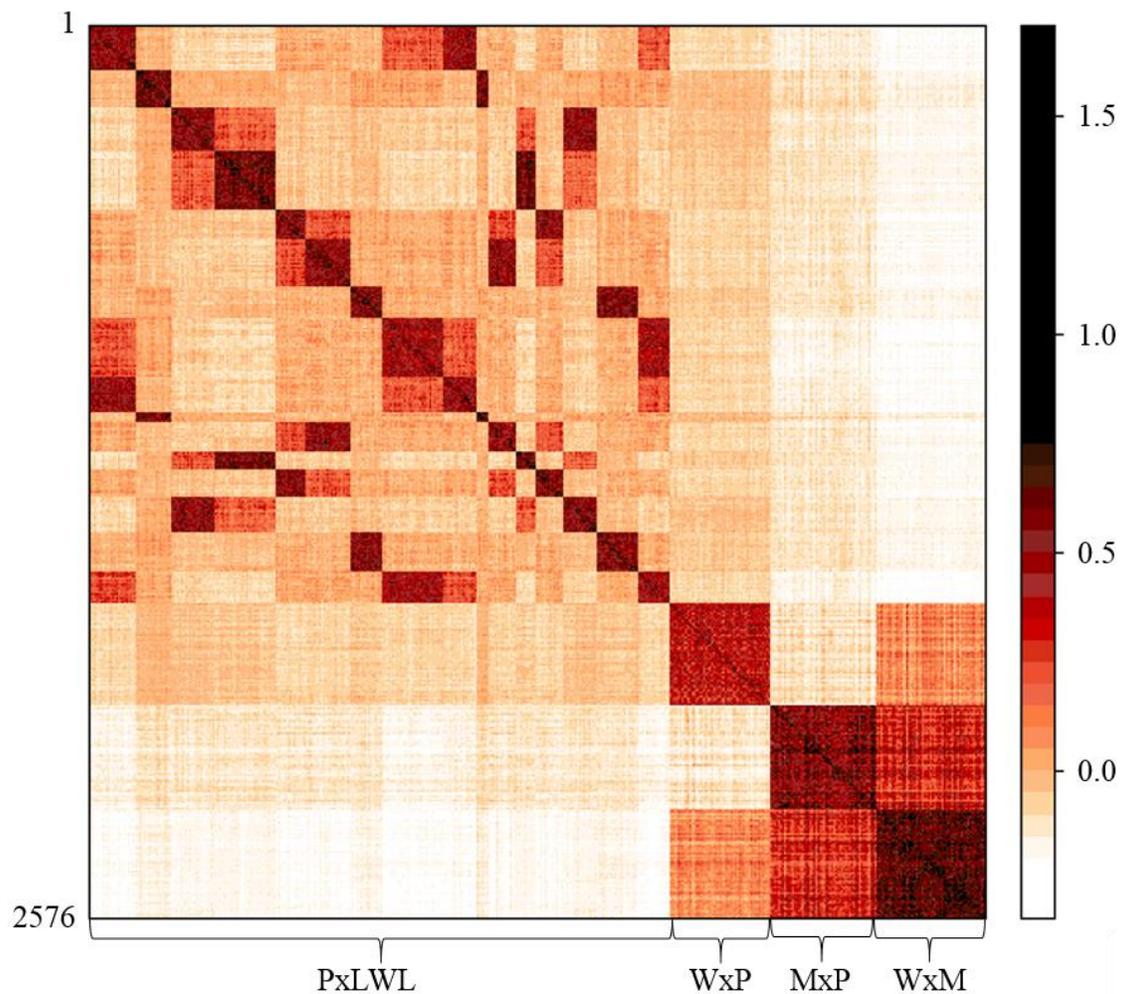
In F2 crosses with long-range LD blocks, the benefit of applying a GWAS instead of linkage analysis is only limited, even if dense marker panels are used. The results of this study showed that pooling data from different F2 crosses reduced the length of the LD blocks substantially, and thus a GWAS is justified. The median breakdown of the LD for the autosomes was fastest using pooled F2 crosses. The huge variation in the LD in 50-kb distance classes over the autosomes (Fig. 3) highlights differences in recombination rates and thus in the mapping resolution across the autosomes. This is an explanation of why we received less precise GWAS results, which is indicated by their large window sizes (Table 3). Further investigations on the impact of the cross, the chromosome and the chromosomal position on the  $r^2$  value could be helpful in determining the optimal window size.

To unravel population substructures, MDS was used, which enabled a good separation of the full-sib families in the PxLWL cross and in the WxP, MxP and WxM crosses. It is obvious that some full-sib families from the PxLWL cross are genetically divergent, which requires a correction for population substructure in the GWAS. This idea is consistent with the high genetic diversity in the current Piétrain reference population (Stratz *et al.* 2014b).

Within this study, the type of cross was pre-corrected prior to the GWAS. However, this could have had an impact on the mapping power, as part of the genetic differences between individuals would have been captured. Therefore, further investigations into how pre-correcting for type of cross affects the variance components and GWAS results are recommended in ongoing studies.

In multi-marker GWASes, the population structure is modelled by the SNPs allocated in the  $t$ -distribution with the smaller variance. In single-marker association analyses, accounting for substructures between F2 individuals of different crosses and full-sib families was achieved by using the GRM (Fig. 7). To increase the mapping power, the autosome harbouring the SNP to be tested for association was excluded from the GRM (Yang *et al.* 2014). The LD surrounding a causal mutation was assumed to be higher in F2 crosses compared to purebred lines. Therefore, a relaxed significance threshold was applied in the single-marker association analysis, which was supported by the low number of false positives among the significant SNPs.

A comparison of the results from the overlapping QTL regions and studies was performed using the pig QTL database (Hu *et al.* 2005). For cond45, QTL were found on SSC6 (De Koning *et al.* 2003; Evans *et al.* 2003) and on SSC15 (Wimmers *et al.* 2006). QTL for dressing out were detected on SSC2 (Liu *et al.* 2007) and on SSC7 (Sato *et al.* 2003; Edwards *et al.* 2008; Liu *et al.* 2008; Choi *et al.* 2011).



**Figure 7** Heatmap of the marker-based relationship matrix realized for the 2572 F2 individuals from the four crosses.

The realized relatedness illustrates the substructures between F2 individuals of different crosses [Wild boar x Piétrain (WxP), Meishan x Piétrain (MxP) and Wild boar x Meishan (WxM)] and FS families [Piétrain x F1 Large White x Landrace (PxLWL)]. The darker the colour the more related the individuals were.

In conclusion, the lowest differentiation of the breeds observed between European-type breeds positively impacted the mapping resolution because of the fastest breakdown of LD within that cross. In contrast, greater differentiation, which was detected between the European- and Asian-type breeds, led to the slowest breakdown of LD and hence a lower mapping resolution. The fastest breakdown of LD was observed when pooling the data, which increases the mapping resolution in a GWAS. To dissect the causal mutation, a combination of single-marker and Bayesian multi-marker models seems to be promising. The results of the present study serve to preselect target regions in the genome for subsequent fine-mapping in a future study. Regions harbouring SNPs that are still segregating in the current Piétrain reference population are especially promising target regions. A GWAS model with breed-specific SNP effects would facilitate the identification of these SNPs. Further research is needed to investigate if dominance is also important in crossbreed pig breeding schemes, for which Piétrain is frequently used as a terminal sire breed.

### **Acknowledgements**

This study was supported by a grant from the German Research Foundation (DFG).

**References**

- Bennewitz J. & Wellmann R. (2014) Mapping resolution in single and multiple porcine F2 populations. Book of Abstracts, 10<sup>th</sup> World Congress on Genetics Applied to Livestock Production, Session 184. WCGALP, Vancouver, BC.
- Bennewitz J., Edel C., Fries R., Meuwissen T.H.E. & Wellmann R. (2017) Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. *Genetics Selection Evolution* 49, 7.
- BLE, Bundesanstalt fuer Landwirtschaft und Ernaehrung (2010) Erhebung von Populations Daten Tiergenetischer Ressourcen in Deutschland: Schweine. Federal Office for Agriculture and Food, Bonn, Germany. <http://download.ble.de/06BE001.pdf>.
- Bosse M., Megens H.-J., Frantz L.A.F. *et al.* (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature Communications* 5, 4392.
- Bosse M., Lopes M.S., Madsen O., Megens H.J., Crooijmans R.P., Frantz L.A.F., Harlizius B., Bastiaansen J.W.M. & Groenen M.A.M. (2015) Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. *Proceedings of Biological Science* 282, 2015-9.
- Boysen T.J., Tetens J. & Thaller G. (2010) Detection of a quantitative trait locus for ham weight with polar overdominance near the ortholog of the callipyge locus in an experimental pig F2 population. *Journal of Animal Science* 88, 3167-72.
- Browning B.L. & Browning S.R. (2008) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* 84, 210-23.
- Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K. & Madden T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.

- Choi I., Steibel J.P., Bates R.O., Raney N.E., Rumph J.M. & Ernst C.W. (2011) Identification of carcass and meat quality QTL in an F2 Duroc x Pietrain pig resource population using different least-squares analysis models. *Frontiers in Genetics* 2, 18.
- De Koning D.J., Pong-Wong R., Varona L., Evans G.J., Giuffra E., Sanchez A., Plastow G., Noguera J.L., Andersson L. & Haley C.S. (2003) Full pedigree quantitative trait locus analysis in commercial pigs using variance components. *Journal of Animal Science* 81, 2155-63.
- Edwards D.B., Ernst C.W., Raney N.E., Doumit M.E., Hoge M.D. & Bates R.O. (2008) Quantitative trait locus mapping in an F2 Duroc x Pietrain resource population: II. Carcass and meat quality traits. *Journal of Animal Science* 86, 254-66.
- Evans G.J., Giuffra E., Sanchez A. *et al.* (2003) Identification of quantitative trait loci for production traits in commercial pig populations. *Genetics* 164, 621-7.
- Fernando R.L. & Garrick D. (2013) Bayesian methods applied to GWAS. *Methods in Molecular Biology* 1019, 237-74.
- Fujii J.J. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* 253, 448-51.
- Groenen M.A.M., Archibald A.L., Uenishi H. *et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393-8.
- Hill W.G. & Robertson A. (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38, 226-31.
- Holsinger K.E. & Weir B.S. (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* 10, 639-50.
- Hu Z., Dracheva S., Jang W., Maglott D., Bastiaansen J., Rothschild M. & Reecy J. (2005) A QTL resource and comparison tool for pigs: PigQTLDB. *Mammalian Genome* 16, 792-800.

- de Leeuw J. & Mair P. (2009) Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software* 31, 1-30.
- Liu G., Jennen D.G.J., Tholen E. *et al.* (2007) A genome scan reveals QTL for growth, fatness, leanness and meat quality in a Duroc-Pietrain resource population. *Animal Genetics* 38, 241-52.
- Liu G., Kim J.J., Jonas E. *et al.* (2008) Combined line-cross and half-sib QTL analysis in Duroc-Pietrain population. *Mammalian Genome* 19, 429-38.
- Lutz V., Stratz P., Preuß S., Tetens J., Grashorn M.A., Bessei W. & Bennewitz J. (2017) A genome-wide association study in a large F2-cross of laying hens reveals novel genomic regions associated with feather pecking and aggressive pecking behavior. *Genetics Selection Evolution* 49, 18.
- Purcell S., Neale B., Todd-Brown K. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81, 559-75.
- Ramos A.M., Crooijmans R.P.M.A., Affara N.A. *et al.* (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE* 4, e6524.
- Rothschild M.F., Hu Z.-L. & Jiang Z. (2007) Advances in QTL mapping in pigs. *International Journal of Biological Sciences* 3, 192-7.
- Rückert C. & Bennewitz J. (2010) Joint QTL analysis of three connected F2-crosses in pigs. *Genetics Selection Evolution* 42,40.
- Sato S., Oyamada Y., Atsuji K. *et al.* (2003) Quantitative trait loci analysis for growth and carcass traits in a Meishan x Duroc F2 resource population. *Journal of Animal Science* 81, 2938-49.
- Schmid M., Wellmann R. & Bennewitz J. (2016) Power and precision of mapping genes in simulated F2 crosses using whole genome sequence data. Book of Abstracts, 67th Annual EAAP Meeting, Session 27, European Association for Animal Production, Belfast.

