Accuracy of Genomic Prediction in Dairy Cattle

Dissertation

for the Doctoral Degree

at the Faculty of Agricultural Sciences,

Department of Animal Sciences,

Georg-August-University Göttingen

presented by

Malena Erbe

born in Roth

Göttingen, May 2013

D 7

1st Referee: Prof. Dr. Henner Simianer

Animal Breeding and Genetics Group

Department of Animal Sciences

Georg-August-University Göttingen

2nd Referee: Prof. Dr. Georg Thaller

Institute of Animal Breeding and Husbandry

Christian-Albrechts-University Kiel

Date of disputation: 16th of May, 2013

TABLE OF CONTENTS

SUMMARY

Prediction of genomic breeding values has become a popular tool for obtaining reliable breeding values of animals without phenotypic information, especially in dairy cattle breeding. The main focus of this thesis was to investigate different factors that influence the accuracy of predicted genomic breeding values in real dairy cattle data sets.

A basic study on cross-validation in dairy cattle is presented in ***Chapter 2***. The study was designed to figure out properties of different cross-validation strategies in real data sets. Cross-validation implies dividing the available data set into training and validation set, masking observations of all individuals in the validation set and predicting this information with a model trained with individuals in the training set. In the context of genomic breeding value prediction, cross-validation can be used to assess the accuracy of genomic breeding values for candidates given a specific reference population. The correlation between masked and predicted values for the validation individuals then reflects the accuracy of prediction. The way of how the data is subdivided may influence the results obtained with cross-validation. Thus, this study tried to find an optimal subdivision strategy for different purposes – describing the accuracy for potential candidates when having the available data set for training on the one hand and differentiating between two proposed models on the other hand. A data set of around 2'300 Holstein Friesian bulls genotyped with the Illumina BovineSNP50 BeadChip (termed 50K Chip in the following) was divided in different ways having around 800 up to 2'200 individuals in the training and the remaining individuals in the validation set. Two BLUP approaches, one containing only a random genomic effect and one containing a random polygenic and a random genomic effect, were applied. The highest accuracies could be obtained with the largest training sets, but this also implicates that the validation set was small and the standard error of the obtained accuracy was inflated. Hence, if the purpose is to find significant differences between approaches, larger validation sets are recommended. A five-fold cross-validation in most cases appears to be a good compromise.

Relationship structure between individuals in the training and in the validation set has a big impact on the accuracy of genomic prediction. At the moment, there are still enough progeny-tested training bulls that are highly related to the validation set. If genomic selection will be consequently applied, there may, however, be a lack of such training individuals. Thus, ***Chapter 3*** comprises investigations on how the relationship and age structure influences the accuracy of genomic breeding values of young bulls. A study using 5'698 Holstein Friesian bulls genotyped with the 50K Chip and born between 1981 and 2005 was designed taking always the 500 youngest bulls as a constant validation set over all scenarios. Different training sets of 1'500 individuals were used to predict genomic breeding values for those candidates: bulls were chosen randomly, were among the oldest or youngest bulls in the remain-

ing data set, had a maximum relationship of less than 0.25 or 0.5 to the candidates or were highest related with the candidates. Compared to the random standard scenario, reduced relationship levels led to an apparent decrease of accuracy in prediction. Accuracy of prediction was higher for the scenarios with the highly related individuals and with the youngest bulls in the training sets. For practical application in strongly related samples like progeny tested Holstein Friesian bulls there is not critical point as long as sires, half or full sibs are in the training set. New progeny tested bulls should therefore be continuously added to the training set. Otherwise a clear decrease of accuracy of prediction will be observable even after only one or two generations.

Chapter 4 deals with two further factors that may influence the accuracy of genomic prediction, namely the marker density and the method used for prediction. Usually 50K SNPs have been used for genomic prediction so far, but lately a new high density (HD) SNP array with 777K has become available. Thus, the question has arisen whether higher marker density will lead to an increase in prediction accuracy. The more markers have to be modeled the more important may be the development of prediction approaches that allow a proportion of SNPs to be uninformative. Therefore, a new and efficient Bayesian method (*BayesR*) was developed assuming SNP effects are derived from a series of normal distributions that have different variances and with the number of SNPs per distribution being not fixed but modeled with a Dirichlet distribution. Furthermore, this chapter also addresses the issue of multi-breed training sets with different marker densities. In dairy cattle, large training sets are necessary to obtain robust estimates of SNP effects, but building large reference sets may be challenging for smaller breeds. Multi-breed training sets can be an option to overcome this problem. With 50K marker sets the increase in accuracy, however, was very limited, probably because of a lack of consistent phases between breeds with this marker density. Having a high density marker set available should thus be beneficial also for the across breed prediction. Data sets of Australian Holstein and Australian Jersey bulls, all genotyped with the 50K Chip and imputed to 777K, were used to investigate the changes in accuracy of genomic prediction within and between breeds with a GBLUP approach and *BayesR*. Using imputed high density data did not lead to a significant increase of accuracy for the within breed situation and led only to a small increase in the multi-breed scenario for the minor breed. *BayesR* always produced comparable or better results than the GBLUP approach. An additional feature of *BayesR* is that one can learn more about the architecture of quantitative traits, e.g. by considering the average number of SNPs in the different distributions.

