



---

# **GENOMIC AND CONVENTIONAL EVALUATIONS FOR FERTILITY TRAITS IN PIGS**

Dissertation

zur Erlangung des Doktorgrades

der Fakultät für Agrarwissenschaften

der Georg-August-Universität Göttingen

vorgelegt von

Anna Maria Fangmann

geboren in Lohne

Göttingen, im September 2018

D 7

1. Referent: Prof. Dr. Henner Simianer

2. Referent: Prof. Dr. Jörn Bennewitz

Tag der mündlichen Prüfung: 27. September 2018

*meiner Familie*



## TABLE OF CONTENTS

<b>SUMMARY</b>	<b>6</b>
<b>ZUSAMMENFASSUNG</b>	<b>9</b>
<b>1<sup>ST</sup> CHAPTER</b>	<b>13</b>
<b>GENERAL INTRODUCTION</b>	<b>13</b>
Preface	14
Pig production in Germany	14
A brief history of animal breeding	15
Methods of breeding value estimation	17
Assessment of predictive ability	24
Breeder's equation	26
Genomic selection in pig breeding	26
Potentials for improving prediction accuracy in pig breeding	28
Fertility traits	30
Objectives of this thesis	32
<b>2<sup>ND</sup> CHAPTER</b>	<b>39</b>
<b>CAN MULTI-SUBPOPULATION REFERENCE SETS IMPROVE THE GENOMIC PREDICTIVE ABILITY FOR PIGS?</b>	<b>39</b>
<b>3<sup>RD</sup> CHAPTER</b>	<b>71</b>
<b>EMPIRICAL COMPARISON BETWEEN DIFFERENT METHODS FOR GENOMIC PREDICTION OF NUMBER OF PIGLETS BORN ALIVE IN MODERATE SIZED BREEDING POPULATIONS</b>	<b>71</b>
<b>4<sup>TH</sup> CHAPTER</b>	<b>94</b>
<b>ESTIMATION OF GENETIC PARAMETERS FOR REPRODUCTION TRAITS IN DAM LINES OF A GERMAN PIG BREEDING ORGANIZATION</b>	<b>94</b>
<b>5<sup>TH</sup> CHAPTER</b>	<b>119</b>
<b>GENERAL DISCUSSION</b>	<b>119</b>
Preface	120
Perspectives and challenges of the genomic selection in pig breeding	120
Genomic selection in pig breeding	121
Improvement of genomic predictability for pigs	127
Single-step genomic BLUP in pig breeding	129
Validation methods for predictions	134
Genetic progress in (genomic) pig breeding	137
Future prospective and concluding remarks	140

## SUMMARY

The aim of genomic selection (GS) is to predict breeding values with high accuracy for young animals (without own phenotypic record) as early as possible. GS can increase the accuracy of the breeding values at the time point of selection, but often the number of available animals for the reference set within an organization (subpopulation) is the limiting factor. One possibility to overcome this problem is to enlarge the reference population by combining closely (or distantly) related subpopulations within a breed, a so called multi-subpopulation reference population. The assessment of predictive ability of genomic breeding values when using single- and multi-subpopulation references sets within a breed for the trait number of piglets born alive (NBA) was conducted in **Chapter 2**. Furthermore, a comprehensive comparison of different genomic relationship matrices (partly accounting for subpopulation structures) was investigated to assess their usefulness for multi-subpopulation approaches.

Superiority of multi-subpopulation predictions in pigs compared to within-subpopulation predictions turned out to be rather small. Although predictions were performed within one breed (i.e. Large White), but different subpopulations, no increase or even a decrease in predictive ability was observed. Anyway, closely related subpopulation reference sets performed better than distantly related subpopulation reference sets. Despite the low differentiation of the subpopulations (low  $F_{ST}$ -values), the genetic connectedness between different subpopulations seems to be too small to improve the prediction accuracy by using multi-subpopulation reference sets, which may be caused by the separate breeding work of different German pig breeding organizations and have led to stratified subpopulations within the breed German Large White. The consideration of possible substructures through the use of different genomic relationship matrices in genomic estimations was also only partially successful. For practical application, resources of pig breeding companies should be used genotyping animals (boars and sows) within organization to create a sufficient large reference population which should be updated continuously.

Since GS is considered to be state-of-the-art in animal breeding, a comprehensive comparison of different genomic models, multi- and single-step, was performed for NBA and two breeds (German Landrace and German Large White) in **Chapter 3**. Multi-step methods consist of mainly three parts with many parameters and multiple assumptions: (i) constructing of a response variable for genotyped animals that integrate all phenotypic information, (ii) exploiting the association between response variable and marker information through

genomic prediction, and (iii) blending the genomic information with parental average. If assumptions in those steps are violated, loss of information, inaccuracies and biases may arise. One possibility to overcome these issues is the single-step method. In single-step methodology, all available information (i.e. pedigree, phenotypic and genetic) is combined within a single model.

Assessment of predictive abilities for young genotyped animals indicated that both genomic methods, multi- and single-step, outperformed conventional predictions, while single-step provided higher reliabilities than multi-step. Bias was assessed by regression of corrected phenotypes on the different genomic breeding values. Predictions were less biased for single-step compared to multi-step. In general, reliabilities and predictive abilities for young animals were relatively small for both breeds, which may be caused by (i) small numbers of genotyped animals in general, (ii) rather moderate reliabilities of pseudo-observations, (iii) low numbers of genotyped progenies per boar and (iv) only few parent-offspring-links between reference and validation set. In order to potentially improve prediction accuracy and reduce bias of genomic predictions, an adjustment of **G** through sophisticated weighting and scaling strategies was performed. However, an increase of predictive ability through adjustment was not successful for both small empirical data sets. For practice, single-step turned out to be useful and conceptually convincing approach for NBA in moderately sized German Large White and German Landrace populations.

Although GS is considered to be the preferred method, accurately estimated conventional breeding values through (consequent) performance testing along with recording phenotypes still remains one of the most important steps in the animal breeding schemes. Fertility traits such as NBA are economically important and included in most breeding schemes. In order to improve efficiency of breeding programs (and efficiency of piglet producers), traits like mothering ability (MA) of a sow, piglet survival (PS) or number of piglets weaned (NOW) from a sow have become more and more important. Therefore, knowledge of genetic parameters of fertility traits is necessary to estimate conventional breeding values accurately, to combine fertility traits in selection and to optimize breeding schemes. In **Chapter 4** estimates of genetic parameters for e.g. heritability, repeatability, genetic and phenotypic variances and correlations between traits were calculated in order to evaluate an appropriate model for the routine breeding value estimation for a German pig breeding organization. The analyzed traits were: NBA, NOW, MA, PS and farrowing interval (FI). Variable selection for

---

fixed effects was performed and different models (bivariate animal or repeatability model) were used to estimate genetic components.

Genetic components were generally close to literature means, although estimated variance components strongly depended on the population structure and data set used, which made a direct comparison (of differences) difficult. However, estimates of additive genetic variance, heritability and genetic correlation indicated that the amount of genetic variation for selection was large enough to improve the traits studied. Trends observed from the data already showed an improvement of NBA (NOW) per sow and year. For the routine breeding value estimation, a bivariate animal model should be used for NBA and NOW, in which the first parity and subsequent parities should be considered as different traits. A repeatability animal model should be used for MA, PS and FI. With regard to animal welfare of sow and piglet, decreasing individual birth weights and biological limitations of reproduction performance of a sow, especially PS and MA are getting more and more important and thus should be further addressed and studied.

## ZUSAMMENFASSUNG

Ziel der genomischen Zuchtwertschätzung (ZWS) ist es, Zuchtwerte mit hoher Genauigkeit für junge (nicht-phänotypisierte) Tiere zu einem möglichst frühen Zeitpunkt vorherzusagen. In der Schweinezucht kann durch die genomische ZWS eine Erhöhung der Genauigkeit der Zuchtwerte zum Selektionszeitpunkt erreicht werden. Jedoch ist häufig die Anzahl der zur Verfügung stehenden Tiere für die Referenzpopulation innerhalb einer Organisation der limitierende Faktor. Eine Möglichkeit dieses Problem zu überwinden besteht darin, die Referenzpopulation durch eng (oder entfernt) verwandte Subpopulationen innerhalb einer Rasse zu einer sogenannten Multi-Subpopulation-Referenzpopulation zu erweitern.

Durch die getrennte Zuchtarbeit verschiedener Schweinezuchtorganisationen sind über Jahrzehnte Subpopulationen in der Rasse Deutsches Edelschwein entstanden. Ziel dieser Untersuchung (Kapitel 2) ist daher die Evaluierung einer gemeinsamen genomischen ZWS, wobei die Daten verschiedener Zuchtorganisationen kombiniert werden sollen, um dadurch eine verbesserte Vorhersagegenauigkeit für die Selektion im Merkmal Anzahl lebend geborene Ferkel (LGF) zu erhalten. Des Weiteren wurde untersucht, ob die Berücksichtigung möglicher Substrukturen von gemischten Subpopulationen in der genomischen ZWS (z.B. durch die Skalierung der genomischen Verwandtschaftsmatrix) zu einer Erhöhung der Vorhersagegenauigkeit führt.

Durch das Zusammenlegen von mehreren Subpopulationen zu einer gemeinsamen Multi-Subpopulations-Referenzstichprobe kommt es generell zu Genauigkeitsverlusten. Obwohl die Berechnungen innerhalb einer Rasse, aber in verschiedenen Subpopulationen durchgeführt wurden, konnte die Vorhersagegenauigkeit nicht verbessert werden. Tendenziell konnten jedoch eng verwandte Subpopulationen besser vorhergesagt werden als entfernt verwandte. Trotz der geringen Differenzierung der Populationen (geringe  $F_{st}$ -Werte) scheinen die Subpopulationen durch die getrennte Zuchtarbeit zu wenig genetische Verknüpfungen zu haben, als dass eine gemeinsame genomische ZWS die Genauigkeit der Vorhersage erhöhen könnte. Die Berücksichtigung möglicher Substrukturen durch den Einsatz verschiedener genomischer Verwandtschaftsmatrizen war ebenfalls nur teilweise erfolgreich. Sofern eine gemeinsame ZWS für die Rasse Deutsches Edelschwein etabliert werden soll, ist es in Zukunft besonders wichtig, genetische Verknüpfungen zwischen den Subpopulationen zu schaffen. Andernfalls sollten die Ressourcen der Schweinezüchter v.a. für die Phäno- und

Genotypisierung von Tieren innerhalb der Organisation genutzt werden, um eine ausreichend große und aktuelle Referenzpopulation zu schaffen.

Da die genomische Selektion als „Status quo“ in der Tierzucht gilt, wurde in Kapitel 3 ein umfassender Vergleich verschiedener genomischer Modelle für das Merkmal LGF durchgeführt. Aktuelle Methoden der genomischen ZWS sind sog. multi-step-Verfahren, die aus mehreren Berechnungsschritten bestehen: (1) Berechnung der Inputvariable für die ZWS wie z.B. dem konventionellen Zuchtwert oder eines sog. Quasi-Phänotypen für die typisierten Tiere, (2) Berechnung der genomischen Zuchtwerte für typisierte Tiere, und (3) Blending der genomischen Zuchtwerte mittels der konventionellen Zuchtwerte bzw. des Elternzuchtwerts, um den finalen geblendeten genomischen Zuchtwert zu berechnen. Eine der Herausforderungen dieses Verfahrens ist die Abhängigkeit von vielen Parametern und Annahmen, was die Methode komplex und anfällig für Fehler machen kann. In diesem Zusammenhang wurde das sog. single-step-Verfahren entwickelt. Hierbei werden alle verfügbaren Informationen wie Pedigree, Rohphänotypen und Genotypen aller Tiere in einem Schritt miteinander verknüpft, was zu einer sichereren Schätzung der genomischen Zuchtwerte beitragen kann. In Kapitel 3 wurden die Vorhersagegenauigkeiten der genomischen Zuchtwerte mit den beiden Verfahren, multi- und single-step, für das Merkmal LGF miteinander verglichen. Für diese Analyse standen Daten der Rassen Deutsche Landrasse und Deutsches Edelschwein für das Merkmal LGF zur Verfügung.

Die Schätzung von sicheren genomischen Zuchtwerten in Populationen mit geringer Anzahl an genotypisierten Tieren ist generell problematisch. Allerdings lieferten genomische Verfahren genauere Vorhersagen für junge Tiere als konventionelle Zuchtwertschätzverfahren. Um eine mögliche Verzerrung der Zuchtwerte zu erfassen, wurde eine Regression der korrigierten Phänotypen auf die verschiedenen genomischen Zuchtwerte durchgeführt. Genomische Vorhersagen mittels single-step waren weniger verzerrt als mit dem multi-step-Verfahren. Die Sicherheit der Zuchtwerte sowie die Vorhersagefähigkeit für junge Tiere waren bei beiden Rassen eher klein, was (i) auf die geringe Anzahl genotypisierter Tiere im Allgemeinen, (ii) die geringen Sicherheiten der Quasi-Phänotypen, (iii) die geringe Anzahl genotypisierter Nachkommen pro Eber sowie (iv) die geringe Anzahl an Eltern-Nachkommen-Verknüpfungen zwischen Referenz- und Validierungstieren zurückzuführen ist. Um die Vorhersagegenauigkeit zu verbessern und die Verzerrung genomischer Vorhersagen zu reduzieren, wurde eine Anpassung der genomischen Verwandtschaftsmatrix durch verschiedene Gewichtungs- und Skalierungsstrategien

durchgeführt. Eine Erhöhung der Vorhersagekraft erwies sich jedoch für keinen der beiden empirischen Datensätze als erfolgreich. Zusammenfassend lässt sich sagen, dass für kleine und strukturierte Schweinezuchtpopulationen das single-step-Verfahren eine robuste Alternative zu dem bisher genutzten multi-step-Verfahren für Fruchtbarkeitsmerkmale darstellt.

Obwohl die genomische Selektion mittlerweile einer der wichtigsten Bestandteile in der Schweinezucht ist, bleibt die sichere Schätzung der konventionellen Zuchtwerte durch die konsequente und genaue Erfassung von Phänotypen nach wie vor eine der wichtigsten Aufgaben und Schritte in Zuchtprogrammen. Fruchtbarkeitsmerkmale wie z.B. LGF sind wirtschaftlich wichtige Merkmale und in den meisten Zuchtprogrammen implementiert. Um die Effizienz der Zuchtprogramme oder die Effizienz der Ferkelproduzenten zu verbessern, gewinnen Merkmale wie die Mütterlichkeit einer Sau (MS), die Überlebensfähigkeit des Ferkels (PS) sowie die Anzahl abgesetzter Ferkel (AGF) zunehmend an Bedeutung. Dementsprechend sind Kenntnisse zu genetischen Parametern notwendig. Für eine deutsche Schweinezuchtorganisation soll ein geeignetes Model für die routinemäßige ZWS für verschiedene Fruchtbarkeitsmerkmale untersucht werden (Kapitel 4). Die folgenden genetischen Parameter werden für fünf verschiedene Fruchtbarkeitsmerkmale berechnet: Erbllichkeit, Wiederholbarkeit, genetische und phänotypische Varianzen und Korrelation zwischen den Merkmalen. Die analysierten Merkmale waren: LGF, AGF, MS, PS und Absetzintervall (AI). Eine Variablenselektion wurde für die fixen Effekte im Model durchgeführt. Zwei verschiedene Modelle (Zwei-Merkmals-Modell und Wiederholbarkeitsmodell) wurden zur Schätzung und Berechnung der genetischen Komponenten verwendet.

Die geschätzten genetischen Komponenten stimmen generell mit Werten aus der Literatur überein. Die berechneten Parameter wie z.B. die Varianzkomponenten sind populations- sowie datenstrukturabhängig, daher ist ein direkter Vergleich mit Literaturwerten nur bedingt möglich. Schätzungen der additiven genetischen Varianz, der Erbllichkeit und der genetischen Korrelation deuten darauf hin, dass die untersuchten reproduktiven Merkmale durch Selektion verbessert werden können. Die aus den Daten beobachteten Trends von 2010 bis 2016 zeigten bereits eine Verbesserung in z.B. den Merkmalen LGF und AGF pro Sau und Jahr. Als routinemäßige ZWS sollte für LGF und AGF das Zwei-Merkmals-Modell verwendet werden, wobei der erste Wurf und die darauffolgenden Würfe je als ein Merkmal betrachtet werden sollten. Wiederholbarkeitsmodelle sollten hingegen bei PS, AI und MS verwendet werden.

Zusammenfassend lässt sich sagen, dass die Überlebensfähigkeit eines Ferkels sowie die Mütterlichkeit der Sau immer wichtiger werden, insbesondere im Hinblick auf den Tierschutz (Sau und Ferkel), die geringen individuellen Geburtsgewichte sowie das Erreichen der biologische Leistungsgrenze einer Sau.

# **1<sup>st</sup> CHAPTER**

## **GENERAL INTRODUCTION**

### **Preface**

The primary focus of this thesis is to explore the potential of genomic prediction of breeding values for fertility traits in practical pig data. A short overview of breeding history, relevant methods and genomic selection in pig breeding will be presented hereafter.

### **Pig production in Germany**

In the past decades, the northwest of Europe has developed into a concentration area of pork production. This development was due to high growth rates in production, a high degree of farm specialization and infrastructure. In 2018, the number of pigs kept in Germany amounts to roughly 27 million, of which 7.9 million are piglets, 1.9 million are breeding sows and 17.0 million are slaughter pigs<sup>1</sup>, respectively.

In Germany, pig breeding is organized by separate herdbook societies and breeding companies. Herdbook organizations usually operate on a regional basis and conduct their own breeding programs with self-defined breeding purpose. In general, breeding organizations work with the same breeds, which are bred separately for several generations and which create subpopulations within a breed. Sire and dam lines are bred specifically for markets where each breeding organization pursuing its own breeding objective. Therefore, availability of uniformly defined phenotypes is highly fragmented which complicates a common evaluation of breeding values within and across breeds. The main goal for every breeding organization, especially for those working with dam lines, is the improvement of reproduction traits, e.g. “number of piglets born alive” and “number of piglets weaned” (Knol et al., 2016; Willam and Simianer, 2017).

Breeding organizations are represented by a nucleus population and are responsible for the breeding of purebred animals. Genetic progress is mainly achieved at this stage of breeding, primarily through extensive performance tests and intensive selection. A typical large pure line pig population counts about 2000 sows with around 50 sires selected by year, but can be much smaller (Knol et al., 2016). The crossbred end product derives from multiple breeds and lines. Classical breeding programs are crossbreeding programs (e.g. three-way cross) which incorporates a F1 sows, mostly a cross of Landrace and Large White, with a sire line. Typical sire lines are Piétrain, Duroc or Hampshire.

---

<sup>1</sup> [https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2018/06/PD18\\_234\\_413.html](https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2018/06/PD18_234_413.html)

### **A brief history of animal breeding**

Animal breeding is based on the selection of the best animals from the current population as parents for the next generation. Estimating reliable breeding values for selection is one of the most important issues for an efficient breeding program. Genetic improvement in livestock species aims to increase sustainability and efficiency of animal products. Pig production centers around the use of crossbred animals (Knol et al., 2016). In the pork industry the systematic breeding started in the 1940's to 60's with the application of quantitative genetics through selection index (Hazel, 1943) and systematic crossbreeding (Dickerson, 1952; 1974). Later, Henderson introduced BLUP (best linear unbiased prediction), which has become the most widely accepted method for genetic evaluations in domestic livestock and provides the fundament for a comprehensive selection (Henderson, 1975). Based on this theory polygenic breeding values with maximum achievable accuracy can be estimated from phenotypic records of an individual itself and/or from records of relatives, which are linked by their relationship (pedigree). This estimation is carried out simultaneously for all individuals in the pedigree. As parents for the next generation, animals with the highest estimated breeding values are chosen. Through the years, the BLUP method evolved in terms of application and a number of extensions like e.g. sire models, sire and maternal grandsire models, reduced animal models or univariate and multivariate models have been introduced (Mrode, 2013).

With the advent of DNA technology and the associated consideration of DNA markers in the selection process, animal breeding has changed substantially. Various methods have been proposed which are briefly outlined.

With the availability of the first genetic molecular markers, the development of a variety of experimental studies to locate genomic regions and Quantitative Trait Loci (QTL) associated with economically interesting traits was encouraged (Mrode, 2013; Ibáñez-Escriche et al., 2014). In the early 1990s marker-assisted selection (MAS) became available and quickly became immense popular at that time (Knol et al., 2016). Expectations that QTL discovery raised in scientific community were not accompanied by a variety of identification of causal polymorphism that could be directly applied in breeding industry (Ibáñez-Escriche et al., 2014). Fernando and Grossman (1989) introduced a method in which marker information was included into conventional BLUP (MA-BLUP) to obtain marker-enhanced breeding values. At that time the number of considered markers – usually microsatellites - was limited up to roughly 300. The incorporation of such markers that are linked to a particular QTL (and phenotype) in a genetic evaluation procedure would increase the accuracy of evaluation and

therefore the expected genetic progress. The benefits of these methods depend largely on the amount of genetic variance explained by the marker (Mrode, 2013; Knol et al., 2016). However, most economically and ecologically important traits are of complex nature and usually influenced by a large number of genes (or QTLs) having only small effects (e.g. Hayes and Goddard, 2001; Knol et al., 2016). As a consequence, the application of gene or marker assisted selection in breeding industry was minor and appeared not as successful as expected (Ibáñez-Escriche et al., 2014; Knol et al., 2016).

In 2001, the general idea of using thousands of markers across the whole genome to predict genetic values instead of looking for particular QTLs with large effects arose and turned out to be a major breakthrough (Meuwissen et al., 2001). With the development of high-throughput SNP (single nucleotide polymorphisms) chips in 2006, large numbers of markers became commercially available (Mrode, 2013; Knol et al., 2016; Weller, 2016). Exploiting linkage disequilibrium (LD) between SNPs and genes that are involved in complex trait variation with the aim to map genes and to predict genomic values became possible (Goddard and Hayes, 2009; Bennewitz et al., 2017). In 2001, Meuwissen et al. (2001) demonstrated in a simulation study how to link all markers to the considered trait (phenotype) simultaneously and that, with a sufficient marker density, genomic values can be properly estimated (Meuwissen et al., 2001). The procedure involves estimating SNP effects simultaneously based on individuals with phenotypic and genotypic records (reference population). Estimated SNP effects are then used to obtain genomic breeding values for genotyped selection candidates that do not yet have phenotypic records (Goddard and Hayes, 2007; Mrode, 2013; Knol et al., 2016). The usage of such genomic breeding values for the selection decision has been referred to as genomic selection (GS).

The implementation of GS has the potential to fundamentally alter the structure of livestock breeding programs caused by e.g. early availability of genotypes of young animals (immediately after birth). Young animals (selection candidates) only need to have marker genotypes and there will be no further need to record phenotypes for the selection which has the potential to save a huge amount of costs at that point of time. The reference population might involve genotyped animals with alternative types of information including single or repeated measures of individual phenotypic performance, information on progeny, estimated breeding values (EBV) from genetic evaluations, or a pooled mixture of more than one of these information sources (Garrick et al., 2009). The reference population could be additionally composed of commercial animals that can be extensively recorded including

traits that cannot be measured from selection candidates such as carcass and meat quality traits, feed consumption or response to disease challenge (Goddard and Hayes, 2007). An appropriate reference population size is required for GS profitability (Meuwissen, 2009), which depends on many parameters such as effective population size, currency of reference population (up to date, a reference population reflecting the actual population structure), genetic architecture of the trait, genome size and SNP chip density (Goddard, 2009). Feasibility of the application of GS is thus breeding scheme and population dependent (Ibáñez-Escriche et al., 2014).

The ultimate method for determining all DNA variation is complete sequencing of the genome which is referred to as “next-generation sequencing” (NGS) or “massively parallel sequencing”. Those high-throughput DNA sequencing methods were developed in the mid to late 1990s and were implemented into commercial DNA sequencers by 2000 (Mrode, 2013). Different to SNP arrays, which rely on already known positions, NGS is based on the re-sequencing of the whole genome. With this procedure it is expected that causative variants (mutations) are present in the whole-genome sequence data and therefore, GS can directly work with causative variants instead of having to rely on LD structure between markers and causative mutations (Meuwissen et al., 2016). Recently, small increases (2-5%) in accuracy of genomic breeding values with sequence data were demonstrated in cattle (Brøndum et al., 2015). However, WGS data is expected to be future’s genotype data. If sequencing costs continue to fall, WGS may become the most effective genotyping method (Gorjanc et al., 2015).

## **Methods of breeding value estimation**

### **Conventional BLUP**

A traditional best linear unbiased prediction (BLUP) animal model as described by Henderson (1975) was used for the prediction of conventional estimated breeding values (EBVs) in this thesis. The basic animal model is as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e},$$

in which  $\mathbf{y}$  represents the vector of observations for the target trait,  $\mathbf{b}$  is the vector of fixed effects,  $\mathbf{a}$  is the vector of random additive genetic effects of all animals which is assumed to be normally distributed with  $\mathbf{a} \sim \mathbf{N}(0, \mathbf{A}\sigma_a^2)$ , and  $\mathbf{e}$  being the vector of residual effects which is

assumed to be normally distributed with  $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ . Matrices  $\mathbf{X}$  and  $\mathbf{Z}$  are incidence matrices relating records to fixed ( $\mathbf{b}$ ) and animal ( $\mathbf{a}$ ) effects, respectively.  $\mathbf{A}$  represents the pedigree-based numerator or additive genetic relationship matrix between individuals,  $\mathbf{I}$  is an identity matrix and  $\sigma_a^2$  and  $\sigma_e^2$  are additive genetic and residual variances, respectively.

In the presented animal model, replacing  $\mathbf{A}$  by a genomic relationship matrix ( $\mathbf{G}$ ) will result in genomic BLUP (GBLUP) (Goddard, 2009; Hayes et al., 2009), while replacing  $\mathbf{A}$  by a mixed pedigree and genomic relationship matrix  $\mathbf{H}$  will lead to single-step genomic BLUP (Legarra et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). The construction of both mentioned matrices will be briefly presented in the following.

In this thesis, conventional BLUP as described above is used to estimate breeding values in Chapter 2, 3 and 4.

### Genomic BLUP

In traditional BLUP, breeding values are estimated using phenotypes and family relationships, which are based on the pedigree of the individuals. In genomic BLUP (GBLUP), genomic breeding values are estimated using quasi-phenotypes and genomic relationships, which are based on genome-wide dense marker data (Meuwissen et al., 2016). In the context of genomic predictions two equivalent systems of predicting genomic breeding values (as shown by Hayes et al., 2009) exist: genomic BLUP and SNP-BLUP (Ridge-Regression (RR)-BLUP). Since SNP-BLUP is not applied in this thesis, only GBLUP is presented. The model behind GBLUP is defined as follows:

$$\mathbf{y} = \mu + \mathbf{W}\mathbf{g} + \mathbf{e},$$

where  $\mathbf{y}$  represents the vector of phenotypes or quasi-phenotypes (e.g. deregressed proofs, DRP) for the target trait,  $\mu$  is the overall mean,  $\mathbf{g}$  is the vector of random genomic effects (DGV), and  $\mathbf{e}$  is the vector of random residual effects.  $\mathbf{W}$  is the corresponding design matrix for the random genomic effects. It is assumed that  $\mathbf{g} \sim N(0, \mathbf{G}_x\sigma_g^2)$  and  $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ , in which  $\mathbf{G}_x$  is the genomic relationship matrix,  $\mathbf{x}$  defines the used G-matrix according to the different approaches and  $\sigma_g^2$  is the corresponding additive-genetic variance.  $\mathbf{I}$  is the corresponding identity matrix with  $\sigma_e^2$  being the residual variance. In GLBUP, all individuals with and without phenotype are taken into account to obtain genomic breeding values directly and simultaneously from MME in one step, which is used in this thesis.

The computational requirements for SNP-BLUP and GBLUP are very different. GBLUP is computationally less demanding than SNP-BLUP as long as the amount of animals is smaller than the number of estimated random effects. For pigs, the number of genotyped animals often is still smaller than the number of markers (SNPs), which makes GBLUP (computationally) preferable compared SNP-BLUP and so GBLUP has become standard procedure (Knol et al., 2016; Meuwissen et al., 2016). However the number of genotyped animals is expected to increase in the future. Especially for cattle, SNP-BLUP might become the method of choice (Meuwissen et al., 2016).

In conventional BLUP the independent variables are phenotypic records of the selection candidates or observations of their relative. In genomic BLUP, the independent variables are quasi-phenotypes, e.g. EBVs estimated from BLUP or deregressed proofs (DRP) as described by Garrick et al. (2009). DRPs are always calculated with removal of parental average effect (Garrick et al., 2009). When EBVs are used as quasi-phenotypes in the analysis, it was assumed that  $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ , with  $\mathbf{I}$  being an identity matrix and  $\sigma_e^2$  being the residual variance. When DRPs were used in the analysis, it was assumed that  $\mathbf{e} \sim N(0, \mathbf{R}\sigma_e^2)$  with

$\text{diag}(\mathbf{R}) = \frac{h^2 \left( \frac{1-r_i^{2*}}{r_i^{2*}} + c \right)}{1-h^2}$ . The reliability of the DRPs for each individual  $i$  ( $r_i^{2*}$ ) was recalculated as described by Garrick et al. (2009). The heritability ( $h^2$ ) corresponds to estimates used in the conventional breeding value estimation. To assess the constant  $c = 1 - \frac{\sigma_g^2}{\sigma_g^2 + \sigma_r^2}$ , some preliminary analyses with ASReml (Gilmour et al., 2009) were performed using the genomic relationship matrix and EBV to obtain estimates of the genetic variance explained by markers ( $\sigma_g^2$ ) and the residual variance ( $\sigma_r^2$ ) (i.e. genetic variance not explained by markers).

In this thesis, mainly DRPs with removal of the parent average (PA) instead of EBVs are used as response variable for genomic predictions for some reasons (Garrick et al., 2009):

- (i) If the parent average is still included in the quasi-phenotypes (e.g. in EBVs) of the reference set and DGVs and EBVs of the validation set are correlated later, predictive ability can easily be overestimated through e.g. double counting.
- (ii) DRPs exclude ancestral information. If both an offspring and its parent are genotyped, the degree of double-counting decreases when using deregressed EBV as the response variable.

- (iii) Animals without individual or progeny information cannot usefully contribute to genomic prediction because they do not add information that would not be already available from their parent's EBV.
- (iv) Using EBVs as the response variable results in double shrinkage of the genomic breeding values, particularly when the reliabilities of the EBV are low. The shrinkage of DGVs is generally smaller using DRPs.

### **Blending to obtain Genomic Enhanced Breeding Values**

Genomic breeding values can be combined with the conventional breeding value that takes no account of genomic information (the polygenic EBV) to obtain the genomically enhanced breeding value (GEBV), a procedure which is called “blending” (VanRaden et al., 2009). To compute those GEBVs, three different breeding values are combined in a selection index and weighted according to their statistical reliability. In this thesis, mainly breeding values (BV) for validation individuals were computed and combined in the blending index as follows: (i) the parental index (PA) of an animal from the BLUP breeding value estimation, (ii) the direct genomic breeding value, which was computed on the basis of the genomic relationship matrix and deregressed proofs (DRP) with removal of the PA (as described in Garrick et al., 2009) and (iii) parental index (PI), which was estimated based on a pedigree relationship matrix using only the genotyped animals and DRPs as pseudo-observations. Subsequently, the variances of estimated breeding values ( $=r_{BV}^2\sigma_g^2$ ) of these three information sources are combined in a selection index to calculate the optimum weighting of each part of the breeding values for each animal resulting in the final GEBV (VanRaden et al., 2009) of an animal for which also the respective reliability can be calculated.

In this thesis, methods including conventional BLUP, GBLUP and blending are referred to as “multi-step method”. This multi-step method mainly consist of three parts: (i) construction of a response variable (e.g. quasi-phenotypes derived from routine BLUP) for genotyped animals that integrate all available phenotypic information, (ii) exploiting the association between response variable and marker information through genomic prediction (genomic BLUP), and (iii) blending genomic predictions with parental average estimated breeding values (Christensen et al., 2012). GBLUP for estimating direct genomic breeding values and blending for estimating genomically enhanced breeding values are used in Chapter 2 and 3.

### Construction of the genomic relationship matrix

The additive genetic relationship matrix uses only pedigree data to compute probabilities that genes are identical by descent (Wright, 1922). Consequently, the expected relationship coefficient (pedigree-based relationship) between two full sib animals is 0.5. The genomic relationship matrix utilizes genomic data to estimate the fraction of total DNA or fraction of alleles at a specific locus (realized relationship coefficient) that two individuals share (VanRaden, 2007). Deviations between expected and realized relationship between full sibs might occur due to the fact that individuals might inherit different alleles from the last generation (parents), which is referred to as the “Mendelian sampling effect”. In contrast to expected values, where all full sibs of a family have the same value as long as no own or progeny information is available, realized (genomic) relationships may differ between full sibs of one family and thus allow selecting within full sib groups based on genomic values. Further, predictions based on genomic relationships may provide more accurate breeding values than pedigree-based ones (VanRaden, 2007).

VanRaden (2008) introduced one of the first and most widely used genomic relationship matrix which will be described as follows:

$$\mathbf{G} = \frac{(\mathbf{M}-\mathbf{P})(\mathbf{M}-\mathbf{P})'}{2 \sum_{j=1}^m \pi_j(1-\pi_j)},$$

where  $\mathbf{M}$  is the marker genotype matrix with genotypes coded 0, 1 and 2 for AA, AB and BB,  $\mathbf{P}$  contains two times the allele frequencies  $\pi_k$  of the B allele at each locus  $j$  such that all entries of column  $j$  of  $\mathbf{P}$  are  $2\pi_j$  and  $m$  is the total number of markers. This standardization (division by  $2 \sum_{j=1}^m \pi_j(1-\pi_j)$ ) makes the pedigree-based relationship and genomic relationship comparable on the same scale (VanRaden, 2008). This genomic relationship matrix is used in Chapter 2 and 3

In order to account for population structure and genetic architecture, different genomic relationship matrices are presented in Chapter 2. For instance, Gengler et al (2007) proposed to use allele frequencies from the base population rather than actual frequencies to scale the genomic relationship matrix (Gengler et al., 2007). Zhou et al. (2014) introduced a genomic relationship matrix that should be weighted according to marker effects and LD phase consistencies (Zhou et al., 2014b). More details regarding their construction can be found in Chapter 2 or corresponding literature.

### Single-step genomic BLUP

In recent years, another approach was introduced by Legarra et al. (2009), Aguilar et al. (2010) and Christensen and Lund (2010) which is referred to as “single-step evaluation” or “single-step genomic BLUP” (ssGBLUP). They propose to do the blending step inside the mixed models equations (MME) system. The key idea is to use all available information such as phenotypes, genomic and pedigree information together in a single model to predict genomic breeding values (ssGEBVS) for all individuals simultaneously.

In a single-step genomic BLUP model, the pedigree-based relationship matrix  $\mathbf{A}$  will be replaced in the conventional BLUP model (described above) with  $\mathbf{H}$ , a mixed pedigree and genomic relationship matrix. It is assumed that  $\mathbf{u}$  ( $\mathbf{a}$  in conventional BLUP) and  $\mathbf{e}$  are normally distributed with  $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$  and  $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ , respectively, where  $\mathbf{u}$  is the vector of single-step genomic breeding values with  $\sigma_u^2$  being the corresponding genetic variance,  $\mathbf{I}$  is an identity matrix and  $\sigma_e^2$  is the corresponding residual variance. The inverse variance-covariance matrix of the genetic effects  $\mathbf{H}^{-1}$  is calculated as a combined relationship matrix as suggested by Aguilar et al. (2010) and Christensen and Lund (2010) and written as follows:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_w^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where  $\mathbf{A}^{-1}$  is the inverse of the pedigree-based relationship matrix  $\mathbf{A}$ ,  $\mathbf{G}^{-1}$  is the inverse of the genomic relationship matrix  $\mathbf{G}$  and  $\mathbf{A}_{22}^{-1}$  is the inverse of the subset of the pedigree-based relationship matrix  $\mathbf{A}$  between genotyped animals only. To make  $\mathbf{G}$  compatible with  $\mathbf{A}$ ,  $\mathbf{G}$  is modified to be on the same scale as  $\mathbf{A}$  (as described by Christensen et al. (2012)). To scale the genomic information and to improve convergence behavior of iterative approaches (Miszta et al., 2013),  $\mathbf{G}_w$  is calculated as follows:

$$\mathbf{G}_w = (\alpha * \mathbf{G} + \beta * \mathbf{A}_{22}),$$

with  $\alpha = 0.95$ ,  $\beta = 0.05$ . For the proportions of  $\alpha$  and  $\beta$ , the default values implemented in the software BLUPF90 were used for breeding value calculation in this thesis (Miszta et al., 2002). Single-step genomic BLUP for estimating single-step genomic breeding values as described above is also referred to as “single-step method” and used in Chapter 3.