- Storey J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 64, 479-98.
- Storey J.D. & Tibshirani R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100, 9440-5.
- Stratz P., Baes C., Rückert C., Preuss S. & Bennewitz J. (2013) A two-step approach to map quantitative trait loci for meat quality in connected porcine F2 crosses considering main and epistatic effects. *Animal Genetics* 44, 14-23.
- Stratz P., Wellmann R., Preuss S., Wimmers K. & Bennewitz J. (2014a) Genome-wide association analysis for growth, muscularity and meat quality in Piétrain pigs. *Animal Genetics* 45, 350-6.
- Stratz P., Wimmers K., Meuwissen T.H.E. & Bennewitz J. (2014b) Investigations on the pattern of linkage disequilibrium and selection signatures in the genomes of German Piétrain pigs. *Journal of Animal Breeding and Genetics* 131, 473-82.
- Verbyla K.L., Hayes B.J., Bowman P.J. & Goddard M.E. (2009) Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research* 91, 307-11.
- Warr A., Hall R., Kim K. *et al.* (2016) Exploiting long read sequencing technologies to establish high quality highly contiguous pig reference genome assemblies. *International Plant & Animal Genome, XXV, Paper No. 25025.* <https://pag.confex.com/pag/xxv/webprogram/Paper25025.html>.
- Weir B.S. & Cockerham C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358-70.
- Wellmann R. (2017) OPTISEL: optimum contribution selection and population genetics. R package version 0.7.1 (<https://cran.r-project.org/web/packages/optiSel/optiSel.pdf>).
- Wellmann R. & Bennewitz J. (2012) Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genetics Research* 94, 21-37.

- Wellmann R., Preuss S., Tholen E., Heinkel J., Wimmers K. & Bennewitz J. (2013) Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics Selection Evolution* 45, 28.
- Wimmer V., Albrecht T., Auinger H. & Schon C. (2012) SYNBREED: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28, 2086-7.
- Wimmers K., Fiedler I., Hardge T., Murani E., Schellander K. & Ponsuksili S. (2006) QTL for microstructural and biophysical muscle properties and body composition in pigs. *BMC Genetics* 7, 15.
- Yang J., Lee S.H., Goddard M.E. & Visscher P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 88, 76-82.
- Yang J., Zaitlen N.A., Goddard M.E., Visscher P.M. & Price A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* 46, 100-6.

**Supporting information**

**Table S1** Chromosomal position of markers on the current genome assembly Sscrofa11.1.

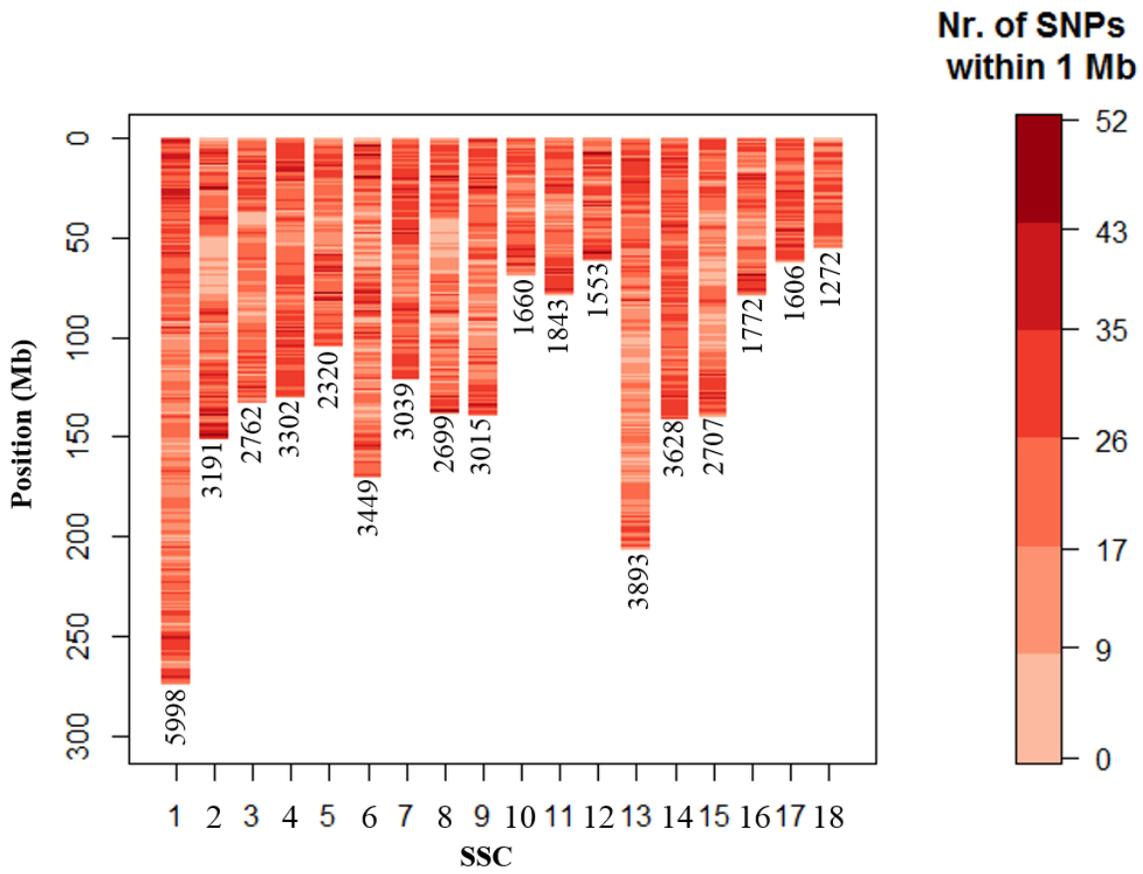
Table S1 can only be provided in digital format.

**Table S2** List of significant SNPs with a  $P$ -value  $\leq 5 \times 10^{-5}$ , their chromosomal position (SSC), Pos (kb) and their  $P$ -value for cond45 and dressing out.

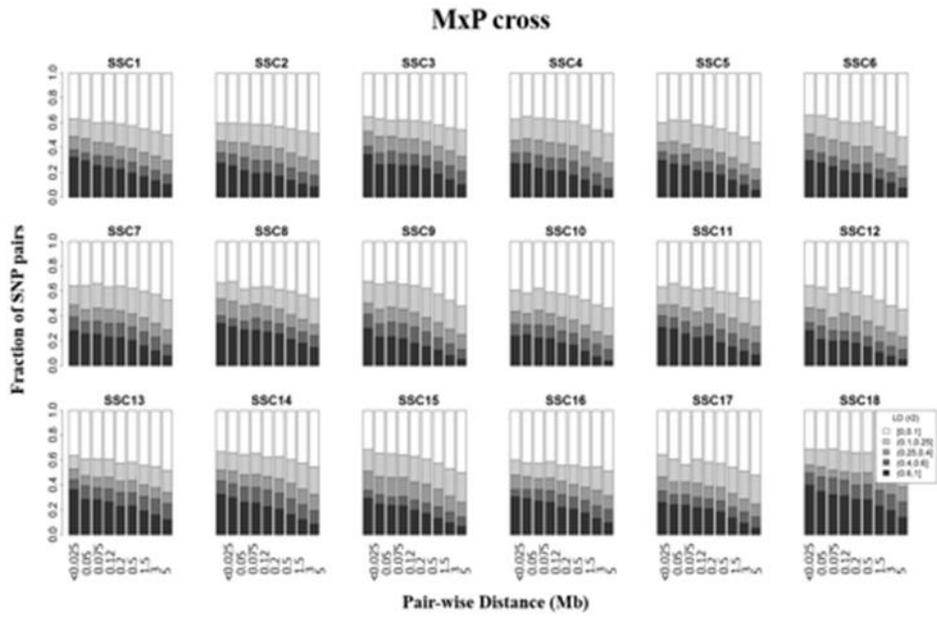
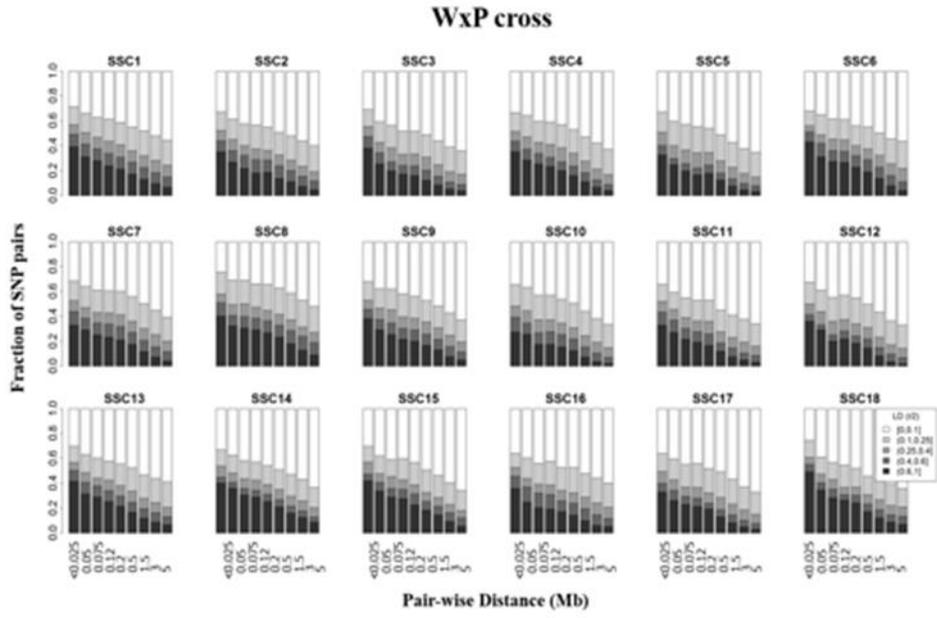
Table S2 can only be provided in digital format.

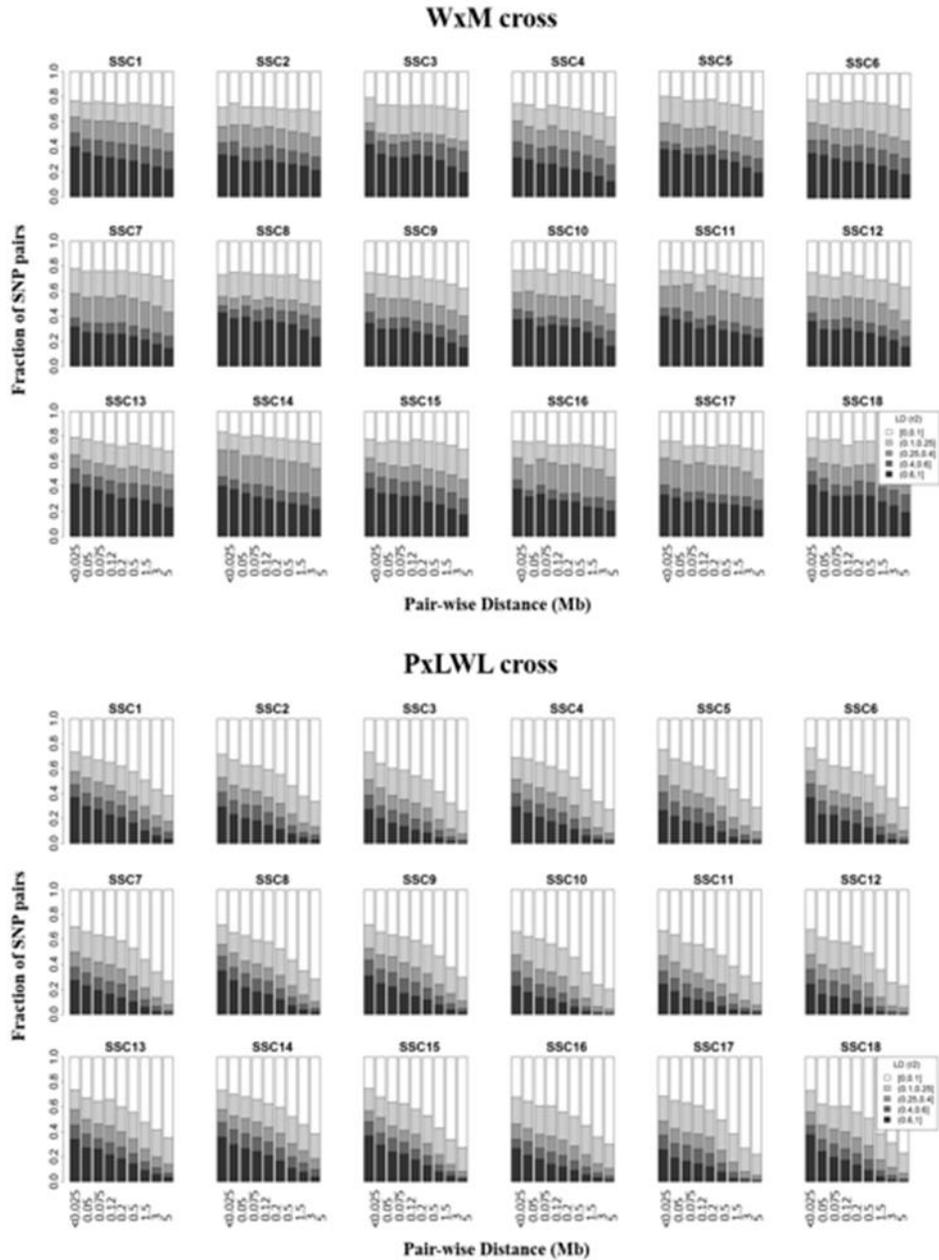
**Table S3** List of sliding windows that exceed the 97.5% quantile of the WPPA, their flanking markers (first marker, last marker) with positions (SSC, Start and End), the number of SNPs in the window (#SNP) and their WPPA.