Several methods of combining  $\mathbf{G}$  and  $\mathbf{A}$  have been proposed in literature: (i) adapt  $\mathbf{G}$  to  $\mathbf{A}$  (Forni et al., 2011; Christensen, 2012) and conversely (ii)  $\mathbf{A}$  to  $\mathbf{G}$  (Christensen, 2012; Legarra

et al., 2015; Garcia-Baccino et al., 2017). Another possibility to ensure convergence of iterative approaches (Miszta et al., 2010) and to reduce inflation of predictions (Koivula et al., 2015), two scaling factors ( $\tau, \omega$ ) have been introduced to calculate  $\mathbf{H}_{\tau, \omega}^{-1}$  (Miszta et al., 2010; Tsuruta et al., 2011):

$$\mathbf{H}_{\tau, \omega}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tau \mathbf{G}^{-1} - \omega \mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Martini et al. (2018) investigated optimal values of  $\tau$  and  $\omega$  in terms of predictive ability, inflation and iterations up to convergence on a publicly available wheat data set. Value used for  $\tau$  and  $\omega$  in this thesis is 1, which will be referred to as “original single-step genomic BLUP”. Further aspects of combining genomic and pedigree relationship matrix for  $\mathbf{H}^{-1}$ ,  $\mathbf{H}_{\tau, \omega}^{-1}$  and values for  $\tau$  and  $\omega$  will be discussed in detail in Chapter 5.

Due to the fact that  $\mathbf{G}$  must be inverted directly, the size of the dataset to which single-step genomic BLUP could be applied to is limited. To overcome this limitation and to expand the application of single-step genomic BLUP to millions of genotyped animals, the ancestor, proven and young bull algorithm (APY) has been introduced (e.g. Fragomeni et al., 2015). APY is a method based on genomic recursion (Miszta et al., 2014)), in which genomic breeding values of new genotyped animals (young) are conditioned on genomic breeding values of all previous genotyped animals (proven). The direct inversion is required for only a small proportion of  $\mathbf{G}$  composed of relationships among animals treated as “proven” comprising those animals with high accuracies and thus containing most genomic information. Further issues on APY will be discussed in Chapter 5.

## Nonlinear methods

In animal breeding and for the prediction of genomic breeding values, other models such as Bayesian methods (often called the Bayesian alphabet) can be applied. Since nonlinear methods e.g. Bayesian methods are beyond the subject of this thesis, this short paragraph will just give a brief outline over a few approaches.

Bayesian linear regression models can e.g. assume different priors (Gianola, 2013). In BayesA (Meuwissen et al., 2001) each SNP is assumed to be sampled from a distribution with a different variance. In BayesB a certain amount of SNPs is assumed to have no effect and the others have a SNP-specific variance (Meuwissen et al., 2001). In BayesR SNP effects are

supposed to be zero or to come from different normal distributions (Erbe et al., 2012). Further models from the Bayesian alphabet are BayesC or BayesD $\pi$  (Habier et al., 2011).

### Assessment of predictive ability

The accuracy of estimated breeding values is an important parameter in livestock genetic improvement. It is used to calculate response to selection and to express the credibility of individual EBVs (Bijma, 2012). Empirical accuracy of prediction can be measured as the correlation between the true breeding values ( $A$ ) and breeding values ( $\hat{A}$ ) estimated from models:  $\rho = \text{Corr}(A, \hat{A})$ .

The accuracy is an important parameter for two reasons. First, response to selection is proportional to accuracy (e.g. Falconer and Mackay, 1996). Further details can be found in the following chapter on the breeder's equation.. Second, the accuracy reflects the credibility of an individual EBV and relates to the risk that this EBV will change over time when more information becomes available (Bijma, 2012). From this perspective, the accuracy is a measure of the standard error (SE) of prediction of an individual EBV (Henderson, 1975; Meyer, 1989). Accuracy of prediction ( $\rho_i$ ) is routinely available from genetic evaluations for each individual  $i$  can be calculated from the framework of BLUP, i.e.

$$\rho_i = \sqrt{1 - \frac{PEV_i}{\text{var}(A_i)}},$$

where  $PEV_i$  is the prediction error variance of the corresponding BV of the  $i^{\text{th}}$  individual ( $\text{var}(A - \hat{A})$ ), and can be obtained from the diagonal elements of the inverse of the coefficient matrix of MME,  $\text{var}(A_i)$  is the genetic variance calculated from the corresponding genomic model (Tier and Meyer, 2004).

Since the true breeding values are unknown, the correlation between phenotype (or pseudo-observation/quasi-phenotype) and estimated breeding value will be calculated which is referred to as “predictive ability”. To assess the accuracy of prediction in (genomic) BLUP, there are two different approaches used in this thesis: (i) the accuracy of prediction obtained from the framework of mixed model equations as theoretical reliability of a breeding value for an individual (Henderson, 1975,), as described above, and (ii) from observed correlations from cross-validation approaches which became quite popular with the advent of genomic selection.

Several different cross-validation strategies can be used for calculating the predictive ability and different ways of splitting the data sets have been developed. In this thesis, k-fold cross-validation and forward prediction (stratified cross-validation version, e.g. sorted by age) are used to assess the predictive ability for selection candidates. In this procedure, the whole data set is equally and randomly divided in k subsets. Through replication, each subset acts as the validation set once and guarantees that each observation is used for validation exactly one time. The validation set is consequently left out in the learning process. The number of iterations as well as the size of reference and validation sets depends on the chosen factors.

A five-fold ( $k=5$ ) cross-validation with 20 random replications is used. Animals are allocated to five folds completely at random, while in each run 80% of the animals (four folds) are used for calibration (reference population) of the model to predict DGVs of the remaining 20 % of the animals (validation population) so that each fold is used as the validation set once. To avoid sampling bias the whole cross-validation procedure is repeated 20 times.

Forward prediction is used to mimic the real situation in animal breeding, where the youngest animals depicted the selection candidates as parents for the next generation. According to that, the data set is divided into a reference and a validation set by year of birth while older animals belong to the reference and younger animal to the validation population. Since only one such split is possible, the procedure cannot be replicated and thus does not provide any empirical standard errors.

As a measure of predictive ability in both approaches, various correlations between the phenotypes (or quasi-phenotypes) and genomic breeding values (e.g. DGV, GEBV, ssGEBV) are calculated in the validation population in Chapter 2 and 3. For the five-fold cross-validation, the predictive ability is calculated for each run and averaged over folds.

In practical breeding, phenotypes or conventional breeding values are not available for the animals for which genomic breeding values should be predicted (selection candidates). For validation purpose, it is important to assess properties of models and to predict the potential accuracy of genomic prediction for those animals. Hence, cross-validation within the set of genotyped and phenotyped animals has become the procedure of choice.

### Breeder's equation

Breeding means selecting the best male and female animals of a generation to produce progenies that are on average superior to the parents. The accuracy of prediction ( $\rho$ ) is a measure of the correlation between true ( $A$ ) and estimated ( $\hat{A}$ ) breeding value which is important for the response to selection per year ( $\Delta G$ ) for a particular trait (Falconer and Mackay, 1996):

$$\Delta G = \frac{i\rho\sigma_a}{L},$$

where  $i$  represents the intensity of selection in the population,  $\rho$  the accuracy of selection,  $\sigma_a$  the additive genetic standard deviation and  $L$  the generation interval (age of parents at birth of progenies used for breeding). The equation provides information on how the mean value of a trait under selection will change from one year to the next, which is also referred to as “genetic trend”. The genetic trend in pig breeding is discussed more detailed in Chapter 5.

### Genomic selection in pig breeding

Genomic selection was first applied in dairy cattle (VanRaden et al., 2009), where the main purpose is to improve performance of purebred animals. With the development of the 60k SNP array for *Sus scrofa* (Ramos et al., 2009), genomic information became available and can be used to augment classical (conventional) breeding value estimation by genomic breeding value estimation in pigs. Implementation of genomic selection in dairy cattle (e.g. Hayes et al., 2009a; Hayes et al., 2009b; VanRaden et al., 2009) has resulted in increased genetic gain, which has been demonstrated by genetic trend analysis in various countries (Meuwissen et al., 2016). The implementation of genomic selection into pig breeding has not been implemented as consistently as in cattle breeding. Peculiarities like e.g. small nucleus sizes, diversity of breeding goals, pyramid system or small pig breeding organizations made genomic evaluation strategies not straightforward to implement in pigs (Ibáñez-Escriche et al., 2014). However, the potential of GS for pig breeding is known, e.g. advantages for complex traits with low heritability, increased accuracy of breeding values, availability of breeding values at an earlier point in time, reduction of test-matings and therefore cost savings in performance testing (e.g. Simianer, 2009).

In recent years, studies on the potential of GS in pigs e.g. FrOGS<sup>2</sup> for German Landrace or pigGS<sup>3</sup> for German Large White and German Landrace (among others) have been conducted in Germany. Today, most international breeding companies e.g. PIC<sup>4</sup>, Topigs Norsvin<sup>5</sup>, DanBred<sup>6</sup>, Hypor<sup>7</sup> and DanAvl<sup>8</sup> have integrated GS into their breeding programs in order to achieve a higher genetic progress. This development is promoted by the reduction of costs for genotyping. However, due to the limited transparency of the activity of breeding companies it often remains unclear to what extent GS is used, e.g. with regard to the number of genotyped animals and selection candidates, respectively.

In Germany and Switzerland, herdbook organizations have also started to implement GS into their breeding programs for some breeds and traits. For instance, the “Erzeugergemeinschaft und Züchtervereinigung für Zucht- und Hybridzuchtschweine”<sup>9</sup> (EGZH) in Bavaria has developed a single-step procedure for Piétrain (sire line) and German Landrace (dam line) to routinely estimate genomic breeding values. Data from the station test (sire and dam line) is used for breeding value estimations. Additionally for dam lines (Landrace and Large White) and their crosses, data collected on practical piglet farms is also taken into account. German Genetic<sup>10</sup> implements GS for Piétrain in a two-step procedure. First, selection candidates are genotyped using low density markers (n=384 SNPs) and in a second step, imputing procedures are used to increase the number of markers up to 60k SNPs to estimate genomic breeding values (Wellmann et al., 2013). SUISAG<sup>11</sup> calculates based on high density SNP data, genomically enhanced breeding values for young, untested full sibs to select the best full sib for breeding purposes from a family. For sire lines, production performance is the most important breeding goal, for dam lines it is the reproduction performance.

The most important selection step in pig breeding is the selection of elite boars in nucleus herds. The improvement of genetic gain on the nucleus herd has an important impact on the large commercial populations and can make GS economically feasible, given the large influence of elite animals (Simianer, 2009). In general, boar test recording occurs before

---

<sup>2</sup> [http://www.lfl.bayern.de/mam/cms07/itz/dateien/schwein\\_genomische\\_selektion\\_endbericht\\_frogs.pdf](http://www.lfl.bayern.de/mam/cms07/itz/dateien/schwein_genomische_selektion_endbericht_frogs.pdf)

<sup>3</sup> <http://www.fbf-forschung.de/aktuelles/piggs-fbf-mitglieder-gemeinsam-auf-dem-weg-zur-g.html?highlight=piggs>

<sup>4</sup> <http://de.pic.com/>

<sup>5</sup> <https://topignorsvin.de/>

<sup>6</sup> <https://danbred.com/de/>

<sup>7</sup> <https://www.hypor.com/de/>

<sup>8</sup> <http://www.danzucht.com/>

<sup>9</sup> <http://www.lfl.bayern.de/itz/schwein/023973/index.php>

<sup>10</sup> <https://www.german-genetic.de/>

<sup>11</sup> <https://www.suisag.ch/>

selection of elite boars, so that extra gains due to a reduction of generation interval are limited, although 23% reduction in generation interval may be realized by introduction of GS (Meuwissen et al., 2016). However, compared to dairy cattle (generation interval of 6-8 years), the generation interval is already relatively short (roughly 2 years). Implementation of GS in pig breeding is thus mainly focused on traits being invasive (e.g. slaughter quality) and which can neither be recorded on boars (e.g. maternal traits) nor on purebred animals (e.g. crossbred performance). Besides this, it is important to have a relationship between purebred production in very good environment and crossbred performance in harsher environment due to the fact that pork is produced by crossbred animals, whereas purebred animals (elite nucleus herd) are being selected (Meuwissen et al., 2016). Esfandiyari et al. (2015) reported that in practice only 40-70% of the genetic improvement realized in the nucleus herd (e.g. pig grow 100g/day faster) will also result in improved crossbred performance (e.g. pig grow 40-70g/day faster). In this situation, an advantage of the GS model can be the estimation of marker effects from data on the purebred and crossbred individual, which could help to improve the performance of purebred nucleus animals with respect to crossbred performance under commercial environment, respectively. Hence, optimal across breed/purebred genomic selection methods need to be developed (Meuwissen et al., 2016).

### **Potentials for improving prediction accuracy in pig breeding**

#### **Enlarging reference sets**

Especially for small breeds or populations, establishing a reference population with sufficient size to obtain a higher accuracy from genomic prediction than from a simple parental average might be difficult (Thomasen et al., 2012). In order to obtain more phenotypes for the prediction equation, pooling references populations for predictions can be a good strategy. Therefore, genomic predictions can be done in several ways: (i) within breed and population, (ii) within breed and multi-population, and (ii) across breeds or lines.

In pig breeding it has been shown (e.g. Nielsen et al., 2010) that incorporating SNP information into BLUP evaluations lead to an increase in reliability of estimated breeding values for genotyped and non-genotyped animals. Several studies that have investigated the different genomic methods reported increases in observed reliability of prediction due to single-step compared to multi-step models e.g. in cattle (Koivula et al., 2012; Guarini et al., 2018) or in pigs (Chapter 3). Single-step model may also account for pre-selection of young

genotyped animals to avoid bias in GEBV, for which multi-steps methods do not account for (Vitezica et al., 2011).

One strategy of enlarging the reference population is the combination of different populations of a breed or different breeds. Due to possible differences between the populations there are at least three factors known that can influence the value of an individual from another population: differences in (i) LD pattern, (ii) allele frequencies of QTL and SNPs and (iii) level of family relationships. The mentioned factors will be discussed in detail in Chapter 5.

In this thesis, Chapters 3 and 4 mainly deal with the estimation of breeding values within a breed and population. Chapter 3 focuses on multi-subpopulations within a breed, whereas chapter 4 compares different methods of genomic breeding value estimation.

### **Expected potential of multi-subpopulation genomic prediction**

The potential of combining populations to enlarge reference sets is first examined with simulations. Most of those simulation studies (e.g. Ibánñez-Escriche et al., 2009; de Roos et al., 2009) showed that combining populations for genomic predictions led to an increase in prediction accuracy, especially when populations are separated only a few generations ago and have a reasonable SNP density to find consistency in LD pattern between QTL and SNP across populations.

Most studies on real data for multi-subpopulations are conducted on large cattle populations (e.g. Hayes et al., 2009; Brøndum et al., 2011; Lund et al., 2011; Pryce et al., 2011; Harris et al., 2014) and started focusing on combining populations from the same breed across countries. In general, those studies show a higher accuracy of genomic prediction when populations across countries are combined in one reference population than within-country reference populations (e.g. VanRaden et al., 2012; Zhou et al., 2014a).

In recent years evaluations on multi-populations started to increase in pig breeding. Hidalgo et al. (2015) investigated the use of across-breed reference populations (Landrace and Yorkshire) to increase the accuracy of genomic selection and reported limited gains in terms of accuracy. Boré et al. (2018) combined different Yorkshire (Landrace) reference populations from different countries in order to evaluate the feasibility of across-country reference populations for pig genomic selection and received promising results. Song et al. (2017) studied the efficiency of genomic prediction using an admixed reference population

comprising three Yorkshire populations with different genetic backgrounds and discovered that the prediction accuracy was slightly improved through enlarging the reference population by admixing different populations. The enlargement of the reference population within-breed by combining subpopulations from herdbook associations are evaluated in Chapter 2.

### **Single-step**

Another possibility to increase the accuracy of prediction is the usage of different information sources in one step as applied e.g. in single-step genomic BLUP. Song et al. (2017) obtained 23 to 31% higher accuracies from single-step genomic BLUP than from GBLUP (GEBVs) for reproduction traits in pigs. Improvements in accuracy of selection with the single-step method are also reported for pigs by Forni et al. (2011) and Christensen et al. (2012). Single-step genomic BLUP has been investigated in various livestock studies e.g. in cattle (e.g. Aguilar et al., 2010; Lourenco et al., 2015b) or broiler (e.g. Chen et al., 2011; Lourenco et al., 2015a) showing that single-step produces higher prediction accuracies than pedigree-based or multi-step methods. Further aspects concerning the comparison of methods for genomic predictions are presented in Chapter 3 and further discussed in Chapter 5.

### **Fertility traits**

Fertility is the ability to produce offspring. Pig breeding attempts to optimize fertility and to improve the selection of the best animals as parents for the next generation through the development of enhanced traits with reliable recording. The overall purpose is to achieve highest possible number of weaned piglets per sow and year over a long lifetime which is influenced by many different factors, e.g. weight gain, puberty age, first pregnancy, litter size, nursing period, number of born alive/weaned piglets, farrowing interval, successful pregnancy after insemination. In addition, these factors are influenced by the following environmental factors such as management (housing and feeding) and season and maternal effects (environment and genetics). Due to the fact that the return of the pig breeders results from the weaned piglets per sow and year, age at the first estrus, litter size at birth, survival rates of piglets after birth and intervals between litters are important, economic reproductive traits and are of special interest.

Besides litter size (e.g. number of piglets born alive), which is a major component of the selection decision and the most important economic trait, other (alternative) traits such as

mothering ability or piglet survival are gaining importance for a sustainable pig production. The Nebraska selection experiments on litter size in pigs (Johnson et al., 1999) showed increases in number of ovulations, number total born, number live born and number of stillborn piglets, but not in the number of weaned piglets and thus piglet survival decreased. The genetic correlation of litter size with survival is negative in pigs (Knol, 2001).

For accurate breeding value estimation, the combination of different maternal traits in the selection process and the optimization of breeding schemes and knowledge of genetic parameters of reproduction traits is necessary. In Table 1.1 heritability of some maternal (fertility) traits is represented. Evaluations of models for improving (alternative) maternal traits are conducted in Chapter 4.

**Tab. 1.1:** Heritabilities for different fertility traits in dam lines

<b>Trait</b>	<b>Heritability</b>	<b>Reference</b>
<b>Number of piglets born alive</b>	0.05/0.06 (both Landrace)	Alfonso et al. (1997)
	0.09 (Large White)	Arango et al. (2005)
	0.10 (Yorkshire)	Chen et al. (2003)
	0.08 (Landrace)	Chen et al. (2003)
	0.08/0.09 (Landrace)	Hanenberg et al. (2001)
<b>Number of piglets stillborn</b>	0.09 (Large White)	Arango et al. (2005)
	0.02/0.05 (both Landrace)	Hanenberg et al. (2001)
	0.03	Knol (2001)
<b>Number of piglets born total</b>	0.06/0.07 (both Landrace)	Alfonso et al. (1997)
	0.09/0.01 (Landrace)	Hanenberg et al. (2001)
<b>Number of piglets weaned</b>	0.05 (Yorkshire)	Chen et al. (2003)
	0.05 (Landrace)	Chen et al. (2003)
<b>Piglet survival</b>	0.05 (Dam line)	Knol et al. (2002)
<b>Farrowing interval</b>	0.04/0.05 (Dam lines)	Knol et al. (2002)
	0.01/0.03 (Landrace)	Hanenberg et al. (2001)
<b>Mothering ability</b>	0.02/0.03 (Landrace)	Hanenberg et al. (2001)

### **Objectives of this thesis**

The general aim of animal breeding is to improve the performance of future generations by selecting the best animals of the current population. For many years, those animals were identified based on conventional estimated breeding values. At the beginning of this century, Meuwissen et al. (2001) proposed to use thousands of genome-wide markers simultaneously for the prediction of genomic estimated breeding values. The process of calculating genomic estimated breeding values, which are used for selection of the best candidates, is known as genomic prediction and widely used. The accuracy of genomic prediction depends on various factors, for instance, the strength of LD between SNP and QTL and the size of reference animals and the number of animals with known phenotype and genotype, respectively.

The overall objective of this thesis is the evaluation of genomic predictions and the improvement of the prediction accuracy in a practical pig breeding data set through (i) enlarging reference populations or (ii) enhanced methods for genomic predictions. As described above, several conventional and genomic methods for predictions are investigated in this thesis:

**Chapter 2** studies if the use of multi-subpopulation reference sets for small pig populations is a suitable option to enlarge the reference size for predictions and to ascertain if multi-subpopulation genomic prediction is superior compared to within-subpopulation prediction in pigs.

**Chapter 3** shows a comprehensive comparison of different methods for genomic predictions with regard to predictive abilities of single-step and multi-step methods for a lowly heritable trait with data from a practical pig breeding program.

**Chapter 4** presents results of a comprehensive evaluation of different conventional approaches with variable selection of components for estimating the variance components for different fertility traits in practical pig data.

**Chapter 5** provides a general discussion on factors affecting genomic selection and predictive accuracy, perspectives and critical issues of methods.

## REFERENCES

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93(2):743–752. doi:10.3168/jds.2009-2730.
- Alfonso, L., J. L. Noguera, D. Babot, and J. Estany. 1997. Estimates of genetic parameters for litter size at different parities in pigs. *Livest. Sci.* 47(2):149–156. doi:10.1016/S0301-6226(96)01401-7.
- Arango, J., I. Misztal, S. Tsuruta, M. Culbertson, and W. Herring. 2005. Threshold-linear estimation of genetic parameters for farrowing mortality, litter size, and test performance of Large White sows. *J. Anim. Sci.* 83(3):499–506. doi:10.2527/2005.833499x.
- Bennewitz, J., C. Edel, R. Fries, T. H. E. Meuwissen, and R. Wellmann. 2017. Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. *Genet. Sel. Evol.* 49(1). doi:10.1186/s12711-017-0284-7.
- Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations: Accuracy and selection. *J. Anim. Breed. Genet.* 129(5):345–358. doi:10.1111/j.1439-0388.2012.00991.x.
- Boré, R., L. F. Brito, M. Jafarikia, A. Bouquet, L. maignel, B. Sullivan, and F. S. Schenkel. 2018. Genomic data reveals large similarities among Canadian and French maternal pig lines. *Can. J. Anim. Sci.* 98(4):809–817. doi:10.1139/CJAS-2017-0103.
- Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *J. Dairy Sci.* 94(9):4700–4707. doi:10.3168/jds.2010-3765.
- Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J. Dairy Sci.* 98(6):4107–4116. doi:10.3168/jds.2014-9005.
- Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci.* 89(9):2673–2679. doi:10.2527/jas.2010-3555.
- Chen, P., T. J. Baas, J. W. Mabry, K. J. Koehler, and J. C. M. Dekkers. 2003. Genetic parameters and trends for litter traits in US Yorkshire, Duroc, Hampshire, and Landrace pigs. *J. Anim. Sci.* 81(1):46–53. doi:10.2527/2003.81146x.
- Christensen, O. F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet. Sel. Evol.* 44(1):37. doi:10.1186/1297-9686-44-37.
- Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42(2):1–8. doi:10.1186/1297-9686-42-2.
- Christensen, O. F., P. Madsen, B. Nielsen, T. Ostensen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. *animal.* 6(10):1565–1571. doi:10.1017/S1751731112000742.

- Dickerson, G. E. 1952. Inbred lines for heterosis tests? Chapter 21 of *Heterosis*, Iowa State University, Ames, Iowa (USA).
- Dickerson, G. E. 1974. Evaluation and utilization of breed differences. In: *Proceedings of the working symposium on breed evaluation and crossing with farm animals*. Zeist, Germany.
- Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95(7):4114–4129. doi:10.3168/jds.2011-5019.
- Esfandyari, H., A. C. Sørensen, and P. Bijma. 2015. A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.* 47(1):76. doi:10.1186/s12711-015-0155-z.
- Falconer, D. S., and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th edition, Longman, Burnt Mill, Harlow, UK.
- Fernando, R. L., and M. Grossman. 1989. Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* 21(4):467. doi:10.1186/1297-9686-21-4-467.
- Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *GSE.* 43(1). doi:10.1186/1297-9686-43-1.
- Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.* 98(6):4090–4094. doi:10.3168/jds.2014-9125.
- Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. C. Cantet. 2017. Metafounders are related to *F<sub>st</sub>* fixation indices and reduce bias in single-step genomic evaluations. *Genet. Sel. Evol.* 49(1):34. doi:10.1186/s12711-017-0309-2.
- Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* 41(1):55. doi:10.1186/1297-9686-41-55.
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *animal.* 1(1):21. doi:10.1017/S1751731107392628.
- Gianola, D. 2013. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics.* 194(3):573–596. doi:10.1534/genetics.113.151753.
- Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica.* 136(2):245–257. doi:10.1007/s10709-008-9308-0.
- Goddard, M. E., and B. J. Hayes. 2007. Genomic selection. *J. Anim. Breed. Genet.* 124(6):323–330. doi:10.1111/j.1439-0388.2007.00702.x.
- Goddard, M., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10(6):381–391. doi:10.1038/nrg2575.

- Gorjanc, G., M. A. Cleveland, R. D. Houston, and J. M. Hickey. 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet. Sel. Evol.* 47(1):12. doi:10.1186/s12711-015-0102-z.
- Guarini, A. R., D. A. L. Lourenco, L. F. Brito, M. Sargolzaei, C. F. Baes, F. Miglior, I. Misztal, and F. S. Schenkel. 2018. Comparison of genomic predictions for lowly heritable traits using multi-step and single-step genomic best linear unbiased predictor in Holstein cattle. *J. Dairy Sci.* 101(9):8076-8086. doi:10.3168/jds.2017-14193.
- Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics.* 12:186. doi:10.1186/1471-2105-12-186.
- Hanenbergh, E., E. F. Knol, and J. W. M. Merks. 2001. Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. *Livest. Prod. Sci.* 69(2):179–186. doi:10.1016/S0301-6226(00)00258-X.
- Harris, B. L., A. M. Winkelman, and D. E. Johnson. 2014. Across-breed genomic prediction in dairy cattle. In: *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Vancouver, Canada.
- Hayes, B. E. N., and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33(3):209. doi:0.1186/1297-9686-33-3-209.
- Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41(1):51. doi:10.1186/1297-9686-41-51.
- Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92(2):433–443. doi:10.3168/jds.2008-1646.
- Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91(1):47. doi:10.1017/S0016672308009981.
- Hazel, L. N. 1943. The Genetic Basis for Constructing Selection Indexes. *Genetics.* 28(6):476–490.
- Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics.* 31:423–447. doi: 10.2307/2529430.
- Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, and D. J. de Koning. 2015. Accuracy of Predicted Genomic Breeding Values in Purebred and Crossbred Pigs. *G3: Genes, Genomes, Genetics* 5(8):1575–1583. doi:10.1534/g3.115.018119.
- Ibáñez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol.* 41(1):12. doi:10.1186/1297-9686-41-12.
- Ibáñez-Escriche, N., S. Forni, J. L. Noguera, and L. Varona. 2014. Genomic information in pig breeding: Science meets industry needs. *Livest. Sci.* 166:94–100. doi:10.1016/j.livsci.2014.05.020.

- Johnson, R. K., M. K. Nielsen, and D. S. Casey. 1999. Responses in ovulation rate, embryonal survival, and litter traits in swine to 14 generations of selection to increase litter size. *J. Anim. Sci.* 77(3):541–557. doi:10.2527/1999.773541x.
- Knol, E. F. 2001. Genetic aspects of piglet survival. Doctoral thesis, Institute for Pig Genetics and Animal Breeding and Genetics Group, Wageningen University, Wageningen, The Netherlands.
- Knol, E. F., B. J. Ducroa, J. A. M. van Arendonk, and T. J. van der Lende. 2002. Direct, maternal and nurse sow genetic effects on farrowing-, pre-weaning- and total piglet survival. *Livest. Prod. Sci.* 73(2–3):153–164. doi:10.1016/S0301-6226(01)00248-2.
- Knol, E. F., B. Nielsen, and P. W. Knap. 2016. Genomic selection in commercial pig breeding. *Anim. Front.* 6(1):15. doi:10.2527/af.2016-0003.
- Koivula, M., I. Strandén, J. Pösö, G. P. Aamand, and E. A. Mäntysaari. 2015. Single-step genomic evaluation using multitrait random regression model and test-day data. *J. Dairy Sci.* 98(4):2775–2784. doi:10.3168/jds.2014-8975.
- Koivula, M., I. Strandén, G. Su, and E. A. Mäntysaari. 2012. Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J. Dairy Sci.* 95(7):4065–4073. doi:10.3168/jds.2011-4874.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92(9):4656–4663. doi:10.3168/jds.2009-2061.
- Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. *Genetics*. 200(2):455–468. doi:10.1534/genetics.115.177014.
- Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015a. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* 47(1):56. doi:10.1186/s12711-015-0137-1.
- Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and others. 2015b. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J. Anim. Sci.* 93(6):2653–2662. doi:10.2527/jas.2014-8836.
- Lund, M. S., A. P. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbbrandtsen, Z. Liu, R. Reents, and others. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.* 43(1):43. doi:10.1186/1297-9686-43-43.
- Martini, J. W. R., M. F. Schrauf, C. A. Garcia-Baccino, E. C. G. Pimentel, S. Munilla, A. Rogberg-Muñoz, R. J. C. Cantet, C. Reimer, N. Gao, V. Wimmer, and H. Simianer. 2018. The effect of the H–1 scaling factors  $\tau$  and  $\omega$  on the structure of H in the single-step procedure. *Genet. Sel. Evol.* 50(1):16. doi:10.1186/s12711-018-0386-x.
- Meuwissen, T. H. E. 2009. Accuracy of breeding values of “unrelated” individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* 41(1):35. doi:10.1186/1297-9686-41-35.
- Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. *Anim. Front.* 6(1):6–14. doi:10.2527/af.2016-0002.

- Meuwissen, T., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157(4):1819–1829.
- Meyer, K. 1989. Approximate accuracy of genetic evaluation under an animal-model. *Livest. Prod. Sci.* 21:87–100. doi:10.1016/0301-6226(89)90041-9.
- Misztal, I., S. E. Aggrey, and W. M. Muir. 2013. Experiences with a single-step genome evaluation1. *Poult. Sci.* 92(9):2530–2534. doi:10.3382/ps.2012-02739.
- Misztal, I., I. Aguilar, A. Legarra, and T. J. Lawlor. 2010. Choice of parameters for single-step genomic evaluation for type. *J. Dairy Sci.* 93 (Suppl. 1) (2010), p. 533(Abstr.)
- Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* 97(6):3943–3952. doi:10.3168/jds.2013-7752.
- Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, D. H. Lee, and others. 2002. BLUPF90 and related programs (BGF90). In: *Proceedings of the 7th world congress on genetics applied to livestock production*. Montpellier, France.
- Mrode, R. A. 2013. *Linear Models for Prediction of Animal Breeding Values*. 3rd Edition. CAB International, UK.
- Nielsen, B., T. Ostersen, G. Su, O. F. Christensen, and M. Henryon. 2010. Use of genomic SNP information in pig breeding. In: *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*. Vancouver, Canada.
- Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. *J. Dairy Sci.* 94(5):2625–2630. doi:10.3168/jds.2010-3719.
- Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z.-L. Hu, H. H. Kerstens, A. S. Law, H.-J. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. V. Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. M. Groenen. 2009. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLOS ONE*. 4(8):e6524. doi:10.1371/journal.pone.0006524.
- de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. *Genetics*. 183(4):1545–1553. doi:10.1534/genetics.109.104935.
- Simianer, H. 2009. The potential of genomic selection to improve litter size in pig breeding programs. In: *Proceedings 60th Annual meeting of the European Association of Animal Production*, Barcelona, Spain.
- Song, H., J. Zhang, Y. Jiang, H. Gao, S. Tang, S. Mi, F. Yu, Q. Meng, W. Xiao, Q. Zhang, and X. Ding. 2017. Genomic prediction for growth and reproduction traits in pig using an admixed reference population. *J. Anim. Sci.* 95(8):3415–3424. doi:10.2527/jas.2017.1656.
- Thomassen, J. R., B. Guldbrandtsen, G. Su, R. F. Brøndum, and M. S. Lund. 2012. Reliabilities of genomic estimated breeding values in Danish Jersey. *animal*. 6(5):789–796. doi:10.1017/S1751731111002035.

- Tier, B., and K. Meyer. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed. Genet.* 121(2):77–89. doi:10.1111/j.1439-0388.2003.00444.x.
- Tsuruta, S., I. Misztal, I. Aguilar, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci.* 94(8):4198–4204. doi:10.3168/jds.2011-4256.
- VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. *Interbull Bull.* No. 37:33.
- VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91(11):4414–4423. doi:10.3168/jds.2007-0980.
- VanRaden, P. M., K. M. Olson, D. J. Null, M. Sargolzaei, M. Winters, and J. B. van Kaam. 2012. Reliability increases from combining 50,000-and 777,000-marker genotypes from four countries. *Interbull Bull.* No. 46.
- VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92(1):16–24. doi:10.3168/jds.2008-1514.
- Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genet. Res.* 93(5):357–366. doi:10.1017/S001667231100022X.
- Weller, J. I. 2016. *Genomic Selection in Animals*. John Wiley & Sons, Inc.
- Willam, A., and H. Simianer. 2017. *Tierzucht*. 2nd Edition, Eugen Ulmer, Stuttgart.
- Wright, S. 1922. Coefficients of inbreeding and relationship. *Am. Nat.* 56(645):330–338.
- Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. H. E. Meuwissen, M. Svendsen, H. Grove, U. S. Nielsen, and M. S. Lund. 2014a. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *J. Dairy Sci.* 97(7):4485–4496. doi:10.3168/jds.2013-7580.
- Zhou, L., M. S. Lund, Y. Wang, and G. Su. 2014b. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *J. Anim. Breed. Genet.* 131(4):249–257. doi:10.1111/jbg.12089.

## 2<sup>nd</sup> CHAPTER

### **Can multi-subpopulation reference sets improve the genomic predictive ability for pigs?**

A. Fangmann\*, S. Bergfelder-Drüing†, E. Tholen†, H. Simianer\*, M. Erbe\*‡

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

† Institute of Animal Science, Group of Animal Breeding and Husbandry, University of Bonn, 53113 Bonn, Germany

‡ Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, 85586 Poing-Grub, Germany

Published in Journal of Animal Science

Vol. 93, No. 12, p. 5618-5630, November 2015

© 2015 American Society of Animal Science. All rights reserved.

doi: 10.2527/jas.2015-9508

### ABSTRACT

In most countries and for most livestock species, genomic evaluations are obtained from within-breed analyses. To achieve reliable breeding values, however, a sufficient reference sample size is essential. To increase this size, the use of multi-breed reference populations for small populations is considered as a suitable option in other species. Over decades the separate breeding work of different pig breeding organizations in Germany has led to stratified subpopulations in the breed German Large White. Due to this fact and the limited number of Large White animals available in each organization, there was a pressing need for ascertaining if multi-subpopulation genomic prediction is superior compared to within-subpopulation prediction in pigs. Direct genomic breeding values (DGVs) were estimated with GBLUP for the trait ‘number of piglets born alive’ using genotype data (Illumina Porcine 60K SNP BeadChip) from 2,053 German Large White animals from five different commercial pig breeding companies. To assess the prediction accuracy of within- and multi-subpopulation reference sets, a random five-fold cross validation with 20 replications was performed. The five subpopulations considered were only slightly differentiated from each other. However, the prediction accuracy of the multi-subpopulations approach was not better than that of the within-subpopulation evaluation, for which the predictive ability was already high. Reference sets composed of closely related multi-subpopulation sets performed better than sets of distantly related subpopulations but not better than the within-subpopulation approach. Despite the low differentiation of the five subpopulations the genetic connectedness between these different subpopulations seems to be too small to improve the prediction accuracy by applying multi-subpopulation reference sets. Consequently, resources should be used for enlarging the reference population within subpopulation, e.g. by adding genotyped females.

**Key words:** genomic selection, multi-subpopulation, pig, predictive ability

## INTRODUCTION

With genomic prediction an individual breeding value can be predicted from genomic markers, using a prediction equation derived in a reference population with known phenotypic and genomic information. To achieve reliable breeding values, a reference sample of reasonable size is essential. If not enough reference individuals are available within a specific breed, one possible way to overcome this problem is to train the model in a breed with a large reference population and use the estimated marker effects to predict genomic breeding values in other breeds (Harris et al., 2008; Hayes et al., 2009). A more promising solution is combining different breeds within one species to a large reference population, a so called multi-breed reference, which might capture most of the genetic variants segregating within and across breeds (Hayes et al., 2009; Pryce et al., 2011) when estimating the marker effects.

In pig breeding, different breeding organizations generally have closed breeding populations of limited size which are only loosely genetically linked. In this study we will use the general idea of combining different groups of animals to form a reference population when a specific group lacks reference individuals. However, the considered structure is not multi-breed, but multi-subpopulation within a breed, which has not been studied extensively in the literature so far. Differing breeding programs of various pig breeding organizations in Germany, Austria and Switzerland have led to stratified subpopulations in the breed German Large White. Due to this fact and the limited number of animals available in each organization, the following research objectives will be addressed: 1. Assessment of prediction ability of genomic breeding values when using single- and multi-subpopulation reference sets within a breed. 2. Comparison of different genomic relationship matrices, partly accounting for subpopulation structure, to assess their usefulness for multi-subpopulation approaches.

## MATERIALS AND METHODS

### Data

Conventional breeding values and genotypes were available from boars and sows of the breed German Large White from five commercial pig breeding organizations from Germany, Austria and Switzerland (also termed subpopulations in the following). The selection of genotyping was done separately by each commercial breeding organization. The trait under consideration was ‘number of piglets born alive’ (NBA), for which a conventional estimated breeding value (EBV) with the corresponding reliability was available for each individual.

Since our study is based on real data provided by different commercial pig organizations and no pedigree links between individuals of different organizations exist, we did not run combined conventional breeding value estimation with individuals of all organizations. Organization 2, 3 and 5 only provided EBVs with corresponding reliabilities for genotyped animals, but no information about raw phenotypes or corresponding parents. For organization 1 and 4, raw phenotypes were available and for genotyped animals EBVs were calculated with a repeatability BLUP-model. The corresponding reliabilities were computed with the software package WOMBAT (Meyer, 2007). Organization 3 and 5 also used a repeatability BLUP-model to calculate EBVs, whereas the corresponding reliabilities were approximated using the method of Tier and Meyer (2004) in organization 3 and using the method of Graser and Tier (1997) in organization 5, respectively. Organization 2 calculated EBVs using a multiple-trait BLUP-model and reliabilities were approximated using the method of Misztal and Wiggans (1988).

In total, 2,365 animals were genotyped with the Illumina Porcine 60K SNP BeadChip (Illumina, Inc., San Diego, CA). Based on a principal component analysis (PCA) calculated on genomic data, 168 animals were removed from the data set due to incorrect labeling of those animals in a first quality step. Markers with unknown positions (7,381) or not called in all animals (781) and monomorphic markers (2,896) were also removed from the data set.

Quality control was performed simultaneously for all individuals of all subpopulations using the software PLINK (Purcell et al., 2007). Quality control involved excluding SNPs based on call rate ( $< 97\%$ ) and frequency of occurrence of the minor allele, which should be observed at least ten times in order to exclude genotyping errors. Accordingly, 2'862 SNPs were removed from the data set. After quality control, 46'064 SNPs on 18 autosomes remained. Individuals with low call rate ( $< 98\%$ ) were removed as well, leaving a data set of 2,053 individuals from 5 subpopulations for further analyses (see Table 2.1 for details). Since no major imputing differences occurred across all subpopulations compared to results within subpopulation, missing genotypes of the remaining animals were imputed chromosome-wise across all subpopulations using BEAGLE (Browning and Browning, 2007).

**Table 2.1.** Total number of animals and number of animals in reference and validation sets for the forward prediction. The reference set for each subpopulation comprised animals with a year of birth before 2010 and the validation set contained animals born in 2010 and 2011, which represented the youngest animals in the data set.

Subpopulation	Total number of animals	Born between	Reference set Born before 2010	Validation set Born in 2010/2011
1	187	2002 – 2011	100	87
2	140	1997 – 2011	122	18
3	155	2001 – 2011	120	35
4	821	1993 – 2011	564	257
5	540	2002 – 2011	295	245

### GBLUP model

EBVs for the trait NBA were used as quasi-phenotypes in the analyses. It was also tested how the usage of deregressed proofs (DRPs) (Garrick et al., 2009) instead of EBVs affected the predictive ability of the DGVs. The estimation of DGVs was performed using a GBLUP animal model and variance components were estimated using the software ASReml (Gilmour et al., 2009). The following model was used:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{g} + \mathbf{e} \quad (1)$$

where  $\mathbf{y}$  was the vector of EBVs or DRPs for NBA,  $\mathbf{X}$  was a design matrix of fixed effects,  $\mathbf{b}$  was a vector containing the fixed effects. In case of within-subpopulation scenarios, the only fixed effect was an overall mean, while a general mean and a fixed effect for subpopulation was modelled for the multi-subpopulation scenarios.  $\mathbf{W}$  was a design matrix for the random genomic effects,  $\mathbf{g}$  was a vector of random genomic effects (DGV) and  $\mathbf{e}$  was a vector of random residual effects. It was assumed that  $\mathbf{g} \sim N(0, \mathbf{G}_x\sigma_g^2)$ , where  $\mathbf{G}_x$  was a genomic relationship matrix, where  $x$  defined the used G-matrix according to the different approaches and  $\sigma_g^2$  was the corresponding additive-genetic variance. As one option, the G-matrix suggested by VanRaden (2007) was used ( $x = VR$ ), with  $\mathbf{G}_{VR}$  calculated as

$$\mathbf{G}_{VR} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2 \sum_{k=1}^n \pi_k (1 - \pi_k)} \quad (2)$$

where  $\mathbf{M}$  was the marker genotype matrix with genotypes coded 0, 1 and 2,  $\mathbf{P}$  contained two times the allele frequencies  $\pi_k$  of the second allele at each locus  $k$  such that all entries of column  $k$  of  $\mathbf{P}$  were  $2\pi_k$  and  $n$  was the total number of markers.

When EBVs were used as quasi-phenotypes, it was assumed that  $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ , with  $\mathbf{I}$  being an identity matrix and  $\sigma_e^2$  being the residual variance. The calculation of the DRPs (including the removal of the parent average effect (PA) for subpopulation 1 and 4) was performed as described by Garrick et al. (2009) in which the residual variance received the weight:

$$\text{diag}(\mathbf{R}) = \frac{h^2 \left( \frac{1 - r_i^{2*}}{r_i^{2*}} + c \right)}{1 - h^2} \quad (3)$$

so that it was assumed that  $\mathbf{e} \sim N(0, \mathbf{R}\sigma_e^2)$ . The reliability of the DRPs for each individual  $i$  ( $r_i^{2*}$ ) was recalculated as described by Garrick et al. (2009). The heritability ( $h^2$ ) corresponded to estimates used in the conventional breeding value estimation. To assess the constant  $c = 1 - \frac{\sigma_g^2}{\sigma_g^2 + \sigma_r^2}$  we ran some preliminary analyses with ASReml (Gilmour et al., 2009) using the genomic relationship matrix (VanRaden, 2007) and EBV as quasi-phenotype to obtain estimates of the genetic variance explained by markers ( $\sigma_g^2$ ) and the residual variance ( $\sigma_r^2$ ) (i.e. genetic variance not explained by markers). Based on these results we obtained  $c = 0.12$ , which we used in the further analyses.

Another genomic relationship matrix was calculated based on VanRaden (2007), but using allele frequencies from the base generation of the respective subpopulation following Gengler et al. (2007) rather than actual frequencies in the whole sample. For scaling of the  $\mathbf{G}_G$ -matrix (Gengler et al., 2007), the so-called base allele frequencies of the founder animals were calculated for all markers within the subpopulation. For the calculation, 276 founder animals for subpopulation 1 born before 1991 and 809 founder animals for subpopulation 4 born between 1976 and 1985, respectively, were used. Since no year of birth was available for the other three subpopulations, we decided to take the individuals with unknown parents and their direct offspring as base animals. To calculate the respective base allele frequency, we used 515, 490, and 407 founder animals for subpopulation 2, 3, and 5, respectively.  $\mathbf{G}_G$  was calculated analogously to equation (2) but with  $\mathbf{P}$  containing base allele frequencies different in each subpopulation rather than actual frequencies, whereas  $2 \sum_{k=1}^n \pi_k (1 - \pi_k)$  was calculated with  $\pi_k$  being the average frequency over all subpopulations considered.

### **G-matrices weighted by marker effects and LD phase consistencies ( $\mathbf{G}_Z$ )**

In addition, another genomic relationship matrix, termed  $\mathbf{G}_Z$  in the following, was calculated according to Zhou et al. (2014). Within and between subpopulation blocks of  $\mathbf{G}_Z$  were

calculated differently based on marker effects and LD phase consistencies to weight  $\mathbf{G}$ .  $\mathbf{G}_Z$  can only be calculated for a maximum of two stratified groups (Zhou et al., 2014). For the within-subpopulation blocks (termed  $\mathbf{G}_{Zw\_i}$ ), the following formula was used:

$$\mathbf{G}_{Zw\_i} = \frac{\sum_{k=1}^n w_{ik} \mathbf{M}_{iw} \mathbf{M}'_{iw}}{\sum_{k=1}^n 2\pi_{ik} (1 - \pi_{ik})} \quad (4)$$

where  $i$  refers to the subpopulation ( $i=1,2$ ),  $\mathbf{M}_{iw}$  was the marker genotype matrix with  $0 - 2\pi_{ik}$ ,  $1 - 2\pi_{ik}$ , and  $2 - 2\pi_{ik}$ , respectively, and  $\pi_{ik}$  was the observed allele frequency at locus  $k$  in subpopulation  $i$ . For within-subpopulation blocks, the weight  $w_{ik}$  was calculated as  $w_{ik} = \alpha'_{ik}{}^2$ , where  $\alpha'_{ik} = \pm \sqrt{\alpha_{ik}^2 / \alpha^2}$  was the scaled marker effect for marker  $k$  for the trait within subpopulation  $i$  and the sign was assigned according to the original sign of the marker effect  $\alpha_{ik}$  (Zhou et al., 2014). The marker effects were extracted from the GBLUP solutions of the reference population in the respective run.

For between-subpopulation blocks, the weights were calculated as  $w_{k12} = \text{cor}_{LD} |\alpha'_{1,k} \alpha'_{2,k}|$ , where  $\text{cor}_{LD}$  describes the LD phase consistency and  $\alpha'_{1,k}$  and  $\alpha'_{2,k}$  were scaled marker effects of marker  $k$  on the trait of subpopulation 1 and subpopulation 2, respectively. For the calculation of the phase consistency SNPs were pooled in intervals of 12 markers and the correlation  $\text{cor}_{LD}$  was calculated for each interval (Zhou et al., 2014). To calculate between subpopulations blocks ( $\mathbf{G}_{Zb}$ ), the following formula was used:

$$\mathbf{G}_{Zb} = \frac{\sum_{k=1}^n \mathbf{M}_b \mathbf{M}'_b w_{k12}}{2\sqrt{\sum_{k=1}^n \pi_{1,k} (1 - \pi_{1,k}) \sum_{k=1}^n \pi_{2,k} (1 - \pi_{2,k})}} \quad (5)$$

where  $\mathbf{M}_b$  was the marker genotype matrix connecting individuals from subpopulations 1 and 2.

### Assessment of prediction accuracy with five-fold-cross validation

Based on the results of a PCA and calculated  $F_{ST}$  values (Wright, 1943) between subpopulations, different subpopulations were combined to investigate the effect of a multi-subpopulation reference population on the predictive ability of DGVs. Here, closely and distantly related subpopulations, respectively, were combined to form several reference populations. For details see Table 2.2.

**Table 2.2.** Close and distant reference populations for subpopulation 1 and 4.

Subpopulation	Close reference populations	Distant reference populations
1	1 + 2	1 + 4 + 5
1	1 + 3	1 + 2 + 3 + 5
1	1 + 2 + 3	all
1	1 + 2 + 4	-
4	4 + 2	4 + 5
4	4 + 2 + 1	4 + 5 + 1
4	-	all

To test the predictability of within- and multi-subpopulation reference sets, a random five-fold cross validation was performed. In the within-subpopulation scenarios, animals were allocated to the five folds completely at random. For multi-subpopulation scenarios, individuals were distributed randomly to folds subpopulation-wise to guarantee that the proportion of individuals per subpopulation in the reference sets was the same in all runs. In each run, 80% (four folds) of the animals were used for calibration (reference set) of the model and DGVs of the remaining 20% of the animals (validation set) were predicted such that each fold was used as the validation set once. As a measure of predictive ability, the correlation between the quasi-phenotypes (EBVs or DRPs) and DGVs was calculated in the validation set for each run and averaged over folds. To avoid sampling bias the whole cross-validation procedure was repeated 20 times. A paired t-test was used to test the difference of the predictive ability between within-subpopulation and multi-subpopulation reference sets. In case of the scenarios with  $\mathbf{G}_Z$ , marker effects were recalculated in each run for each replicate using only individuals from the reference set.

Since organizations 1 and 4 provided raw phenotypes for a genomic prediction, only individuals of these subpopulations were used as validation individuals in multi-subpopulation runs throughout this study. The predictive ability was calculated only with those individuals in the validation set belonging to subpopulation 1 or 4, respectively.

Furthermore, subsets were constructed to assess whether sows of the same subpopulation or boars and sows of another subpopulation contributed more heavily to the predictive ability of DGVs of subpopulation 1. For details see Table 2.3.

**Table 2.3.** Predictive ability of the DGVs for subpopulation 1 with  $G_{VR}$ , expressed as correlation between EBV and DGV, using a five-fold cross validation with 20 replicates and different reference populations. To assess whether sows of the same subpopulation or another subpopulation contributed more to the predictive ability of DGVs, different subpopulations were combined. For the displayed predictability of the DGVs, the average ( $\pm$  standard error) over the 20 replicates of each reference population was calculated.

Reference population				Total number of animals	Predictive ability of DGVs
Subpopulation 1		Subpopulation 2	Subpopulation 3		
♂	♀	♂/♀	♂/♀		
192				192	0.63 ± 0.005
192	146			338	0.76 ± 0.002
192		140		332	0.63 ± 0.005
192	146	140		478	0.77 ± 0.002
192		140	155	487	0.62 ± 0.005
192	146	140	155	633	0.76 ± 0.002

### Forward Prediction

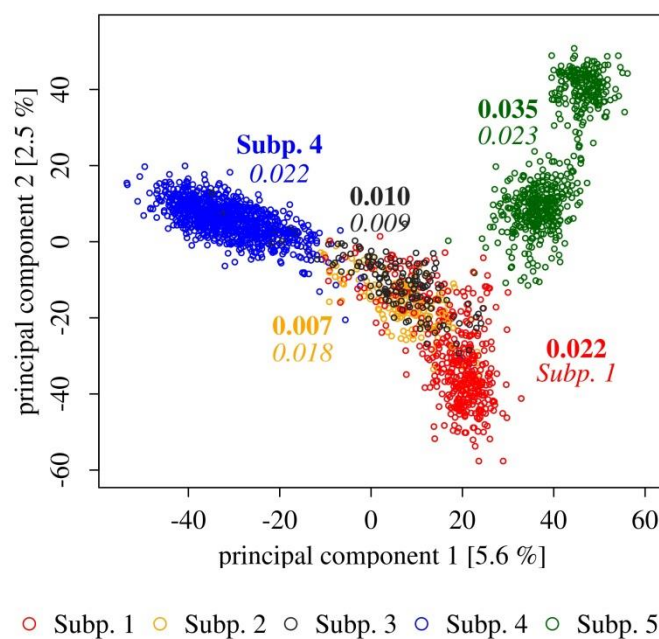
Additional to the random five-fold cross validation, the data was divided into reference and validation set by year of birth of the animals, defined as forward prediction. The reference set contained animals born before 2010 whereas the validation set was composed of the youngest animals born in 2010 and 2011. For this approach, EBVs for the reference populations 1 and 4 were calculated based on available information until 2009, so that no information of validation individuals was used to estimate EBVs of the reference individuals. The number of animals used as reference and validation set, respectively, is shown in Table 2.1. Throughout this paper, the predictive ability is always defined as the correlation between the quasi-phenotypes EBVs or DRPs and DGVs in the validation set.

## RESULTS

### Stratification of the subpopulations

When different subpopulations are to be used in a multi-subpopulation reference set, it is important to analyze the population structure first. To visualize the relationship between the five subpopulations, a plot of the first two principal components for all subpopulations is presented in Fig. 2.1. In addition, pairwise  $F_{ST}$  values of each population to population 1 and 4

are given in Fig. 2.1. The first and second principal components explain 5.6% and 2.5% of the total variance, respectively. In general, the  $F_{ST}$  values turned out to be very small and ranged from 0.010 to 0.035 for subpopulation 4 with all other subpopulations (Fig. 2.1; bold colored numbers) and from 0.007 to 0.023 for subpopulation 1 with all other subpopulations (Fig. 2.1; italic colored numbers), respectively. Subpopulation 1 was closely related to subpopulation 2 and 3 and less related to subpopulation 5, but  $F_{ST}$  values were also small for distantly related subpopulations. As with subpopulation 1, subpopulation 4 was more closely related to subpopulations 2 and 3 and seemed to be more distant from subpopulation 1 and 5, respectively. For subpopulation 5, stratification into two subgroups was observed.



**Figure 2.1.** Principal components analysis for Large White subpopulations and  $F_{ST}$  values between subpopulation 1 and 4 and all other subpopulations, respectively. The bold colored numbers represent the  $F_{ST}$  values between subpopulation 4 and another subpopulation, the italic colored numbers between subpopulation 1 and another subpopulation.

One goal of this study was to verify if a multi-subpopulation reference set can help to overcome the problem of limited genomic prediction accuracy due to small reference sets within the breeding populations of different pig breeding organizations. The best basis for a two-step genomic breeding value prediction would thus be a common conventional breeding value scheme among all organizations which was however not possible in this study. The present study was conducted using real data provided by five commercial pig breeding companies. Only organization 2 and 4 provided raw phenotypes for ‘number of piglets born alive’ such that EBVs could be estimated with the same model assumptions, whereas the

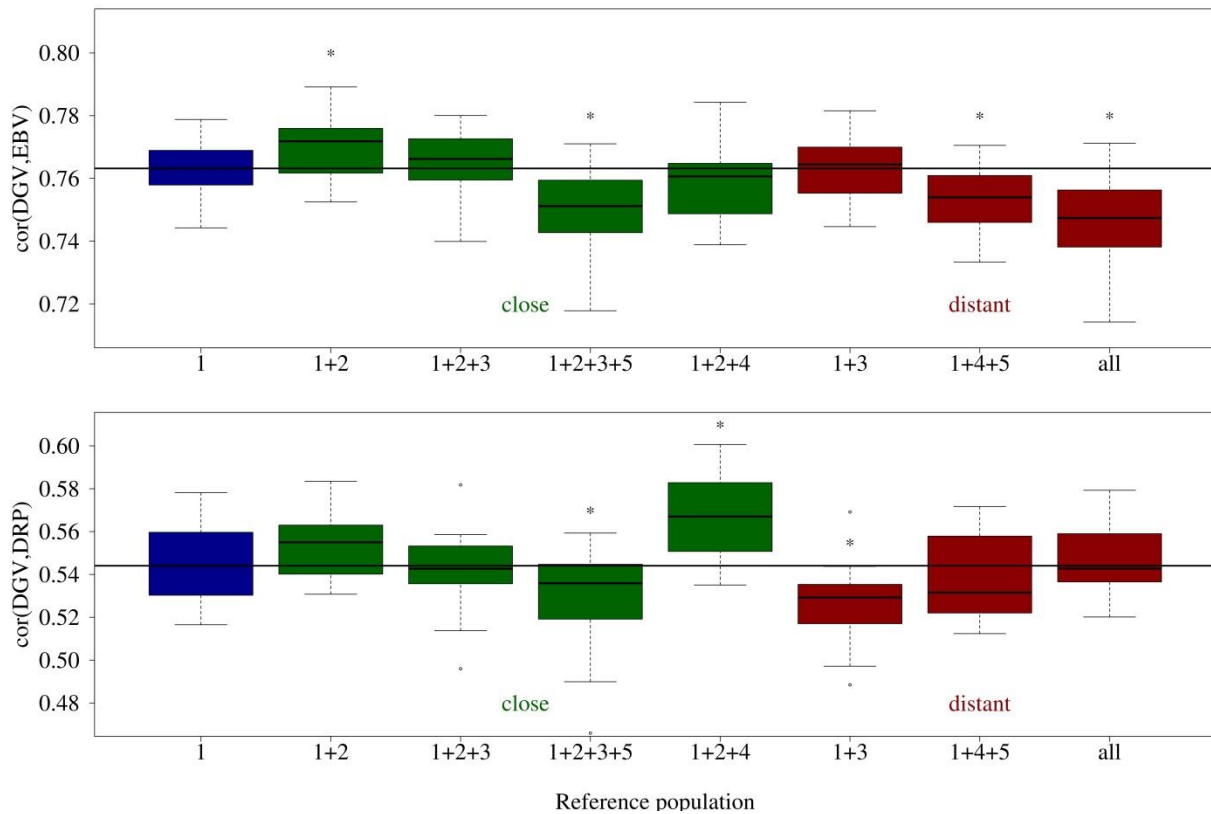
other organizations only provided calculated EBVs with the corresponding reliabilities (for descriptive statistics see Appendix 2.1).

As no pedigree links existed between individuals of different organizations in the data set used, disconnected subsets would nevertheless exist in a joint conventional estimation. We further checked the genomic relationships coefficients between individuals of the different subpopulations to inspect if substantial relationships exist that would have remained undetected with pedigree analyses. Appendix 2.2 shows that genomic relationships were relatively small on average with only a few exceptions, thus confirming that there was hardly any genetic exchange between subpopulations. Thus we do not expect that a joint conventional breeding value estimation, if it were possible, would have provided more accurate or less biased conventional breeding values.

### **Assessment of prediction accuracy**

To evaluate the predictive ability of DGVs, a five-fold cross validation with a quasi-phenotype for different reference populations with individuals of subpopulation 1 as validation population was performed (Fig. 2.2). The upper plot shows the results for the quasi-phenotype EBV and the lower plot for DRP. Within subpopulation 1, the recorded predictive ability with EBV was 0.77. Adding closely related subpopulations to subpopulation 1 as multi-subpopulation reference population, the predictive ability remained constant. Adding distantly related subpopulations to subpopulation 1 resulted in a slight decrease of the predictive ability for subpopulation 1. The difference between within- and multi-subpopulation reference sets was significant ( $P < 0.05$ ) for two closely related reference populations as well as two distantly related reference populations.

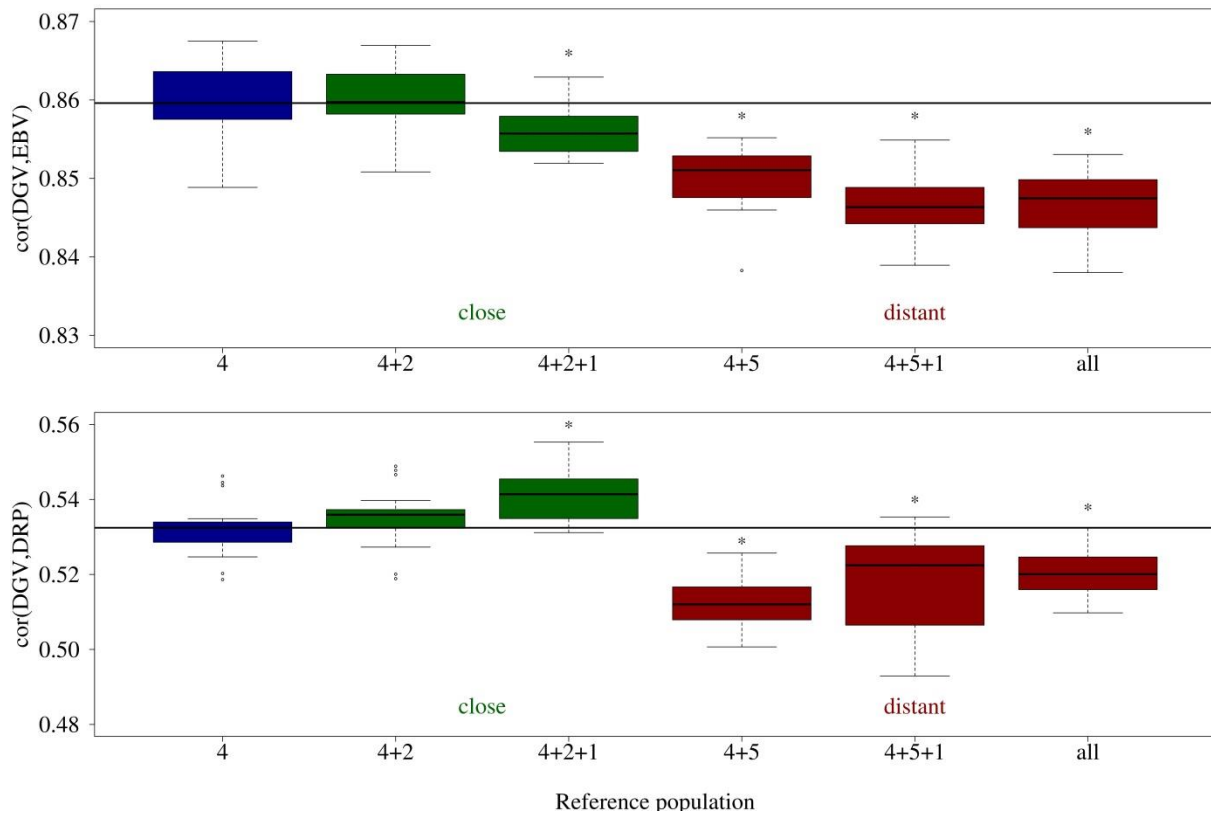
Within subpopulation 1, the recorded predictive ability with DRP was 0.54 (lower plot). Using a multi-subpopulation reference, the predictive ability slightly increased to 0.56 and significantly to 0.57 by adding subpopulation 2 and subpopulations 2 and 4, respectively. By adding subpopulation 3 to the reference population, the predictability decreased slightly. For distantly related multi-subpopulations, the correlation between DGVs and DRPs declined to 0.53 (significantly), 0.54 and 0.54 (Fig. 2.2, lower plot).



**Figure 2.2.** Predictive ability of the DGV prediction for subpopulation 1 using the G-matrix introduced by VanRaden (2007), expressed as correlation between DGV and EBV (top) and DGV and DRP (bottom), using a five-fold cross validation with 20 replicates for different Large White reference populations. A paired t-test was used to test the difference between within-subpopulation and multi-subpopulation reference sets ( $P < 0.05$ ). Blue = within-subpopulation 1 approach, green = closely related reference populations, red = distantly related reference population.

The predictive ability of DGVs for subpopulation 4 using a five-fold cross validation with quasi-phenotype EBV and DRP is displayed in Fig. 2.3, respectively. Within subpopulation 4, the correlation between DGVs and EBVs was 0.86 (upper plot). Adding closely related subpopulations resulted in a predictive ability of 0.86 by combining subpopulations 4 and 2 as well as combining subpopulations 4, 2 and 1 as the multi-subpopulation reference set, respectively. Using distantly related multi-subpopulation reference sets resulted in a significant decline in predictability for subpopulation 4 of around 0.01 for all three distant reference populations. Within subpopulation 4, the predictive ability of DGVs with DRP (lower plot) was 0.53. Adding closely related subpopulations led to a slight increase in predictive ability for subpopulation 4 of 0.03 with the multi-subpopulations 4 and 2 and 0.08 with 4, 2 and 1, respectively. A decrease in the predictability for subpopulation 4 of 0.02 occurred with the inclusion of distantly related subpopulations to the multi-subpopulation

reference set. Except for the reference subpopulations 4 and 2, all differences between within- and multi-subpopulation reference set for both plots were significant ( $P < 0.05$ ).

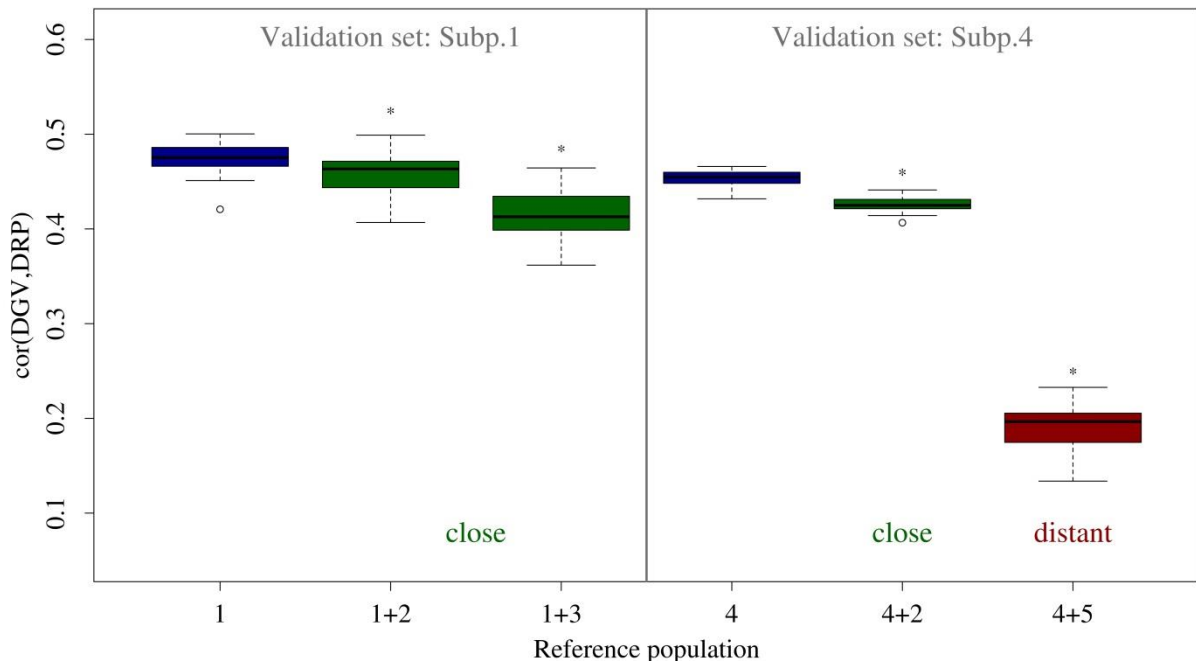


**Figure 2.3.** Predictive ability of the DGV prediction for subpopulation 4 using the G-matrix introduced by VanRaden (2007), expressed as correlation between DGV and EBV (top) and DGV and DRP (bottom), using a five-fold cross validation with 20 repetitions for different Large White reference populations. A paired t-test was used to test the difference between within-subpopulation and multi-subpopulation reference sets ( $P < 0.05$ ). Blue = within-subpopulation 4 approach, green = closely related reference populations, red = distantly related reference population.

Results for the predictive ability of DGVs with  $\mathbf{G}_Z$  and with quasi-phenotype DRP are displayed in Fig. 2.4 for subpopulation 1 (left plot) and 4 (right plot). Within subpopulation 1, the correlation between DGVs and DRPs reached 0.47 and decreased significantly by adding closely related subpopulations. Within subpopulation 4, the predictive ability of DGVs was 0.49 and significantly decreased to 0.45 for closely and 0.2 for distantly related multi-subpopulations as reference populations.

Comparing the five-fold cross validation results of subpopulation 1 with  $\mathbf{G}_{VR}$  (Fig. 2.2) to the results with  $\mathbf{G}_Z$  (Fig. 2.4), the variation in predictive ability for the respective reference set for the DGVs seemed to be the same, likewise for subpopulation 4 (Fig. 2.3). The level of the

predictive ability of the DGV prediction with  $G_Z$  was slightly lower than the level with  $G_{VR}$  for both within subpopulation approaches.



**Figure 2.4.** Predictive ability of the DGV prediction for subpopulation 1 (left) and 4 (right) using the G-matrix introduced by Zhou et al. (2014), expressed as correlation between DGV and DRP, using a five-fold cross validation with 20 repetitions for different Large White reference populations. A paired t-test was used to test the difference between within-subpopulation and multi-subpopulation reference sets ( $P < 0.05$ ). Blue = within-subpopulation 1 or 4 approach, green = closely related reference populations, red = distantly related reference populations.

### Assessment of prediction accuracy (subset)

We compared the combination of boars of subpopulation 1 with sows of subpopulation 1 and with another subpopulation (2 or/and 3) to assess the predictive ability of DGVs (Table 2.3). If only male animals of subpopulation 1 were used as the reference set, a predictability of 0.63 was obtained. Adding sows of subpopulation 1 to the reference population (from 0.63 to 0.76) contributed more to the predictive ability of DGVs than adding animals from another subpopulation (0.62 or 0.63).

### Forward Prediction

Predictability of DGVs for the youngest animals of subpopulation 1 and 4 is shown in Table 2.4 for different multi-subpopulation reference sets as well as with  $G_{VR}$ ,  $G_G$  and  $G_Z$  being used to describe the genomic covariance structure. The predictive ability of DGVs of the

youngest animals of subpopulation 1 was relatively high and ranged from 0.42 to 0.51 (EBV) with  $G_{VR}$  and  $G_G$  and between 0.49 and 0.52 (EBV) with  $G_Z$ . By adding close or distant subpopulations to the reference population, the predictability remained constant or decreased. The lowest value was reached for the multi-subpopulation reference set with subpopulations 1, 4 and 5 (0.42) and the greatest value with subpopulations 1 and 2 (0.51), respectively. Using DRPs as the input-variable, the highest value (0.31) in predictive ability was observed when using a multi-subpopulation with all subpopulations, whereas only 0.25 could be reached with subpopulation 1 as the reference population. The predictive abilities of DGVs using  $G_Z$  were slightly better than with  $G_{VR}$  or  $G_G$  in most of the considered scenarios.

The predictive ability of DGVs of subpopulation 4 were high with quasi-phenotype EBVs and ranged from 0.55 to 0.57 with  $G_{VR}$ , from 0.57 to 0.70 with  $G_G$  and from 0.60 to 0.61 with  $G_Z$ . The greatest predictive abilities were observed within subpopulation 4 for the prediction with  $G_G$ , quasi-phenotype EBV and multi-subpopulation reference set 4 and 2 (0.70) as well as 4, 2 and 1 (0.70), respectively. With quasi-phenotype DRP, the greatest predictive ability was achieved by using subpopulations 4 and 2 as a multi-subpopulation reference set and declined from closely to distantly related subpopulations. The lowest value for DRP and  $G_{VR}$  was obtained using subpopulation 4 and 5 as the reference population (0.17). Consequently and for all approaches, the more distantly related the reference population to the validation set the poorer was the predictability of DGVs.

To account for the fact that quasi-phenotypes originated from different runs, we modeled a fixed subpopulation effect when using a multi-subpopulation reference set. For comparison, the predictive ability with quasi-phenotype DRP with and without the fixed effect of the subpopulation (Appendix 2.3) was calculated for subpopulation 1 with two multi-subpopulations (closely and distantly related). For the random five-fold cross-validation as well as the forward prediction approach, no differences in the predictive ability for subpopulation 1 were obtained by treating the two subpopulations as one common population (i.e. no subpopulation effect modeled) and there was hardly any re-ranking of individuals.

Since no parent breeding values were available for subpopulations 2, 3 and 5, we were not able to remove the parental average (PA) for these subpopulations. Thus, the weighting factors used for the residual variance in the genomic prediction model when DRP was used as quasi-phenotype may be biased. To check the influence on our results, we also calculated DRPs without removing of PA (DRP\*) for subpopulations 1 and 4, i.e. EBVs divided by their

reliabilities and respective weights were also calculated with the corresponding EBV reliabilities. Next, the correlation between the DGVs calculated with DRPs (removal of PA) and DGVs calculated with DRPs\* (without removal of PA) were computed for all animals being in the reference set and for those animals being in the respective validation set (Appendix 2.4). Except for two values, all correlations were above 0.9, suggesting that the DGVs were highly correlated and there was no strong impact of removing or not removing the PA on DGV prediction.

**Table 2.4.** Predictive abilities of DGVs for the youngest animals for subpopulation 1 and 4 and different G matrices (introduced by VanRaden (2007), Gengler et al. (2007) and Zhou et al. (2014)) were presented. To assess the predictability of the DGVs of the youngest animals with multi-subpopulation reference sets, correlations between DGV and EBV as well as DGV and DRP within the subpopulation were calculated. Greatest value in each column and subpopulation are printed in bold.

Reference population	Validation set	$G_{VR}$		$G_G$		$G_Z$	
		Cor(DGV and EBV)	Cor(DGV and DRP)	Cor(DGV and EBV)	Cor(DGV and DRP)	Cor(DGV and EBV)	Cor(DGV and DRP)
Subpopulation 1	Subpopulation 1	0.45	0.25	0.45	0.25	<b>0.52</b>	0.30
Subpopulation 1 and 2	Subpopulation 1	<b>0.51</b>	0.29	<b>0.51</b>	0.29	0.49	0.31
Subpopulation 1 and 3	Subpopulation 1	0.44	0.26	0.44	0.26	<b>0.52</b>	<b>0.32</b>
Subpopulation 1, 2 and 4	Subpopulation 1	0.45	0.30	0.45	<b>0.30</b>	-	-
Subpopulation 1, 4 and 5	Subpopulation 1	0.42	0.27	0.42	0.27	-	-
Subpopulation 1, 2, 3 and 5	Subpopulation 1	0.46	0.26	0.46	0.26	-	-
All Subpopulations	Subpopulation 1	0.44	<b>0.31</b>	0.44	<b>0.30</b>	-	-
Subpopulation 4	Subpopulation 4	<b>0.57</b>	<b>0.19</b>	0.57	0.19	<b>0.61</b>	<b>0.38</b>
Subpopulation 4 and 2	Subpopulation 4	<b>0.57</b>	<b>0.19</b>	<b>0.70</b>	<b>0.30</b>	<b>0.61</b>	0.29
Subpopulation 4, 2 and 1	Subpopulation 4	<b>0.57</b>	<b>0.19</b>	<b>0.70</b>	0.28	-	-
Subpopulation 4 and 5	Subpopulation 4	0.54	0.17	0.68	0.28	0.60	0.30
Subpopulation 4, 5 and 1	Subpopulation 4	0.54	0.18	0.68	0.27	-	-
All Subpopulations	Subpopulation 4	0.55	0.18	0.67	0.28		

## DISCUSSION

### Stratification of the subpopulation

Stratification was observed in our data set, but the recorded  $F_{ST}$  values between subpopulations were small. Thus, it seems reasonable to expect the prediction of more accurate DGVs based on multi-subpopulation rather than within-subpopulation reference sets. Different to results of studies with multi-breed reference populations in cattle (e.g. Pryce et al., 2011; Schrooten et al., 2013), predicting DGVs using a combination of subpopulations within a breed did not show the desired increase in the predictive ability of genomic selection in pigs. In cattle, due to e.g. exchange of semen and a high rate of artificial insemination, genetic links between individuals even from different countries or managed by different breeding associations are often high and clear population stratification into different groups is rare. In pig breeding, different commercial breeding companies have relatively closed breeding stocks and different breeding goals. One common objective is the protection of the 'breeding product'. Usually, there is hardly any exchange of animals or semen so that few genetic links between animals of the same breed but from different breeding organizations exist.