Table S3 can only be provided in digital format.



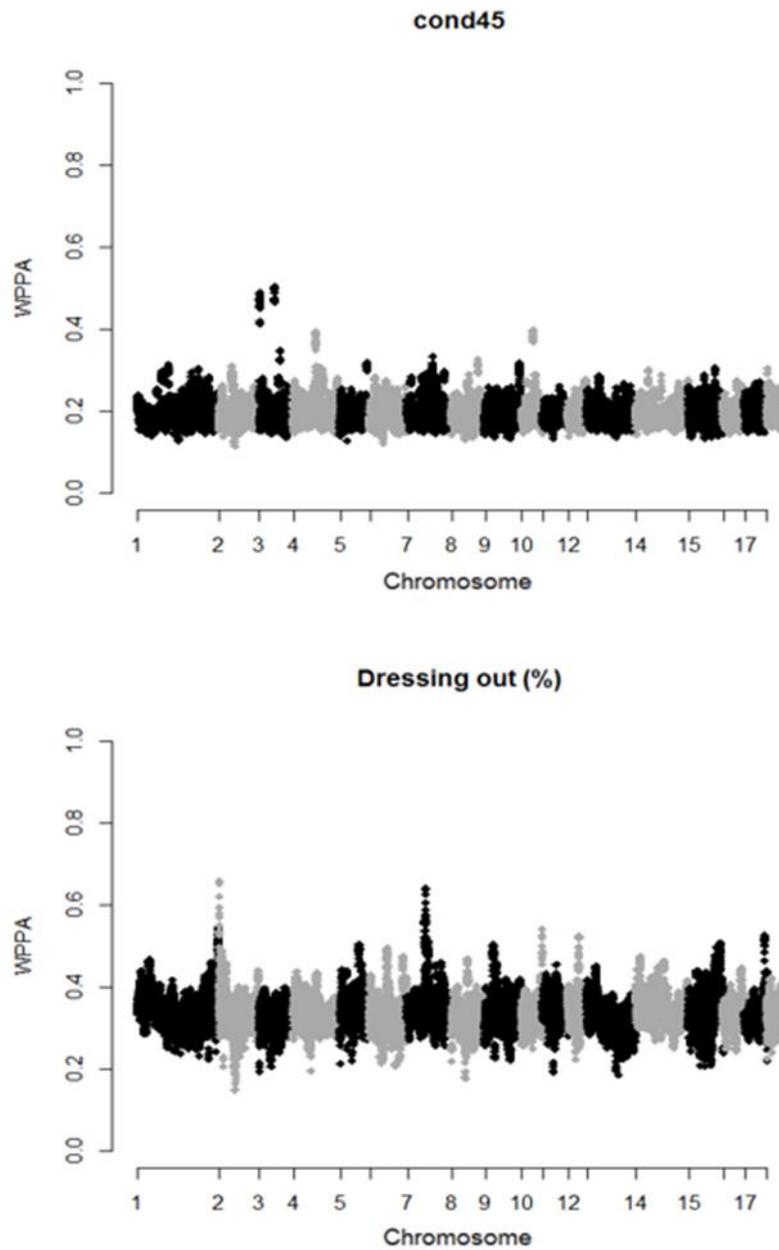
**Figure S1** Distribution of the autosomal SNPs in the porcine genome sequence (*Sus scrofa* Build 11.1 assembly). The SNP density in 1-Mb intervals is shown for 49 690 SNPs.





**Figure S2** LD pattern

Level of linkage disequilibrium decay in the crosses WxP, MxP, WxM and PxLWL as a function of distance between pairs of SNPs up to 5 Mb for the autosomes (SSC1 – SSC18). The fraction of marker pairs with different  $r^2$  levels is shown for different distances between loci (in Mb) for the following bins (0, 0.025], (0.025, 0.05], (0.05, 0.075], (0.075, 0.12], (0.12, 0.2], (0.2, 0.5], (0.5, 1.5], (1.5, 3], (3, 5).



**Figure S3** BayesC WPPA calculated for 1-cM windows for cond45 and dressing out over the autosomes.

The x-axis denotes the 18 chromosomes, represented as alternating black and dark grey colours, and the y-axis the WPPAs. In the top (bottom) results for cond45 (dressing out) are shown.



## CHAPTER FOUR

### **Mapping QTL for production traits in segregating Piétrain pig populations using GWAS results of F2 crosses**

**Markus Schmid\*, Maria Maushammer\*, Siegfried Preuß\*, Jörn Bennewitz\***

\*Institute of Animal Science, University of Hohenheim, 70599 Stuttgart, Germany

Corresponding author: markus\_schmid@uni-hohenheim.de

Published in:

**Animal Genetics** (2018) 49: 317-320.

doi: 10.1111/age.12663.

## Summary

In this study, genome-wide association study (GWAS) results of porcine F2 crosses were used to map QTL in outcross Piétrain populations. For this purpose, two F2 crosses (Piétrain x Meishan, n = 304; Piétrain x Wild Boar, n = 291) were genotyped with the PorcineSNP60v2 BeadChip and phenotyped for the dressing yield, carcass length, daily gain and drip loss traits. GWASs were conducted in the pooled F2 cross applying single marker mixed linear models. For the investigated traits, between two and five (in total 15) QTL core regions, spanning 250 segregating SNPs around a significant trait-associated peak SNP, were identified. The SNPs within the QTL core regions were subsequently tested for trait association in two outcross Piétrain populations consisting of 771 progeny-tested boars and 210 sows with their own performance records. In the sow (boar) dataset, five (eight) of the 15 mapped QTL were validated. Hence, many QTL mapped in the F2 crosses (with Piétrain as a common founder breed) are still segregating in the current Piétrain breed. This confirms the usefulness of existing F2 crosses for mapping QTL that are still segregating in the recent founder breed generation. The approach utilizes the high power of an F2 cross to map QTL in a breeding population for which it is not guaranteed that they would be found using a GWAS in this population.

## Keywords

mapping power, pooled F2 cross, QTL core regions, QTL validation, SNP validation

Before SNP panels were available, QTL mapping was frequently done by linkage mapping using mainly sparse microsatellite marker information, which required informative experimental population structures. In pig breeding, several F2 crosses were established in the past. The QTL mapping power was usually high in these crosses; however, the mapping precision was low due to limited population sizes and the use of only a few generations, limiting the number of useable meioses. In many cases, distantly related founder breeds were chosen to maximise the difference in QTL allele frequencies between the founder breeds. Thereby, European commercially used sire- or dam-line pig breeds (e.g. Piétrain) were frequently chosen as one of the founder breeds (Grosse-Brinkhaus *et al.* 2010; Rückert & Bennewitz 2010). Ledur *et al.* (2009) as well as

Schmid *et al.* (2016) showed that it is worthwhile in some situations to genotype the individuals of an F2 cross with dense SNP panels for genome-wide association studies (GWASs). This holds true especially if the founder breeds are not too distantly related and the number of individuals is sufficiently large. Although the former cannot be changed for existing F2 crosses, the number of individuals can be increased by pooling several F2 crosses. Schmid *et al.* (2016) showed that this is a suitable strategy for precisely mapping QTL segregating in a founder breed. In commercial breeding, these QTL are most important, because breeding takes place within the founder breeds. However, until now it has been largely unknown how many of the QTL segregating in a F2 cross are also segregating in a founder breed, because this requires a founder breed validation dataset containing a sufficient number of individuals that are genotyped at the markers showing significant trait associations in the F2 cross.

The aim of the present study was to map QTL in founder breed datasets using results of GWASs in a pooled F2 cross using dense SNP genotype information. GWASs were initially conducted in a pooled F2 cross with Piétrain as a common founder breed. Subsequently, the QTL identified in this pooled cross were tested for trait association in two Piétrain populations consisting of several hundred sires and sows, genotyped with the same SNP panel.

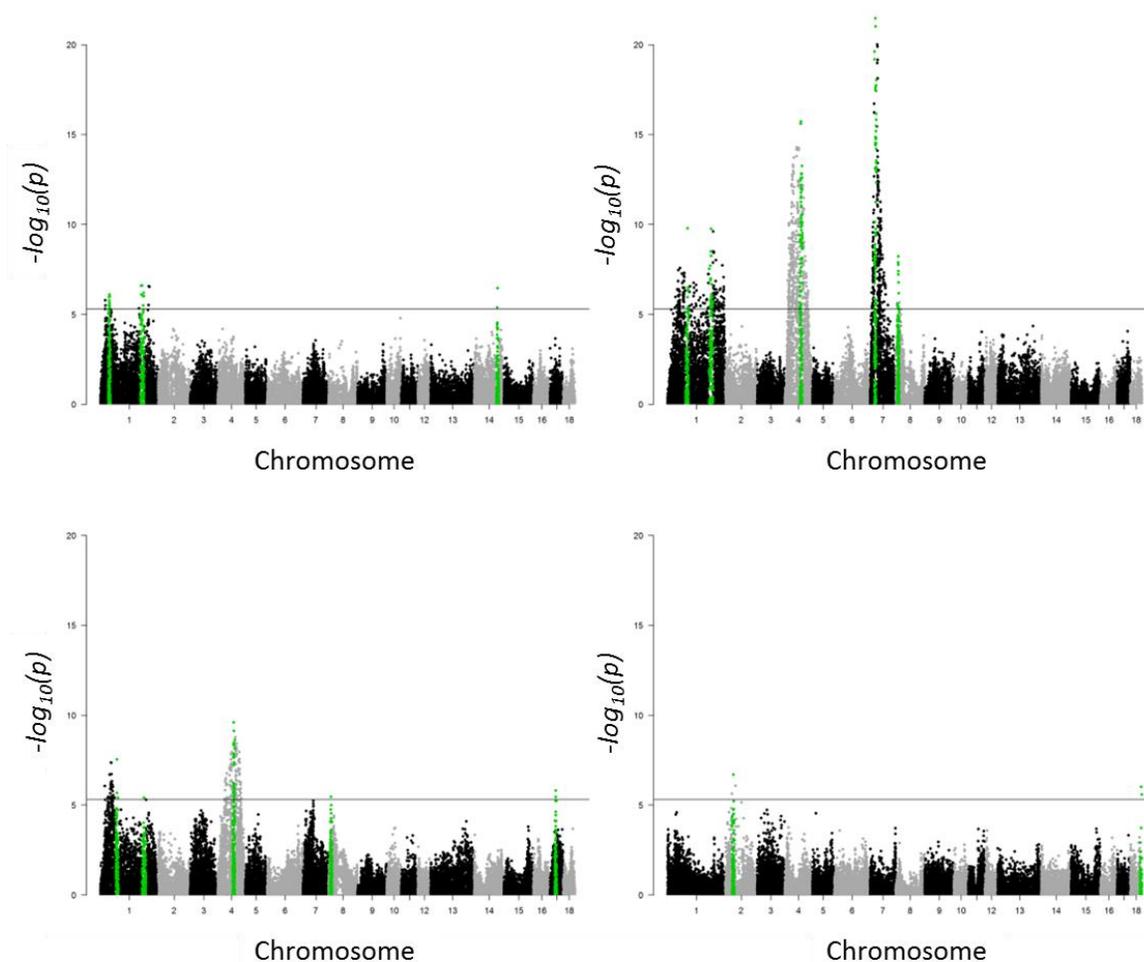
The F2 crosses resulted from mating Piétrain with Meishan ( $n = 304$ ) and Piétrain with Wild Boar ( $n = 291$ ). All individuals were housed on one experimental farm at the same time in mid 1990s (Müller *et al.* 2000). The traits investigated in this study were dressing yield, carcass length, daily gain and drip loss. The phenotypes were pre-corrected for the effects of sex, litter, season and age at slaughtering as well as the effect of the *RYRI* locus (Fujii *et al.* 1991). All individuals were genotyped with the Illumina PorcineSNP60v2 BeadChip, and the marker positions were taken from the recent genome assembly Sscrofa11.1. Only annotated autosomal SNPs with a minor allele frequency greater than 0.05, a call rate greater than 0.95, and a call frequency greater than 0.95 remained in the datasets. Sporadic missing genotypes were imputed using BEAGLE version 3.3.2 (Browning & Browning 2007). GWASs were conducted in the pooled F2 dataset ( $n = 595$ ) for each trait using GCTA software (Yang *et al.* 2014) and applying a single marker regression model that included a fixed cross effect and a random animal effect. To capture population stratification effects, a random polygenic effect was added. The covariance structure of the polygenic effects was modelled using a genomic relationship matrix (GRM). To avoid double fitting the SNP to be tested as a fixed and also as a random effect in the regression