### Assessment of prediction accuracy

Given the small number of individuals in the within-subpopulation reference set, the recorded predictive abilities for the within-subpopulation prediction with EBVs and  $G_{VR}$  were high. When using DRPs as quasi-phenotypes, predictive abilities were generally much lower. If the parent average is still included in the quasi-phenotypes (e.g. in EBVs) of the reference set and DGVs and EBVs of the validation set are correlated later, predictive ability can easily be overestimated through e.g. double counting (Garrick et al., 2009). DGVs will then be a good proxy for the parent average part of the validation individual's EBV while they may still not describe the Mendelian sampling properly. Generally, we are interested in the best prediction of the true breeding value. Thus the predictive ability is more relevant when using DRPs with the PA removed used as quasi-phenotypes than with EBVs. For constructing information for genomic predictions, Garrick et al. (2009) argued for the removal of parent average effects. Furthermore, animals without individual or progeny information cannot usefully contribute to genomic prediction because they do not add information that would not be already available from their parent's EBV (Garrick et al., 2009).

In our data sets predictive abilities of genomic breeding values from the five-fold cross-validation were equally or only slightly better than those of a conventional parent average (results not shown) which is not surprising regarding the small number of genotyped individuals. However, we were most interested in relative differences between predictive ability results obtained with different reference sets.

Hayes et al. (2009) and Pryce et al. (2011) found that for small cattle breeds it is advantageous to form a multi-breed reference by combining them with others. Our results do not confirm this finding. The reference population of subpopulation 1 (n=150) was quite small, but still performed better than using multi-subpopulation reference sets, even if the multi-subpopulation reference sets were quite large (up to n=1,201). The same tendency to decrease was shown for subpopulation 4 which provided a major part of individuals in multiple-subpopulation reference sets tested.

To assess a possible bias of DGV, we performed linear regressions of DRP on DGV within random five-fold cross-validation validation sets. The regression coefficients from random five-fold cross-validation indicated that the prediction bias of DGV slightly increased from 1.10 to 1.46 (decreased from 1.35 to 1.21) when adding subpopulations to the reference set of subpopulation 1 (4). This inconsistency in results partly must be attributed to the limited sample size. Wu et al. (2015) reported that the prediction bias diminished with increasing relationships between training and test animals, which agreed with our results for subpopulation 4 and was also stated by previous studies (e.g. Gao et al., 2013). To prevent increases of bias and decreases of reliability, continuous updates of the reference set with animals from recent generations are required (Habier et al., 2010; Pszczola et al., 2012).

In general, the improvement of the prediction accuracy depends on the amount of useful information that is added on top of the already available information. Calus et al. (2014) stated for layer chickens that multi-line prediction may be effective when lines are closely related and accuracy may be slightly improved. This conclusion is based on a study of two closely related and one unrelated chicken layer line where an increase in accuracy by using closely related multi-subpopulations as reference population was observed. We also tested different levels of relatedness by building multi-subpopulation reference sets with closely and distantly related subpopulations, respectively. The increase of predictive ability by using closely related multi-subpopulations as the reference population was poor and stayed almost at the same level as within-subpopulation in our data set. Furthermore, Zhou et al. (2014)

reported that close genomic relationships were required to increase the accuracies with two-breed prediction. Even though  $F_{ST}$  values were small in our study, average genomic relationships for individuals from subpopulation 1 (4) with individuals from other subpopulations were basically zero or even slightly negative (Appendix 2.2), reflecting the lack of genetic ties (even lower than theoretically expected) and possibly explaining why no advantage was observed for closely related multi-subpopulation references.

Hidalgo et al. (2014) obtained accuracies, expressed as correlation between GEBV and deregressed EBV, of 0.16 and 0.24 for ‘number of piglets born alive’ within a Large White and within a Dutch Landrace population (training and prediction data were subsets from the same population), respectively. The obtained accuracies across the two populations were close to zero when one population was used for training to predict the other population. Hidalgo et al. (2014) mentioned as a possible explanation for the poor multi-population performance that the QTL and markers might not be in the same linkage phase for sow-dependent traits and that the GEBV accuracies may depend on the genetic architecture of the trait and of the population stratification. Furthermore, Hidalgo et al. (2014) mentioned that they observed modest predictive ability for the traits within populations, which were nevertheless lower than expected. They suggested that due to the fact that reproductive traits are rather polygenic and are a result of complex genetic and biological mechanisms, a less accurate prediction in a multi-population setting is to be expected. Moreover, reproductive traits are affected by a large number of QTL (138 QTL for NBA, PigQTLdb, <http://www.animalgenome.org/cgi-bin/QTLdb/SS/index>, June 2015; Bergfelder-Drüing et al., 2015) with low to moderate effects (Rothschild and Ruvinsky, 1998) or may have an incomplete LD between marker and QTL (de Roos et al., 2009; Hidalgo et al., 2014; Zhou et al., 2014). For traits with larger heritability, e.g. growth rate or feed intake, or traits with a smaller number of affecting QTL, an increase of predictive ability of DGVs would generally be possible. Since the genetic links between individuals of different subpopulations were low on average (Appendix 2.2), we do however not expect significantly greater predictive abilities for those traits.

De Roos et al. (2009) showed with a simulation study that a marker might be in high LD with QTL in one population only. This resulted in poor predictions for another population, in addition different QTL might segregate in different populations and the genetic background in the other line might change the effect of a specific QTL allele. With increasing marker density, the probability increases that a marker is in high LD with a QTL. Ibáñez-Escriche et

al. (2009) and de Roos et al. (2009) concluded that multi-population evaluations were preferable to within-population evaluations when the populations were closely related, the marker density was high or the number of phenotypic records was small. Since the marker spacing in our study was modest (49 kbp on average), the reasons mentioned by de Ross et al. (2009) and Hidalgo et al. (2014) might explain the results obtained with the multi-subpopulations as the reference set.

Daetwyler et al. (2012) suggested that exploiting population structure arising from relatedness could increase the accuracy of prediction in some cases, for example when the selection candidates were offspring of the reference population from which the prediction equation was derived. For the multi-subpopulation scenarios, no direct relatedness of the selection candidates and reference animal was given, which might explain the poor performance. Adding sows of the same population to the boar reference population contributed more to the DGV prediction than adding boars and sows from another subpopulation to enlarge the reference set (Table 2.3), which again confirms the importance of relatedness. Even accounting for population stratification by using relationship matrices e.g. based on Zhou et al. (2014), did not yield considerably greater predictive ability of DGVs for reasons mentioned above (for details see ‘Different G-matrix approaches’).

Since GBLUP is strongly based on genomic relationships between the reference and validation set, it might not be an optimal model for genomic predictions across breeds (Zhou et al., 2014). Variable selection models make different assumptions on the contribution of marker effects to the total variance. Some studies applied Bayesian methods for across-breed predictions in dairy and beef cattle (Hayes et al., 2009; Pryce et al., 2011; Erbe et al. 2012; Saatchi and Garrick, 2013) or chicken (Calus et al., 2014). In summary, they found partly a small increase in accuracy when multi-breed or across-breed reference sets were used. However, no consistent increase in accuracy across different traits was obtained. Variable selection methods have indeed some utility, e.g. if a common QTL with large effects (as e.g. DGAT1 for fat yield) is segregating among different subpopulations (Pryce et al., 2011; Saatchi and Garrick, 2013). However, in our study with stratified pig subpopulations, we would not expect significantly better results from variable selection models.

Other possible reasons for not observing any advantage of multi-subpopulation reference sets compared to within-breed models could be genotype by environment interactions and differences in methods of phenotypic and genetic evaluation (Pryce et al., 2011). Another

possibility to predict accurate DGVs within multi-breed models was suggested by Harris et al. (2014). They suggested that the use of high density SNP data to form haplotypes blocks for usage in a multi-breed genomic analyses could improve the accuracy and remove the need for homogeneity of QTL and marker phase and effective size. They suggested that the formation of breed-specific haplotypes blocks that were associated with the QTL alleles segregating within a given breed can explain a larger proportion of the genetic variance within multi-breed models.

### **Forward prediction**

In practice we are interested in DGVs with a high accuracy for the youngest animals, which at the time of selection do not have performance values of their own or for their progeny. Since the reference set of subpopulation 1 was substantially smaller in the forward prediction scenario than in the five-fold cross-validation scenarios, the lower predictive ability observed in the forward prediction scheme may be partly explained by this. Adding information from closely related multi-subpopulations, especially subpopulation 2, slightly increased the accuracy, whereas adding distantly related reference populations to the reference set led to a slight decrease of the predictive ability.

### **Different G-matrix approaches**

The results of the cross validation and forward prediction indicated that for multi-subpopulation sets G should be scaled in a different way to capture the structure and relationship for different subpopulations as also stated by Hayes et al. (2009) and Calus et al. (2014). Therefore, the main interest was to predict DGVs more realistically through an improved adjustment of G. Using a genomic relationship matrix based on VanRaden (2007) and treating all subpopulations in the same manner largely ignores a possible substructure, reflected by different allele frequencies. Vallée et al. (2014) predicted the accuracy using crossbred performances within and across beef cattle breeds and indicated that the loss of accuracy using admixed or multi-population sets might be due to different allele frequencies between populations. They concluded that building G with breed-specific allele frequencies might lead to improved accuracies.

To test the impact of allele frequencies when accounting for substructures, we used base allele frequencies of each subpopulation according to Gengler et al. (2007) instead of actual

frequencies derived from multi-subpopulation sets. Another way to adjust for subpopulation-specific allele frequencies is to calculate the genomic relationship matrix introduced by VanRaden (2007) with allele frequencies present in the respective population. This was done exemplary for the multi-subpopulations 4 and 2, and we used the mean allele frequency of both subpopulations for standardization.

Predictive ability of the random five-fold cross validation and forward prediction did not increase for subpopulation 1 with base allele frequencies. Predictive ability with EBVs used as quasi-phenotypes was +0.06 greater than with the genomic relationship matrix of VanRaden (2007) (results not shown). However, the predictive ability with allele frequencies present in the respective population was slightly greater than with the standard genomic relationship matrix introduced by VanRaden (2007) (results not shown). The goal of using base allele frequencies was to account for subpopulation structures. It was thus not so important to use base allele frequencies (Gengler et al., 2007) of the respective subpopulation, but more relevant to use a combination of population-specific allele frequencies which were present in the jointly analyzed subpopulations.

Another possibility to account for the subpopulation structure is by using the genomic relationship matrix suggested by Zhou et al. (2014). With this approach, lower predictive abilities were obtained and predictive ability deteriorated with multi-subpopulation reference sets (Fig. 2.4), which might be due to the low correlations between the marker effects for different subpopulations (Appendix 2.5). In addition, the present results agree with findings of Daetwyler et al. (2012) that accounting for LD structure or fitting of principle components to account for population structure decreased the accuracy of multi-breed genomic predictions. With the adjustment for marker effects and LD structure, the predictive ability increased only slightly, which agrees with the results of Zhou et al. (2014) that weighing the G-matrices by LD phases consistency or marker effects did not improve accuracies of the two-breed genomic predictions.

## CONCLUSIONS

Using multi-breed populations to predict genomic breeding values is still considered an attractive option to increase the reference set size used for improving prediction equations in genomic selection in dairy cattle. In this study we showed that combining different subpopulations of German Large White pigs to form a multi-subpopulation reference

population does not lead to a better predictive ability for individuals within a specific subpopulation. Adding individuals from distantly related subpopulations actually decreased the predictive ability of DGVs. However, adding sows of the same subpopulation to the reference set yielded in an improved predictability of DGVs. Implementation of genomic breeding value prediction itself into a pig breeding program might have the potential to improve the selection response for the trait ‘number of piglets born alive’. However, the accuracy of genomic prediction with different subpopulations in the reference set might only be improved after creating more concurrent links between subpopulations, e.g. by using the same boars across populations.

### **ACKNOWLEDGEMENT**

The authors gratefully acknowledge the financial support of the project ‘pigGS’ by the Europaeischen Fonds fuer regionale Entwicklung (EFRE), the state of North Rhine-Westphalia, the project management Juelich and the pig breeding organizations for providing the data.

## REFERENCES

- Bergfelder-Drüing, S., C. Grosse-Brinkhaus, B. Lind, M. Erbe, K. Schellander, H. Simianer and E. Tholen. 2015. A Genome-Wide Association Study in Large White and Landrace Pig Populations for Number Piglets Born Alive. *PLoS ONE* 10(3): e0117468. doi:10.1371/journal.pone.0117468
- Browning, S.R., and B.L. Browning. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* 81(5):1084–97. doi:10.1086/521987
- Calus, M.P.L., H. Huang, Y.C.J. Wientjes, J. Ten Napel, J.W.M. Bastiaansen, M.D. Price, R.F. Veerkamp, J. J.A. Vereijken, and J. J. Windig. 2014. (A)cross-breed Genomic Prediction. *Proc. 10th World Congr. Genet. Appl. Livest. Prod.*, Vancouver, Canada. [https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/064\\_paper\\_10274\\_manuscript\\_1278\\_0.pdf?sfvrsn=2](https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/064_paper_10274_manuscript_1278_0.pdf?sfvrsn=2)
- Calus, M. P., H. Huang, A. Vereijken, J. Visscher, J. Ten Napel, and J. J. Windig. 2014. Genomic prediction based on data from three layer lines: a comparison between linear methods. *Genet. Sel. Evol.* 46:57.
- Daetwyler, H.D., K.E. Kemper, J.H.J. van der Werf, and B.J. Hayes. 2012. Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep Population. *J. Anim. Sci.* 90(10):3375–84. doi:10.2527/jas.2011-4557
- De Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. *Genetics* 183(4). doi:10.1534/genetics.109.104935
- Erbe, M., B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, B.A. Mason, and M.E. Goddard. 2012. Improving Accuracy of Genomic Predictions within and between Dairy Cattle Breeds with Imputed High-Density Single Nucleotide Polymorphism Panels. *J. Dairy Sci.* 95(7): 4114–29. doi:10.3168/jds.2011-5019.
- Gao, H., G. Su, L. Janss, Y. Zhang, and M. S. Lund. 2013. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. *J. Dairy Sci.* 96:4678–4687.
- Garrick, D.J, J.F Taylor, and R.L Fernando. 2009. Deregressing Estimated Breeding Values and Weighting Information for Genomic Regression Analyses. *Genet. Sel. Evol.* 41(1):55. doi:10.1186/1297-9686-41-55
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A Simple Method to Approximate Gene Content in Large Pedigree Populations: Application to the Myostatin Gene in Dual-Purpose Belgian Blue Cattle. *Animal* 1(1). doi:10.1017/S1751731107392628
- Gilmour, Arthur R., B.J. Gogel, B.R. Cullis, R. Thompson, D. Butler, and others. 2009. ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK, 2009. <http://vsni.de/downloads/asreml/release3/UserGuide.pdf>
- Harris, B.L., D.L Johnson, R.J. Spelman. 2009. Genomic selection in New Zealand and implications for national genetic evaluation. *Proc. of the Interbull Meeting*. Niagara, Canada

- Harris, B.L., F.E. Creagh, A.M. Winkelman, and D.L. Johnson. 2011. Experiences with the Illumina High Density Bovine BeadChip. *Interbull Bulletin*, no. 44. <https://journal.interbull.org/index.php/ib/article/view/1182>
- Harris, B.L., A.M. Winkelman, and D.E. Johnson. 2014. Across-Breed Genomic Prediction in Dairy Cattle. *Proc. 10th World Congr. Genet. Appl. Livest. Prod.*, Vancouver, Canada. [https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/066\\_paper\\_10257\\_manuscript\\_1267\\_0.pdf?sfvrsn=2](https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/066_paper_10257_manuscript_1267_0.pdf?sfvrsn=2)
- Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009. Accuracy of Genomic Breeding Values in Multi-Breed Dairy Cattle Populations. *Genet. Sel. Evol.* 41(1):51. doi:10.1186/1297-9686-41-51
- Hidalgo, A. M., J.W.M. Bastiaansen, M.S. Lopes, B. Harlizius, M.A.M. Groenen and D.J. de Koning. 2014. Accuracy of Genomic Breeding Values Predicted Within and Across Breeds in Pig Populations. *Proc. 10th World Congr. Genet. Appl. Livest. Prod.*, Vancouver, Canada. [https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/076\\_paper\\_8546\\_manuscript\\_158\\_0.pdf?sfvrsn=2](https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/076_paper_8546_manuscript_158_0.pdf?sfvrsn=2)
- Ibáñez-Escriche, N., R.L. Fernando, A. Toosi, and J.C.M. Dekkers. 2009. Genomic Selection of Purebreds for Crossbred Performance. *Genet. Sel. Evol.* 41(1):12. doi:10.1186/1297-9686-41-12
- Kachman, S.D., M.L. Spangler, G.L. Bennett, K.J. Hanford, L.A. Kuehn, W.M. Snelling, R.M. Thallman, et al.. 2013. Comparison of Molecular Breeding Values Based on within-and across-Breed Training in Beef Cattle. *Genet. Sel. Evol.* 45:30. doi:10.1186/1297-9686-45-30
- Kizilkaya, K., R.L. Fernando, and D.J. Garrick. 2010. Genomic Prediction of Simulated Multibreed and Purebred Performance Using Observed Fifty Thousand Single Nucleotide Polymorphism Genotypes. *J. Anim. Sci.* 88(2):544–51. doi:10.2527/jas.2009-2064
- Makgahlela, M.L., I. Strandén, U.S. Nielsen, M.J. Sillanpää, and E.A. Mäntysaari. 2014. Using the Unified Relationship Matrix Adjusted by Breed-Wise Allele Frequencies in Genomic Evaluation of a Multibreed Population. *J. Dairy Sci.* 97(2):1117–27. doi:10.3168/jds.2013-7167
- Meyer, K.. 2007. WOMBAT—A Tool for Mixed Model Analyses in Quantitative Genetics by Restricted Maximum Likelihood (REML). *J Zhejiang Univ Sci B* 8 (11):815–21. doi:10.1631/jzus.2007.B0815
- Ostensen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genet. Sel. Evol.* 43:38.
- Pryce, J.E., B. Gredler, S. Bolormaa, P.J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M.E. Goddard, and B.J. Hayes. 2011. Short Communication: Genomic Selection Using a Multi-Breed, across-Country Reference Population. *J. Dairy Sci.* 94(5):2625–30. doi:10.3168/jds.2010-3719
- Pszczola, M., T. Strabel, H.A. Mulder, M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95(1):389-400. doi:10.3168/jds.2011-4338

- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81(3):559–75. doi:10.1086/519795
- Rothschild, M.F. and A. Ruvinsky. 1998. *The genetics of the pig*. CAB INTERNATIONAL, Wallingford, UK
- Saatchi, M. and D. J. Garrick. 2013. Accuracies of genomic predictions in US beef cattle. *Proc. Assoc. Adv. Anim. Breed. Genet. Conf.*, 20: 207–210
- Schrooten, C., G.C.B. Schopen, A. Parker, A. Medley, and P. Beatson. 2013. Across-Breed Genomic Evaluation Based on Bovine High Density Genotypes and Phenotypes of Bulls and Cows. *Proc. Assoc. Advmt. Anim. Breed. Genet.*, 20:138–41
- Strandén, I., O.F. Christensen, and others. 2011. Allele Coding in Genomic Evaluation. *Genet. Sel. Evol.* 43(1). doi:10.1186/1297-9686-43-25
- Tiezzi, F., and C. Maltecca. 2014. Genomic Prediction Using a Weighted Relationship Matrix to Account for Trait Architecture in US Holstein Cattle. *Proc. 10th World Congr. Genet. Appl. Livest. Prod.*, Vancouver, Canada
- Tribout, T., C. Larzul, and F. Phocas. 2012. Efficiency of Genomic Selection in a Purebred Pig Male Line. *J. Anim. Sci.* 90(12):4164–76. doi:10.2527/jas.2012-5107.
- Vallee, A., J.A.M. van Arendonk, and H. Bovenhuis. 2014. Accuracy of Genomic Prediction When Combining Two Related Crossbred Populations. *J. Anim. Sci.* 92(10):4342–48. doi:10.2527/jas.2014-8109
- VanRaden, P. M.. 2007. Genomic Measures of Relationship and Inbreeding. *Interbull Bulletin*, no. 37:33
- VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited Review: Reliability of Genomic Predictions for North American Holstein Bulls. *J. Dairy Sci.* 92(1):16–24. doi:10.3168/jds.2008-1514
- Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2013. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics* 193(2):621–31. doi:10.1534/genetics.112.146290
- Wright, Sewall. 1943. Isolation by Distance. *Genetics* 28(2):139–156
- Zhou, L., M.S. Lund, Y. Wang, and G. Su. 2014. Genomic Predictions across Nordic Holstein and Nordic Red Using the Genomic Best Linear Unbiased Prediction Model with Different Genomic Relationship Matrices. *J. Anim. Breed. Genet.* 131(4):249–57. doi:10.1111/jbg.12089

**APPENDIX****APPENDIX 2.1**


---

Summary for all animals in the subpopulation

	EBV			DRP		
	min	mean	max	min	mean	max
Subpopulation 1*	-1.81	0.87	3.44	-3.87	0.92	8.37
Subpopulation 2	-2.06	-0.24	1.67	-3.27	-0.35	2.83
Subpopulation 3	-1.51	0.17	2.28	-1.86	0.26	4.43
Subpopulation 4*	-1.45	1.22	3.57	-3.80	1.28	5.88
Subpopulation 5	-2.93	0.43	3.28	-4.23	0.98	7.92

---

\* With removal of PA in the DRP calculation

**APPENDIX 2.2**Genomic Relationship between different subpopulations ( $G_{VR}$ )

Subpopulations	Summary					Number of combinations between animals of different subpopulations	
	Min	1 <sup>st</sup> Qu.	Mean	3 <sup>rd</sup> Qu.	Max	$G_{VR} > 0.25$	$G_{VR} > 0.5$
1	-0.17	-0.05	-0.001	0.03	0.59	240	10
1 and 2	-0.14	-0.05	-0.02	0.004	0.46	4	0
1 and 3	-0.17	-0.05	-0.02	0.007	0.49	8	0
1, 2 and 3	-0.13	-0.04	-0.01	0.01	0.55	50	7
1, 4 and 5	-0.18	-0.10	-0.07	-0.05	0.17	0	0
1, 2, 3 and 5	-0.08	-0.05	-0.02	0.01	0.59	57	11
4	-0.17	-0.03	-0.001	0.02	0.68	1875	185
4 and 2	-0.13	-0.05	-0.02	0.003	0.31	1	0
4, 2 and 1	-0.07	-0.05	-0.02	0.003	0.43	9	0
4 and 5	-0.20	-0.10	-0.08	-0.06	0.11	0	0
All	-0.07	-0.03	-0.08	0.014	0.58	55	10

**APPENDIX 2.3**

Assessment of the predictive ability calculated with and without effect of the line included in the model for the validation set subpopulation 1

Cor (DGV, DRP)	Random 5-fold cross-validation		Forward Prediction		Correlation between DGVs with and without line effect
	Effect of line	No effect of line	Effect of line	No effect of line	
Subpopulation 1 and 3	0.527	0.530	0.343	0.346	0.998
Subpopulation 1, 4 and 5	0.540	0.540	0.308	0.309	0.991

Rank of Breeding Value (DGV) over the whole population

Breeding Value without effect of line compared to with effect of line	Reference and validation set					
	Top 10		Top 20		Top 40	
	In	Out	In	Out	In	Out
Subpopulation 1 and 3*	8	2	19	1	33	7
Subpopulation 1, 4 and 5*	9	1	19	1	39	7

\*Consideration as common population without subpopulation structure

Rank of Breeding Value (DGV) within each subpopulation

Breeding Value without effect of line compared to with effect of line		Reference and validation set						Validation set*					
		Top 10		Top 20		Top 40		Top 10		Top 20		Top 40	
		In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
Subpopulation 1 and 3	Subpopulation 1	8	2	19	1	33	7	4	6	10	10	21	19
	Subpopulation 3	9	1	19	1	40	0			§			
Subpopulation 1, 4 and 5	Subpopulation 1	10	0	19	1	36	4	5	5	9	11	17	23
	Subpopulation 4	9	1	18	2	39	1			§			
	Subpopulation 5	9	1	19	1	39	1			§			

\* Animals belonging to the group with year of birth being 2010 and 2011

§ Not considered as validation set

**APPENDIX 2.4**

Correlation between DGV calculated with DRP and DGV calculated with DRP\*

Reference population	Validation set	Correlation for all animals in reference set	Correlation for Validation set
1	1	0.91	0.91
1 and 2	1	0.90	0.87
1 and 3	1	0.91	0.88
1, 2 and 4	1	0.91	0.94
1, 4 and 5	1	0.94	0.94
1, 2, 3 and 5	1	0.94	0.93
All	1	0.94	0.95
4	4	0.93	0.93
4 and 2	4	0.93	0.93
4, 2 and 1	4	0.91	0.93
4 and 5	4	0.94	0.93
4, 5 and 1	4	0.94	0.93
All	4	0.94	0.93

\* without removal of PA

**APPENDIX 2.5**

Correlation between SNP effects for different populations

Combination	Correlation	Rank-based correlation according to spearman	Correlation with sliding window (10 SNPs)	Correlation with Top 10 SNPs
Subpopulation 1 and 2	0.085	0.078	0.074	0.211
Subpopulation 1 and 3	0.090	0.086	0.075	0.001
Subpopulation 1 and 4	0.067	0.072	0.052	-0.167
Subpopulation 1 and 5	0.039	0.035	0.053	-0.059
Subpopulation 4 and 2	0.046	0.043	0.056	0.214
Subpopulation 4 and 3	0.054	0.058	0.057	-0.016

### **3<sup>rd</sup> CHAPTER**

#### **Empirical comparison between different methods for genomic prediction of number of piglets born alive in moderate sized breeding populations**

A. Fangmann\*, R.A. Sharifi\*, J. Heinkel†, K. Danowski†, H. Schrade†, M. Erbe‡, H. Simianer\*

\* Department of Animal Sciences, Animal Breeding and Genetics Group, University of Goettingen, 37075 Goettingen, Germany

† Bildungs- und Wissenszentrum Boxberg- Schweinehaltung, Schweinezucht-, 97944 Boxberg, Germany

‡ Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, 85586 Poing-Grub, Germany

Published in Journal of Animal Science

Vol. 95, No. 4, p. 1434–1443, April 2017

© 2017 American Society of Animal Science. All rights reserved.

doi: 10.2527/jas2016.0991

### ABSTRACT

Currently used multi-step methods to incorporate genomic information in the prediction of breeding values (BV) implicitly involve many assumptions which, if violated, may result in loss of information, inaccuracies and bias. To overcome this, single-step genomic best linear unbiased prediction (ssGBLUP) was proposed combining pedigree, phenotype and genotype of all individuals for genetic evaluation. Our objective was to implement ssGBLUP for genomic predictions in pigs and to compare the accuracy of ssGBLUP with that of multi-step methods with empirical data of moderately sized pig breeding populations. Different predictions were performed: conventional parent average (PA), direct genomic value (DGV) calculated with genomic BLUP (GBLUP), a genomic enhanced breeding value (GEBV) obtained by blending the DGV with PA, and ssGBLUP. Data comprised individuals from a German Landrace (LR) and Large White (LW) population. The trait ‘number of piglets born alive’ (NBA) was available for 182,054 litters of 41,090 LR sows and 15,750 litters from 4,534 LW sows. The pedigree contained 174,021 animals, of which 147,461 (26,560) animals were LR (LW) animals. In total, 526 LR and 455 LW animals were genotyped with the Illumina PorcineSNP60 BeadChip. After quality control and imputation, 495 LR (424 LW) animals with 44,368 (43,678) SNPs on 18 autosomes remained for the analysis. Predictive abilities, i.e. correlations between deregressed proofs and genomic BVs, were calculated with a five-fold cross validation and with a forward prediction for young genotyped validation animals born after 2011. Generally, predictive abilities for LR were rather small (0.08 for GBLUP, 0.19 for GEBVs and 0.18 for ssGBLUP). For LW, ssGBLUP had the greatest predictive ability (0.45). For both breeds, assessment of reliabilities for young genotyped animals indicated that genomic prediction outperforms PA with ssGBLUP providing greater reliabilities (0.40 for LR and 0.32 for LW) than GEBVs (0.35 for LR and 0.29 for LW). Grouping of animals according to information sources revealed that genomic prediction had the highest potential benefit for genotyped animals without own phenotype. Although, ssGBLUP did not generally outperform GBLUP or GEBVs, the results suggest that ssGBLUP can be a useful and conceptually convincing approach for practical genomic prediction of NBA in moderately sized LR and LW populations.

**Key words:** genomic selection, multi-step, number of piglets born alive, pig, single-step

## INTRODUCTION

Initial methods for using genomic information to predict genomic breeding values were multi-step methods (VanRaden, 2008; Hayes et al., 2009a). These methods mainly consist of three parts: (i) construction of a response variable (also called ‘pseudo-observation’) for genotyped animals that integrate all available phenotypic information, (ii) exploiting the association between response variable and marker information through genomic prediction, and (iii) blending genomic predictions with parental average estimated breeding values (Christensen et al., 2012). Those steps involve many parameters and multiple assumptions, which, if violated, may result in loss of information, inaccuracies and biases which strongly depends on species and data set (Legarra et al., 2014). A strategy to overcome these problems is the ‘single-step’ approach. In general, single-step methodology combines pedigree, phenotypic and genomic information of all breeding individuals for genetic evaluation in one model (Christensen et al., 2012). Single-step methods have been proposed by Legarra et al. (2009), Aguilar et al. (2010) and Christensen and Lund (2010) and were found to be simple, fast and accurate (Miszta et al., 2013). Christensen et al. (2012) suggest that the single-step method is at least as accurate as a multi-step method for prediction of genotyped animals and that both methods perform better than purely pedigree-based methods. Studies in cattle (e.g. Lourenco et al., 2015b), chickens (e.g. Misztal et al., 2013; Lourenco et al., 2015a) or pigs (e.g. Guo et al., 2015) confirmed that single-step methods are at least as accurate as multi-step methods. Our main objective was to implement single-step methodology for genomic predictions for a lowly heritable trait with data from a practical pig breeding program and to conduct a comprehensive comparison between accuracy of single-step and multi-step methods. For this purpose, we studied the trait ‘number of piglets born alive’ recorded in moderately-sized commercial German Landrace and Large White breeding populations.

## MATERIALS AND METHODS

### Animals and Data

Real data from purebred and crossbred animals from German Landrace (**LR**) and German Large White (**LW**) pigs were provided from a commercial pig breeding organization in Germany. The analyzed trait was ‘number of piglets born alive’ (**NBA**). The data set comprised 182,054 litters from 41,090 LR sows, 15,750 litters from 4,534 LW sows and 23,143 litters from 4,725 multiplier sows (LW x LR). Pedigree data consisted of 174,021

animals, of which 147,461 (26,560) animals were LR (LW) animals. Records of litters were available from 2000 to 2014. Because we had no genotypes of multiplier sows, we only used their phenotypic information for the conventional and single-step breeding value estimation.

In total, 526 Landrace and 455 Large White animals were both genotyped with the Illumina PorcineSNP60 BeadChip (Illumina, Inc., San Diego, CA). Uncalled markers and markers with unknown position were removed from the data set. Quality control (QC) was performed with the software PLINK (Purcell et al., 2007) requesting that each animal had a call rate > 98% and each marker had a call rate > 97%. In order to exclude genotyping errors, the minor allele of the marker had to be observed at least ten times. Missing genotypes were subsequently imputed with the software BEAGLE (Browning and Browning, 2007). After QC and imputation, 495 LR (424 LW) animals with 44,368 (43,678) SNPs on 18 autosomes remained for the analysis.

### Multi-step approach

The multi-step approach consisted of the following three steps:

**BLUP model.** A traditional best linear unbiased prediction (**BLUP**) animal model with repeated measurements was used to predict conventional breeding values (**EBV**) for NBA of LW and LR using the software ASReml (Gilmour et al., 2009). The model was very similar to the routine breeding value evaluation of the organization. Fixed effects included herd-year-season, breed of sow, breed of sire, number of litter, age at first farrowing (only for first parity) and interval between farrowing (linear and quadratic, not for first parity). In a second step, deregressed proofs (**DRP**) were calculated from EBVs with removal of the parent average effect (**PA**) as described by Garrick et al. (2009). Due to missing parent average values for genotyped animals and genotyped animals only with PA as breeding value, we were not able to calculate DRPs for all genotyped animals and had to remove those animals from analysis. Table 3.1 shows the remaining number of genotyped animals for genomic evaluation for which DRPs could be calculated.

**Table 3.1.** Total number of genotypes for genomic predictions for ‘number of piglets born alive’ (NBA) for Landrace (LR) and Large White (LW) animals

Total number of	LR	LW
Genotypes	526	455
Genotypes after quality control	495	424
Genotypes after calculating deregressed proofs	443	287
Validation individuals for forward predictions	97	53

**Genomic BLUP model.** Estimation of direct genomic breeding values (**DGV**) was performed using a genomic BLUP (**GBLUP**) animal model for each breed separately. The following model was used:

$$\mathbf{y} = \mu + \mathbf{W}\mathbf{g} + \mathbf{e},$$

where  $\mathbf{y}$  was the vector of DRPs for trait NBA,  $\mu$  was the overall mean,  $\mathbf{g}$  was a vector of random genomic effects, and  $\mathbf{e}$  was a vector of random residual effects.  $\mathbf{W}$  was the corresponding design matrix. We assumed that  $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ , where  $\mathbf{G}$  was a genomic relationship matrix as proposed by VanRaden et al. (2007) and  $\sigma_g^2$  was the corresponding genomic variance. Genomic BVs obtained from this model were denoted DGVs. Variance components were computed using the software ASReml (Gilmour et al., 2009).

**Blending to obtain Genomic Enhanced Breeding Values.** In a routine breeding program, the DGV is not used directly but is combined with EBV to obtain a genomically enhanced breeding value (**GEBV**), a procedure which is called ‘blending’ (VanRaden et al., 2009). For the blending index as defined in VanRaden et al. (2009) three different breeding values (**BV**) were combined for the validation individuals in the forward prediction scheme: (i) the PA of an animal from the BLUP breeding value estimation, (ii) the DGV, which was computed on the basis of the genomic relationship matrix and DRPs with removal of the PA (Garrick et al., 2009) and (iii) parental index (**PI**), which was estimated based on a pedigree relationship matrix using only the genotyped animals and DRPs as pseudo-observations. Subsequently, the reliabilities of these three information sources are combined in a selection index to calculate the optimum weighting of each part of the breeding values for each animal resulting in the

final GEBV (VanRaden et al., 2009) of an animal for which also the respective reliability can be calculated.

### Single-step genomic BLUP

A single-step genomic BLUP (**ssGBLUP**) model for each breed separately was used by replacing the pedigree-based relationship matrix **A** with **H**, a mixed pedigree and genomic relationship matrix. It was assumed that  $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$  and  $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ , where **u** was the vector of single-step genomic BVs (**ssGEBV**) with  $\sigma_u^2$  being the corresponding genetic variance, **I** was an Identity matrix and  $\sigma_e^2$  was the corresponding residual variance.

The inverse variance-covariance matrix of the genetic effects ( $\mathbf{H}^{-1}$ ) was calculated as a combined relationship matrix suggested by Aguilar et al. (2010) and Christensen and Lund (2010) as:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_w^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix},$$

where  $\mathbf{A}^{-1}$  was the inverse of the pedigree-based relationship matrix **A**,  $\mathbf{G}^{-1}$  was the inverse of the genomic relationship matrix **G** and  $\mathbf{A}_{22}^{-1}$  was the inverse of the subset of the pedigree-based relationship matrix **A** between genotyped animals only. To make **G** compatible with **A**, **G** was modified to be on the same scale as **A** (Christensen et al., 2012). To scale the genomic information and to avoid singularity problems, we calculated  $\mathbf{G}_w$  as:

$$\mathbf{G}_w = (\alpha * \mathbf{G} + \beta * \mathbf{A}_{22}),$$

with  $\alpha = 0.95$ ,  $\beta = 0.05$ . The proportions of  $\alpha$  and  $\beta$  were used from the default values in the software BLUPF90 (Misztal et al., 2002).