model, we applied a leave-one-chromosome-out approach. Thereby, the SNP to be tested and all other SNPs located on the same chromosome were excluded from calculating the GRM. Because Bonferroni-type correction for multiple testing is too conservative in an F2 cross due to the high level of LD, we declared SNPs with  $P$ -values  $< 5 * 10^{-6}$  as significantly trait-associated SNPs.

Schmid *et al.* (2016) showed by means of simulation that the support interval for a QTL is generally large in F2 designs established from distantly related founder breeds, as was the case in our study. This is due to the long-range LD blocks in these populations. In contrast to QTL linkage mapping (Visscher *et al.* 1996), no straightforward methods exist to estimate QTL confidence intervals in a GWAS setting. Therefore, we defined QTL core regions in a simplified manner by defining windows of 250 segregating SNPs surrounding a peak SNP (i.e. a significant SNP with the highest test statistic).

Two Piétrain populations were used for the validation of mapped QTL. The first population consisted of 771 progeny-tested boars (boar validation set), which were born between 2007 and 2010. Progeny testing was conducted based on five to 10 progeny per boar, using standardised recording schemes on one experimental farm (a different one compared to the F2 individuals). From the progeny records, yield deviations for the four traits considered in this study were calculated, as described in detail in Wellmann *et al.* (2013). The yield deviations were adjusted for the effect of the *RYRI* locus (Fujii *et al.* 1991). The second validation population consisted of 210 sows with their own performance records (sow validation set) born in 2014. All sows were housed and slaughtered on the same farm as the progeny from the boar validation set. The phenotypes were pre-corrected for the effect of the day of slaughter and the fixed effect of the *RYRI* locus (Fujii *et al.* 1991). Both validation populations were genotyped with the Porcine-SNP60v2 BeadChip, applying the same filter criteria as in the F2 cross dataset. For QTL validation, the QTL core region SNPs discovered in the F2 GWASs were tested for association separately in the boar validation and in the sow validation set using the model described above. The number of tests depended on the number of QTL detected and the number of segregating SNPs within the QTL core region. Naturally, this number was much lower than the number of genome-wide segregating SNPs. The problem of multiple testing was much less evident (only the QTL core SNPs were tested for association). Therefore, a QTL was validated in the respective validation set if at least one SNP in a QTL core region showed a nominal  $P$ -value  $< 5 * 10^{-3}$ .

Additionally, the sign of the effect had to be the same in the F2 data set and in the validation datasets.



**Figure 1** Error probabilities [ $(-\log_{10}(P\text{-values}))$ ] in the F2 population for the dressing yield (top left), carcass length (top right), daily gain (bottom left) and drip loss (bottom right) traits. The green dots highlight the QTL core regions containing 250 SNPs with the most significant SNP as the centre. The black solid line corresponds to a significance level of  $P = 5 \times 10^{-6}$ .

Genome-wide association study results of the F2 cross are shown in Table 1 and in Fig. 1. Two to five QTL regions were mapped per trait. The number of QTL core region SNPs that segregated in the validation sets are shown in Table 1. The proportion of segregating SNPs was about 75-95 %, and no systematic difference could be observed between the two validation sets. Thus, a large number of F2-segregating SNPs also segregated in the validation sets. The number of validated

SNPs for each QTL in each data set is also given in Table 1. In the sow validation dataset, five out of 15 mapped QTL were validated; however, the number of significant SNPs within the core regions was small. In the boar validation set, 11 QTL could be confirmed, whereby the number of significant SNPs per validated QTL was much higher in most cases compared with the sow validation set. This results from the boosted power in the boar validation set. Please note that only a few validated SNPs would have passed stringent multiple testing criteria, which would be required in a full GWAS setting. The validated SNPs are listed in Tables S1 and S2.

The results show that many QTL mapped in the 25-year-old Piétrain-derived F2 crosses are still segregating within the current Piétrain breed. This confirms the usefulness of existing F2 crosses to map QTL that are still segregating in the recent founder breed generation. Hereby, this approach utilizes the high power of F2 crosses to map QTL in a commercial breeding population for which it is not guaranteed that they would be found in GWASs conducted in this population.

**Table 1** Results of the validation studies.

Trait	QTL number <sup>1</sup>	Centre SNP <sup>2</sup>	Reference ID	SSC <sup>3</sup>	Position (bp)	Sow validation set		Boar validation set	
						SNPs in the QTL core region segregating (in %)	Significant core SNPs	SNPs in the QTL core region segregating (in %)	Significant core SNPs
Dressing yield	1	<i>ALGA0102587</i>	rs81329420	1	41 975 747	240 (96)	-	239 (95)	24
	2	<i>INRA0005724</i>	rs325729300	1	198 450 596	195 (78)	-	192 (76)	-
	3	<i>MARC0014536</i>		14	110 994 062	190 (76)	2	215 (86)	41
Carcass length	1	<i>MARC0070002</i>		1	94 198 690	228 (91)	-	231 (92)	-
	2	<i>ASGA0093811</i>	rs81312801	1	206 498 445	202 (81)	-	205 (82)	17
	3	<i>ASGA0020383</i>	rs80847280	4	76 776 035	218 (87)	2	197 (78)	-
	4	<i>DRGA0007396</i>		7	25 565 917	229 (91)	1	232 (92)	1
	5	<i>H3GA0024312</i>	rs81401838	8	12 664 341	217 (86)	-	219 (87)	3
Daily gain	1	<i>ALGA0004473</i>	rs80853077	1	77 914 431	230 (92)	3	222 (88)	2
	2	<i>MARC0071934</i>		1	206 525 376	202 (80)	-	205 (82)	-
	3	<i>INRA0014994</i>	rs329076226	4	75 608 806	220 (88)	-	197 (78)	3
	4	<i>H3GA0024295</i>	rs81401432	8	11 805 802	218 (87)	-	218 (87)	-
	5	<i>ALGA0111568</i>	rs81340011	17	27 581 342	226 (90)	-	234 (93)	-
Drip loss	1	<i>ASGA0010012</i>	rs81357600	2	39 217 661	238 (95)	4	233 (93)	1
	2	<i>ALGA0098868</i>	rs81471303	18	52 259 665	151 (88)	-	147 (86)	-

<sup>1</sup> QTL core regions, spanning 250 SNPs, derived from the GWAS results of the pooled F2 cross.<sup>2</sup> Midpoint of the defined QTL core region, significant SNP with the highest test statistic.<sup>3</sup> Sus scrofa chromosome.

## Acknowledgements

This study was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

## Conflict of Interest

The authors declare that they have no conflict of interest.

## References

- Browning S.R. & Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* 81, 1084–97.
- Fujii J., Otsu K., Zorzato F., de Leon S., Khanna V.K., Weiler J.E., O'Brien P.J. & MacLennan D.H. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* 253, 448–51.
- Grosse-Brinkhaus C., Jonas E., Buschbell H., Phatsara C., Tesfaye D., Jüngst H., Looft C., Schellander K. & Tholen E. (2010) Epistatic QTL pairs associated with meat quality and carcass composition traits in a porcine Duroc x Pietrain population. *Genetics Selection Evolution* 42, 39.
- Ledur M.C., Navarro N. & Pérez-Enciso M. (2009) Large-scale SNP genotyping in crosses between outbred lines: how useful is it? *Heredity* 105, 173–82.
- Müller E., Moser G., Bartenschlager H. & Geldermann H. (2000) Trait values of growth, carcass and meat quality in Wild Boar, Meishan and Pietrain pigs as well as their crossbred generations. *Animals* 117, 189–202.
- Rückert C. & Bennewitz J. (2010) Joint QTL analysis of three connected F2-crosses in pigs. *Genetics Selection Evolution* 42, 40.
- Schmid M., Wellmann R. & Bennewitz J. (2016) Power and precision of mapping genes in simulated F2 crosses using whole-genome sequence data. Book of Abstracts, 67 annual EAAP meeting, Session 27.
- Visscher P.M., Thompson R. & Haley C.S. (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* 143, 1013–20.
- Wellmann R., Preuß S., Tholen E., Heinkel J., Wimmers K. & Bennewitz J. (2013) Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics Selection Evolution* 45, 28.

Yang J., Zaitlen N.A., Goddard M.E., Visscher P.M. & Price A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* 46, 100–6.

### Supporting Information

**Table S1** Results of SNP validation in the sow validation set. Associated trait, validated QTL core region and chromosomal position of significant SNPs ( $P < 5 \times 10^{-3}$ ) are given.

Trait	QTL number	SNP	Reference ID	SSC <sup>1</sup>	Position (bp)
Dressing yield	3	<i>DIAS0000134</i>	rs342166357	14	105,797,768
	3	<i>ALGA0080788</i>	rs80986601	14	106,887,624
Carcass length	3	<i>ALGA0026064</i>	rs80985095	4	76,376,049
	3	<i>MARC0022657</i>		4	76,946,230
	4	<i>ALGA0039770</i>	rs81397551	7	25,014,857
Daily gain	1	<i>ASGA0003633</i>	rs80978823	1	80,871,555
	1	<i>ASGA0003643</i>	rs80895673	1	81,450,582
	1	<i>H3GA0002156</i>	rs80812751	1	81,563,939
Drip loss	1	<i>ASGA0010206</i>	rs81358221	2	41,793,635
	1	<i>H3GA0006664</i>	rs81358211	2	41,860,911
	1	<i>DIAS0000957</i>	rs329454916	2	41,991,741
	1	<i>DIAS0001086</i>	rs327433958	2	42,031,934

<sup>1</sup>*Sus scrofa* chromosome.

**Table S2** Results of SNP validation in the boar validation set. Associated trait, validated QTL core region and chromosomal position of significant SNPs ( $P < 5 \times 10^{-3}$ ) are given.

Trait	QTL number	SNP	Reference ID	SSC <sup>1</sup>	Position (bp)
Dressing yield	1	<i>ALGA0002800</i>	rs81353682	1	39,641,167
	1	<i>DRGA0000650</i>		1	39,655,284
	1	<i>ASGA0002445</i>	rs80809795	1	39,942,193
	1	<i>ASGA0002447</i>	rs80938019	1	39,984,957
	1	<i>ASGA0090189</i>	rs81308453	1	40,077,222
	1	<i>ALGA0106451</i>	rs81333947	1	40,403,712
	1	<i>ALGA0002842</i>	rs81353699	1	40,842,099
	1	<i>ALGA0002853</i>	rs80792861	1	41,058,097
	1	<i>MARC0027915</i>		1	42,542,130
	1	<i>ASGA0104733</i>	rs81304762	1	42,772,655
	1	<i>CASI0010130</i>	rs338789128	1	43,092,094
	1	<i>ALGA0002938</i>	rs80994400	1	43,257,183
	1	<i>H3GA0001563</i>	rs80836620	1	43,378,688
	1	<i>ALGA0002946</i>	rs81353739	1	43,450,949
	1	<i>ASGA0002548</i>	rs81353773	1	43,699,076
	1	<i>ASGA0002551</i>	rs80974726	1	43,713,569
	1	<i>H3GA0001574</i>	rs80836362	1	43,843,133
	1	<i>ALGA0002991</i>	rs80909355	1	44,450,710
	1	<i>ALGA0002996</i>	rs80883735	1	44,498,050
	1	<i>ALGA0003003</i>	rs81353795	1	44,623,901
	1	<i>ALGA0003008</i>	rs81353800	1	44,655,399
	1	<i>INRA0002055</i>		1	44,733,990
	1	<i>ASGA0002601</i>	rs81353848	1	46,303,340
	1	<i>ASGA0002612</i>	rs80792564	1	46,578,307
	3	<i>H3GA0041818</i>	rs80936691	14	105,039,074
	3	<i>ASGA0065826</i>	rs80820662	14	105,071,764
	3	<i>ALGA0114211</i>	rs80921467	14	105,072,756
	3	<i>MARC0055120</i>		14	105,076,407
	3	<i>H3GA0041827</i>	rs80866686	14	105,195,568
	3	<i>ASGA0065830</i>	rs80849604	14	105,256,566
	3	<i>MARC0027573</i>		14	105,312,199
	3	<i>ASGA0065838</i>	rs80879250	14	105,395,775
	3	<i>ALGA0080719</i>	rs80879912	14	105,408,647
	3	<i>H3GA0041837</i>	rs80885050	14	105,428,138
	3	<i>ALGA0080726</i>	rs80936072	14	105,467,022
	3	<i>H3GA0041842</i>	rs80804294	14	105,514,250

3	<i>ASGA0092332</i>	rs80847089	14	105,533,878	
3	<i>ASGA0065840</i>	rs80813006	14	105,592,400	
3	<i>DIAS0000134</i>	rs342166357	14	105,797,768	
3	<i>ALGA0080732</i>	rs80818067	14	105,884,062	
3	<i>ASGA0065848</i>	rs80914034	14	105,988,669	
3	<i>ALGA0117921</i>	rs80968709	14	106,065,100	
3	<i>DRGA0014459</i>		14	106,676,616	
3	<i>ALGA0080776</i>	rs80866873	14	106,750,452	
3	<i>ALGA0080788</i>	rs80986601	14	106,887,624	
3	<i>MARC0040955</i>		14	106,905,089	
3	<i>MARC0040736</i>		14	106,927,384	
3	<i>ALGA0080802</i>	rs80898271	14	106,969,141	
3	<i>H3GA0041895</i>	rs80951405	14	107,017,326	
3	<i>H3GA0041897</i>	rs80968475	14	107,031,142	
3	<i>ASGA0065917</i>	rs80871416	14	107,086,741	
3	<i>ASGA0065921</i>	rs80817091	14	107,106,912	
3	<i>ALGA0080825</i>	rs80950421	14	107,125,722	
3	<i>H3GA0041906</i>	rs80827099	14	107,162,911	
3	<i>ALGA0080831</i>	rs80938898	14	107,377,294	
3	<i>DRGA0014464</i>		14	107,506,648	
3	<i>ALGA0080841</i>	rs80828846	14	107,621,253	
3	<i>ASGA0065939</i>	rs80886408	14	107,677,777	
3	<i>ALGA0080877</i>	rs80976446	14	107,925,126	
3	<i>ALGA0080883</i>	rs80984312	14	107,948,513	
3	<i>INRA0046548</i>	rs319394733	14	108,115,777	
3	<i>ALGA0080896</i>	rs80871613	14	108,134,979	
3	<i>H3GA0041938</i>	rs80822812	14	108,252,178	
3	<i>ASGA0066239</i>	rs80922390	14	114,701,575	
3	<i>ASGA0066251</i>	rs80844511	14	114,964,705	
Carcass length	2	<i>ALGA0007613</i>	rs80932851	1	203,432,031
	2	<i>MARC0052458</i>		1	203,877,649
	2	<i>MARC0046138</i>		1	204,652,721
	2	<i>INRA0005794</i>	rs338971680	1	204,700,013
	2	<i>ALGA0007638</i>	rs81350271	1	204,756,911
	2	<i>ASGA0005594</i>	rs81350272	1	204,807,320
	2	<i>H3GA0003574</i>	rs81350285	1	204,836,887
	2	<i>ALGA0007641</i>	rs81350275	1	204,878,750
	2	<i>H3GA0003575</i>	rs81350288	1	204,904,961
	2	<i>MARC0055516</i>		1	204,978,538
	2	<i>H3GA0003611</i>	rs80908701	1	207,290,739
	2	<i>MARC0111426</i>		1	207,430,247
	2	<i>ASGA0005653</i>	rs81350381	1	207,558,207

	2	<i>ALGA0007745</i>	rs81350405	1	207,742,839
	2	<i>MARC0072754</i>		1	207,902,349
	2	<i>MARC0112103</i>		1	208,148,958
	2	<i>MARC0025913</i>		1	208,619,780
	4	<i>INRA0024524</i>	rs333021601	7	26,069,284
	5	<i>ASGA0037721</i>	rs81400678	8	9,720,242
	5	<i>ASGA0104632</i>	rs81304635	8	9,863,002
	5	<i>DRGA0008295</i>		8	9,920,747
Daily gain	1	<i>ALGA0004384</i>	rs80796868	1	74,913,110
	1	<i>ALGA0004392</i>	rs80791631	1	74,972,431
	3	<i>H3GA0013176</i>	rs80924250	4	78,168,852
	3	<i>ALGA0026197</i>	rs80901394	4	79,142,579
	3	<i>ASGA0020462</i>	rs80965943	4	80,408,198
Drip loss	1	<i>H3GA0006633</i>	rs81357983	2	41,082,587

---

<sup>1</sup>*Sus scrofa* chromosome.

## GENERAL DISCUSSION

The present studies were conducted to investigate the potential of existing F2 data, established for linkage analysis and updated towards single nucleotide polymorphism (SNP) chip data, to map quantitative trait loci (QTL) in the genomic era. **Chapter 1** is a review article of methods to conduct genome-wide association studies (GWASs) and frequently used mapping populations. In **chapter 2**, GWASs mapping power and precision was investigated in different simulated F2 and purebred populations. **Chapter 3** included linkage disequilibrium (LD) structure analyses of different existing F2 populations and GWASs for economically relevant meat quality and carcass traits in a pooled F2 cross. **Chapter 4** addressed the suitability of F2 data to map genes that also segregate within the sire line Piétrain.

This general discussion provides additional GWAS results for seldom investigated traits (e.g. organ weights), debates the use of F2 data for GWAS and the possibilities for the application in a purebred breed, and proposes future research directions.

### Mapping power and precision

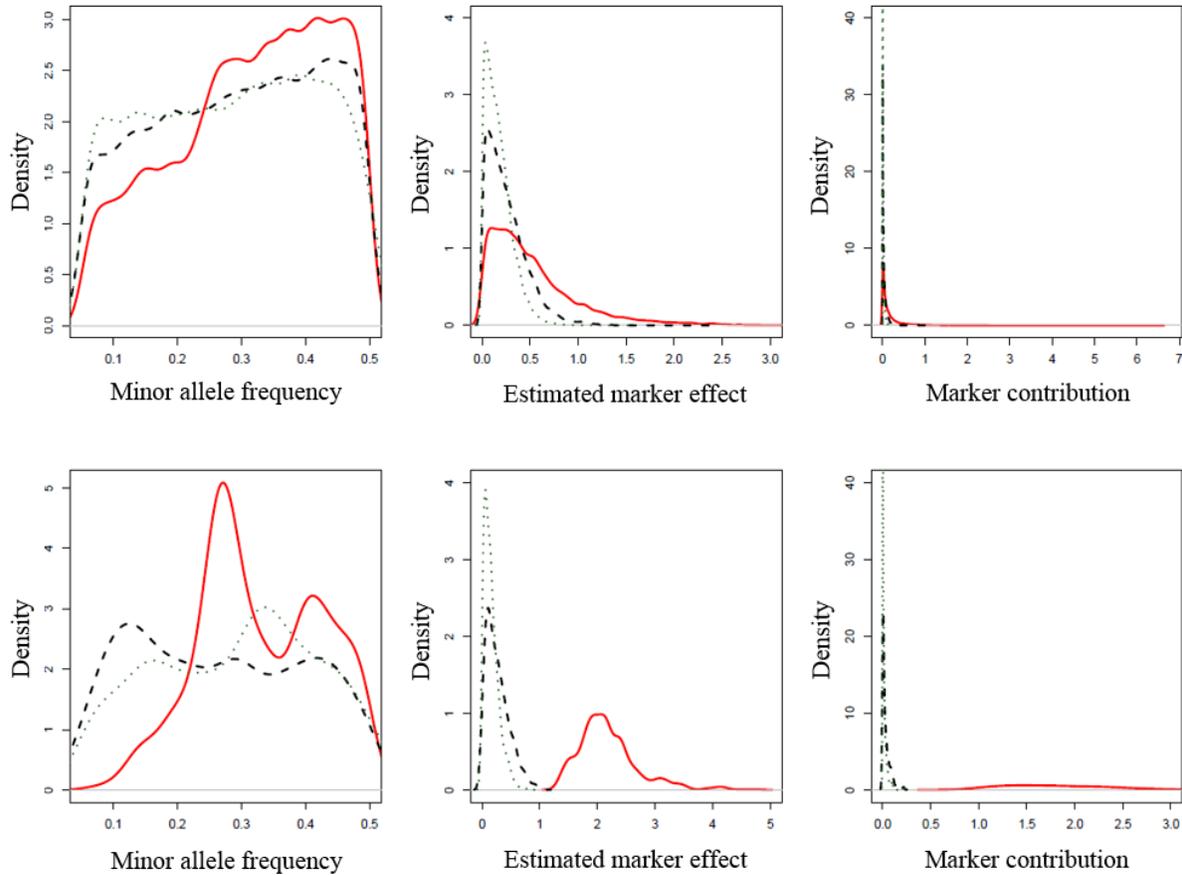
GWAS basically aims to detect SNPs causing genetic variation in a quantitative trait (causative mutations, quantitative trait nucleotide (QTN)). Since available SNP chips do not capture all polymorphic sites in the genome, associations of SNPs with the phenotype of a trait rely on LD information. Consequently, most of the associated SNPs detected in GWAS indicate a causative SNP due to LD and hence determine chromosomal regions affecting quantitative traits (QTL). The QTN and QTL mapping success is limited by two parameters, power and precision. Both are strongly influenced by GWAS design (e.g. marker density, mapping population, sample size), the genetic architecture of quantitative traits (e.g. distribution of QTN and QTL, effect sizes), and population specific parameters (e.g. variance components including heritability, extend of LD) (Ledur et al. 2009, Toosi et al. 2010).

F2 crosses were established to create informative mapping populations. Crossing divergent lineages results in many segregating SNPs that are divergently fixed within the founder breeds which in turn increase the absolute number of segregating SNPs in F2 populations compared to their founder breeds (see **chapter 2**). This is especially valid for F2 crosses derived from

genetically distantly related founder breeds like Asian and European breeds, since they have more private alleles ( $F_{ST}$  values see **chapter 2 and 3**). Pooling data can additionally increase the number of segregating SNPs, particularly the number of SNPs that also segregate within a founder breed (see **chapter 2** and Bennewitz & Wellmann (2014)).