Prediction of genomic BVs and estimation of variance components were performed with the programs implemented in the software BLUPF90 (Misztal et al., 2002). BVs obtained from single-step procedure were denoted ssGEBVs.

### Assessment of predictive ability

**Five-fold cross validation.** To assess the predictive ability, a five-fold cross validation with 20 random replicates was performed. As a measure of accuracy, the correlation between

DRPs and genomic BVs being either DGVs, GEBVs or ssGEBVs, was calculated in the validation set for each run and averaged over folds.

**Forward prediction.** To mimic a real situation in animal breeding, the data set was divided into a reference and validation set by the year of birth. Animals that were born before 2011 belonged to the reference set, whereas the youngest animals born from 2011 to 2014 belonged to the validation set. Predictive ability was calculated based on the correlation between DRPs and genomic BVs, which were either DGVs, GEBVs or ssGEBVs. Genomic BVs for the validation set were calculated based on reference individuals only using information that were available for animals that were born before 2011.

Additionally, we conducted a forward prediction to assess the predictive ability, expressed as correlation between corrected phenotype ( $y_c$ ) and different genomic BVs (DGV, GEBV or ssGEBV) for young genotyped validation animals. The reference set size remained as before. Due to the small number of genotyped LW animals, we performed this forward prediction only for young LR animals.

### MME reliabilities in forward prediction scenarios

Theoretical reliabilities of the corresponding BVs for the youngest animals (from forward prediction) were calculated from the inversion of the mixed model equation (**MME**; Henderson, 1975; Tier and Meyer, 2004) and averaged over the youngest animals for the different genomic methods. The theoretical reliability ( $r^2$ ) of the corresponding BV for an individual (i) was calculated as

$$r_i^2 = 1 - \frac{SE(g_i)^2}{\sigma_g^2},$$

where  $g_i$  was the corresponding BV of the  $i^{\text{th}}$  individual,  $SE(g_i)$  was the standard error of the corresponding BV and  $\sigma_g^2$  was the corresponding genetic variance calculated from the genomic model (Tier and Meyer, 2004; Hayes et al., 2009b).

Furthermore, to evaluate the quality of prediction for different groups of animals, theoretical reliabilities for EBV and ssGEBV were calculated and averaged within groups for all available animals. Used animals were classified according to their available information: For groups with phenotypes (**P**; own and/or progeny), BVs were calculated based on all available information. For groups of young genotyped boars and sows, BVs were calculated based on

reference animals only. Those young animals had no own or progeny performance (**nP**) and were denoted as selection candidates. Within each group, the correlation between EBVs and ssGEBVs was calculated separately.

## RESULTS

### Descriptive analysis

Descriptive analysis of the raw data for LR and LW sows is show in Table 3.2. On average, 11.3 (13.8) LR (LW) piglets were born alive per sow and litter, while LR (LW) had on average 2.8 (2.6) parities. LR sows farrowed the first time with 360 days of age, LW sows with 363.1 days of age. Heritability for NBA calculated from the whole data set (ASReml; Gilmour et al., 2009) was  $0.19 \pm 0.01$ . This heritability was also used in the de-regression procedure for EBVs (Garrick et al., 2009).

**Table 3.2.** Structure of the raw data for Landrace (LR) and Large White (LW) sows

	LR				LW			
	$\bar{x}$	std	min	max	$\bar{x}$	std	min	max
<b>Number of piglets born alive (NBA)</b>	11.3	2.6	3.0	24.0	13.8	3.1	3.0	25.0
<b>Age at first farrowing (days)</b>	360.0	23.6	274.0	420.0	363.1	19.2	282.0	420.0
<b>Interval between farrowing (days)</b>	202.4	89.5	136.0	420.0	212.7	97.5	136.0	420.0
<b>Parity</b>	2.8	1.2	1.0	4.0	2.6	1.2	1.0	4.0

### Assessment of predictive ability

**Five-fold cross-validation.** Predictive ability for NBA was calculated with five-fold cross validation for a total of 443 genotyped LR animals and a total of 287 all genotyped LW animals (Table 3.3). Predictive ability, expressed as correlation between DRP and DGV, was on a moderate level with  $0.43 \pm 0.02$  ( $0.39 \pm 0.02$ ) for LR (LW). Correlations between DRP and GEBV turned out to be greatest with for LR (LW) animals, while correlations between DRP and ssGEBV were slightly higher than with DGV.

**Table 3.3.** Assessment of predictive ability ( $\pm$  standard error of correlation coefficient, SE), i.e. correlation (cor) between deregressed proofs (DRP) and different breeding values (DGV, GEBV and ssGEBV), for all genotyped (five-fold cross-validation) and young genotyped (forward prediction) Landrace (LR) and Large White (LW) animals. Breeding values for young genotyped animals were calculated based on reference animals and compared to DRP, which were calculated based on all genotyped animals.

	Five-fold cross validation		Forward prediction	
	LR	LW	LR	LW
<b>COR(DRP<sup>1</sup>, DGV<sup>2</sup>)</b>	0.43 $\pm$ 0.02	0.39 $\pm$ 0.02	0.08 $\pm$ 0.14	0.43 $\pm$ 0.11
<b>COR(DRP, GEBV<sup>3</sup>)</b>	0.52 $\pm$ 0.01	0.52 $\pm$ 0.01	0.19 $\pm$ 0.13	0.40 $\pm$ 0.12
<b>COR(DRP, ssGEBV<sup>4</sup>)</b>	0.43 $\pm$ 0.02	0.40 $\pm$ 0.02	0.18 $\pm$ 0.13	0.45 $\pm$ 0.11

<sup>1</sup>DRP = deregressed proofs.

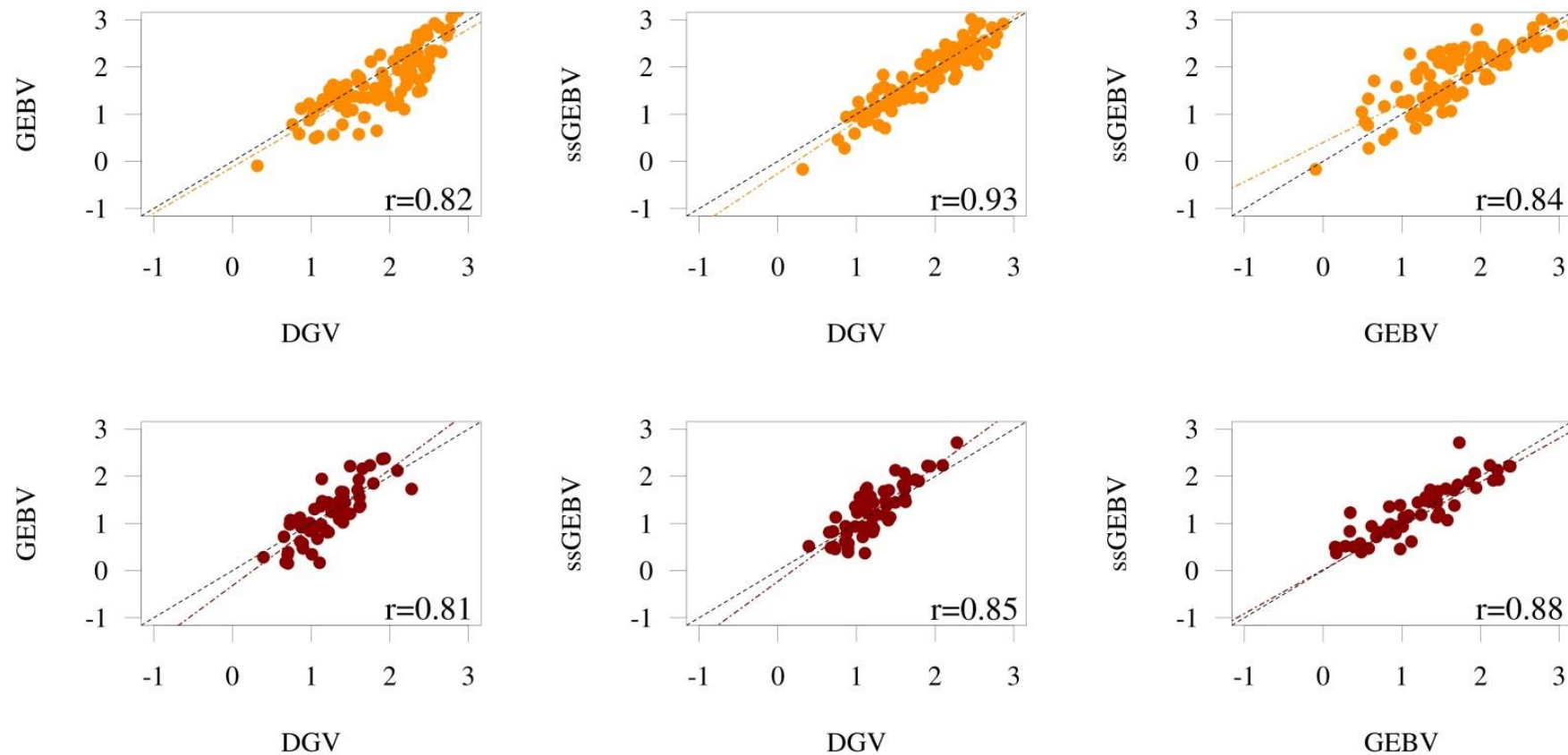
<sup>2</sup>DGV = direct genomic breeding value.

<sup>3</sup>GEBV = genomically enhanced breeding value.

<sup>4</sup>ssGEBV = single-step genomically enhanced breeding value.

**Forward prediction.** Predictive abilities for the youngest genotyped animals for LR and LW are shown in Table 3.3. In general, predictive abilities turned out to be substantially higher for young genotyped LW animals than for young genotyped LR animals. For young genotyped LR animals, correlation between DRP and GEBV was superior to the correlation between DRP and DGV and ssGEBV, respectively. In LW, the greatest correlation was measured for DRP with ssGEBV, followed by DGV and GEBV.

Figure 3.1 shows correlations ( $r$ ) between different genomic BVs, for young genotyped LR (upper plots) and LW (lower plots) animals. For young genotyped LR animals, the correlation between DGV and ssGEBV was superior to GEBV and ssGEBV and DGV and GEBV, respectively. For young genotyped LW animals, the correlation between ssGEBV and GEBV was superior to DGV and GEBV and DGV and ssGEBV, respectively.



**Figure 3.1.** Correlation (r) between different breeding values (direct genomic value, DGV; genomically enhanced breeding value, GEBV and single-step genomically enhanced breeding value, ssGEBV) for young genotyped animals. Breeding values were calculated based on reference animals. DGV were calculated with quasi-phenotype deregressed proof (DRP). Results are presented for Landrace (LR, upper plot) and Large White (LW, lower plot). The blue lines display the angle bisector while the orange and red lines show regression lines.

Predictive abilities, expressed as correlation between corrected phenotype  $y_c$  and different genomic BVs (DGV, GEBV or ssGEBV) for a total of 59 young genotyped LR sows are displayed in Table 3.4. BVs were calculated based on reference animals and compared to corrected phenotypes, which were calculated based on all animals. Predictive abilities turned out to be greatest with ssGEBV, which emphasizes that single-step is superior to GEBVs obtained after blending (decreased by 0.05 points) and DGVs from a GBLUP model (decreased by 0.14 points).

Among predictive abilities, Table 3.4 displays regression coefficients of corrected phenotypes on different genomic BVs (DGV, GEBV or ssGEBV) for young genotyped LR sows, which turned out to be close to 1 with ssGEBV indicating that single-step estimates appear to be less biased than GEBV estimates obtained by blending DGVs with EBVs (0.08 points) or DGVs (0.20 points).

**Table 3.4.** Assessment of predictive ability, expressed as correlation between corrected phenotype ( $y_c$ ) and different breeding values (DGV, GEBV or ssGEBV), and regression coefficients (slope  $\pm$  standard error) for 59 young genotyped Landrace (LR) sows. Breeding values (BV) were calculated based on reference animals and compared to corrected phenotypes, which were calculated based on all genotyped animals.

<b>Forward Prediction      young genotyped LR sows</b>		
<b>BV</b>	<b>Cor(<math>y_c^1</math>, BV)</b>	<b>Regression of <math>y_c</math> on BV</b>
<b>DGV<sup>2</sup></b>	0.30	$0.80 \pm 0.33$
<b>GEBV<sup>3</sup></b>	0.39	$0.92 \pm 0.29$
<b>ssGEBV<sup>4</sup></b>	0.44	$0.99 \pm 0.27$

<sup>1</sup> $y_c$  = corrected phenotype.

<sup>2</sup>DGV = direct genomic breeding value.

<sup>3</sup>GEBV = genomically enhanced breeding value.

<sup>4</sup>ssGEBV = single-step genomically enhanced breeding value.

### **MME reliabilities in forward prediction scenarios**

Theoretical reliabilities calculated from MME for different BVs for young genotyped LR and LW animals are displayed in Table 3.5. Reliabilities were calculated based on reference animals that were born before 2011. The highest reliability for young genotyped animals was calculated for ssGEBV with 0.40 (0.32) for LR (LW), while reliabilities for GEBVs decreased

by 0.05 (0.03) points and for DGVs by 0.12 (0.09) points. Reliabilities of genomic BV were always higher than those of conventional PAs, and increased by 0.13 (0.11) points for LR (LW) animals when applying ssGBLUP.

**Table 3.5.** Theoretical reliability from MME ( $\pm$  standard error, SE) for different breeding values (PA, DGV, GEBV and ssGEBV) for young genotyped Landrace (LR) and Large White (LW) validation animals calculated based on reference animals.

Forward Prediction	LR	LW
<b>PA</b> <sup>1</sup>	0.27 $\pm$ 0.014	0.21 $\pm$ 0.030
<b>DGV</b> <sup>2</sup>	0.28 $\pm$ 0.009	0.23 $\pm$ 0.010
<b>GEBV</b> <sup>3</sup>	0.35 $\pm$ 0.011	0.29 $\pm$ 0.013
<b>ssGEBV</b> <sup>4</sup>	0.40 $\pm$ 0.006	0.32 $\pm$ 0.012

<sup>1</sup>PA = parental average.

<sup>2</sup>DGV = direct genomic breeding value.

<sup>3</sup>GEBV = genomically enhanced breeding value.

<sup>4</sup>ssGEBV = single-step genomically enhanced breeding value.

Table 3.6 shows reliabilities for EBV and ssGEBV for groups of genotyped LR sows and boars. Animals belonging to group **P** provided genomic and phenotypic information (**P**) for the BV estimation. Due to the considered trait, females provided own and/or progeny performance of NBA, while for males only progeny performances were available. BVs and corresponding reliabilities of **P** were calculated based on all available information. Animals belonging to group **nP** were young selection candidates with genomic but no phenotypic information. Those BVs and reliabilities were calculated based on reference animals. Reliabilities of EBV and ssGEBV in group **P** turned out to be on the same level for phenotyped females (0.73 for EBV and 0.74 for ssGEBV). For phenotyped males, a moderate increase in reliability from EBV to ssGEBV was observed (by 0.05 points), while the correlation compared to phenotyped females slightly decreased (by 0.07 points). The greatest increase in reliability from EBV to ssGEBV was observed for young candidates in group **nP** (both approx. by 25 per cent), showing that genomic prediction has the highest potential benefit for genotyped animals without own phenotype. Consequently the correlations between EBVs and ssGEBVs for young candidates were also low (**nP**; 0.80 for females and 0.67 for males), while it was 0.98 in the group with phenotyped sows (**P**).

**Table 3.6.** Theoretical reliability ( $\pm$  standard error, SE) of conventional estimated breeding value (EBV) and single-step genomically enhanced breeding value (ssGEBV) for different groups of genotyped Landrace (LR) animals. Animals were grouped according to provided information. For groups with phenotypes (P, own and/or progeny), breeding values were calculated based on all available information. For groups of young genotyped boars and sows, breeding values were calculated based on reference animals. Those young animals had no own or progeny performance (nP).

Genotyped groups		Number of animals	Reliability of		Cor(EBV and ssGEBV)
			EBV <sup>1</sup>	ssGEBV <sup>2</sup>	
P	Sows with own and/or progeny phenotype	232	0.73 ± 0.01	0.74 ± 0.01	0.98
	Boars with progeny phenotype	263	0.60 ± 0.01	0.65 ± 0.02	0.91
nP	Young sows	59	0.33 ± 0.01	0.41 ± 0.01	0.85
	Young boars	38	0.32 ± 0.02	0.40 ± 0.01	0.67

<sup>1</sup>EBV = conventional estimated breeding value.

<sup>2</sup>ssGEBV = single-step genomically enhanced breeding value.

For some genotyped animals, no parent information and corresponding PAs were available so that we could not calculate DRPs for those animals. Thus, these genotyped animals were removed from the analysis. Single-step can ignore filter criteria for calculating pseudo-observations and includes all data for the prediction of BVs. Thus, we were able to estimate EBV, GEBV and ssGEBV with corresponding theoretical reliability from the inverse of MME. Because most of the LR animals were included in our analysis, this approach was performed for a total of 137 excluded LW animals, which were mainly boars without progeny phenotypes. For excluded LW animals, EBVs obtained from BLUP had an average reliability of 0.12. GEBVs obtained after blending and ssGEBVs obtained from single-step were on average 0.36 (both) with an increased accuracy of 0.22 for GEBV and 0.20 for ssGEBV compared to EBV.

## DISCUSSION

In the current study for German Landrace and Large White populations, we basically compared four methods for evaluation of NBA: conventional parent average (PA), which is the conventional breeding value for a young animal without own or progeny performance, the

direct genomic value (DGV) calculated with genomic BLUP (GBLUP), a genomic enhanced breeding value (GEBV) obtained by blending the DGV with the conventional EBV, and genomically enhanced breeding values obtained from a single-step GBLUP (ssGEBV).

A comparison with other studies shows that the estimated heritability of NBA  $h^2 = 0.19$  was close to literature means (Rothschild and Ruvinsky, 1998; Huby et al., 2003; Putz et al., 2015). Compared to Huby et al. (2003), average NBA was larger (approx. 1.6 piglets born alive more) for both breeds in our study, reflecting the recent genetic trend in this trait. Compared to our study, Putz et al. (2015) recorded nearly the same average performance for NBA with 11.47 (10.86) piglets born alive for LR (LW).

In general, multi-step evaluations are rather complex and susceptible to errors, which have been observed in publications (e.g. in Mistzal et al., 2013). Single-step provides a unified approach which, in theory, has the following advantages: (i) it does not rely on pseudo-observations, (ii) it weighs all information properly, (iii) it accounts for preselection bias of genomically selected parents without phenotypes and (iv) it provides more accurate genomic BVs than e.g. multi-step approaches (Aguilar et al., 2010; Legarra et al., 2014).

Similar to results in the literature (Christensen et al., 2012; Guo et al., 2015) we also found that predictive abilities, measured as correlation between DRP and BV, showed that in general methods accounting for genomic information were more accurate than methods based on pedigree information alone. In genomic prediction, realized relationships among individuals are captured by marker information and might explain why genomic methods provided better predictabilities (Guo et al., 2015). Single-step and selection-index blending mainly uses the same information but single-step method still performed similarly or slightly better than selection-index blending in forward prediction, which is the case of highest practical relevance (Table 3.3). This might be attributable to the use of raw data in single-step instead of pseudo-observations like in multi-step. Also, selection-index blending involves two steps, where in the second step possible uncertainty from the first step is not properly taken into account (Christensen and Lund, 2010) and therefore might cause prediction problems (Table 3.3). Predictive ability for young LW candidates decreases with blending compared to GBLUP, which is rather surprising due to the fact that blending in principle uses more information for predictions than GBLUP. This could be partly explained by the quality of the empirical data like e.g. low and/or heterogeneous reliabilities of traditional EBVs. In any

case, blending is based on various assumptions which might not hold in the present data constellation.

In five-fold cross validation (Table 3.3), predictive abilities turned out to be higher for GEBV than for ssGEBV, a pattern that we did not observe in the forward prediction scheme. One possible reason might be that the set-up of the single-step system was not optimal in the present data (see further discussion about that below) and this becomes more obvious in the case a breeding value of an individual in the validation set contains information from the individual's genotype but potentially also from phenotypic observations of progeny which can be the case in a cross-validation scheme as individuals are not ordered by age.

Reliabilities of EBVs for individuals were rather small in this data set. Subsequently, deregression might not work properly and corresponding DRPs might not be accurate enough. To evaluate whether it would be a better option to use EBVs as response variable directly, we exemplarily calculated predictive abilities, expressed as correlation between EBV and genomic BV, for a five-fold cross validation. As expected, the correlations were on a higher absolute level than with DRP, but the overall pattern of predictive abilities was nearly the same as with DRP (results not shown). Since using EBVs and DRPs leads to basically the same conclusions, we only report DRP-based results to avoid redundancy.

Predictive ability for NBA for young genotyped animals calculated with forward prediction was generally low, especially for young genotyped LR animals. This could be due to the low heritability of NBA and the low number of genotypes available for LR and LW. Ibáñez-Escriche and Gonzalez-Recio (2011) suggested that genomic prediction could be of special interest when the accuracy of selection was generally low, as well as for low heritability traits or traits that can only be recorded directly in one sex, as is the case for NBA here. Prediction of young female candidates was generally more accurate than for young male candidates (results not shown), for which we also observed some outliers, which generally might lead to a poorer prediction for this breed (especially for LR). Besides the low heritability of the considered trait, another reason for poor predictions could be the generally low number of genotyped progenies per boar (2.74 for LW and 2.87 for LR) in combination with low reliabilities of EBVs for boars. Especially for traits with low heritability, larger numbers of genotypes and phenotypes are required to obtain an acceptable accuracy for genomic predictions (Calus and Veerkamp, 2007).

Since reliabilities for pseudo-observations (DRP) of reference animals were also rather moderate (0.27 for LW and 0.40 for LR), obtaining acceptable reliabilities might become difficult. It is also rather surprising that predictive abilities for young LW were greater than for young LR animals, although the number of genotyped animals for LW was lower than for LR. This might be an artifact of the small data size and/or might be due to the structure of reference and validation set. Only a few parent-offspring-links between reference and validation set in both breeds existed, but large quantities of really close relationships between individuals of reference and validation set were missing. The genotyping strategy of the small pig breeding organization mainly focused on representing the whole population of LW and LR animals and secondly to create close genetic links between reference and validation set. Thus, genomic relationships between reference and validation set were on average rather small for LW animals (0.0050) and even smaller for LR (0.0004). The composition of the reference population is as important as the reference size (Lourenco et al., 2015b) and there is a potential for improvement of both in this study.

Due to the small population size of LR and LW and to assess the variation of predictive abilities of the different genomic methods, we calculated bootstrap predictive abilities for randomly sampled candidates as described in Cuyabano et al. (2015). Predictive abilities, expressed as correlation between DRPs and genomic BVs, were calculated for each of the randomly sampled sets of candidates and averaged over sets (Appendix 3.1). As expected, the calculated predictabilities varied between sampled sets. Predictive abilities resulting from bootstrapping turned out to be on average slightly lower than predictive abilities of the original genomic models for LW and LR, while the ranking of the three models did not change compared to the original genomic models. The maximum sampled predictive abilities for both breeds were higher than those from the original genomic models indicating that with sufficient more individuals the predictive ability might increase. The ranges for predictive abilities ( $\pm$  standard deviation) for LR (LW) varied between  $0.25 \pm 0.057$  ( $0.35 \pm 0.071$ ) points for GBLUP,  $0.28 \pm 0.045$  ( $0.33 \pm 0.072$ ) for blending and  $0.22 \pm 0.048$  ( $0.35 \pm 0.069$ ) for ssGBLUP. Overall, bootstrap sampling did not increase the predictive ability of genomic prediction but could be a helpful tool to determine the variation of predictive ability for the different genomic methods (Mikshowsky et al., 2016).

In this multi-step methods, the dependent variable for predictions are pseudo-observations (DRP), while in single-step the independent variable is the raw-phenotype or a corrected phenotype. Correlations (predictive abilities) between DRP and genomic BVs from forward

predictions (Table 3.3) were generally smaller than obtained between  $y_c$  and genomic BV for young genotyped LR animals (Table 3.4). Both results indicated that predictions with ssGBLUP were preferable to blending and GBLUP and the calculated regression coefficient of  $y_c$  on ssGEBV (close to 1) supported this conclusion.

In this study, young genotyped animals showed higher theoretical reliabilities for blending and ssGBLUP than PA and DGV (Table 3.5). Usually, the average of the theoretical reliability is similar to the prediction accuracy, unless the trait is under strong selection (Bijma, 2012). In this study, (young) genotyped animals were selected based on PA and then genotyped. BVs of those animals are only expected to be unbiased, if those (young) animals were a random sample (Su et al., 2012), which is not the case here. In these cases, the predictive ability is lower because it accounts for selection. Table 3.5 showed that forward predictions, especially for LR (Table 3.3), were less accurate than expected, which could be due to the following reasons: (i) candidates are selected, (ii) DRPs might not be well estimated and/or (iii) the small amount of data caused inflation of prediction ability, which is quite common in small genotyped populations. Single-step GBLUP can avoid a bias due to preselection of young animals on Mendelian sampling variations and thus should be less biased (Su et al., 2012).

Since there might be a scope to improve predictive accuracy and reduce bias of genomic predictions with such small empirical data sets by various means, we exemplarily performed analyses putting **G** and **A** on a comparable scale by sophisticated weighting and scaling strategies (Christensen et al., 2012; Su et al., 2012; Misztal et al., 2013) as e.g. described in Vitezica et al. (2011) or Meuwissen et al. (2011). Furthermore, we used base allele frequencies (Gengler et al., 2007) instead of actual allele frequencies estimated from markers for calculating **G** to be analogous to **A**. In this approaches, we used both frequencies (base and actual) for calculating **G** and afterwards adjusted **G** according to Meuwissen et al. (2011). Predictive abilities from five-fold cross validation and forward prediction turned out to be almost equal to the ones obtained with **G** according to VanRaden et al. (2007) (Table 3.3). Vitezica et al. (2011) showed that tuning **G** helped to account for selective genotyping of animals and therefore should provide more accurate predictions, which we cannot confirm with these empirical data sets. According to Meuwissen et al. (2011), the correction for base level inbreeding does provide an improved genomic relationship matrix **G**. Since there was hardly any difference between predictive abilities obtained with the approach suggested by Meuwissen et al. (2011) compared to predictive abilities from the original calculation of **G**

(VanRaden et al., 2007), irrespective of which allele frequencies we used, scaling **G** as described in Meuwissen et al. (2011) led to no improvement. Accordingly, we also checked for correlations between off diagonal elements of **G** (VanRaden et al., 2007) and **A**, which turned out to be rather high for LW (0.82) and LR (0.88), which indicated a reasonable pedigree quality and provided no indication of major genotyping errors. Values for genomic inbreeding and pedigree-based inbreeding of genotyped individuals from the corresponding relationship matrices were both on average  $0.003 \pm 0.025$  (for LW) and  $0.009 \pm 0.023$  (for LR). In conclusion, an adjustment of **G** in order to substantially increase the predictive ability proved to be not successful for both small empirical data sets.

Single-step GBLUP has been applied in various livestock studies (e.g. Aguilar et al., 2010; Chen et al., 2011; Guo et al., 2015; Lourenco et al., 2015a; Lourenco et al., 2015b). In general, these studies showed that single-step methods produced higher predictabilities than pedigree-based and multi-step methods for varying production and fertility traits of different species. In summary, our results were largely in agreement with these previous studies.

## CONCLUSION

Estimation of accurate genomic BVs in populations with low number of genotyped animals is generally problematic for highly selected traits (Lourenco et al., 2014). The realized accuracy of genomic evaluation is dependent on many factors, including the quality of genomic data and the structure of the population (Miszta et al., 2013). Although, using the single-step approach to calculate more accurate breeding values is not clearly outperforming the blending approach in our data set, it might be a good alternative when a reasonable number of genotypes become available. In moderately sized pig breeding organizations, the decision which genomic method should be used in future needs to be monitored by ongoing validation assessments. Especially when some results do not conform with expectations, it is particularly important to regularly update the reference set. Single-step is relatively easy to apply with available software (Miszta et al., 2002) without a need to calculate DRPs or to construct an approximate selection-index for blending. Thus, it is recommended as the more consistent and robust approach to be used with typical data structures in moderately sized pig breeding organizations.

### ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of the project ‘pigGS’ by the Europaeischen Fonds fuer regionale Entwicklung (EFRE), the state of North Rhine-Westphalia, the project management Juelich, and the pig breeding organization for providing the data.

## REFERENCES

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. *J. Dairy Sci.* 93:743–752. doi: 10.3168/jds.2009-2730.
- Browning, S. R., and B. L. Browning. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* 81:1084–1097. doi: 10.1086/521987.
- Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations: Accuracy and selection. *J. Anim. Breed. Genet.* 129:345–358. doi: 10.1111/j.1439-0388.2012.00991.x.
- Calus, M. P. L., and R. F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J. Anim. Breed. Genet.* 124:362–368. doi: 10.1111/j.1439-0388.2007.00691.x.
- Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. *J. Anim. Sci.* 89:2673–2679. doi:10.2527/jas.2010-3555.
- Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:1–8. doi:10.1186/1297-9686-42-2.
- Christensen, O. F., P. Madsen, B. Nielsen, T. Ostensen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. *animal* 6:1565–1571. doi:10.1017/S1751731112000742.
- Cuyabano, B. C. D., G. Su, G. J. M. Rosa, M. S. Lund, and D. Gianola. 2015. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. *J. Dairy Sci.* 98:7351–7363. doi: 10.3168/jds.2015-9360.
- Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* 41:55. doi: 10.1186/1297-9686-41-55.
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *animal* 1:21. doi: 10.1017/S1751731107392628.
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, R. Thompson, D. Butler, M. Cherry, D. Collins, G. Dutkowski, S. A. Harding, K. Haskard, A. Kelly, S. G. Nielsen, A. Smith, A. P. Verbyla, S. J. Welham, and I. M. S. White. 2009. ASReml user guide release 3.0. VSN Int. Ltd Hemel Hempstead UK. <http://vsni.de/downloads/asreml/release3/UserGuide.pdf>. (Accessed February 19, 2014).
- Guo, X., O. F. Christensen, T. Ostensen, Y. Wang, M. S. Lund, and G. Su. 2015. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. *J. Anim. Sci.* 93:503–512. doi:10.2527/jas2014-8331.

- Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433–443. doi:10.3168/jds.2008-1646.
- Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41:51. doi: 10.1186/1297-9686-41-51.
- Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423–447.
- Huby, M., L. Canario, T. Tribout, J. C. Caritez, Y. Billon, J. Gogu  , and J. P. Bidanel. 2003. Genetic correlations between litter size and weights, piglet weight variability and piglet survival from birth to weaning in Large White pigs. In: *Proc. 54th Annu. Meet. Eur. Assoc. Anim. Prod.*, Roma, Italy. p. 362.
- Iba  ez-Escriche, N., and O. Gonzalez-Recio. 2011. Review. Promises, pitfalls and challenges of genomic selection in breeding programs. *Span. J. Agric. Res.* 9:404–413. doi: 10.5424/sjar/20110902-447-10.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656–4663. doi: 10.3168/jds.2009-2061.
- Legarra, A., Christensen, O.F., Aguilar, Ignacio, and Misztal, Ignacy. 2014. Single Step, a general approach for genomic selection. *Livest. Prod. Sci.* 166:54–65. doi:10.1016/j.livsci.2014.04.029.
- Lourenco, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J. I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. *J. Dairy Sci.* 97:1742–1752. doi: 10.3168/jds.2013-6916.
- Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015a. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* 47:56. doi: 10.1186/s12711-015-0137-1.
- Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and others. 2015b. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J. Anim. Sci.* 93:2653–2662. doi:10.2527/jas2014-8836.
- Meuwissen, T. H. E., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited: Unified approach to genomic selection. *J. Anim. Breed. Genet.* 128:429–439. doi: 10.1111/j.1439-0388.2011.00966.x.
- Mikshowsky, A. A., D. Gianola, and K. A. Weigel. 2016. Improving reliability of genomic predictions for Jersey sires using bootstrap aggregation sampling. *J. Dairy Sci.* 99:3632–3645. doi: 10.3168/jds.2015-10715.
- Misztal, I., S. E. Aggrey, and W. M. Muir. 2013. Experiences with a single-step genome evaluation1. *Poult. Sci.* 92:2530–2534. doi: 10.3382/ps.2012-02739.
- Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, D. H. Lee. 2002. BLUPF90 and related programs (BGF90). In: *Proc. 7th World Congr. Genet. Appl. Livest. Prod.*, Montpellier, France. p. 21–22.

- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81:559–575. doi: 10.1086/519795.
- Putz, A. M., F. Tiezzi, C. Maltecca, K. A. Gray, and M. T. Knauer. 2015. Variance component estimates for alternative litter size traits in swine. *J. Anim. Sci.* 93:5153–5163. doi:10.2527/jas2015-9416.
- Rothschild, M. F., and A. Ruvinsky. 1998. *The Genetics of the Pig*. CAB International, Oxon, UK.
- Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, O. F. Christensen, and M. S. Lund. 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J. Dairy Sci.* 95:909–917. doi: 10.3168/jds.2011-4804.
- Tier, B., and K. Meyer. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed. Genet.* 121:77–89. doi: 10.1111/j.1439-0388.2003.00444.x.
- VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91:4414–4423. doi: 10.3168/jds.2007-0980.
- VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24. doi:10.3168/jds.2008-1514.
- Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genet. Res.* 93:357–366. doi: 10.1017/S001667231100022X.

**APPENDIX****APPENDIX 3.1: RANGE OF PREDICTIVE ABILITIES FROM BOOTSTRAPPING CANDIDATES (FORWARD PREDICTION;  $\pm$  SE OF CORRELATION COEFFICIENT)**

	Landrace				Large White			
	Bootstrapped samples			Full set	Bootstrapped samples			Full set
	Min	Mean	Max		Min	Mean	Max	
<b>GBLUP<sup>1</sup></b>	0.00 $\pm$ 0.11	0.08 $\pm$ 0.11	0.25 $\pm$ 0.11	0.08 $\pm$ 0.14	0.25 $\pm$ 0.15	0.43 $\pm$ 0.13	0.60 $\pm$ 0.10	0.43 $\pm$ 0.11
<b>Blending</b>	0.05 $\pm$ 0.11	0.18 $\pm$ 0.10	0.33 $\pm$ 0.10	0.19 $\pm$ 0.13	0.24 $\pm$ 0.15	0.39 $\pm$ 0.13	0.57 $\pm$ 0.11	0.40 $\pm$ 0.12
<b>ssGBLUP<sup>2</sup></b>	0.00 $\pm$ 0.11	0.07 $\pm$ 0.11	0.22 $\pm$ 0.11	0.18 $\pm$ 0.13	0.27 $\pm$ 0.15	0.44 $\pm$ 0.13	0.62 $\pm$ 0.09	0.45 $\pm$ 0.11

<sup>1</sup>GBLUP = genomic BLUP.<sup>2</sup>ssGBLUP = single-step genomic BLUP.

## **4<sup>th</sup> CHAPTER**

### **Estimation of genetic parameters for reproduction traits in dam lines of a German pig breeding organization**

A. Fangmann\*, R.A. Sharifi\*, J. Heinkel†, H. Schrade†, H. Simianer\*

\* University of Goettingen, Department of Animal Sciences, Center for Integrated Breeding Research, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany

†Bildungs- und Wissenszentrum Boxberg- Schweinehaltung, Schweinezucht-, 97944 Boxberg, Germany

**ABSTARCT**

Records of a German pig breeding organization comprising 78,749 litters from 22,116 dams from four different dam lines born from 2010 to 2016 were used for the analysis (data set 1). Records of six consecutive parities of 3,673 dams with 22,038 litters beginning with the first parity were included in data set 2. A third data set was based on piglets and comprised of 1,780,753 piglets from 36,612 dams (data set 3). A comparison of different models with variable selection was conducted for estimation of variance components. Animal models with repeated measurements as well as bivariate models were used to estimate variances and covariances for the animal genetic, phenotypic and permanent environmental component. Based on this genetic parameters, heritabilities and genetic correlations were derived for the following reproductive traits: number born alive (NBA), farrowing interval (FI), number weaned (NOW), mothering ability (MA) and piglet survival (PS). The goodness of the fit of the corresponding models can be determined by their Akaike Information Criterion (AIC). Based on different single-trait repeatability animal models, estimates of heritability range between 0.11 and 0.14 for NBA; between 0.10 and 0.15 for FI; between 0.06 and 0.12 for NOW; between 0.08 and 0.15 for MA and between 0.08 and 0.10 for PS, respectively. Estimates of  $h^2$  obtained from the bivariate analysis were 0.12 for NBA; between 0.11 and 0.15 for FI; between 0.06 and 0.14 for NOW and between 0.09 and 0.12 for MA, respectively. Genetic correlations between NBA and NOW were on average 0.80, while genetic correlations varied between -0.30 and -0.45 for NBA and MA, 0.01 and 0.05 for FI and NOW, 0.09 and 0.10 for FI and MA and -0.05 and 0.04 for FI and NBA, respectively. Although the estimates of  $h^2$  for reproductive traits were low, traits are heritable and genetic variances are viable for selection.

**Key words:** genetic parameter, pig, reproduction traits

## INTRODUCTION

Improvement of litter size is an economical important breeding goal in swine and included in most breeding schemes (Rydhmer, 2000). Under commercial management, the number of piglets born alive is reliably recorded and the trait of interest for selection in dam lines (Irgang et al., 1994). However, number of piglets born alive or born alive per litter and sow often is still the only reproduction trait in the selection process. Another interesting trait could be piglet survival, which can be determined in two ways: first, piglet survival itself, where it is independent of which sow the piglet was nursed or weaned and second, mothering ability or number of weaned piglets of a sow, which include the number of adopted piglets but exclude the number of piglets transferred to another sows.

Therefore, knowledge of genetic parameters of reproduction traits is necessary to estimate accurate breeding values, to combine different fertility traits in selection and to optimize breeding schemes (Roehe and Kennedy, 1995). The objective of this study is to provide estimates of genetic parameters for reproductive traits and to evaluate an appropriate model for those traits for the routine breeding value estimation of a German pig breeding organization.

## MATERIAL AND METHODS

### Animals and Data

Farrowing records of four dam lines from were obtained from a German pig breeding organization. Records of litters born from 2010 to 2016 were used in this study. Traits of interest were number of piglets born alive per litter of a sow (NBA), farrowing interval (FI; calculated according to ten Napel et al., 1995) defined as the interval between weaning and successful insemination, number of weaned piglets per litter of a sow (NOW; including the number of adopted piglets and excluding the cross-fostered piglets), mothering ability (MA) of a sows defined as NOW divided by the sum of weaned and potentially weaned piglets (ptNOW, equals sum of weaned and not weaned piglets of a sow) and piglet survival (PS; related to biological litter) referred to as piglet of a sow, that survived the nursing period, independent of which sow (nurse or biological dam) it was nursed, respectively. A quality control was performed using the following limits: pregnancy time: 81-149d, farrowing interval: 18-39d, age at first farrowing: 271-549d, interval between farrowing: 111-279d, time between insemination and fertilization: 0-59d, age of sow at farrowing: 271-2799d, number of piglets born alive: 1-30 and number of nursed piglets: 3-20.

In total, 78,749 litters from 22,116 dams from four different dam lines remained for analysis (data set 1): German Landrace (n=45,069), German Large White (n=32,856), Schwaebisch Haellisch (n=607) and Leicoma (n=217). Records of six consecutive parities of sows (3,673 dams) beginning with the first parity formed data set 2, comprising only those sows that had complete records for the first sixth parities: German Landrace (n= 12,678), German Large White (n= 9162), Schwaebisch Haellisch (n=138) and Leicoma (n=60). As boar of litter, 13 different breeds were used to produce pure-bred piglets within dam line and crossbred piglets with sire breeds such like Pietrain (n=14,667 in data set 1 and n=3,934 in data set 2) and Duroc (n=2,335 in data set 1 and n=570 in data set 2). Data set 3 was based on piglet data and was used to estimate the piglet survival. In total, 1,780,753 piglets from 36,612 dams (German Landrace (n=20,696), German Large White (n=15,362), Schwaebisch Haellisch (n=346) and Leicoma (n=208), respectively) were used for this analysis.

### Statistical Analysis

Variance components were estimated using the software ASReml (Gilmour et al., 2009 ; Version 3.0 and 4.1). All analyses were completed with single and bivariate trait models. Bivariate analyses comprised two different approaches: first, to partly account for bias by involuntary or directional selection from parity to parity, the first parity of a certain trait was considered as different trait compared to the following five parities (Rothschild et al., 1979; Alfonso et al., 1997) and secondly, two different traits were used in the analysis for all parities.

An animal model with repeated measurements was used in which all traits were treated as trait of the dam. The statistical model used to describe the data was as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \mathbf{e} \quad (\text{I})$$

where  $\mathbf{y}$  represent the vector of observations for the different dam traits (NBA, MA, NOW, FI) for all parities (I, data set 1), for the first parity and for the second to sixth parity (data set 2), respectively,  $\mathbf{b}$  the vector of fixed effects,  $\mathbf{a}$  the vector of random additive genetic effects of animals  $\sim N(0, \mathbf{A}\sigma_a^2)$ ,  $\mathbf{p}$  the vector of permanent environment effects  $\sim N(0, \mathbf{I}\sigma_{pe}^2)$  and  $\mathbf{e}$  the vector of residual effects  $\sim N(0, \mathbf{I}\sigma_e^2)$ . Matrices  $\mathbf{X}$ ,  $\mathbf{Z}$  and  $\mathbf{W}$  are incidence matrices,  $\mathbf{A}$  is the additive genetic relationship matrix and  $\sigma_a^2$ ,  $\sigma_{pe}^2$  and  $\sigma_e^2$  are direct, permanent environment and residual variances, respectively.

The following two bivariate animal models with repeated measurements were used:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{p}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \quad (\text{II})$$

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \quad (\text{III})$$

where  $\mathbf{y}_1$  and  $\mathbf{y}_2$  represent observations in the different parities for the different traits (II,  $\mathbf{y}_1$  = first parity and  $\mathbf{y}_2$  = second to sixth parity, data set 2) or different combinations of the different traits for all parities (III, NBA, NOW, FI, MA, data set 2). The vectors  $\mathbf{b}_1$  and  $\mathbf{b}_2$  represent the fixed effects,  $\mathbf{a}_1$  and  $\mathbf{a}_2$  are the random additive genetic effects,  $\mathbf{p}_1$  and  $\mathbf{p}_2$  are the permanent environment effects and  $\mathbf{e}_1$  and  $\mathbf{e}_2$  the residual effect for trait 1 and trait 2, respectively. Matrices  $\mathbf{X}_f$ ,  $\mathbf{Z}_f$  and  $\mathbf{W}_f$  ( $f=1, 2$ ) are the incidence matrices. It assumed that permanent environmental effect, animal and error are independently distributed with mean zero and variance:

$$\mathbf{V} \begin{bmatrix} \mathbf{a} \\ \mathbf{pe} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \otimes \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \otimes \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \otimes \mathbf{I} \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} \sigma_{ait}^2 & \sigma_{aij} \\ \sigma_{aji} & \sigma_{ajj}^2 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} \sigma_{peti}^2 & \sigma_{peij} \\ \sigma_{peji} & \sigma_{pejj}^2 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \sigma_{eii}^2 & \sigma_{eij} \\ \sigma_{eji} & \sigma_{ejj}^2 \end{bmatrix},$$

where  $\otimes$  is the Kronecker product,  $\mathbf{I}$  is the identity matrix equal to number of observations,  $\mathbf{A}$  the pedigree relationship matrix,  $\mathbf{G}$  ( $\mathbf{Q}$  and  $\mathbf{R}$ ) is the variance covariance matrix of random animals effect (permanent environmental and residual effects). Variances and covariances for trait 1 ( $i$ ) and trait 2 ( $j$ ) are defined as follows: additive genetic variance for trait 1 ( $\sigma_{ait}^2$ ) and 2 ( $\sigma_{ajj}^2$ ) and additive genetic covariance between traits ( $\sigma_{aji} = \sigma_{ajit}$ ); permanent environmental variance for trait 1 ( $\sigma_{peti}^2$ ) and 2 ( $\sigma_{pejj}^2$ ) and permanent environmental covariance between traits ( $\sigma_{peji} = \sigma_{pejit}$ ), and residual variance for trait 1 ( $\sigma_{eii}^2$ ) and 2 ( $\sigma_{ejj}^2$ ) and residual covariance between traits ( $\sigma_{eji} = \sigma_{ejit}$ ), respectively. Due to computational limitations and no convergence, we did not run or represent results of a multivariate analysis. For the estimation of variance components of PS in data set 3, a piglet-based logistic animal model with repeated measurements (I) and binary coding for piglet survival at weaning (0=not survived and 1=survived) was used.

### **Fixed and random effects and covariables**

For all reproduction traits, various fixed effects and covariables were used to evaluate the appropriate model for each trait. All these effects were adopted from the used models of the pig breeding company. All models included a fixed effect for herd year season (HYS) and breed of the dam (breed,  $n=4$ ) as fixed effects. Additionally, the following fixed effects could be used as variable in the model: boar of the litter ( $n=2,625$  for data set 1,  $n=1,756$  for data set 2) and parity. Age of sow at farrowing (linear and quadratic), farrowing interval, nursing time (linear and quadratic), number of weaned piglets, litter size, competing piglets (number of competing piglets at the nurse sow/teat) and number of cross-fostered piglets can be used or selected as covariable for the models.

Selected variables and corresponding models for estimating the variance components of NBA are shown in Appendix 4.1 and for FI (NOW, MA, PS) in Appendix 4.2 (4.3, 4.4, 4.5), respectively. For the estimation of genetic correlations between the reproductive traits, various two-trait animal models were used as shown in Appendix 4.6.

### **Model comparison**

To receive an estimator of the relative quality of statistical models for a given data set, we calculated the Akaike Information Criterion (AIC; Akaike, 1974; Gilmour et al., 2009) from ASReml. AIC was used to assess the quality of each model and goodness of the fit, respectively. The model with the lowest AIC was considered optimal.

## **RESULTS AND DISCUSSION**

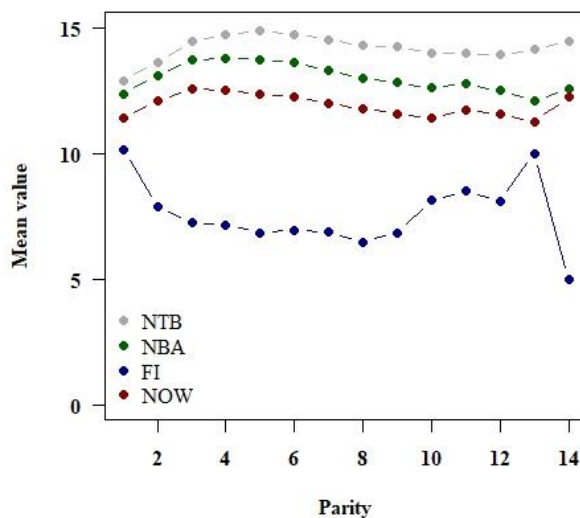
### **Descriptive statistics**

Means and standard deviations of the different traits and two data sets are presented in Table 4.1. Compared to literature, means for NBA (13.19 and 13.52) and NOW (12.07 and 12.39) were as expected for both data sets. FI was as desired low and MA on an acceptable high level (0.92).

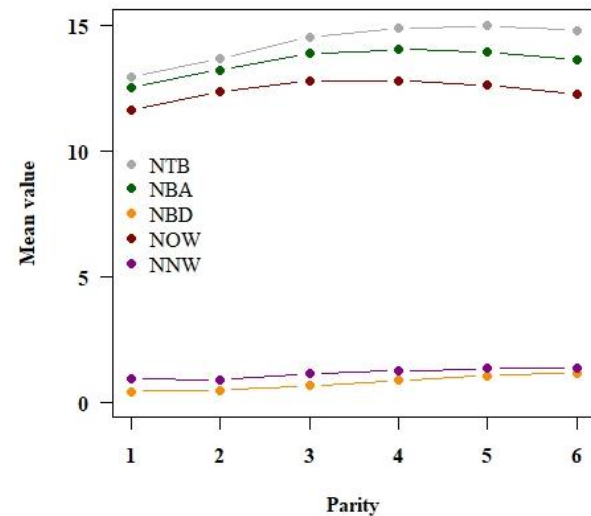
**Table 4.1.** Number of observation (n), means ( $\bar{x}$ ) and standard deviations ( $\pm$  SD) for number of piglets born alive (NBA), farrowing interval (FI), piglet survival (PS) and number of weaned piglets of a sow (NOW) and mothering ability (MA) for data set 1 and data set 2.

Trait	Data set 1 (n= 78,749)		Data set 2 (n=22,038)	
	$\bar{x}$	$\pm$ SD	$\bar{x}$	$\pm$ SD
<b>NBA</b>	13.19	2.92	13.52	2.82
<b>FI</b>	6.20	3.23	5.87	2.77
<b>NOW</b>	12.07	2.72	12.39	2.63
<b>MA</b>	0.92	0.11	0.92	0.10
<b>PS</b>	12.08	2.72	12.38	2.62

On average, total number of born piglets (TNB), NBA and NOW increased from parity to parity (approx. until fifth parity) and slightly decreased afterwards. FI between litters decreased until the eight litter and increased afterwards (Figure 1). From the first to the second parity, there was a greater improvement for every trait relative to the other parities (Figure 1), which lead among other things to analyze and compare the first and the following parities as different traits in some models, such as Boesch et al. (1999) or Hanenberg et al. (2001).



**Figure 1:** Mean value for different maternal traits as a function of parity for data set 1.



**Figure 2:** Mean value for different maternal traits as a function of parity for data set 2.

In the first parity, roughly one piglet less was born and weaned than in the following five parities (data set 2, Figure 2). The amount of stillborn piglets (NBD) increased with parity number from 3.3% (first parity) to 7.8 % (sixth parity). The number of losses at weaning (NNW) increases from 0.91 piglet in the first parity up to 1.34 piglets in the sixth parity.

However, the number of weaned piglets increased on average by approx. 1 piglet per litter from first to sixth parity (Figure 2), because of the fact that in total more piglets were born.

### Variance components and heritability

Estimates of genetic and phenotypic variance, permanent environment, heritability, repeatability and genetic and phenotypic correlation are reported for the different traits and models in Table 4.2 to 4.6, respectively.

For NBA (Table 4.2), heritabilities were generally small (0.14 for model 1) but in the range of literature reports (Southwood and Kennedy, 1990; Roehe and Kennedy, 1995; Chen et al., 2003; Huby et al., 2003; Putz et al., 2015). Splitting NBA into two different traits led to a different magnitude of heritability for NBA1 (approx. 0.14) and NBA2-6 (approx. 0.11), irrespective of which model 2 was used. Different magnitudes have also been reported by Alfonso et al. (1997). Genetic correlations between NBA1 and NBA2-6 in the two-trait analysis (model 3) are low (0.55), which suggested treating NBA1 and NBA2-6 as genetically different traits, as suggested by Rothschild et al. (1979).

**Table 4.2:** Genetic variance ( $\sigma_a^2$ ), phenotypic variance ( $\sigma_p^2$ ), permanent environment ( $\sigma_{pe}^2$ ), heritability ( $h^2$ ), repeatability ( $w^2$ ), genetic ( $r_g$ ) and phenotypic ( $r_p$ ) correlation of different models for number of piglets born alive (NBA).

Model	Trait	$\sigma_a^2$	$\sigma_p^2$	$\sigma_{pe}^2$	$h^2$	$w^2$	$r_g$	$r_p$
<b>Model 1</b>								
1.1	NBA	1.111	7.853	0.364	0.141	0.180		
1.2	NBA	0.109	7.853	0.365	0.141	0.180		
<b>Model 2</b>								
2.1	NBA1	0.949	6.711		0.141			
2.2	NBA2-6	0.782	7.347	0.558	0.106	0.182		
<b>Model 3</b>								
3.1	NBA1	0.796	6.835		0.116			
	NBA2-6	0.885	7.322	0.397	0.121	0.175	0.550	0.086

Heritabilities for FI (Table 4.3) were low (0.11 to 0.14) but in agreement with literature (ten Napel et al., 1995; Hanenberg et al., 2001). Hanenberg et al. (2001) found heritabilities for FI of 0.14 (first parity) and 0.07 (following parities). Compared to results of ten Napel et al. (1995), heritability estimates of FI in our study were lower. The difference might arise due to the use of practical on-farm data in this study compared to the selection experiment data of

ten Napel et al. (1995). Genetic correlations between FI1 and FI2-6 are moderate (0.72). Genetic and phenotypic variances were generally low but even lower for FI2-6.

**Table 4.3:** Genetic variance ( $\sigma_a^2$ ), phenotypic variance ( $\sigma_p^2$ ), permanent environment ( $\sigma_{pe}^2$ ), heritability ( $h^2$ ), repeatability ( $w^2$ ), genetic ( $r_g$ ) and phenotypic ( $r_p$ ) correlation of different models for farrowing interval (FI).

Model	Trait	$\sigma_a^2$	$\sigma_p^2$	$\sigma_{pe}^2$	$h^2$	$w^2$	$r_g$	$r_p$
<b>Model 1</b>								
1.1	FI	1.311	10.051	0.236	0.130	0.154		
1.2	FI	1.316	10.039	0.218	0.131	0.153		
<b>Model 2</b>								
2.1	FI1	1.397	11.273		0.124			
2.2	FI2-6	0.711	6.457	0.000	0.110	0.110		
<b>Model 3</b>								
3.1	FI1	1.576	11.212		0.141			
	FI2-6	0.718	6.460	0.000	0.111	0.111	0.715	0.089

Heritabilities for NOW (Table 4.4) for the first parity (NOW1) were higher (approx. 0.12) than for the following parities (NOW2-6, approx. 0.06). One reason could be due to the low genetic variances in the subsequent parities. Estimates of heritability for NOW in this study were in agreement with estimates reported by Southwood and Kennedy (1990), Roehe and Kennedy (1995) and Putz et al. (2015) but slightly higher (approx. +0.07) than estimates reported by Chen et al. (2003) and Huby et al. (2003), respectively. Genetic correlations between NOW1 and NOW2-6 from two-trait analysis (model 3) were moderate with 0.58.

**Table 4.4:** Genetic variance ( $\sigma_a^2$ ), phenotypic variance ( $\sigma_p^2$ ), permanent environment ( $\sigma_{pe}^2$ ), heritability ( $h^2$ ), repeatability ( $w^2$ ), genetic ( $r_g$ ) and phenotypic ( $r_p$ ) correlation of different models for number of weaned piglets (NOW).

Model	Trait	$\sigma_a^2$	$\sigma_p^2$	$\sigma_{pe}^2$	$h^2$	$w^2$	$r_g$	$r_p$
<b>Model 1</b>								
1.1	NOW	0.725	6.685	0.261	0.109	0.148		
1.2	NOW	0.729	6.686	0.258	0.109	0.148		
<b>Model 2</b>								
2.1	NOW1	0.707	6.121		0.116			
2.2	NOW2-6	0.354	6.063	0.487	0.059	0.139		
<b>Model 3</b>								
3.1	NOW1	0.715	6.124		0.117			
	NOW2-6	0.357	6.083	0.405	0.059	0.125	0.575	0.064

For all models for MA (Table 4.5), genetic (approx. 0.001) and phenotypic (approx. 0.01) variances and heritability estimates (range between 0.08 and 0.15) were low, respectively. Compared to Hanenberg et al. (2001), values of MA in this study was higher. Hanenberg et al. (2001) concluded that mothering ability is a combination of real mothering ability and piglet vitality (quality). In most herds, sows with better mothering abilities nurse weak piglets of sows with low mothering abilities. Since no correction for quality of the piglet was made in Hanenberg et al. (2001) analysis, estimates of heritability for MA are expected to have a downward bias (Hanenberg et al., 2001). Genetic correlations between MA1 and MA2-6 from two-trait analysis were high (0.73 for model 3.1 and 0.82 for model 3.2).

**Table 4.5:** Genetic variance ( $\sigma_a^2$ ), phenotypic variance ( $\sigma_p^2$ ), permanent environment ( $\sigma_{pe}^2$ ), heritability ( $h^2$ ), repeatability ( $w^2$ ), genetic ( $r_g$ ) and phenotypic ( $r_p$ ) correlation of different models for mothering ability (MA).

Model	Trait	$\sigma_a^2$	$\sigma_p^2$	$\sigma_{pe}^2$	$h^2$	$w^2$	$r_g$	$r_p$
<b>Model 1</b>								
1.2	MA	0.0009	0.0090	0.0003	0.0996	0.1333		
<b>Model 2</b>								
2.1	MA1	0.0010	0.0103		0.1012			
2.2	MA2-6	0.0013	0.0092	0.0001	0.1456	0.1522		
2.3	MA2-6	0.0007	0.0080	0.0003	0.0813	0.1250		
<b>Model 3</b>								
3.1	MA1	0.00122	0.01038		0.1177		0.7372	0.0919
	MA2-6	0.00094	0.00869	0.00009	0.1076	0.1185		
3.2	MA1	0.00108	0.00980		0.1098		0.8205	0.0871
	MA2-6	0.00076	0.00809	0.00017	0.0936	0.1150		

For PS (Table 4.6 and 4.7), different univariate variants of model 1 were compared. Due to large amount of data, data structures and computational limitations and convergence problems, bivariate models were not considered. Therefore, genetic and phenotypic correlations could not be estimated. Additionally, possible effects of cross-fostering were taken into account and each model was calculated with a cross-fostering effect as covariables (Table 4.7). Due to the fact, that we had approx. 4,000 cross-fostered piglets (roughly 4% of the data set), no major differences between models considering cross-fostering or not arose. Irrespective of cross-fostering, heritabilities (ranging from 0.089 to 0.103), genetic (varying between 0.304 and 0.402) and phenotypic (varying between 3.749 and 3.939) variances did not vary greatly between corresponding models. Putz et al. (2015) obtained estimates of heritabilities for the number of piglets born alive according to the biological litter of 0.11

(0.12) for Large White (Landrace), which are slightly higher than estimates of PS reported from this data. Obtained variance are also higher (Putz et al., 2015) than presented here.

Another restriction, which might be considered and recorded for future analyses, was the aspect of which piglet (light or heavy piglet) was cross-fostered, which was reported to play an important role (Knol et al., 2002). To conclude, using cross-fostering as fixed effect in the model did not improve the estimations. Putz et al. (2014) suggested to account for cross-fostering if it occurs in large quantities. However, PS was lowly heritable and its genetic variance was large enough to enable selection on that trait.

**Table 4.6:** Genetic variance ( $\sigma_a^2$ ), phenotypic variance ( $\sigma_p^2$ ), permanent environment ( $\sigma_{pe}^2$ ), heritability ( $h^2$ ) and repeatability ( $w^2$ ) for piglet survival (PS) without cross-fostering effect.

Model	$\sigma_a^2$	$\sigma_p^2$	$\sigma_{pe}^2$	$h^2$	$w^2$
<b>Model 1</b>					
<b>1.1</b>	0.402	3.939	0.247	0.102	0.165
<b>1.3</b>	0.323	3.835	0.222	0.084	0.142
<b>1.5</b>	0.304	3.783	0.189	0.080	0.130
<b>1.6</b>	0.335	3.749	0.125	0.089	0.123

**Table 4.7:** Genetic variance ( $\sigma_a^2$ ), phenotypic variance ( $\sigma_p^2$ ), permanent environment ( $\sigma_{pe}^2$ ), heritability ( $h^2$ ) and repeatability ( $w^2$ ) for piglet survival (PS) with cross-fostering effect.

Model	$\sigma_a^2$	$\sigma_p^2$	$\sigma_{pe}^2$	$h^2$	$w^2$
<b>Model 1</b>					
<b>1.1</b>	0.402	3.939	0.247	0.102	0.165
<b>1.3</b>	0.322	3.835	0.223	0.084	0.142
<b>1.5</b>	0.299	3.780	0.191	0.079	0.130
<b>1.6</b>	0.338	3.750	0.123	0.090	0.123

In general, estimated variance components depend on the population structure and data set which aggravates comparison of differences in variance components with literature estimates difficult. Estimates of the heritabilities of sow reproductivity might also be biased due to selection. For all traits, lower estimates of heritabilities were obtained in the subsequent parities (second to sixth), which might have resulted from reduced additive genetic variances due to culling sows with low prolificacy, as was also found by Irgang et al. (1994). Selection and culling occurs especially after the first farrowing. This might explain the absolute difference in the heritabilities between first parity and second to sixth parity. However,

absolute differences were small ( $<0.1$ ), which indicates that preselection and culling occur but have a minor impact.

Another point was the maternal genetic effect. Southwood and Kennedy (1990), Alfonso et al. (1997), Crump et al. (1997) and Chen et al. (2014) reported minor differences in estimates of heritability with and without accounting for maternal genetic effects. Chen et al. (2003) reported that the percentage of the variance accounted for maternal genetic effect ranged from 0.00 to 0.02 for NBA and NOW. Furthermore, they reported that only little changes in the ranking of the sows on estimated breeding values occurred. Boesch et al. (1999) estimated a small portion of maternal genetic variance compared to the total variance and concluded to exclude maternal genetic effects from the model. Thus, an animal model with repeated measurements for estimation of permanent environmental effects was recommended for litter size (NBA) (Boesch et al., 1999). However, Boesch et al. (1999) and Roehe (1998) recommended including maternal traits (effects) like individual birth weight or weight at weaning into the analysis, especially when cross-fostering was practiced. With respect to the analyzed traits, considering birth weights or weights at weaning in our analysis would have been desirable, but unfortunately those weights were not available. Rydhmer (2000) included that biological (maternal) and fostering (nurse) dam should be included in the model, when cross-fostering is heavily applied. Due to the fact, that the proportion of cross-fostering in this data set was rather minor (approx. 4 %), maternal (or nurse) genetic effects are not considered in the presented models.

The size of genetic correlations between the first and second to sixth parity was important to find appropriate models for selection programs for improvement of reproductive traits (Irgang et al., 1994). Generally, genetic correlation of approx. 1 between first and second to sixth parity suggest that a single-trait model with repeated measurements (I) should be used for estimation of breeding values. Contrary to this, low genetic correlations suggest that first and second to sixth parity should be treated as two different traits and that bivariate animal-models (like model III) should be used for estimation. Accordingly, for FI ( $r_g = 0.71$ ) and MA ( $r_g > 0.73$ ) with moderate to high genetic correlations a single-trait models should be used. For NBA ( $r_g = 0.55$ ) and NOW ( $r_g = 0.58$ ), bivariate models in which first and second to sixth parity are treated as different traits should be considered as appropriate. However, a current limitation of the bivariate model could be increasing computing costs (Irgang et al., 1994) which should be taken into account for commercial use.

**Model comparison**

Results for AIC for each trait and corresponding models are reported in Appendix 4.1-4.5. To assess which model and corresponding effects were optimal for which trait; AIC had to be compared within each trait, model and data set, respectively. For all models, irrespective of single-trait or two-trait model, HYS and Breed are always considered as a fixed effect and for this reason are not discussed in detail below. Only additional effects (from variable selection) are mentioned in the following. For all models and traits, including boar as random effect did not improve the goodness of the model (results not shown).

For NBA (Appendix 4.1), a single-trait model considering parity and age of sow at farrowing turned out to be optimal. In general, considering boar did not improve the fit of the model (results not shown). Incorporating farrowing interval into the two-trait analysis did not improve the fit of the model (result not shown), while age of sow at farrowing seems to be important for the analysis.

For FI (Appendix 4.2), considering nursing time (linear and quadratic) as fixed effect improved the fit of the model more than including age of the sow at farrowing (quadratic), especially for the bivariate analysis (result not shown). For the single-trait models, NOW should be included in the analysis to improve the fit of the model.

For an estimation of NOW (Appendix 4.3), boar, age of farrowing and nursing time should be included as fixed effects in presented models.

In the single-trait models and for the two-trait model for MA (Appendix 4.4), nursing time and potentially weaned piglets, parity and boar improve the fit. Age of sow at farrowing seems to be important for the analysis of second to sixth parity (model 2.2).

Considering cross-fostering for the estimation of variance components for PS turn out to be generally slightly better. Within each analysis, irrespective of cross-fostered or not cross-fostered, model 1.1 was the optimal model. Including litter size or competing piglets into the model, estimation of variance components did not benefit, which was rather surprisingly considering the biological background of those two effects for PS.

**Genetic correlation between traits**

Estimates of genetic ( $\sigma_a^2$ ) and residual ( $\sigma_e^2$ ) variance, heritability ( $h^2$ ) and genetic correlations ( $r_g$ ) between two traits are shown in Table 4.9, respectively. For both models, genetic

correlation between NBA and NOW were high ( $r_g > 0.78$ ) and favorable. An increase in piglets born alive led to an increase in piglets weaned. High genetic correlations between NBA and NOW are also reported by McCarter et al. (1987), Kaplon et al. (1991a) and Huby et al. (2003). Compared to this study, Chen et al. (2003) and Putz et al. (2015) reported lower correlations between NBA and NOW but concluded that this was due to the fact of cross-fostering.

Genetic correlations between NBA and MA were negative and moderate with -0.45 and -0.30, respectively. An increase in piglets born alive led to a decrease in the mothering abilities of a sow, which is unfavorable. Thus, the more pigs a sow had to suckle and feed the harder for the sow to raise all piglets. However, this moderate negative correlation could also occur due to higher piglet mortality, which is quite difficult to investigate (Knol et al., 2002). Along with that, individual birth weight is one of the most important factors for piglet survival (Knol et al., 2002). For instance, individual birth weights might be rather small for piglets in large litters compared to piglets in smaller litters (Knol, 2001). The Pearson correlation between total born (born alive) and not weaned piglets was 0.39 (0.38), indicating that piglet mortality should not be ignored in this context. However, this undesirable negative genetic correlation was also found by Hanenberg et al. (2001).

Genetic correlations between FI and NOW (0.01 and 0.05), and MA (0.09 and 0.10) were marginal positive for all models and economically unfavorable. Genetic correlations between FI and NBA were around zero (0.04 and -0.05). An increase in born alive and weaned piglets might increase the farrowing interval between weaning and successfully gestation of the sow. Accordingly, fewer litters per sow and year can be realized and less piglets will be weaned. However, the farrowing interval between parities was quite acceptable ranging between 6.46 and 10.12 days.

**Table 4.9:** Genetic variance ( $\sigma_a^2$ ), residual variance ( $\sigma_e^2$ ), heritability ( $h^2$ ) and corresponding standard error (SE) and genetic correlation ( $r_g$ ) and corresponding standard error (SE) between traits for different bivariate models.

Model	Trait	$\sigma_a^2$	$\sigma_e^2$	$h^2$ (SE)	$r_g$ (SE)
NOW and NBA					
1.1	NOW	0.384	5.413	0.063 (0.011)	0.781 (0.042)
	NBA	0.563	5.958	0.081 (0.012)	
1.2	NOW	0.475	5.503	0.075 (0.012)	0.829 (0.031)
	NBA	0.735	6.102	0.101 (0.013)	
MA and NBA					
1.1	MA	0.001	0.008	0.095 (0.012)	-0.449 (0.088)
	NBA	0.575	5.960	0.083 (0.013)	
1.2	MA	0.002	0.008	0.155 (0.015)	-0.288 (0.078)
	NBA	0.712	6.103	0.098 (0.013)	
NOW and FI					
1.1	NOW	0.371	5.406	0.061 (0.011)	0.011 (0.112)
	FI	0.835	6.529	0.113 (0.013)	
1.2	NOW	0.369	5.404	0.060 (0.011)	0.049 (0.112)
	FI	0.822	6.525	0.112 (0.013)	
MA and FI					
1.1	MA	0.001	0.008	0.088 (0.012)	0.093 (0.095)
	FI	0.835	6.517	0.145 (0.013)	
1.2	MA	0.001	0.008	0.107 (0.013)	0.102 (0.091)
	FI	0.768	6.593	0.104 (0.128)	
NBA and FI					
1.1	NBA	0.544	5.952	0.078 (0.012)	0.041 (0.104)
	FI	0.821	6.526	0.112 (0.013)	
1.2	NBA	0.751	6.103	0.103 (0.014)	-0.053 (0.094)
	FI	0.780	6.520	0.106 (0.013)	

### Implications and final remarks

This report of genetic estimations for different reproductive traits was based on data from dam lines of a German pig breeding organization. Estimates of additive genetic variance, heritability and genetic correlation for the analyzed traits indicate that the amount of additive genetic variation available for selection is large enough and that the investigated reproductive traits can be improved through selection. Trends observed from the data show that total number of born piglets increased by approx. 0.41 piglets per year and number of born alive (weaned) piglets by approx. 0.29 (0.25) piglets per year (results not shown).

When genetic correlation between parities is clearly distinct from 1, bivariate animal models should be used for estimation of breeding values, in which the first parity and subsequent parities should be considered as different traits (e.g. for the traits NBA and NOW). However, computing time should be taken into account in commercial use. Alternatively, single-trait models with repeated measurements can be used for routine breeding value estimations on this data set. Single-trait models should also be used for piglet survival, farrowing interval and mothering ability.

Taken the increasing amount of piglets to be weaned into account, piglet survival and mothering ability of a sow becomes more important and should be further addressed. Future investigations on reproductive traits can also focus on including other genetic effects into the models like maternal genetic effect, a genetic effect for the piglet itself and/or a genetic effect of the nurse sow for estimation of variance components. Another interesting issue is the consideration of individual birth weights or weights at different stages during the nursing period to further evaluate different aspects of reproduction traits. However, the above mentioned studies will require precise recording of cross-fostering, individual weights, a large data set and powerful statistical analysis.

## APPENDIX

## Appendix 4.1. Different models for estimating the variance components of NBA

Mode l	Trait	Fixed effect				Covariable		AIC
		HYS <sup>1</sup>	Breed	Boar	Parity	Age of sow at farrowing	Farrowin g interval	
Model 1								
1.1	NBA	x	x		x	x	x	15,439.62
1.2	NBA	x	x		x	x		15,429.70
Model 2								
2.1	NBA1	x	x			x	x	10,663.14
2.2	NBA2-6	x	x		x	x	x	13,950.22
Model 3								
3.1	NBA1	x	x	x		x		4,541.76
	NBA2-6	x	x	x	x	x		

<sup>1</sup> HYS = herd year season.<sup>2</sup> NBA1-6 = number of piglets born alive for all available parities.<sup>3</sup> NBA1 = number of piglets born alive for the first parity.<sup>4</sup> NBA2-6 = number of piglets born alive for second to sixth parity.

**Appendix 4.2.** Different models for estimating the variance components of FI

Model	Trait	Fixed effect			Covariable				NOW	AIC
		HYS	Breed	Parity	Age of sow at farrowing		Nursing time			
					linear	quadratic	linear	quadratic		
Model 1										
1.1	FI <sup>1</sup>	x	x	x	x	x		x		15,929.60
1.2	FI	x	x	x	x		x	x		15,834.66
Model 2										
2.1	FI1	x	x	x	x		x	x	x	12,492.32
2.2	FI2-6	x	x		x		x	x	x	12,197.72
Model 3										
3.1	FI1	x	x		x		x	x		574.08
	FI2-6	x	x	x	x		x	x		

<sup>1</sup> FI1-6 = farrowing interval for all available parities.<sup>2</sup> FI1 = farrowing interval for the first parity.<sup>3</sup> FI2-6 = farrowing interval for second to sixth parity.

**Appendix 4.3.** Different models for estimating the variance components of NOW

Model	Trait	Fixed effect				Covariable		AIC
		HYS	Breed	Boar	Parity	Age of sow at farrowing	Nursing time	
Model 1								
1.1	NOW <sup>1</sup>	x	x	x	x		x	2,636.70
1.2	NOW	x	x	x	x	x	x	2,650.50
Model 2								
2.1	NOW1 <sup>2</sup>	x	x	x		x	x	8,631.50
2.2	NOW2-6 <sup>3</sup>	x	x	x	x	x	x	8,671.50
Model 3								
3.1	NOW1	x	x	x		x	x	17,266.1
	NOW2-6	x	x	x		x	x	

<sup>1</sup> NOW1-6 = number of weaned piglets for all available parities.

<sup>2</sup> NOW1 = number of weaned piglets for the first parity.

<sup>3</sup> NOW2-6 = number of weaned piglets for second to sixth parity.

**Appendix 4.4.** Different models for estimating the variance components of MA

Model	Trait	Fixed effects				Covariables Age of sow at farrowing	ptNOW <sup>4</sup>	Nursing time	AIC
		HYS	Breed	Boar	Parity				
Model 1									
1.1	MA <sup>1</sup>	x	x	x	x			x	-11,276.60
1.2	MA	x	x	x	x		x	x	-18,050.10
Model 2									
2.1	MA1 <sup>2</sup>	x	x	x					-8,993.30
2.2	MA2-6 <sup>3</sup>	x	x	x	x	x		x	-15,247.10
2.3	MA2-6	x	x	x	x		x	x	-941.98
Model 3									
3.1	MA1	x	x	x				x	-8,851.70
	MA2-6	x	x	x				x	
3.2	MA1	x	x	x			x	x	9,748.54
	MA2-6	x	x	x			x	x	

<sup>1</sup> MA1-6 = mothering ability of a sow for all available parities.<sup>2</sup> MA1 = mothering ability of a sow for the first parity.<sup>3</sup> MA2-6 = mothering ability of a sow for second to sixth parity.<sup>4</sup> ptNOW = number of potentially weaned piglets equals sum of weaned and not weaned piglets.