Figure 1 compares the minor allele frequency (MAF), estimated SNP effects and the marker contributions of a pooled F2 cross and samples of the Piétrain population. Obviously, the MAF in F2 designs is on average higher than in purebred samples since SNPs that were divergently fixed (close to fixation) within the founder populations show a MAF around 0.5 (a strongly boosted MAF) in the F2 generation. Similar allele frequency distributions were also reported by Ledur et al. (2009). As the broad allele frequency spectrum in F2 designs makes a strong impact on the marker contributions (Ledur et al. 2009), the additive genetic variance tends to be enhanced (also visualized in Figure 1) which might result in a larger heritability and thus in an increased power to map such SNPs. This is also supported by the simulation study in **chapter 2** showing an extraordinary power to map QTL in F2 crosses derived from distantly related founder breeds. Pooling data lead to a further increase in power due to larger sample sizes.

For the majority of regressions single-marker models were applied since they are straightforward to implement and provide  $p$ -values which are required for using standard procedures in terms of significance testing and controlling false positives as reviewed in **chapter 1**. However, estimating effects simultaneously accounts better for population structure and can decrease false positive associations and increase power (Goddard et al. 2014). In addition, there are Bayesian models available including the consideration of dominance. These may lead to a further power increase on the part of the model (Bennewitz et al. 2017). Both approaches revealed reasonable GWAS results in F2 data conducted in **chapter 3**. For traits where dominance plays an important role, the inclusion of dominance effects resulted in a higher power and a lower rate of false positives.



**Figure 1.** Distributions of the marker minor allele frequency (MAF > 0.05) (left), estimated marker effects (middle) and additive genetic variances of the markers ( $2pqa^2$ ) (right) in a pooled F2 population (red solid line) and two Piétrain populations Pit210 (black dashed line) and Pit771 (green dotted line). The data refer to the populations investigated in **chapter 4**. The density functions are plotted for all markers segregating in both, the F2 and the respective Piétrain population ( $n = 38,505$ ; top line), and for significant F2 markers ( $n = 1,238$ ; bottom line) for the trait carcass length.

Different observations were made in regards to precision. Investigations in LD structure demonstrated that F2 crosses derived from phylogenetically divergent founder breeds showed long ranging LD blocks which resulted in a low mapping resolution (Bennewitz & Wellmann 2014). This might be due to various divergently fixed SNPs being in high LD with a causative locus despite large distances. In this case recombination events in F2 cross designs are not sufficient to break down LD blocks. LD blocks in F2 crosses from two European breeds were

even shorter than in purebred populations because the founder breeds have many segregating markers in common and thus show a high mapping resolution (Toosi et al. 2010, Bennewitz & Wellmann 2014). These findings confirm the information about LD structure in F2 designs obtained from the analyses in **chapter 3**. Pooling F2 data led to the fastest breakdown of LD. The mapping precision evaluations conducted in **chapter 2** showed the impact of LD structure on precision resulting in a low (high) precision for F2 crosses derived from distantly (closely) related founder breeds.

According to these findings regarding the mapping power and precision, F2 datasets represent powerful mapping populations, however, the mapping resolution is compromised when the founder breeds are distantly related. The pooling of F2 data can improve mapping success. Both, single-marker and multi-marker models are applicable to achieve reasonable results.

### **Pooling data**

For pooling data in genomic studies, two approaches are frequently used: (1) Combining the phenotypic and the genotypic information from different datasets (see **chapter 2, 3, and 4**) and (2) combining the resulting *p*-values or effect estimates of studies in so called meta analyses (Mao et al. 2016, Lutz et al. 2017). Studies of Rückert & Bennewitz (2010) showed that pooling F2 data results in an increased mapping power in F2 designs. Based on the simulations of Toosi et al. (2010), who stated that the LD block length in F2 crosses can be even shorter than in a outbred population, and their own findings, Bennewitz & Wellmann (2014) suggested to pool data from F2 designs with closely related founder breeds in order to maximize mapping resolution. The results of the simulation study in **chapter 2** are in agreement with these expectations. It also showed that the mapping precision was higher in such F2 crosses compared with the founder breed at equal sample size. Further, the low precision of F2 designs derived from distantly related founder breeds was improved when F2 designs established from two European breeds were added, and QTL mapping power was highest in the pooled datasets. Consequently, the studies in **chapter 3 and 4** were based on pooled F2 data in order to maximize mapping success.

### Suitability for use of existing F2 crosses in genome-wide QTL mapping analyses

Besides production traits investigated in **chapter 3** and **4**, the existing F2 cross datasets provided phenotypic data for seldom investigated traits playing a minor role in terms of meat production, but are important respecting reproduction and management, or, if the pig is seen as a model animal to humans, medical advance. To take advantage of the broad phenotypic information of the datasets including such traits, additional GWASs were conducted and presented in the following.

The dataset analyzed was created by pooling the Wild boar x Piétrain (WxP), Meishan x Piétrain (MxP) and Wild boar x Meishan (WxM) F2 populations described in detail by Rückert & Bennewitz (2010). The pooled F2 dataset consisted of 907 animals (291 WxP individuals, 304 MxP individuals, 312 WxM individuals). All individuals were SNP-genotyped using the PorcineSNP60v2 BeadChip (Illumina San Diego, CA; Ramos et al. 2009). After the exclusion of SNPs that had not passed the filter criteria (see **chapter 3** and **4**), genotypic information of 44,385 annotated autosomal SNPs with a MAF > 0.05 was available for all individuals.

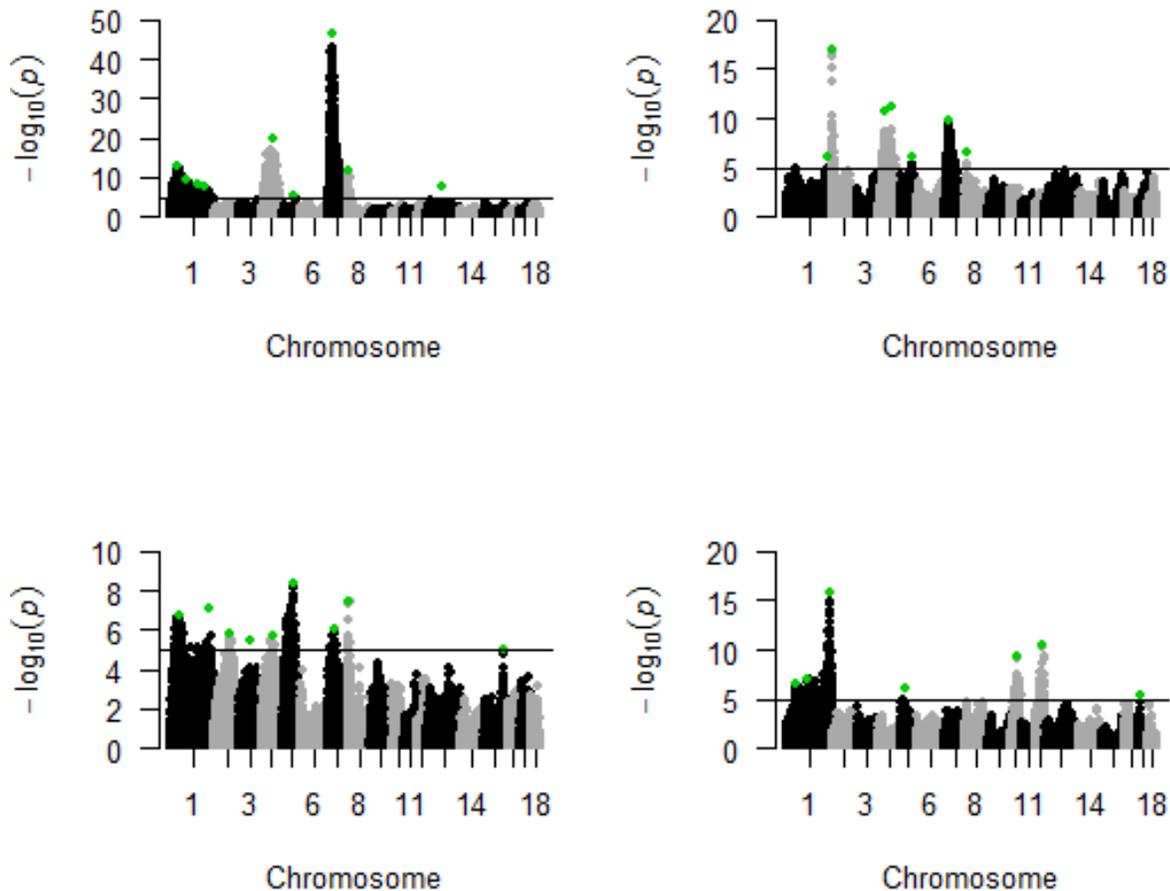
For the analyses, the traits head weight (HEW), heart weight (HTW), liver weight (LIW) and teat number (TEN) were considered, which phenotypic records were collected according to Müller et al. (2000) and Geldermann et al. (2003), and pre-corrected the same way as presented in **chapter 3** and **4**. The phenotypic correlations between the bodyweight of the individuals and the traits investigated ranged from 0.750 to 0.864 and implied a strong dependence. Genetic correlations were calculated by applying a bivariate GREml analysis using the software GCTA (Yang et al. 2014) which resulted in genetic correlations between 0.759 and 0.827. Due to these strong correlations, and to map QTL that directly affect the traits and not body weight, additional GWASs were performed considering HEW, HTW and LIW relative to the body weight of the individuals by taking the ratio of the phenotypes of the traits and the body weight as observations.

GWASs were conducted using the Software Genome-wide Complex Trait Analysis (GCTA) of Yang et al. (2014) and applying the following single-marker regression model:

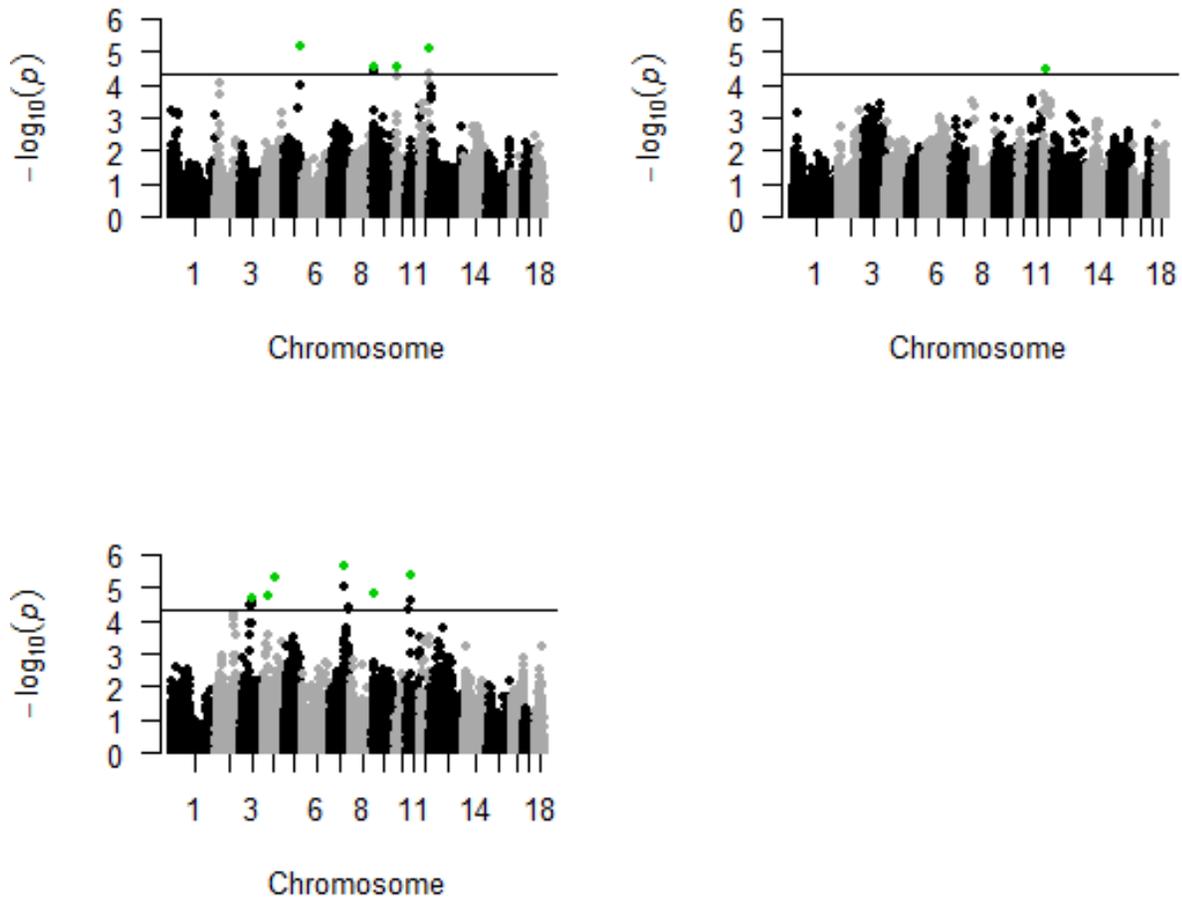
$$y_i = \mu + b_j x_{ij} + g_i + e_i$$

Here,  $y_i$  denotes the pre-corrected phenotypic record of individual  $i$ ;  $\mu$  describes the overall population mean;  $x_{ij}$  denotes the number of copies of a randomly chosen allele of SNP  $j$

( $x_{ij} = 0, 1, \text{ or } 2$ ) and  $b_j$  is the regression coefficient for SNP  $j$ . To capture population stratification effects, a random polygenic effect of the individual ( $g_i$ ) was included. The covariance structure of the polygenic effects was modeled using a genomic relationship matrix (GRM). A leave-one-chromosome-out approach was applied to avoid a double-fitting of the SNP to be tested as a fix and also as a random effect in the regression model. Thereby, the SNP to be tested and all other SNPs located on the same chromosome were excluded from calculating the GRM. To infer trait associations, SNPs with  $p_{(nominal)} < 5 * 10^{-5}$  were significant. The most significant SNP within a certain peak derived from Manhattan plots, was assumed to be a putative QTL.