**Appendix 4.5.** Different models for estimating the variance components of PS

Model	Trait	Fixed effects				Covariables					Random Effects	AIC <sup>7</sup>	
		HYS	Breed	Boar	Parity	Age of sow at farrowing	Nursing time	Litter size	Competing piglets <sup>1</sup>	Cross-fostered piglets <sup>2</sup>		Boar	I
Model 1													
1.1	PS <sup>3</sup>	x	x	x	x					(x) <sup>6</sup>		2,850.62	4,472.46
1.2	PS	x	x	x	x	x	x			(x)		6,719.34	6,670.40
1.3	PS	x	x	x	x	x	x	x		(x)		15,276.36	16,150.20
1.4	PS	x	x	x	x	x	x		x	(x)		5,872.32	5,381.66

<sup>1</sup> Competing piglets = number of competing piglets at nurse sow; equals number of weaned piglets (NOW).

<sup>2</sup> Binary coding, 0 = not cross-fostered and 1 = cross-fostered.

<sup>3</sup> PS = Piglet survival.

<sup>4</sup> PS1 = piglet survival for first parity. <sup>5</sup> PS2ff = piglet survival from second parity upwards.

<sup>6</sup> Equal Model with additional covariable (cross-fostered).

<sup>7</sup> I = AIC for models without cross-fostering effect, II = AIC for models with cross-fostering effect.

**Appendix 4.6.** Different models for estimating the genetic correlation between reproductive traits

Model	Trait	Fixed effects				Covariables				
		HYS	Breed	Boar	Parity	Age of sow at farrowing	Nursing time	Farrowing Interval	ptNOW <sup>1</sup>	NOW
NOW and NBA										
1.1	NOW	X	X	X	X		X			
	NBA	X	X	X	X	X		X		
1.2	NOW	X	X	X	X	X	X			
	NBA	X	X		X	X		X		
MA and NBA										
1.1	MA	X	X	X	X		X		X	
	NBA	X	X	X	X	X		X		
1.2	MA	X	X	X			X			
	NBA	X	X		X	X		X		
NOW and FI										
1.1	NOW	X	X	X	X		X			
	FI	X	X		X	X	X			X
1.2	NOW	X	X	X	X	X	X			
	FI	X	X		X	X	X			
MA and FI										
1.1	MA	X	X	X	X		X		X	
	FI	X	X		X	X	X			X
1.2	MA	X	X	X			X			
	FI	X	X		X	X	X			

**NBA and FI**

<b>1.1</b>	<b>NBA</b>	x	x	x	x	x		x	
	<b>FI</b>	x	x		x	x	x		x
<b>1.2</b>	<b>NBA</b>	x	x		x	x		x	
	<b>FI</b>	x	x		x	x	x		

<sup>1</sup>ptNOW = number of potentially weaned piglets equals sum of weaned and not weaned piglets.

## REFERENCES

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control.* 10:716–723. doi:10.1109/TAC.1974.1100705.
- Alfonso, L., J. L. Noguera, D. Babot, and J. Estany. 1997. Estimates of genetic parameters for litter size at different parities in pigs. *Livest. Sci.* 47:149–156.
- Boesch, M., R. Roehe, H. Looft, and E. Kalm. 1999. Selection for litter size in swine - Die Selektion auf Wurfgröße beim Schwein. *Arch Tierz Dummerstorf.* 42:555–570.
- Chen, P., T. J. Baas, J. W. Mabry, K. J. Koehler, and J. C. M. Dekkers. 2003. Genetic parameters and trends for litter traits in US Yorkshire, Duroc, Hampshire, and Landrace pigs. *J. Anim. Sci.* 81:46–53.
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, R. Thompson, D. Butler, and others. 2009. ASReml user guide release 3.0. VSN Int. Ltd Hemel Hempstead UK. Available from: <http://vsni.de/downloads/asreml/release3/UserGuide.pdf>
- Hanenberg, E., E. F. Knol, and J. W. M. Merks. 2001. Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. *Livest. Prod. Sci.* 69:179–186.
- Huby, M., L. Canario, T. Tribout, J. C. Caritez, Y. Billon, J. Gogu  , and J. P. Bidanel. 2003. Genetic correlations between litter size and weights, piglet weight variability and piglet survival from birth to weaning in Large White pigs. *Proc 54th Annu Meet Eur Assoc Anim Prod Roma Italy Wagening.* Wagening, Neth. 362.
- Irgang, R., J. A. F  vero, and B. W. Kennedy. 1994. Genetic parameters for litter size of different parities in Duroc, Landrace, and large white sows. *J. Anim. Sci.* 72:2237–2246.
- Kaplon, M. J., M. F. Rothschild, P. J. Berger, and M. Healey. 1991. Population parameter estimates for performance and reproductive traits in Polish Large White nucleus herds. *J. Anim. Sci.* 69:91–98.
- Knol, E. F. 2001. Genetic aspects of piglet survival. Wageningen University, Wageningen, The Netherlands [PhD Thesis]. Ph. D. dissertation.
- Knol, E. F., B. J. Ducroa, J. A. M. van Arendonk, and T. J. van der Lendea. 2002. Direct, maternal and nurse sow genetic effects on farrowing-, pre-weaning- and total piglet survival. *Livest. Prod. Sci.* 73:153–164.
- McCarter, M. N., J. W. Mabry, J. K. Bertrand, and L. L. Benyshek. 1987. Components of Variance and Covariance for Reproductive Traits in Swine Estimated from Yorkshire Field Data. *J. Anim. Sci.* 64:1285–1291.
- ten Napel, J., A. G. De Vries, G. A. Buiting, P. Luiting, J. W. Merks, and E. W. Brascamp. 1995. Genetics of the interval from weaning to estrus in first-litter sows: distribution of data, direct response of selection, and heritability. *J. Anim. Sci.* 73:2193–2203.
- Putz, A. M., F. Tiezzi, C. Maltecca, K. A. Gray, and M. T. Knauer. 2015. Variance component estimates for alternative litter size traits in swine. *J. Anim. Sci.* 93:5153–5163.
- Roehe, R., and B. W. Kennedy. 1995. Estimation of genetic parameters for litter size in Canadian Yorkshire and Landrace swine with each parity of farrowing treated as a different trait. *J. Anim. Sci.* 73:2959–2970.

- Rothschild, M. F., C. R. Henderson, and R. L. Quaas. 1979. Effects of Selection on Variances and Covariances of Simulated First and Second Lactations. *J. Dairy Sci.* 62:996–1002. doi:10.3168/jds.S0022-0302(79)83361-5.
- Rydhmer, L. 2000. Genetics of sow reproduction, including puberty, oestrus, pregnancy, farrowing and lactation. *Livest. Prod. Sci.* 66:1–12.
- Southwood, O. I., and B. W. Kennedy. 1990. Estimation of direct and maternal genetic variance for litter size in Canadian Yorkshire and Landrace swine using an animal model. *J. Anim. Sci.* 68:1841–1847.

## **5<sup>th</sup> CHAPTER**

### **GENERAL DISCUSSION**

### **Preface**

This work studies and evaluates the predictability of genomic evaluations of fertility traits for practical pig breeding. Different approaches were used, namely, evaluating the effects of enlarging the reference population within breed and multi-subpopulations for genomic predictions in order to improve the prediction accuracy (Chapter 2) and secondly, a comprehensive comparison of different genomic methods for predictions in practical pig data (Chapter 3). A comprehensive study of genetic components for fertility traits is presented in Chapter 4. The following discussion is devoted to issues arising from these studies and the need for further investigation.

### **Perspectives and challenges of the genomic selection in pig breeding**

The success of a genomic (pig) breeding program depends on many factors, e.g. availability of a sufficiently large reference population, a reasonable (i.e. cost-efficient) genotyping strategy, sustainability of the genomic program and control of inbreeding. GS implementation in pig breeding suffers from small reference (and validation) sets. Although breeding companies breed with the same breeds, cooperation with each other is scarce. In addition, the early specialization in production (e.g. pig breeder, piglet producer and fattener) encourages the development of competing breeding companies and sometimes also replaces traditional structures. For a sufficient large reference population in pigs, genotypes of about 500 to 1000 boars per line with reliable estimated breeding value are required for genomic predictions (personal communication, BHZP). The number of boars used in nucleus pig breeding is rather small. Boars are only used for a short time and are, compared to bulls, less valuable in breeding. Further aspects to be considered are e.g. that recording of phenotypes must be standardized across lines and farms, genetic relationships must exist and estimated breeding values have to be on the same scale.

To implement GS, initial financing is crucial; this is only realistic for the nucleus herds. A comprehensive genotyping of all male piglets of a litter is desirable, but too expensive in regard to cost management. One possibility of a cost-effective alternative is low-density genotyping of animals followed by imputing of genotypes to high density genotypes. Wellmann et al. (2013) developed an imputation method for a situation in which selection candidates are genotyped with a low-density SNP panel but have high-density genotyped sires. This method is suggested as a promising strategy for the implementation of GS at acceptable costs.

## **Genomic selection in pig breeding**

### **Potential of multi-populations in practice**

The combination of subpopulations within one breed in order to enlarge the size of the reference set as described in Chapter 2 is expected to increase the accuracy of genomic prediction. Some studies confirm this expectation, e.g. in cattle (Brøndum et al., 2011; Zhou et al., 2014), while other studies show no increase or even a decrease by combining breeds for e.g. cattle (Erbe et al., 2012; Olson et al., 2012) or pigs (Boré et al., 2018). Some possible factors why the enlargement of the reference set in this thesis (Chapter 2) and in general does not lead to the desired effects are discussed below.

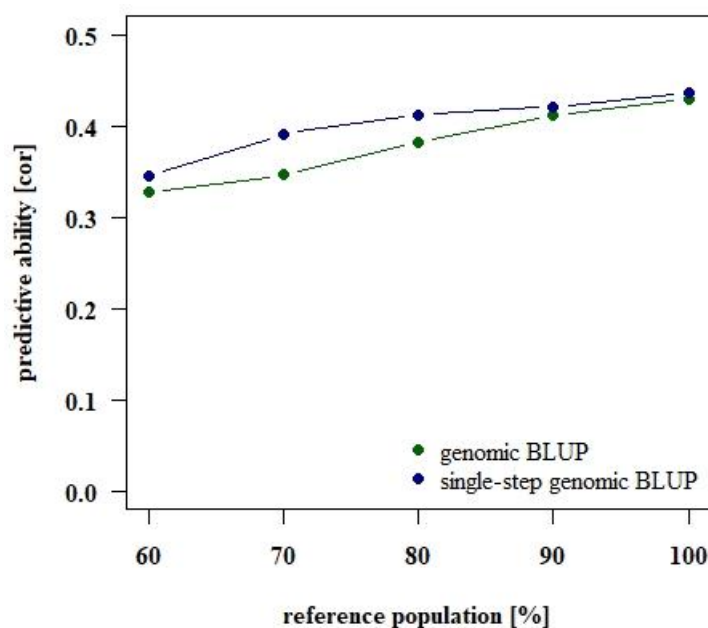
### **Level of family relationships**

Combining populations into one reference population is in theory expected to be beneficial if the combined populations are closely related and the number of individuals added to the reference set from another population is very large. The combination of closely related animals is in accordance with these mentioned theoretical expectations and is presented in Chapter 2 in Table 2.3. Adding sows of the same population to enlarge the reference set within the population improves the predictive ability of the population due to the high level of family relationships between those animals, this effect was also shown for cattle (Plieschke et al., 2018). The assumption that adding a large number of individuals from another subpopulation would also increase the predictive ability (Figure 2.2 and 2.4) could not be confirmed in this thesis, regardless of the fact whether closely or distantly related animals were added to the reference set.

Another aspect is the possible existence of substructures within the (sub-)population, which may not be taken into consideration correctly. Using a genomic relationship matrix based on VanRaden (2007) and treating all subpopulations in the same manner largely ignores possible substructures. Thus, the main interest of this thesis is predicting genomic breeding values more realistically through an improved adjustment of the genomic relationship matrix (Chapter 2). To balance the effects of possible subpopulation structures, different genomic relationship matrices were calculated to assess their usefulness for multi-subpopulation approaches. Results for the youngest animals (Table 2.4) indicate that using population-specific allele frequencies present in the jointly analyzed subpopulations to account for subpopulation structures (instead of using the standard genomic relationship matrix introduced by VanRaden (2007) seems to be more relevant for predictive ability. Predictive

abilities are also raised slightly if closely related subpopulations are added to the reference population (Table 2.4), rather than more distantly related ones. Especially for the largest pig population (subpopulation 4,  $n=821$ ), adding individuals from different subpopulations did not increase the predictive ability, while for smaller pig populations (subpopulation 1,  $n=187$  in Chapter 2) adding individuals from different populations seemed to be generally beneficial. However, the number of animals used in this study is rather small and general conclusions should be drawn very cautiously.

To summarize, subpopulations should be enlarged within their own subpopulation to increase the level of family (and overall) relationships among the reference (and validation) set. To support this, a forward prediction ( $n=53$  validation animals) with a differing number of reference animals ( $n=287$  (100%) reference animals) for Large White animals (Chapter 3) was conducted and correlations between quasi-phenotype DRP and DGV (GBLUP) and ssGEBV (ssGBLUP) have been calculated, respectively (Figure 5.1).



**Figure 5.1:** Forward prediction for differing size of reference population (in %) within one Large White population (Chapter 3). Predictive ability, assessed as correlation between DRP and DGV (GBLUP) and ssGEBV (ssGBLUP), have been calculated for 53 validation animals, respectively.

Although the number of Large White animals used in this analysis is rather small ( $n=287$  reference animals), predictive abilities were increased for both GBLUP and ssGBLUP by enlarging the reference population size from 60% ( $n=172$ ) to 100% ( $n=287$ ) of the original

number of animals. Thus, increasing the reference population size within a population seems to be a reasonable methodological choice.

### **Consistency of LD across populations**

In different populations, the QTL may be in high LD with a different SNP or the linkage phase between QTL and SNP may be reversed; this was shown in studies on chicken (e.g. Heifetz, 2005) or cattle (e.g. Gautier et al., 2007; de Roos et al., 2008). De Roos et al. (2008) reported that a high SNP density of roughly 300k SNPs equally spread across the genome may be able to overcome the differences in LD pattern between cattle populations. In the present study (Chapter 2), a smaller SNP panel of 60k was used for genomic evaluations. Further, the effect of weighing the genomic relationship matrices by e.g. LD phase consistency or calculated marker effects (Zhou et al., 2014a) in genomic predictions was studied, but no substantial increase in predictive ability was found. One possible explanation is the low correlation between the marker effects for different subpopulations (Appendix 2.5). However, accounting for LD structure to account for population structure decreases the accuracy of multi-breed genomic predictions (Daetwyler et al., 2012), which largely agrees with findings presented in this thesis (Figure 2.4).

Reproductive traits (e.g. number of piglets born alive) are polygenic, affected by a large number of QTL having low to moderate effects (Rothschild and Ruvinsky, 1998), and may have an incomplete LD between SNP and QTL (Hidalgo et al., 2014; Zhou et al., 2014b) and thus be the result of complex genetic and biological mechanisms. Due to these facts, in a multi-population setting a less accurate prediction is to be expected (Hidalgo et al., 2014). One QTL may segregate only in one of the studied populations, which is the case for e.g. young mutations (Kemper et al., 2015). Consequently, for this specific QTL one population is not going to improve the prediction (Kemper et al., 2015), which may also explain the poor performance of enlarging reference sets which was found in this study (Chapter 2). Appendix 2.5 shows the correlation between SNP effects for different populations. With respect to the top 10 SNPs, correlations between closely related populations like subpopulation 1 and 2 (0.211) or subpopulation 4 and 2 (0.214) turned out to be rather high and positive, indicating that these SNPs may have a similar effect for both subpopulations and thus, genomic predictions may work better. Correlations between distantly related populations such as subpopulation 4 and 3 (-0.016) or subpopulation 1 and 4 (-0.167) were negative and therefore genomic prediction may be poor.

**Impact of the reference and validation population size**

One of the most important aims, especially in pig breeding, is to increase the reference population. To ensure close relationships between the animals of a population, the composition of the reference set is as important as its size (Lourenco et al., 2015b). In this thesis (Chapter 2 and 3), the genotyping strategy mainly focused on representing the whole population of nucleus animals used for breeding and secondly to create close genetic links between reference and validation set. Therefore, the presented studies have a potential for improvement.

Furthermore, differences in the size of reference animal sets exist between pig breeding organizations. In this thesis, data from small German herdbook organizations was used, which in general manage small nucleus herds and subsequently have rather small reference (and validation) sets. Internationally operating breeding companies such as PIC (<http://de.pic.com/>) or Topigs Norsvin (<https://topignorsvin.de/>) usually have larger nucleus herds. Thus, reference set for genomic predictions are larger making more accurate predictions possible.

At present, the landscape of GS is changing: animals are being routinely genotyped and phenotypes for most traits are being collected (Howard et al., 2018). Genotypes within a species continue to increase in number and the relationship of recent selection candidates to the majority of the historic population is becoming more distant. Moreover, improvements in phenotype collections, changes in genetic architecture and/or changes in models across time create a situation, in which information from an older animal (in the reference set) has the potential to negatively impact the accuracy of selection candidates (Howard et al., 2018). The numbers of animals used for genomic predictions in this thesis were rather small (2,053 Large White animals in Chapter 2 and 495 (424) Landrace (Large White) animals in Chapter 3). Thus, removing older data (truncating data) does not primarily have any particular practical relevance for the presented data (Chapter 2 and 3), but should be addressed with respect to the impact of using all available data on the accuracy for selection candidates.

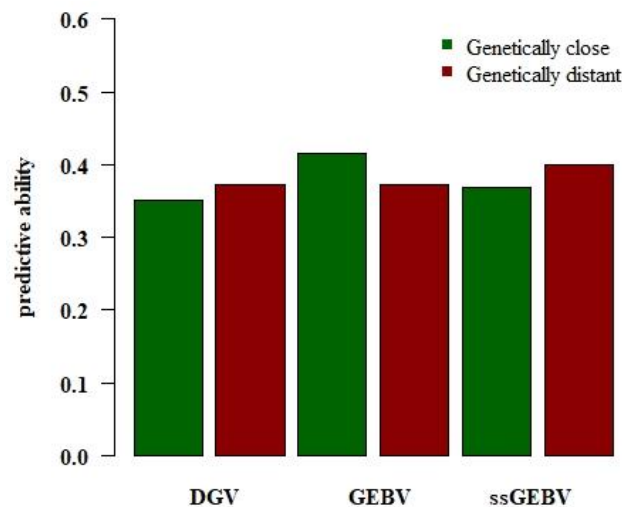
Howard et al. (2018) assessed the impact of removing older data on simulation data (under idealized conditions without any pedigree or genotyping errors) and empirical pig data. For the simulation data, truncating was performed based on the ancestral generation number. For the empirical data it was based on the year an animal was born, which is also the scenario with practical relevance. Howard et al. (2018) had data on genotypes for Yorkshire (n=5,783) and Duroc (n=12,180) available from 2011 to 2017. Across the simulated and the empirical data, removing data from predictions resulted in no change (or a slight increase) in accuracy

for the selection candidates. Pocrnic et al. (2017) have also investigated the effect of pedigree truncation on the solution of ssGBLUP for lowly heritable traits in empirical pig data. They reported that including more than 2 generations of ancestors does not increase accuracy of prediction for young animals, but is sufficient to obtain the same GEBV relative to using full pedigree. On the whole, this is mainly in agreement with findings of Lourenco et al. (2014) for Holstein cattle and Yang and Su (2016) for pig-like simulation data. Both authors observed that generally 3 generations of phenotypic records plus 2 ancestral generations in pedigree were enough for genomic predictions of breeding value, although numbers are dependent on data-structure and heritability (Lourenco et al., 2014).

To conclude: at a high level of heritability, small numbers of reference animals are sufficient to obtain high accuracies. However, since this thesis is concerned with reproductive traits, which have a low heritability, enlarging reference sets is crucial for success.

Putz et al. (2018) reported that the size of the validation population has an impact on accuracy, especially when correlations between estimated breeding values and different types of corrected phenotypes are calculated in empirical pig data. They also reported fluctuating accuracies with large amounts of variation for Landrace animals compared to simulated data. In both of the present studies (Chapter 2 and 3), small numbers of validation animals were used. The number of Large White animals used for the forward prediction in Chapter 2 varied between 35 (Subpopulation 3) and 257 (Subpopulation 4) individuals (Table 2.1), while in Chapter 3 (Table 3.1), 53 Large White and 97 Landrace individuals were used as validation animals. When using small numbers of validation animals, the size and average reliability of the input variable (e.g. DRP or EBV) and corresponding genomic breeding values are of great importance. Having outliers in the data set leads to an emphasized impact and/or results can be biased due to a non-representative sample (VanRaden et al., 2009). Especially for Landrace animals in Chapter 3 (Table 3.3), some outliers were observed in the forward prediction analysis, which generally may lead to poorer predictions. In addition, since reliabilities of pseudo-observations (DRP) of reference animals are small to moderate (0.27 for Large White and 0.40 for Landrace, Chapter 3), obtaining acceptable reliabilities from genomic predictions may become difficult. In order to prevent a future increase of bias and decrease of reliability, the reference and validation sets should be continuously updated with animals from recent generations. However, future research will be needed to assess the design of validation populations.

To encourage the continuous updates of the reference and validation set by the herdbook organization, a forward prediction was performed with two different validation sets for empirical data of Large White animals (Chapter 3) (Figure 5.2). Validation sets were divided according to generation, which was defined based on pedigree data. Generation 1 to 16 (n=114) was used as the reference set in genomic evaluations (multi-step or single-step), generation 17 to 20 (genetically close to reference, n=86) and generation 21 to 24 (genetically distant to reference, n=56) functioned as validation sets. Predictive abilities were assessed as correlation between quasi-phenotype DRP and genomic breeding value obtained from multi-step methods (e.g. DGV or GEBV) or ssGBLUP (ssGEBV). Predictive ability for GEBV turned out to be generally higher for the genetically close validation set than for the genetically distant. For DGV and ssGEBV, predictive abilities turned out to be higher for the distantly related validation sets. Genomic relationships between reference animals and distantly related validation animals turned out to be higher (0.003 from G-matrix and 0.047 from H-matrix) than between reference and closely related validation animals (-0.001 from G-matrix and 0.046 from H-matrix), which may partly explain the higher predictive abilities.



**Figure 5.2:** Predictive ability expressed as correlation between DRP and different breeding values (DGV, GEBV and ssGEBV) obtained from multi-step or single-step method for genetically close and distant validations set.

### **Improvement of genomic predictability for pigs**

#### **Using crossbred information for genomic prediction**

In pigs, crossbreeding is predominantly and intensively used in meat production systems (Xiang et al., 2016). GS offers the potential of selecting purebreds for crossbred performance by using combined information from purebreds and crossbreds (Ibáñez-Escriche et al., 2009; Zeng et al., 2013) or using only purebred data (Esfandyari et al., 2015). Xiang et al. (2016) studied three different ssGBLUP scenarios in which purebred and crossbred data for total number piglets born (TNB) are used to explore the impact of (crossbred) genomic information on prediction accuracy for crossbred performance. For their approach, breed-specific combined relationship matrices are used to incorporate genomic information into the ssGBLUP model. For the construction of breed-specific marker-based relationship matrices, Xiang et al. (2016) assumed that the breed of origin of phased alleles in crossbred animals is known, which was referred to as “crossbred allele tracing”. They reported that methods with genomic marker information are powerful for genetic evaluation for crossbred performance with regard to predictive ability and unbiasedness. They demonstrated that using crossbred genomic information is useful in addition to purebred genotypes. Genetic correlations between purebred and crossbred performance for TNB were 0.75 (0.63) for Landrace (Yorkshire), indicating that selection on purebred performance results in an increased performance of crossbreds (Dekkers, 2007; Xiang et al., 2016). However, two assumptions could diminish and limit the interpretation of the results presented by Xiang et al. (2016): first, the assumption that the breed origin of phased marker genotypes are known (crossbred allele tracing) for crossbred animals, which does not hold in practice and may induce errors, and second, crossbred genotypes are imputed from 8k to 60k prior to allele tracing. Xiang et al. (2016) reported high imputation accuracies (>99%), however, a risk of using incorrect crossbred genotypes cannot be totally eliminated.

#### **Using sequence data for genomic prediction**

In recent years, the availability of whole-genome-sequence (WGS) data has increased rapidly due to the decreasing cost of next-generation sequencing. By using sequence data, the dependency on LD between QTL and SNPs is removed, because it is assumed, that sequence data contains all variants (e.g. causal mutations or causal QTL). Moreover, sequence-based haplotype blocks should be stable across populations and generations. Therefore, using whole-genome-sequence data may be preferable for reference population, especially for those

combining multiple populations as presented in Chapter 2. A first empirical application with a small set of sequenced lines of *Drosophila melanogaster* cannot confirm these expectations (Ober et al., 2012). Brøndum et al., (2015) showed for cattle that current data on WGS do not result in substantial improvements in reliability of genomic breeding values due to the fact that genomic relationships with 777k are already accurately estimated and WGS will neither improve genomic relationships nor GEBV. In addition, current WGS data are not very accurate (Brøndum et al., 2015) either due to imperfect genotype calling (extensive reliance on SNP imputation) or structural genomic variation, which is difficult to assess by short reads. Ni et al. (2017) conducted a study comparing predictive ability obtained with high-density (HD) array data or WGS data in a commercial chicken with various GBLUP models using different approaches to weight SNPs in the genomic relationship matrix. Using all imputed WGS data to perform genomic predictions compared to using HD data did not lead to an improved predictive ability, irrespective of the weighting factors used. Only using genic SNPs from WGS had a positive effect on predictive ability (Ni et al., 2017).

### **Using different information sources in genomic predictions**

Andonov et al. (2017) investigated the accuracy of evaluation on a small population with different approaches for incorporating information from a large related population in a Holstein-like simulation study. Different ways of incorporating information into predictions were utilized: First, data from multiple countries can be combined and all information stemming from all relatives of (non-)genotyped animals and their performance can be used in ssGBLUP without losing any information. And second, GEBVs for genotyped animals in a small population can be computed by using SNP effects derived from a large population (Lourenco et al., 2015b; Andonov et al., 2017).

Andonov et al. (2017) recommended increasing daughter groups per sire in the small population to improve the accuracy, especially for external animals, which are commonly used as sires in the small population. Since the number of genotyped progenies is in general much smaller for boars e.g. Chapter 3; 2.74 (2.87) genotyped progenies for Large White (Landrace) than for bulls, incorporating all available information from multiple countries may not be the best option to use in pig breeding. Thus, Andonov et al. (2017) suggested calculating SNP effects from the large population to derive GEBV for the small population as the best option and obtained highly accurate genomic predictions.

However, high genetic gains can only be expected when individuals of different (pig) breeding programs have a close genetic link to each other e.g. through import or exchange of breeding animals (or semen), which was also proposed and discussed in Chapter 2. If pig breeding companies want to enlarge their reference population in order to obtain higher reliabilities for selection candidates, exchanging animals or semen is unavoidable. Nevertheless, a common genetic breeding value estimation is also complicated due to the specialization of a breeding company in the market. One possibility to more accurately estimate genomic breeding values in pigs could be ssGBLUP, which was studied in detail in Chapter 3 and will be discussed in the following. For successful genomic breeding value estimation, however, large numbers of genetically closely related animals are always essential, irrespective of multi-step or single-step method.

### **Single-step genomic BLUP in pig breeding**

#### **Single-step method**

Single-step genomic BLUP was derived to utilize genotyped and non-genotyped individuals in the same BLUP framework by blending the pedigree-based and genomic relationship matrix (Legarra et al., 2009; Christensen and Lund, 2010). Results from various studies showed that GS models outperform pedigree-based predictions with little variability among genomic models (Chapter 3). Currently used multi-step methods to incorporate genomic information implicitly involve many assumptions which, if violated, may result in loss of information, inaccuracies and bias. In Chapter 3, a comprehensive comparison in terms of accuracy and bias for single-step and multi-step methods was conducted. In the following, some of issues arising from the study will be outlined.

Superiority of ssGBLUP over GBLUP (multi-step) seems to be stronger when the size of the reference population is small, because in this case, the relative contribution of the phenotypic information of non-genotyped animals will be more relevant (Song et al., 2017). However, non-genotyped animals also benefit from genomic information of genotyped animals (Lourenco et al., 2015a; Xiang et al., 2016). Especially in pig breeding, the size of reference and validation set is small, as also presented in this thesis. In Chapter 2, the effect of improving the predictive ability by enlarging the reference size within breed but across subpopulations was studied, a method which unfortunately was not successful. Song et al. (2017) stated that genomic prediction with a small reference population might benefit more from ssGBLUP than genomic prediction with a large mixed reference population. Their

findings are in agreement with the results presented in this thesis. In chapter 3, we showed that with ssGBLUP the accuracy of breeding value estimation (Table 3.3) can be increased despite small reference and validation populations for both, Large White and Landrace (Table 3.1).

Superiority of single-step versus multi-step was also investigated in a simulation study by Schwarzenbacher (2017) with regard to potential bias due to preselection based on genomic information, since most countries use multi-step methods. Single-step produced significantly higher reliabilities, consequently larger genetic gain and evaluations were less biased because preselection was accounted for (Schwarzenbacher, 2017).

Shabalina et al. (2017) investigate the effect of including genotypes from culled bulls on the reliability of genomic predictions from ssGBLUP. Four scenarios with a constant amount of phenotypic information and increasing number of genotypes from culled bulls were simulated and compared regarding prediction reliability. Increases in reliabilities were observed mainly due to the quantitative gain from additional information in the genomic relationship matrix. Shabalina et al. (2017) concluded that improvements in reliability from adding genotypes of culled animals may possibly be lower in real genomic breeding programs with large reference populations. Since the reference populations in pig breeding are generally small with low levels of reliabilities, adding genotypes from culled animals could be a good option to enlarge the reference population within a breed (or population) on the one hand and to create genetic links between reference and validation animals for more reliable predictions on the other hand.

### **Reliabilities in single-step genomic BLUP**

In general, relatively high improvements in theoretical reliabilities were obtained in Chapter 3. Theoretical reliabilities of parental average compared to genomic breeding value obtained from multi-step improved by 29% (Landrace) and 38% (Large White) and compared to single-step genomic breeding values by 48% (Landrace) and 52% (Large White), respectively. Putz et al. (2018) reported that theoretical accuracies improved by 37 % from conventional to genomic predictions. (Theoretical) reliability of genomic predictions (e.g. ssGBLUP) can be calculated by matrix inversion. Especially for large data sets, matrix inversion is not feasible. Misztal et al. (2013) developed a method approximating reliability based on decomposition into contributions from records, pedigrees and genotypes which

involved an inversion of a matrix that contains inverse of genomic and pedigree relationship matrix for genotyped animals. Approximation of reliabilities of predictions by ssGBLUP are accurate and computationally feasible for populations up to 100k genotypes (Misztal et al., 2013), where quality control of information from SNP and proper scaling of  $\mathbf{G}$  is critical.

Recently, a novel approach for approximating genomic reliabilities for national genomic evaluation for Holstein was presented by Liu et al. (2017). The new method was developed to ensure national genomic reliabilities being comparable between countries. Liu et al. (2017) introduced a genomic reliability method which adjusts theoretical genomic reliabilities based on genomic results and is applicable to single-step and multi-step genomic models. The proposed reliability method applies a SNP based genomic model, which is equivalent to GBLUP, but does not rely on genomic relationship matrix. Further, possible changes in genomic reliabilities reflects the changes in GEBV and ensure candidates genomic reliabilities from an early evaluation being consistent with later genomic reliabilities (when they received phenotypic data ; Liu et al., 2017). Although the method is efficient and feasible (Liu et al., 2017), further research on proper validation and verification for official implementation needs to be done.

### **Construction of $\mathbf{H}^{-1}$ matrix**

In ssGBLUP, there seems to be no general agreement between researchers on the selection of a  $\mathbf{G}$  matrix and blending  $\mathbf{G}$  and  $\mathbf{A}$  matrices to become relative to the same base population. Default values used in BLUPF90 programs (Misztal et al., 2002) for scaling and weighting  $\mathbf{G}$  and  $\mathbf{A}$  were used in this thesis (Chapter 3). Further details regarding the construction  $\mathbf{H}^{-1}$  can be found in Chapter 1.

Martini et al. (2018) investigated optimal values of  $\tau$  and  $\omega$  in terms of predictive ability and inflation on publicly available wheat data set. To assess the predictive ability with different  $\mathbf{H}_{\tau,\omega}^{-1}$  matrices, cross-validation was performed splitting the data set randomly into 10 groups and predictive ability was calculated as correlation between phenotype and estimated breeding value for the validation set. The inflation/deflation was calculated as regression coefficient ( $b$ ) of the phenotype on estimated breeding value. Martini et al. (2018) showed that increasing  $\tau$  or decreasing  $\omega$  tends to reduce the empirical variance of the estimated breeding values. Optimal values in terms of maximum predictive ability are given by  $\tau = 1.8$  and  $\omega = 0.2$  for scenario 1, which reflects the closest scenario to the presented empirical data used in this

thesis (Chapter 3). In terms of lowest inflation,  $\tau = 2.0$  and  $\omega = -1.0$  for scenario 1 (Martini et al., 2018). Martini et al. (2018) concluded that a reduction of inflation through increasing  $\tau$  or decreasing  $\omega$  was mainly influenced by decreasing the variance of estimated breeding values in the data set. However, Martini et al. (2018) suggested that consistency problems of **G** and **A** with respect to scaling (and weighting) should be addressed.

The weighting factor  $\beta$ , which displays the proportion of variance that cannot be captured and explained by markers and therefore might be population- and trait-specific, was compared in various studies (e.g. Christensen et al., 2012; Guo et al., 2015; Song et al., 2017). Relatively high accuracies and low bias with  $\beta$  being 0.5 ( $\mathbf{G}_w = (\alpha * \mathbf{G} + \beta * \mathbf{A}_{22})$ ), were found in studies by Guo et al. (2015) and Song et al. (2017) for Danish, American and British Landrace and Danish Yorkshire populations, while values of  $\beta$  being 0.25 were found to be ideal for Danish Duroc in Christensen et al. (2012). An analysis with  $\beta$  being 0.5, which is referred to as “adjusted single-step genomic BLUP”, was performed for the empirical data set of Chapter 3 (results not shown). To summarize: predictive abilities, expressed as the correlation between genomic breeding values from multi-step compared to single-step, do not increase through adjustments of weighting factors ( $\beta$ ) for both breeds compared to original ssGBLUP.

The effect of different calculations of **G** as described in e.g. Meuwissen et al. (2011), Vitezica et al. (2011) or the use of base allele frequencies according to Gengler et al. (2007) to calculate **G** (VanRaden, 2008) was also studied in Chapter 3. However, different **G** calculations did not lead to an improvement in predictive ability of ssGBLUP compared to those obtained with **G** according to VanRaden (2008). Checking for correlations between off diagonal elements of genomic (VanRaden, 2008) and pedigree-based relationship matrix turned out to be rather high for both breeds used in Chapter 3 (0.82 for Large White and 0.88 for Landrace) indicating a reasonable pedigree quality without major genotyping errors and thus, no further need of adjustment for this data set.

### **Algorithm for proven and young animals (APY)**

The computing limit of ssGBLUP is currently up to about 150k genotypes of progeny-tested animals (Aguilar et al., 2011) with no limitation on the number of animals comprising pedigree or number of considered traits. In 2014, a method based on genomic recursion was proposed by Misztal et al. (2014) in which genomic breeding values of new genotyped

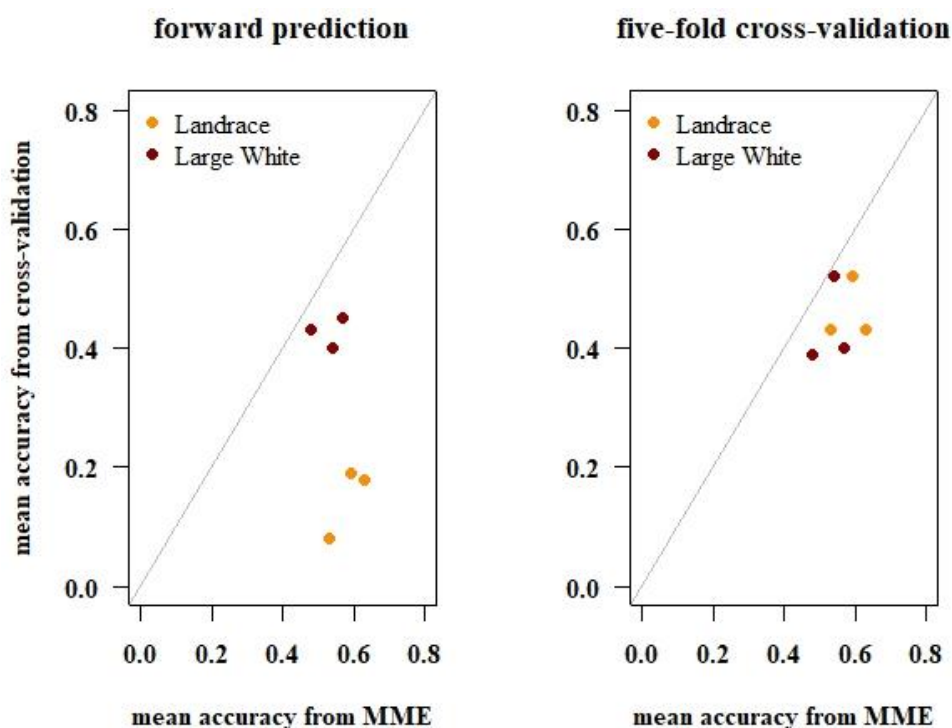
animals (young) are conditioned on genomic breeding values of all previous genotyped animals (proven) which is referred to as “algorithm for proven and young animals” (APY). The direct inversion is required for only a small proportion of **G** composed of relationships among animals treated as “proven” comprising those animals with high accuracies and thus containing most genomic information. Especially for breeds with large quantities of genotypes like Holstein, the development of APY is important. In this thesis, but also in general in pig breeding, numbers of genotypes used for genomic evaluation are rather small (e.g. 2,053 in Chapter 2) and thus, APY will not be applied in pig breeding in the near future. Therefore, only a short discussion of APY will be given below.

In simulations, accuracies with APY were close to those with direct inverted **G** even if some animals with records were treated as “young” (Fragomeni et al., 2015). However, compared to regular inversion of  $\mathbf{G}^{-1}$ , APY results in the same genomic breeding value for GBLUP, for ssGBLUP APY leads to an approximation because young genotyped animals may provide ties to ungenotyped ancestors (Fragomeni et al., 2015).

The optimal composition of core animal sets and how to choose animals for the core set are still some of the most critical issues of APY. Fragomeni et al. (2015) reported for Holstein that the choice of proven (or core) animals for calculating high accuracies is mostly arbitrary but the number matters. With a sufficiently large subset of proven animals in the recursion (>10k animals), the composition may no longer be relevant (Fragomeni et al., 2015). In contrast, Pocrnic et al. (2016) concluded that the number of core animals is not critical since the reduction in accuracy of genomic breeding value is minimal when using half the optimal number (roughly  $n=3k$ ). Ostersen et al. (2016) reported that the choice of core animals is important for the accuracy of APY e.g. because accuracies for genomic breeding values are known to decrease as the prediction and predicted populations become more distantly related (Muir, 2007). They recommended choosing (i) core animals from all generations to ensure an equal representation of genotyped animals in each generation (e.g. increasing the number of independent chromosome segments, because each generation is expected to generate new cross-overs) and (ii) core animals that have the largest number of genotyped offspring. In empirical pig data, the optimal number of core animals was found to be 2k to 6k (Osteren et al., 2016; Pocrnic et al., 2016b). However, further investigation on different core animal definitions/compositions needs to be carried out to quantify accuracy changes (Fragomeni et al., 2015) in practical implementation.

### Validation methods for predictions

In the last years, several authors have considered the question of how selection influences the accuracy of genomic predictions. Bijma (2012) reported that (theoretical) accuracies based on prediction error variance (PEV) obtained from BLUP framework are not valuable in populations under selection, since they overestimate the actual accuracy, while Edel et al. (2012) argued that the accuracy obtained in a forward prediction scheme will underestimate the true prediction accuracy in the presence of selection. Smaller values obtained with forward prediction (Table 3.3) than with averaged accuracies from MME (Table 3.5) were observed for both breeds in Chapter 3 (Figure 5.3). Theoretical reliabilities observed in this thesis (Chapter 3) agreed with findings in literature (e.g. Putz et al., 2018), although the number of validation and reference animals was significantly smaller in this study.



**Figure 5.3:** Mean accuracy for number of piglets born alive calculated from MME averaged over animals and from forward prediction (left) and five-fold cross-validation (right) for different genomic breeding values of Chapter 3 (Table 3.3 and 3.5) for Landrace and Large White.

Usually, the averaged theoretical reliability is similar to the prediction accuracy, unless the trait is under selection (Bijma, 2012). Genotyped animals in Chapter 2 and 3 were (pre)selected based on parental average. Their breeding values are only expected to be unbiased if those animals were randomly chosen (Su et al., 2012), while preselection tends to decrease (realized) accuracies of breeding values (Bijma, 2012). In such preselection cases, predictive abilities obtained from cross-validation scenarios are lower compared to theoretical

accuracies from MME (Figure 5.3) due to the following reasons: (i) preselection is taken into account, (ii) small quantity of data causes inflation of predictive abilities which is quite common in small genotyped populations as presented in Chapter 2 or 3 and/or (iii) the quasi-phenotype (DRP) used may not be well calculated.

Theoretical reliabilities obtained from BLUP framework are a direct function of PEV, where the values of PEV are mostly determined by the numbers of offspring (progeny tested) per boar (Dufrasne et al., 2011). In this thesis (Chapter 3), the number of genotyped offspring per boar is rather small with 2.74 (2.87) for Large White (Landrace) and may also led to uncertainty of prediction. Accounting for inbreeding by calculating theoretical reliabilities

$$\rho_i = \sqrt{1 - \frac{PEV_i}{g_{ii}\text{var}(A_i)}}$$

with  $g_{ii}$  being the diagonal element of the genomic relationship matrix for individual  $i$ , does not substantially change (increase or decrease) the accuracy (results not shown).

However, the main difference between theoretical reliabilities obtained from MME and predictive ability from cross-validation is that theoretical reliabilities give a measure of the individual's genomic breeding value, while the predictive ability is indicative of the average accuracy of genomic breeding values in the population. With increased relatedness between reference and validation animals the empirical accuracy (from cross-validation) of genomic breeding value increases (Badke et al., 2012) as is to be expected.

Putz et al. (2018) recently conducted a comparison of validation methods for pig litter traits to determine the optimal method for comparing pedigree BLUP to ssGBLUP in terms of accuracy and consistency. Empirical data and simulated data, which was created to mimic the field data and to calculate true accuracies, were used. Putz et al. (2018) used different methods to calculate and evaluate the accuracy of prediction within the validation set: (i) theoretical accuracy derived from BLUP framework, (ii) approximated accuracies from the *accf90* or *accf90(GS)* program in the *BLUPF90* family package (Miszta and Wiggans, 1988; Miszta et al., 2013), (iii) correlation between predictions and GEBV from full data set and (iv) correlation between (a) predictions and corrected phenotypes from the full data set, (b) predictions and corrected phenotypes divided by the square root of heritability and (c) sire predictions and the average of their daughters' corrected phenotype, respectively. For the simulated data, Putz et al. (2014) reported relatively small difference between (i) accuracy of

TBV and theoretical accuracy calculated from MME. In empirical pig data, theoretical accuracy (calculated from MME, i) was underestimated compared to the true accuracy.

A general problem in pig breeding is that pure-bred lines suffer from small numbers of sires used in the validation set. It is hardly possible to calculate the accuracy as correlation between (c) sire predictions and the average of their daughters' corrected phenotype in pig empirical data, although the method performed quite well in simulated data (Putz et al., 2018). Cross-validation accuracy, expressed as the correlation between predictions from full data set (iii), turned out to be poor for accuracy validation in empirical pig data due to the fact of extensive overlapping information (Putz et al., 2018).

Further research on validation strategies will be necessary also with respect to future changes in genetic architectures of a trait or species (population structure), and especially in cases, where validation groups will consist entirely of animals preselected according to genomic breeding values (Edel et al., 2012). Masuda et al. (2018) recently conducted a study on preselection bias in Holstein and stated that trends for BLUP indicate a bias due to genomic preselection for genotyped sires and cows.

Estimating how accurate genomic predictions are is important for at least three reasons (Daetwyler et al., 2013): first, response to selection (e.g. Falconer and Mackay, 1996) which will be discussed further on, second, the accuracy of an estimated breeding value expresses the credibility of an individual (e.g. Chapter 2 and 3) which is relevant for a selection decision, and third, the estimation of the prediction accuracy of models is useful for model comparison (e.g. Chapter 3). Daetwyler et al. (2013) suggested reporting the slope of the regression on estimated breeding values, especially when pedigree and genomic information is combined to produce one breeding value. Regressions of observations on predictions differing strongly from 1 point to deficiencies of the model and bias. A comparison of multi-step to single-step was conducted in Chapter 3 with regressions of corrected phenotypes on different genomic breeding values (Table 3.4) indicating that ssGBLUP is preferable in terms of bias compared to multi-step methods (such as GBLUP and blending). But, as long as the information level for pigs is the accurately estimated genomic breeding value, irrespective of its being calculated from GBLUP or ssGBLUP, the ranking of individual is the most important selection criterion which is not influenced by the regressing coefficient (slope). Thus, accuracy of prediction as the correlation between observed and predicted value within the examined group remains the more important selection criterion.

### Genetic progress in (genomic) pig breeding

#### Genetic progress

The main purpose of (pig) breeding is to obtain as much genetic progress as possible for the trait of interest. Genetic gain ( $\Delta G$ ) is a result of various factors, e.g. accurately estimated breeding values and available additive genetic variance (Aasmundstad et al., 2015). Meuwissen et al. (2001) suggested the inclusion of genomic information into estimations in order to improve the accuracy of breeding value estimation. The potential of GS, expressed as genetic trend, is displayed for fertility traits and production traits (Haberland et al., 2010; Haberland et al., 2014) in Table 5.1, respectively. In genomic breeding schemes,  $\Delta G$  for litter size is about 38% (Table 5.1), which is 10% higher compared to conventional breeding schemes ( $\Delta G=28\%$ ). Higher genetic trends are mainly attributable to higher (genomic) reliabilities (e.g. higher  $\rho$ ) and to greater selection intensity for sires (Simianer, 2009). Lillehammer et al. (2011) reported even higher genetic trends for fertility traits in pigs (Table 5.1).

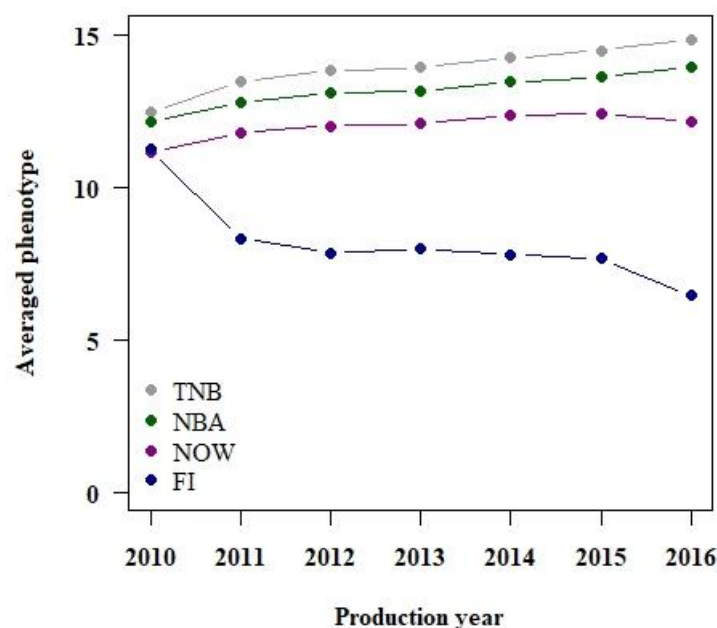
**Tab. 5.1:** Genetic trend for different fertility and production traits realized by implementation of GS in pigs

Trait	Genetic trend	Reference
<b>Litter size</b>	38 %	Simianer (2009)
<b>Production traits</b>	47%	Haberland et al. (2010)
<b>Fertility traits</b>	35-69%	Lillehammer et al. (2011)
<b>Boar taint</b>	10-12%	Haberland et al. (2014)

Haberland et al. (2010) showed that young boars can be selected more accurately by integrating genomic information into selection. Reliabilities of genomic breeding values were 15% (29%) higher compared to conventional estimations for reference sizes of 500 (1000) genotyped animals, which was also shown in Chapter 3. With increasing reference size higher reliabilities are obtained. However, for the moderately sized pig breeding organizations used in this thesis, a reference size of 500 progeny-tested boars currently seems to be more realistic than having 1000 animals in the reference set.

Observed reliabilities of breeding values (conventional and genomic) in this thesis are moderate (Chapter 2 and 3, e.g. Table 3.5), although an improvement in fertility traits is feasible, as was also reported in other studies on litter traits. Genetic variation between and within breeds triggered substantial genetic increase in prolificacy in recent years (Canario et al., 2005).

Figure 5.4 shows the averaged phenotype for different fertility traits (Chapter 4) from 2010 to 2016. For total number of piglets born (TNB), born alive (NBA) and weaned (NOW), a desired increase in performance of +2.39 piglets (+1.77 and +1.00) for TNB (NBA and NOW) from 2010 to 2016 is achievable, respectively. The observed numbers for NBA and NOW are smaller than for TNB, which is mainly due to piglets dying during the nursing period. For farrowing interval (FI), a decrease of 4.78 days from 2010 to 2016 is observed, which is also a desirable effect because sows return to the production cycle earlier and can produce more piglets per year.



**Figure 5.4:** Averaged phenotype for total number of born piglets (TNB), born alive (NBA), weaned (NOW) and farrowing interval (FI) with respect to the production year.

Poulsen et al. (2018) investigated the effects of transferring genetic merit of nucleus selection for litter size to production herds (sows) and defining the impact of realized improvement in production due to purebred selection. Selection for litter size in purebred sows, as it was also studied in this thesis, increases litter size in production sows, but the impact strongly depends on the production herd (GxE interaction; Poulsen et al., 2018).

### Improvement of sows' prolificacy and piglets' survival

Litter size is one of the most important traits for productivity of a sow. In addition to NBA or NOW, alternative fertility traits such as mothering ability (MA) or piglet survival (PS) were studied in Chapter 4 in order to evaluate genetic parameters for selection. Genetic and phenotypic variation between parities of e.g. MA or NOW exist (Chapter 4) and thus, can be

used for the breeding decisions. Instead of evaluating “typical” alternative fertility traits, Freyer (2018) examined the individual variation in litter size based on the level of individual TNB per parity, which turned out to be quite large and heritable. Selection for homogeneous litters of a sow, based on selection on sow’s individual capacity in litter size and the corresponding individual variation in parities (e.g. “range” as second trait with  $h^2 = 0.06$  to  $0.10$ ), is possible if at least five parities are available for estimations. Homogeneous litters also lead to improved animal-related conditions for piglets (Freyer, 2018) and thus commercial management (e.g. large litter size lead to small individual birth weights). Freyer (2018) showed that using individual maximum TNB as a “new trait” in genetic analysis reveals considerably higher heritability ( $h^2 \sim 0.3$ ) than the reported estimates for heritability ( $h^2 \sim 0.12$ ) and repeatability ( $w^2 \sim 0.18$ ) for NBA in this thesis (Chapter 4). The highest TNB arose within parity five (Freyer, 2018), which is in agreement with results in this thesis (Chapter 4). However, variability of performances of sows is a keystone of genetic improvement and needs to be maintained to guarantee sustainability and success in long-term selection, although homogeneity in litter size is desired for commercial use.

Exclusively selecting on prolificacy may initially lead to an increased litter size, but may also result in reduced sow and piglet vitality (e.g. large litter size lead to small individual birth weights). Including a piglet vitality index into routine genetic evaluation, which could consist of litter homogeneity and piglet vitality, would generally be possible (Pfeiffer et al., 2018). Heritabilities and genetic variance for litter quality traits such as individual birth or litter weights in Austrian maternal lines were observed in a study of Pfeiffer et al. (2018) showing a negative relationship between TNB and litter quality traits. In Chapter 4, mothering ability (MA) and piglet survival were studied. The negative relationship between NBA and MA ( $r_g \sim 0.3 - 0.5$ ) which was observed is largely in agreement with findings of Pfeiffer et al. (2018). Unfortunately, individual birth or litter weights were not available and could therefore not be studied in this thesis (Chapter 4).

### **Validity of repeatability model for fertility traits**

In Chapter 4 besides the variable selection for models two different models, bivariate and repeatability model, were examined for fertility traits. The assumptions of repeatability (as presented by Falconer and Mackay, 1996) can be summarized as: (i) variance of different measurements are equal and their components are present in the same proportions and (ii)

different measurements reflect what is genetically the same character. The results presented (Chapter 4) indicated that records of different parities (first and second to sixth) are not entirely controlled by the same genes. E.g. heritabilities of NBA (Table 4.2), NOW (Table 4.4) or FI (Table 4.3) are higher in the first parity than for the following parities. Moreover, estimates of heritability for NOW and FI (Chapter 4) from repeatability (0.116 and 0.124) and bivariate models (0.117 and 0.141) for the first parity were generally higher than those obtained for subsequent parities (0.059 and 0.110). Thus, records of young and adult animals may not be controlled by the same genes, which can at least be partly explained by differences in the stage of maturity (Serenius et al., 2003). Thus, those traits should be considered as different traits and calculated using bivariate models.

### **Future prospective and concluding remarks**

The costs of DNA testing are falling rapidly which will lead to (i) a higher proportion of animals being genotyped or even sequenced and (ii) help to generate larger, more-updated reference populations, which is particularly important for pig breeding. Along with that, single-step methods are likely to become routine due to more accurate breeding values. However, GS will be used for an increasing number of traits, while some will be difficult to record and measure. Especially in pig breeding, one perspective can be to train the prediction equation on commercial animals to select purebred animals on crossbred performance, which has been measured for commercially relevant traits such as meat quality, disease resistance and performance under harsher environment (Meuwissen et al., 2016). Although GS appears to be the method of choice, a continuous recording of phenotypic data and enlarging the reference population for retraining is crucial.

The main conclusions from previous chapters can be summarized as:

1. GS can be successfully implemented in practical pig breeding (Chapter 2 and 3).
2. Enlarging the reference population by combining closely and/or distantly related subpopulations of one breed to form a multi-subpopulation reference set does not necessarily lead to higher prediction accuracy for individuals of a specific breed (Chapter 2).
3. Accounting for subpopulation structures of combined reference populations does not improve the predictive ability (Chapter 2).

4. Enlarging the reference population with female individuals of the studied subpopulation yields an improved predictive ability (Chapter 2).
5. Despite the limited size of the reference population, GS can produce more reliable genomic breeding values than conventional estimations (Chapter 2 and 3).
6. Singel-step GBLUP might be preferable with regard to unbiasedness and produce higher predictive abilities compared to multi-step methods (Chapter 3). In moderately sized pig breeding organizations, however, the decision which genomic method should be used needs to be monitored by ongoing validation assessments.
7. Estimates of genetic components for the fertility traits analyzed indicate that the amount of additive genetic variation is large enough for selection (Chapter 4).

## REFERENCES

- Aasmundstad, T., I. Andersen-Ranberg, Ø. Nordbø, T. Meuwissen, O. Vangen, and E. Grindflek. 2015. The effect of including genomic relationships in the estimation of genetic parameters of functional traits in pigs. *J. Anim. Breed. Genet.* 132(5):386–391. doi:10.1111/jbg.12156.
- Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation: Matrix computation genomic selection. *J. Anim. Breed. Genet.* 128(6):422–428. doi:10.1111/j.1439-0388.2010.00912.x.
- Andonov, S., D. A. L. Lourenco, B. O. Fragomeni, Y. Masuda, I. Pocrnic, S. Tsuruta, and I. Misztal. 2017. Accuracy of breeding values in small genotyped populations using different sources of external information—A simulation study. *J. Dairy Sci.* 100(1):395–401. doi:10.3168/jds.2016-11335.
- Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations: Accuracy and selection. *J. Anim. Breed. Genet.* 129(5):345–358. doi:10.1111/j.1439-0388.2012.00991.x.
- Boré, R., L. F. Brito, M. Jafarikia, A. Bouquet, L. maignel, B. Sullivan, and F. S. Schenkel. 2018. Genomic data reveals large similarities among Canadian and French maternal pig lines. *Can. J. Anim. Sci.* 98(4):809–817. doi:10.1139/CJAS-2017-0103.
- Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *J. Dairy Sci.* 94(9):4700–4707. doi:10.3168/jds.2010-3765.
- Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J. Dairy Sci.* 98(6):4107–4116. doi:10.3168/jds.2014-9005.
- Canario, L., T. Tribout, J. Gogue, and J. P. Bidanel. 2005. Estimation of realized genetic trends in French Large White pigs from 1977 to 1988 using frozen semen: Farrowing and early lactation periods. In: *Proceedings of the 56th Annual Meeting of the European Association for Animal Production*. Uppsala, Sweden.
- Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42(2):1–8. doi:10.1186/1297-9686-42-2.
- Christensen, O. F., P. Madsen, B. Nielsen, T. Ostensen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. *animal.* 6(10):1565–1571. doi:10.1017/S1751731112000742.
- Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics.* 193(2):347–365. doi:10.1534/genetics.112.147983.
- Dekkers, J. C. M. 2007. Marker-assisted selection for commercial crossbred performance. *J. Anim. Sci.* 85(9):2104–2114. doi:10.2527/jas.2006-683.

- Dufresne, M., M. Rustin, V. Jaspard, J. Wavreile, and N. Gengler. 2011. Using test station and on-farm data for the genetic evaluation of Piétrain boars used on Landrace sows for growth performance. *J. Anim. Sci.* 89(12):3872–3880. doi:10.2527/jas.2010-3816.
- Edel, C., S. Neuner, R. Emmerling, and K.-U. Goetz. 2012. A note on using “forward prediction” to assess precision and bias of genomic predictions. *Interbull Bull.* No. 46:16–19.
- Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95(7):4114–4129. doi:10.3168/jds.2011-5019.
- Esfandyari, H., A. C. Sørensen, and P. Bijma. 2015. A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.* 47(1). doi:10.1186/s12711-015-0155-z.
- Falconer, D. S., and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th ed. Longmans Green, Harlow, Essex, UK.
- Fangmann, A., S. Bergfelder-Drüing, E. Tholen, H. Simianer, and M. Erbe. 2015. Can multi-subpopulation reference sets improve the genomic predictive ability for pigs? *J. Anim. Sci.* 93(12):5618–30. doi: 10.2527/jas.2015-9508.
- Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.* 98(6):4090–4094. doi:10.3168/jds.2014-9125.
- Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J.-G. Garnier, D. Boichard, G. M. Lathrop, I. G. Gut, and A. Eggen. 2007. Genetic and Haplotypic Structure in 14 European and African Cattle Breeds. *Genetics*. 177(2):1059–1070. doi:10.1534/genetics.107.075804.
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *animal*. 1(1):21. doi:10.1017/S1751731107392628.
- Guo, X., O. F. Christensen, T. Ostensen, Y. Wang, M. S. Lund, and G. Su. 2015. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. *J. Anim. Sci.* 93(2):503–512. doi:10.2527/jas2014-8331.
- Haberland, A. M., H. Luther, A. Hofer, E. Tholen, H. Simianer, B. Lind, and C. Baes. 2014. Efficiency of different selection strategies against boar taint in pigs. *animal*. 8(1):11–19. doi:10.1017/S1751731113001857.
- Haberland, A. M., F. Ytournal, H. Luther, and H. Simianer. 2010. Evaluation of selection strategies including genomic breeding values in pigs. In: *Proceedings of the 61st Annual Meeting of the European Federation of Animal Science*. Heraklion, Greece.
- Heifetz, E. M. 2005. Extent and Consistency Across Generations of Linkage Disequilibrium in Commercial Layer Chicken Breeding Populations. *Genetics*. 171(3):1173–1181. doi:10.1534/genetics.105.040782.
- Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, and D. J. de Koning. 2014. Accuracy of Genomic Breeding Values Predicted Within and

- Across Breeds in Pig Populations. In: Proceedings of the 10th world congress on genetics applied to livestock production. Vancouver, Canada.
- Howard, J. T., T. A. Rathje, C. E. Bruns, D. F. Wilson-Wells, S. D. Kachman, and M. L. Spangler. 2018. The impact of truncating data on the predictive ability for single-step genomic best linear unbiased prediction. *J. Anim. Breed. Genet.* 135(4):251–262. doi:10.1111/jbg.12334.
- Ibáñez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol.* 41(1):12. doi:10.1186/1297-9686-41-12.
- Kemper, K. E., B. J. Hayes, H. D. Daetwyler, and M. E. Goddard. 2015. How old are quantitative trait loci and how widely do they segregate? *J. Anim. Breed. Genet.* 132(2):121–134. doi:10.1111/jbg.12152.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92(9):4656–4663. doi:10.3168/jds.2009-2061.
- Lillehammer, M., T. H. E. Meuwissen, and A. K. Sonesson. 2011. Genomic selection for maternal traits in pigs. *J. Anim. Sci.* 89(12):3908–3916. doi:10.2527/jas.2011-4044.
- Liu, Z., P. M. VanRaden, M. H. Lidauer, M. P. Calus, H. Benhajali, H. Jorjani, and V. Ducrocq. 2017. Approximating genomic reliabilities for national genomic evaluation. *Interbull Bull.* No. 51:75-85.
- Lourengo, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015a. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* 47(1):47-56. doi:10.1186/s12711-015-0137-1.
- Lourengo, D. A. L., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor, S. Forni, and J. I. Weller. 2014. Are evaluations on young genotyped animals benefiting from the past generations? *J. Dairy Sci.* 97(6):3930–3942. doi:10.3168/jds.2013-7769.
- Lourengo, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and others. 2015b. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J. Anim. Sci.* 93(6):2653–2662. doi: 10.2527/jas.2014-8836.
- Martini, J. W. R., M. F. Schrauf, C. A. Garcia-Baccino, E. C. G. Pimentel, S. Munilla, A. Rogberg-Muñoz, R. J. C. Cantet, C. Reimer, N. Gao, V. Wimmer, and H. Simianer. 2018. The effect of the H–1 scaling factors  $\tau$  and  $\omega$  on the structure of H in the single-step procedure. *Genet. Sel. Evol.* 50(1):16. doi:10.1186/s12711-018-0386-x.
- Masuda, Y., P. M. VanRaden, I. Misztal, and T. J. Lawlor. 2018. Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. *J. Dairy Sci.* 101(6):5194–5206. doi:10.3168/jds.2017-13310.
- Meuwissen, T. H. E., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited: Unified approach to genomic selection. *J. Anim. Breed. Genet.* 128(6):429–439. doi:10.1111/j.1439-0388.2011.00966.x.
- Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. *Anim. Front.* 6(1):6–14. doi:10.2527/af.2016-0002.

- Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* 97(6):3943–3952. doi:10.3168/jds.2013-7752.
- Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013. Methods to approximate reliabilities in single-step genomic evaluation. *J. Dairy Sci.* 96(1):647–654. doi:10.3168/jds.2012-5656.
- Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, D. H. Lee, and others. 2002. BLUPF90 and related programs (BGF90). In: *Proceedings of the 7th world congress on genetics applied to livestock production*. Montpellier, France.
- Misztal, I., and G. R. Wiggans. 1988. Approximation of Prediction Error Variance in Large-Scale Animal Models. *J. Dairy Sci.* 71:27–32. doi:10.1016/S0022-0302(88)79976-2.
- Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J. Anim. Breed. Genet.* 124(6):342–355. doi: 10.1111/j.1439-0388.2007.00700.x.
- Ni, G., D. Cavero, A. Fangmann, M. Erbe, and H. Simianer. 2017. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Sel. Evol.* 49(1):8. doi:10.1186/s12711-016-0277-y.
- Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster*. N. R. Wray, editor. *PLoS Genet.* 8(5):e1002685. doi:10.1371/journal.pgen.1002685.
- Olson, K. M., P. M. VanRaden, and M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.* 95(9):5378–5383. doi:10.3168/jds.2011-5006.
- Ostersen, T., O. F. Christensen, P. Madsen, and M. Henryon. 2016. Sparse single-step method for genomic evaluation in pigs. *Genet. Sel. Evol.* 48(1):48. doi:10.1186/s12711-016-0227-8.
- Pfeiffer, C., B. Fuerst-Waltl, P. W. Knap, A. Willam, C. Leeb, and C. Winckler. 2018. Genetic parameters for litter quality traits of Austrian Large White and Landrace sows. In: *Proceedings of the 69th Annual meeting of the European Association of Animal Production*. Dubrovnik, Croatia.
- Plieschke, L., C. Edel, E. C. G. Pimentel, R. Emmerling, J. Bennewitz, and K.-U. Götz. 2018. Genotyping of groups of cows to improve genomic breeding values of new traits. *J. Anim. Breed. Genet.* 135(4):286–292. doi:10.1111/jbg.12348.
- Pocrnic, I., D. A. L. Lourenco, H. L. Bradford, C. Y. Chen, and I. Misztal. 2017. Technical note: Impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population. *J. Anim. Sci.* 95(8):3391. doi:10.2527/jas2017.1581.
- Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2016. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. *Genet. Sel. Evol.* 48(1):82. doi:10.1186/s12711-016-0261-6.
- Putz, A. M., F. Tiezzi, C. Maltecca, K. A. Gray, and M. T. Knauer. 2018. A comparison of accuracy validation methods for genomic and pedigree-based predictions of swine

- litter size traits using Large White and simulated data. *J. Anim. Breed. Genet.* 135(1):5–13. doi:10.1111/jbg.12302.
- de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. *Genetics*. 179(3):1503–1512. doi:10.1534/genetics.107.084301.
- Rothschild, M. F., and A. Ruvinsky. 1998. *The Genetics of the Pig*. CAB International, Oxon, UK.
- Schwarzenbacher, H. 2017. Single versus two step genomic evaluations over many generations. In: *Proceedings of 68th Annual meeting of the European Association of Animal Production*. Tallinn, Estonia.
- Serenius, T., M.-L. Sevón-Aimonen, and E. A. Mäntysaari. 2003. Effect of service sire and validity of repeatability model in litter size and farrowing interval of Finnish Landrace and Large White populations. *Livest. Prod. Sci.* 81(2–3):213–222. doi 10.1016/S0301-6226(02)00300-7.
- Simianer, H. 2009. The potential of genomic selection to improve litter size in pig breeding programs. In: *Proceedings 60th Annual meeting of the European Association of Animal Production*. Barcelona, Spain.
- Song, H., J. Zhang, Y. Jiang, H. Gao, S. Tang, S. Mi, F. Yu, Q. Meng, W. Xiao, Q. Zhang, and X. Ding. 2017. Genomic prediction for growth and reproduction traits in pig using an admixed reference population. *J. Anim. Sci.* 95(8):3415–3424. doi:10.2527/jas.2017.1656.
- Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, O. F. Christensen, and M. S. Lund. 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J. Dairy Sci.* 95(2):909–917. doi:10.3168/jds.2011-4804.
- VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91(11):4414–4423. doi:10.3168/jds.2007-0980.
- VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92(1):16–24. doi:10.3168/jds.2008-1514.
- Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genet. Res.* 93(5):357–366. doi:10.1017/S001667231100022X.
- Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen. 2016. Application of single-step genomic evaluation for crossbred performance in pig 1. *J. Anim. Sci.* 94(3):936–948. doi:10.2527/jas2015-9930.
- Zeng, J., A. Toosi, R. L. Fernando, J. C. Dekkers, and D. J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol.* 45(11). doi: 10.1186/1297-9686-45-11.
- Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T. H. E. Meuwissen, M. Svendsen, H. Grove, U. S. Nielsen, and M. S. Lund. 2014a. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *J. Dairy Sci.* 97(7):4485–4496. doi:10.3168/jds.2013-7580.

Zhou, L., M. S. Lund, Y. Wang, and G. Su. 2014b. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *J. Anim. Breed. Genet.* 131(4):249–257. doi:10.1111/jbg.12089.

## **APPENDIX**

## ACKNOWLEDGEMENTS

Ich möchte mich ganz herzlich bei meinem Doktorvater **Prof. Dr. Henner Simianer** für die Möglichkeiten, die er mir eröffnet hat, und die vielen spannenden Themen, an denen ich arbeiten durfte, bedanken. Vielen Dank für die große Unterstützung und Motivation in den letzten Jahren.

Bei **Prof. Dr. Jörn Bennewitz** bedanke ich mich für die Übernahme der Zweitprüferschaft meiner Doktorarbeit.

Bei **Prof. Dr. Jens Tetens** möchte ich mich für die Übernahme des Prüfungsbeisitzes bedanken.

Ganz herzlich bedanken möchte ich mich auch bei **Frau Döring**, die immer ein offenes Ohr für mich hatte. Vielen Dank für Ihre große Unterstützung und Hilfe in jeglicher Hinsicht.

Bei meinen **Kolleginnen** und **Kollegen** aus der Arbeitsgruppe, die mich immer unterstützt haben und hilfsbereit zur Seite standen. Besonders möchte ich mich bei **Malena** bedanken, die immer eine offene Tür für mich hatte.

I would like to thank **Jack C. M. Dekkers** and all **people from his lab** for their warm welcome, their support and friendship with various social activities during my visit in Ames (Iowa).

Meiner **Familie** und meinen **Freunden** danke ich für die unendliche Unterstützung und ihr großes Verständnis. Ganz besonders möchte ich mich bei meinen **Eltern**, meinem **Mann** und meiner **Tochter** bedanken - für ihre Liebe, dass sie immer an mich glauben und immer für mich da sind. Vielen lieben Dank!

## CURRICULUM VITAE

**Name** **Anna Maria Fangmann**

Birth name große Holthaus

Date of Birth 14 December 1988

Place of Birth Lohne (Oldenburg), Germany

Nationality German

### School education

September 2008 General qualification for university entrance

### University degrees

November 2013 - today	PhD student, Animal Breeding and Genetics Group Georg-August-University Goettingen, Germany Thesis Topic: “Genomic and conventional evaluations for fertility traits in pigs” under supervision of Prof. Dr. Simianer
April 2012 – October 2013	M. Sc. Agriculture - Animal Sciences, University of Goettingen, Germany Master Thesis: “Einfluss differenter Brutverfahren auf das Reproduktionsgeschehen bei Legehennen” (The influence of different incubation regime on hatchability traits in Layers) under the supervision of Prof. Dr. Knorr and Dr. Sharifi
April 2009 – March 2012	B. Sc. Agriculture - Animal Sciences, University of Goettingen, Germany Bachelor Thesis: “Statistische und genetische Analyse ausgewählter Merkmale der Schlupffähigkeit und embryonalen Mortalität bei unselektierten Legehennenpopulationen” (Statistical and genetic analysis of the traits Hatchability and Embryonic Mortality in Layers) under the supervision of Prof. Dr. Knorr and Dr. Sharifi

### Research stays

June 2014 – August 2014	Department of Animal Science, Institute of Animal Breeding and Genetics, Leader: Jack C. M. Dekkers Iowa State University Ames, Iowa (USA) Topic of research stay: “Including genomic predictions for health into terminal sire lines in pigs”
-------------------------	--

**Courses**

June 2013	Course of the Society for Animal Sciences (GfT) “Statistische Methoden in Quantitativer Genetik und Tierzüchtung“ (Statistical Methods in Quantitative Genetics and Animal Breeding), Kitzingen, Germany
November 2013	Gene2Farm winter school 2013, Piacenza, Italy
August 2015	Workshop “Writing and Presenting Scientific Papers”

**Scholarships and grants**

August 2015	EAAP Conference scholarship 2015 for the society’s annual meeting in Warsaw, Poland, and the associated course “Writing and presenting papers“
August 2016	Schaumann Stiftung Conference scholarship 2015 for the society’s annual meeting in Belfast, Ireland

## FURTHER PUBLICATIONS

**Große Holthaus A.**, Sharifi A. R., Preisinger R., Cavero D., Schmutz M., Weigend S., Sitzenstock F., Knorr C., Simianer H. (2013). The influence of Different Incubation Regime on the Hatchability Traits in Layers. European workshop of Fundamental Physiology and Perinatal Development in Poultry. Book of abstract of the 6<sup>th</sup> Combined Workshop on Fundamental Physiology and Perinatal Development in Poultry. Göttingen, Germany.

Dematawewa C.M.B., **Grosse Holthaus A.**, Simianer H., Dekkers J.C.M. (2015). Genetic and economic effects of incorporating genomic predictions on health in swine breeding schemes. Book of abstracts of Joint Annual Meeting ADSA-ASAS. Orlando, Florida.

Dematawewa C.M.B., **Grosse Holthaus A.**, Simianer H., Dekkers J.C.M. (2015). Economic gain associated with genomic selection for health in a terminal sire line in pigs. Book of abstracts of Joint Meeting of ASAS Midwestern Section and ADSA Midwest Branch. Des Moines, Iowa.

**Fangmann A.**, Bergfelder-Drüing S., Tholen E., Simianer H., Erbe M. (2015). Can multi-subpopulation reference sets improve the genomic predictive ability? Book of abstracts of the Annual Meeting of The European Federation of Animal Science. Warschau, Polen.

**Fangmann A.**, Sharifi A.R., Erbe M., Simianer H. (2016). Empirical comparison of different methods for genomic evaluation in pigs. Book of abstracts of Annual Meeting of The European Federation of Animal Science. Belfast, Ireland.

Ni G., Cavero D., **Fangmann A.**, Erbe M., Simianer H. (2016). Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. Submitted to Genetics Selection Evolution.

Zhang, Y., Weigend, S., Weigend, A., Preisinger, R., Schmutz, M., **Fangmann, A.**, Simianer, H., Sharifi, R.A. (2018). Statistical and Genetic Analysis of hatchability Traits in Laying Hens. Submitted to Poultry Science.