**Figure 2.** Manhattan plots of the investigated traits head weight (top line, left), heart weight (top line, right), liver weight (bottom line, left) and teat number (bottom line, right) for the absolute weights. The horizontal solid line corresponds to a nominal significance level of  $p < 5 * 10^{-5}$ . Putative QTL are highlighted in green.



**Figure 3.** Manhattan plots of the investigated traits head weight (top line, left), heart weight (top line, right) and liver weight (bottom line, left) for the relative weights. The horizontal solid line corresponds to a nominal significance level of  $p < 5 \times 10^{-5}$ . Putative QTL are highlighted in green.

Table 1 shows the number of QTL (number of most significant SNPs within peaks that reached the significance threshold) and their chromosomal positions for the investigated traits. The results are visualized as Manhattan plots in Figure 2 and 3. Seven QTL were mapped for TEN and between 7 and 9 (1 and 6) QTL were associated with the absolute (relative) organ weights. The reduced number and obviously smaller power in the GWASs for the relative weights may be attributed to the drastic loss in genetic variation within the traits when adjusting for the highly correlated body weight (results not shown). Compared with the findings of Geldermann et al. (2003) who performed linkage mapping using 129 microsatellite markers in the same F2 crosses,

additional QTL for the trait HEW (HTW) were found on SSC 5, 8 and 13 (SSC 1, 5, 7 and 8). For the relative organ weights of HEW (HTW), QTL on SSC 10 and 12 (SSC 12) were detected that were neither mapped in Geldermann et al. (2003) nor in our studies using the absolute weights.

**Table 1.** Number of QTL mapped in the pooled F2 cross and their chromosomal positions (*Sus scrofa* chromosome, SSC) at a significance level of  $p_{nominal} < 5 * 10^{-5}$ .

Trait	Number of mapped QTL	Chromosomal position (SSC)
Head weight	9	1, 4, 5, 7, 8, 13
Relative head weight	5	5, 9, 10, 12
Heart weight	7	1, 2, 4, 5, 7, 8
Relative heart weight	1	12
Liver weight	9	1, 2, 3, 4, 5, 7, 8, 15
Relative liver weight	6	3, 4, 7, 9, 11
Teat number	7	1, 5, 10, 12, 17

As described above and shown in Figure 1, the allele frequencies and marker effects tend to be increased in the F2 population which results in greater additive variance fractions explained by single SNPs. Ledur et al. (2009) also found that QTL effects in F2 populations strongly depend on allele frequencies due to a much broader allele frequency spectrum. This in turn may contribute to the generally large number of significant associations showing small false discovery rates for the traits investigated (see also **chapter 3**) and implies that the existing F2 crosses provide a suitable database for GWAS in the era of genomics.

These findings might also hold true for further traits not investigated in the present studies and imply that the use of consisting F2 crosses with genomic era genotypes (SNP chip information) as a pooled dataset is a suitable approach to map QTL. In order to take advantage of both, boosted power and a higher precision, pooling F2 data derived from closely related founder breeds might lead to the most promising results. Since private QTL solely segregate in one of the two populations, it seems to be expedient to first separately analyze the datasets and subsequently extend the analyses towards a joint analysis of both (or more) datasets in order to capture the maximal number of QTL given the data. However, this should be investigated in detail in future studies.

### **Transfer of F2 data mapping results into segregating breeding populations**

Since genomic selection has been introduced in livestock breeding, it is possible to base selection decisions on genomic breeding values. Knowledge of the genetic architecture of quantitative traits and the distribution of gene effects can increase accuracy of genomic selection. Especially identified causative mutations, and the possibility to select individuals carrying the desired alleles, would lead to a substantially higher breeding progress. Consequently, mapping genes in the breeding populations is still required to improve genomic selection (Goddard et al. 2016). As genomic selection is implemented in segregating breeding populations like the economically relevant sire line Piétrain (Wellmann & Bennewitz 2012) and not in mapping populations, a transfer or an inclusion of GWAS results from F2 data into the breeding population is required. According to Bennewitz & Wellmann (2014) and the results of the simulation study in **chapter 2**, mapping genes in one of the founder breeds is particularly beneficial when using F2 data. It was shown that the mapping power was generally high for genes segregating in both populations and the signals, compared with the founder breed, were even more precise when the F2 cross was derived from closely related founder breeds.

Veroneze et al. (2014) investigated LD patterns and the persistence of phase in purebred and crossbred populations and could show that correlations of phases is high if a crossbred and a purebred population share a great proportion of the genome. This might be the case when mapping genes in a founder breed of an F2 design. Investigations in F2 data with Piétrain as one founder breed (**chapter 4**) supported these assumptions and showed that (1) most of the SNPs in significant regions mapped via GWAS in the F2 cross dataset also segregated within samples of the recent German Piétrain population, and (2) GWAS results from F2 datasets supported the mapping of QTL that contribute to the additive genetic variance in both populations (i.e. F2 cross and Piétrain). Nevertheless, inferences about the location of QTL across populations can be compromised by population specific LD (Sved 2009, Tenesa et al. 2007) which strongly relies on allele frequencies and thus may play an important role since MAF can extremely differ between an F2 cross and its founder breed (see Figure 1).

**Future research directions to map QTL using F2 data**

The combination of microarray and genotyping technologies enabled the inclusion of biological information in GWASs and has led to a constantly increasing number of so called expression QTL (eQTL) studies. Thereby, gene expression levels are measured using microarrays, e.g. Affymetrix Snowball Array (Affymetrix, Santa Clara, CA; Freeman et al. 2012), and treated as phenotypes of complex traits in gene mapping experiments. eQTL provide the possibility to infer physiological pathways via post-GWAS network analyses or to confirm QTL mapped in classical GWASs (Cookson et al. 2009). In F2 crosses, differential gene expression studies to functionally annotate genes (Ponsuksili et al. 2008) and eQTL studies, successfully implemented in GWASs (e.g. Ponsuksili et al. 2010, Steibel et al. 2011), were conducted.

However, gene expression data solely can be integrated in GWASs when setting up new F2 mapping populations or if the samples of the respective tissue are still available for existing F2 crosses, which is not applicable for the datasets used in this study.

Another opportunity to improve GWASs is the use of whole-genome sequence data, which can be applied in a follow-up study of the research presented here. Recent studies showed that an increased marker density elevated the power, decreased the FDR and had an positive effects on the mapping precision (Ledur et al. 2009, Toosi et al. 2010, Pérez-Enciso et al. 2015). As reviewed by Visscher et al. (2017), a remarkable proportion of the phenotypic variance of a quantitative trait may rely on rare SNP variants. Because the design of SNP chips (e.g. Ramos et al. 2009) aims to be suitable and informative in various populations, the MAF of pre-selected SNPs on the chips is biased towards intermediate allele frequencies (Ledur et al. 2009). Hence, SNP chips are not able to capture the rare SNP variants ( $MAF < 0.01$ ) of a population. The costs to generate such genotype datasets are still extremely high and not affordable in many situations. An alternative is the application of imputed sequence data, however, the imputation accuracy may be compromised for SNPs with a low MAF (van Binsbergen et al. 2014). Especially in F2 designs, this approach is a favorable way to maximize marker density since the costs can be drastically reduced (whole-genome sequencing of only the few founder individuals) and the imputation accuracy might be very high for F2 individuals since they descend from a small number of common ancestors. However, whole-genome sequence data is challenging across many points. Besides the high computational time and disk space requirements (Meuwissen et al. 2016), statistical analysis may suffer from the limited number of observations to estimate an

enormous number of effects. As shown in **chapter 2**, the power to map QTN, i.e. to directly map the causative SNP, was small and generally below the QTL mapping power in all datasets analyzed although whole-genome sequence data was available. This is attributed to multiple testing problems which increase with the marker density. For WGS data, Bonferroni corrections seem to be too stringent and more robust significance thresholds like FDR would be more expedient to map genes using single-marker regression models. Further, many SNPs that were not QTN showed a more significant  $p$ -value than the causative SNP itself, although one could assume that QTN effects in whole-genome sequence data must show the most significant  $p$ -values. A two-step approach appears to be suitable: GWAS should initially be conducted using SNP chip data to map regions that significantly contribute to the genetic variance of a trait. The detected genomic areas in turn should be used as target regions for subsequent fine mapping experiments applying whole-genome sequence data to pinpoint the causative SNPs. This could overcome the computational time and multiple testing issues since the number of tests is substantially reduced.

**References**

- Bennewitz J., Edel C., Fries R., Meuwissen T.H.E. & Wellmann R. (2017). Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. *Genet Sel Evol* 49: 7.
- Bennewitz J. & Wellmann R. (2014). Mapping Resolution in Single and Multiple F2 Populations using Genome Sequence Marker Panels. In: *Proceedings, 10th World Congress of Genetics Applied to Livestock Production, 17–22 August 2014, Vancouver*, [https://asas.org/docs/default-source/wcgalp-proceedings-oral/184\\_paper\\_8399\\_manuscript\\_106\\_0.pdf?sfvrsn=2](https://asas.org/docs/default-source/wcgalp-proceedings-oral/184_paper_8399_manuscript_106_0.pdf?sfvrsn=2).
- Cookson W., Liang L., Abecasis G., Moffatt M. & Lathrop M. (2009). Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10: 184–194.
- Freeman T.C., Ivens A., Baillie J.K., Beraldi D., Barnett M.W., Dorward D., Downing A., Fairbairn L., Kapetanovic R., Raza S., Tomoiu A., Alberio R., Wu C., Su A.I., Summers K.M., Tuggle C.K., Archibald A.L. & Hume D.A. (2012). A gene expression atlas of the domestic pig. *BMC Biol* 10: 90.
- Geldermann H., Müller E., Moser G., Reiner G., Bartenschlager H., Cepica S., Stratil A., Kuryl J., Moran C., Davoli R. & Brunsch C. (2003). Genome-wide linkage and QTL mapping in porcine F2 families generated from Pietrain, Meishan and Wild Boar crosses. *J Anim Breed Genet* 120: 363–393.
- Goddard M.E., Kemper K.E., MacLeod I.M., Chamberlain A.J. & Hayes B.J. (2016). Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc Biol Sci* 283: 1835.
- Goddard M.E., MacLeod I.M., Kemper K.E., van der Jagt C.J., Savin K., Schrooten C. & Hayes B.J. (2014). A Research Plan for the Identification of QTL. In: *Proceedings, 10th World Congress of Genetics Applied to Livestock Production, 17–22 August 2014, Vancouver*, [https://asas.org/docs/default-source/wcgalp-proceedings-oral/199\\_paper\\_10348\\_manuscript\\_1635\\_0.pdf?sfvrsn=2](https://asas.org/docs/default-source/wcgalp-proceedings-oral/199_paper_10348_manuscript_1635_0.pdf?sfvrsn=2).
- Ledur M.C., Navarro N. & Pérez-Enciso M. (2009). Large-scale SNP genotyping in crosses between outbred lines: how useful is it? *Heredity (Edinb)* 105: 173–182.

- Lutz V., Stratz P., Preuß S., Tetens J., Grashorn M.A., Bessei W. & Bennewitz, J. (2017). A genome-wide association study in a large F2-cross of laying hens reveals novel genomic regions associated with feather pecking and aggressive pecking behavior. *Genet Sel Evol* 49: 18.
- Mao X., Sahana G., de Koning D.J. & Guldbbrandtsen B. (2016). Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data. *J Anim Sci* 94: 1426–1437.
- Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2016). Genomic selection : A paradigm shift in animal breeding. *Anim Front* 6 (1): 6–14.
- Müller E., Moser G., Bartenschlager H. & Geldermann H. (2000). Trait values of growth, carcass and meat quality in Wild Boar, Meishan and Pietrain pigs as well as their crossbred generations. *J Anim Breed Genet* 117: 189–202.
- Pérez-Enciso M., Rincón J.C. & Legarra A. (2015). Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol* 47: 43.
- Ponsuksili S., Murani E., Phatsara C., Jonas E., Walz C., Schwerin M., Schellander K. & Wimmers K. (2008). Expression profiling of muscle reveals transcripts differentially expressed in muscle that affect water-holding capacity of pork. *J Agric Food Chem* 56 (21): 10311–10317.
- Ponsuksili S., Murani E., Schwerin M., Schellander K. & Wimmers K. (2010). Identification of expression QTL (eQTL) of genes expressed in porcine *M. longissimus dorsi* and associated with meat quality traits. *BMC Genom* 11: 572.
- Ramos A.M., Crooijmans R.P.M.A., Affara N.A., Amaral A.J., Archibald A.L., Beever J.E., Bendixen C., Churcher C., Clark R., Dehais P., Hansen M.S., Hedegaard J. Hu, Z.L., Kerstens H.H., Law A.S., Megens H.J., Milan D., Nonneman D.J., Rohrer G.A., Rothschild M.F., Smith T.P.L., Schnabel R.D., van Tassell C.P., Taylor J.F., Wiedmann R.T., Schook L.B. & Groenen M.A.M. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* 4: e6524.

- Rückert C. & Bennewitz J. (2010). Joint QTL analysis of three connected F2-crosses in pigs. *Genet Sel Evol* 42: 40.
- Steibel J.P., Bates R.O., Rosa G.J.M., Tempelman R.J., Rilington V.D., Ragavendran A., Raney N.E., Ramos A.M., Cardoso F.F., Edwards D.B. & Ernst C.W. (2011). Genome-Wide Linkage Analysis of Global Gene Expression in Loin Muscle Tissue Identifies Candidate Genes in Pigs. *PLoS One* 6: e16766.
- Sved J.A. (2009). Linkage disequilibrium and its expectation in human populations. *Twin Res Hum Genet* 12: 35–43.
- Tenesa A., Navarro P., Hayes B.J., Duffy D.L., Clarke G.M., Goddard M.E. & Visscher P.M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17: 520–526.
- Toosi A., Fernando R.L. & Dekkers J.C.M. (2010). Genomic selection in admixed and crossbred populations. *J Anim Sci* 88: 32–46.
- van Binsbergen R., Bink M.C.A.M., Calus M.P.L., van Eeuwijk F.A., Hayes B.J., Hulsegge I. & Veerkamp R.F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* 46: 1–13.
- Veroneze R., Bastiaansen J.W.M., Knol E.F., Guimarães S., Silva F.F., Harlizius B., Lopes M.S. & Lopes P.S. (2014). Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations. *BMC Genet* 15: 126.
- Visscher P.M., Wray N.R., Zhang Q., Sklar P., McCarthy M.I., Brown M.A. & Yang J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101: 5–22.
- Wellmann R. & Bennewitz J. (2012). Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genet Res (Camb)* 94: 21–37.
- Yang J., Zaitlen N.A., Goddard M.E., Visscher P.M. & Price A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 46: 100–6.

## ACKNOWLEDGMENTS

Am Ende dieser Arbeit ist es mir ein Anliegen, mich zu bedanken.

Bereits während meiner Masterarbeit hat mir das Forschen im Bereich der Genetik und Züchtung sehr viel Spaß gemacht. Prof. Dr. Jörn Bennewitz hat mir die Möglichkeit gegeben dies fortzusetzen. Deshalb geht ein ganz großes Dankeschön an ihn! Danke für die super Zusammenarbeit und Betreuung. Ich schätze sehr, dass ich an internationalen Kursen und Kongressen teilnehmen durfte.

Bei Herrn Prof. Dr. Georg Thaller bedanke ich mich für die Übernahme des Koreferats.

Nach manchmal tagelanger Fehlersuche beim Programmieren hatte ich immer ein Ass im Ärmel: Dr. rer. nat. Dr. sc. agr. Robin Wellmann! Ich möchte mich herzlich für seine Hilfsbereitschaft und gute Betreuung bedanken.

Vielen Dank auch an Dr. sc. agr. Patrick Stratz für die vielen Gedankenspaziergänge durch die quantitative Genetik, allerlei Methoden oder verrückte Pläne für neue Forschungsprojekte. Er war mir immer ein guter Ansprechpartner.

An dieser Stelle möchte ich mich bei allen weiteren Co-Autoren für ihren Beitrag an den Veröffentlichungen bedanken.

Mein Dank gilt auch der Deutschen Forschungsgemeinschaft (DFG) für die finanzielle Unterstützung in diesem Projekt.

Ohne Dr. rer. nat. Siegfried Preuß und sein Team Gabi, Heidi, Beate und Marisa hätte ich nicht viel auswerten können. Vielen Dank für die Zusammenarbeit und alle Laborarbeiten die dazu geführt haben, dass ich Datensätzen zur Verfügung hatte.

I gratefully thank Dani and Chris for their friendship, support, and the language editing.

Für die Unterstützung in organisatorischen Dingen bedanke ich mich herzlich bei Christina und Karin. Obwohl dabei oft ihre Spontanität gefragt war, nahmen sie sich immer Zeit für mich.

Besonders bedanken möchte mich bei Maria und Annik, die sich mit mir ein Büro geteilt haben. Die Gemeinschaft in unserer Büro-WG habe ich immer sehr geschätzt (genauso wie das Glaselfleisch oder die Minisalamis ;)). Sie haben die Höhen und Tiefen während der PhD-Zeit miterlebt und haben mich immer unterstützt oder mir auch mal den Kopf gewaschen, wenn es

sein musste. Das gilt auch für unsere Stammgäste Philipp und Basti, denen ich ebenfalls dafür danken möchte.

Herzlichen Dank auch allen Kolleginnen und Kollegen. Sie haben mir ein angenehmes Arbeitsumfeld geschaffen. Das gute Miteinander und die schönen Gespräche in den Kaffeepausen haben dazu beigetragen, dass ich immer gerne am Institut war.

Auf meine Familie und Freunde ist immer Verlass, das war auch während der Promotion so. Deshalb widme ich ihnen das letzte Zitat in dieser Dissertation:

Denkt an die Tage die hinter uns liegen  
Wie lang wir Freude und Tränen schon teilen  
Hier geht jeder für jeden durchs Feuer  
Im Regen stehen wir niemals allein  
Und solange unsre Herzen uns steuern  
Wird das auch immer so sein  
Ein Hoch auf das was vor uns liegt  
Dass es das Beste für uns gibt  
Ein Hoch auf das was uns vereint  
Auf diese Zeit

Andreas Bourani

To God be the glory!

**LIST OF PUBLICATIONS**

## PUBLICATIONS INCLUDED IN THE DISSERTATION

Stratz P., Schmid M., Wellmann R., Preuß S., Blaj I., Tetens J., Thaller G. & Bennewitz, J. (2018). Linkage disequilibrium pattern and genome-wide association mapping for meat traits in multiple porcine F2-crosses. *Anim Genet*: doi: 10.1111/age.12684.

Schmid M., Maushammer M., Preuß S. & Bennewitz J. (2018). Mapping QTL for production traits in segregating Piétrain pig populations using GWAS results of F2 crosses. *Anim Genet* 49: 317-320.

Schmid M., Wellmann R. & Bennewitz J. (2018). Power and precision of QTL mapping in simulated multiple porcine F2 crosses using whole-genome sequence information. *BMC Genet* 19: 22.

Schmid M. & Bennewitz J. (2017). Invited Review: Genome-wide analysis for quantitative traits in livestock – a selective review of statistical models and experimental designs. *Arch Anim Breed* 60: 335-346.

## PUBLICATIONS

Stratz P., Schmid M., Wellmann R., Preuß S., Blaj I., Tetens J., Thaller G. & Bennewitz, J. (2018). Linkage disequilibrium pattern and genome-wide association mapping for meat traits in multiple porcine F2-crosses. *Anim Genet*: doi: 10.1111/age.12684.

Schmid M., Maushammer M., Preuß S. & Bennewitz J. (2018). Mapping QTL for production traits in segregating Piétrain pig populations using GWAS results of F2 crosses. *Anim Genet* 49: 317-320.

Schmid M., Wellmann R. & Bennewitz J. (2018). Power and precision of QTL mapping in simulated multiple porcine F2 crosses using whole-genome sequence information. *BMC Genet* 19: 22.

- Schmid M. & Bennewitz J. (2017). Invited Review: Genome-wide analysis for quantitative traits in livestock – a selective review of statistical models and experimental designs. *Arch Anim Breed* 60: 335-346.
- Schmid M., Wellmann R., Preuß S., Stratz P., Tetens J., Thaller G. & Bennewitz J. (2016). Genomweite Assoziationskartierungen in multiplen porcinen F2-Kreuzungen mit Bayes-Modellen. Proceedings Vortragstagung der DGfZ und GfT, 20 - 21 September 2016, Hannover.
- Schmid M., Wellmann R. & Bennewitz J. (2016). Power and precision of mapping genes in simulated F2 crosses using whole-genome sequence data. Proceedings 67th annual EAAP meeting, 29 August – 2 September 2016, Belfast, Session 27: 282.
- Schmid M., Wellmann R. & Bennewitz J. (2015). Power- und Präzisionsstudie zur Kartierung von Genen in porcinen F2-Designs – eine Simulationsstudie. Proceedings Vortragstagung der DGfZ und GfT, 16 - 17 September 2015, Berlin.
- Schmid M., Wellmann R., Thaller G. & Bennewitz J. (2014). Lokalisierung der additiv-genetischen Varianz für Assoziationskartierung. Proceedings Vortragstagung der DGfZ und GfT, 17 – 18 September 2014, Dummerstorf.

# CURRICULUM VITAE

## PERSONAL DETAILS

Name Markus Schmid  
Date of Birth June 26, 1988  
Place of Birth Böblingen, Germany

## EDUCATION

2014 - 2018 **PhD, Department of Animal Genetics and Breeding**  
University of Hohenheim, Stuttgart, Germany

2012 - 2014 **MSc Agricultural Science**  
University of Hohenheim, Stuttgart, Germany

2009 - 2012 **BSc Agricultural Science**  
University of Hohenheim, Stuttgart, Germany

2007 - 2009 **Agricultural training**  
  
Biolandhof Bodemer, Ehningen, Germany  
  
Friedhelm Spengler GbR, Dagersheim, Germany

2004 – 2007 **High school (Biotechnology)**  
Mildred Scheel Schule, Böblingen, Germany

## COURSES / INTERNSHIP

2017 **Breeding Program Course: Design of breeding programs with Genomic Selection.**  
Universität Wageningen, Wageningen, The Netherlands

2014 **Genomic Selection Course: Introduction to theory and implementation of Genomic Selection.**  
Universität Wageningen, Wageningen, The Netherlands

2013 **Internship: Rinderunion Baden-Württemberg e.V.**  
Cattle breeders' association, Herberlingen, Germany