The accuracy of genomic prediction when having data available can be calculated using different validation procedures. However, in some situations, it may be useful to assess the expected accuracy of prediction in advance of a genomic breeding value prediction study,

e.g. because one wants to know the required size of the training set or the SNP density to achieve a predefined level of accuracy. Different deterministic equations to predict the accuracy level have been suggested in the literature and all rely more or less on the same parameters. One of these parameters is the number of independently segregating chromosome segments (M_e) that is normally determined based on theoretical population parameters like effective population size (N_e) of the underlying population. In ***Chapter 5***, a maximum likelihood approach is presented that allows determining the number of M_e empirically based on a systematic multi-level cross-validation. Based on this, various deterministic prediction equations were compared and modified to fit best to the data. 5'698 Holstein bulls genotyped with the 50K Chip and 1333 Brown Swiss bulls genotyped with the 50K Chip and imputed to 777K SNPs were used for cross-validation studies with different k-fold scenarios (k=2, 3, …, 10, 15, 20) in a genomic BLUP framework. It was thus possible to mimic genomic prediction with different sizes of training sets based on different subsets of SNPs (10'000, 20'000, 30'000, 42'551 SNPs for Holstein Friesian and 2'451, 4'901, …, 627'306 SNPs for Brown Swiss) to study the influence of the SNP density. A maximum likelihood approach was then used to estimate the best value for the number of M_e based on the empirical observed data. The highest likelihood was obtained when using a modified form of the deterministic equation of Daetwyler *et al.* (2010, Genetics **185:**1021-1031) as expected accuracy. The most likely values for M_e using all available markers were 1'241 (412) and 1'046 (197) for the traits somatic cell score and milk yield in Holstein Friesian (Brown Swiss), respectively. Values of M_e were different in Brown Swiss and Holstein Friesian, while N_e of both populations calculated from pedigree and linkage disequilibrium structure was very similar. Having those results at hand it seems that M_e is not a parameter that can be easily modeled by the effective population size and the genome length deterministically since it varies between traits within population and even between populations with similar structure. The modification of the formula of Daetwyler *et al.* (2010) consists of adding a weighting factor based on the assumption that the maximal achievable accuracy with a given SNP set is not one. This was assumed due to the fact that not all of the genetic variance can be captured by the available SNP set. Values for the squared weighting factor, i.e. the percentage of genetic variance captured, were also empirically determined and were between 76% and 82% with SNP subsets of 10'000 up to 42'551 SNPs for Holstein Friesian and between 63% and 75% with SNP subsets of 2'451 up to 627'306 SNPs for Brown Swiss. There is a linear relationship between the weighting factor and the logarithm of the marker density up to a population specific marker density (e.g. ~ 20'000 in Brown Swiss) which ends in a plateau, i.e. adding more SNPs will not change the proportion of genetic variance captured.

ZUSAMMENFASSUNG

Die genomische Zuchtwertschätzung ist vor allem im Bereich der Milchrinderzucht in den letzten Jahren zu einer beliebten Methode geworden, um sichere Zuchtwerte von Tieren ohne phänotypische Information zu erhalten. Das Ziel dieser Arbeit war es, verschiedene Einflussfaktoren auf die Genauigkeit der genomischen Zuchtwertschätzung in realen Rinderdatensätzen genauer zu untersuchen.

In ***Kapitel 2*** findet sich eine grundlegende Arbeit zur Kreuzvalidierung, in der die Eigenschaften verschiedener Kreuzvalidierungsstrategien in realen Datensätzen untersucht wurden. Kreuzvalidierung bedeutet, dass die verfügbaren Daten in eine Trainings- und eine Validierungsstichprobe aufgeteilt werden, wobei für die Individuen in der Validierungsstichprobe alle Beobachtungswerte als nicht vorhanden angenommen werden. Die Werte der Individuen in der Validierungsstichprobe werden dann mit einem Modell, das mit Hilfe der Beobachtungswerte der Individuen in der Trainingsstichprobe angepasst wird, vorhergesagt. Im Kontext der genomischen Zuchtwertschätzung werden Kreuzvalidierungsstrategien benutzt, um die Genauigkeit der genomischen Zuchtwertschätzung mit einer bestimmten Trainingspopulation abzubilden. Die Korrelation zwischen maskierten und vorhergesagten Werten der Tiere in der Validierungsstichprobe spiegelt die Genauigkeit der genomischen Zuchtwertschätzung wider. Die Art und Weise, wie der Datensatz in Trainings- und Validierungsstichprobe unterteilt wird, kann die Ergebnisse einer Kreuzvalidierung beeinflussen. Das Ziel dieser Studie war es deshalb, optimale Strategien für unterschiedliche Zwecke – Beschreibung der Genauigkeit der genomischen Vorhersage für mögliche Selektionskandidaten mit dem vorhandenen Datensatz oder Vergleich von zwei Methoden zur Vorhersage – zu finden. Ein Datensatz von etwa 2‘300 Holstein Friesian-Bullen, die mit dem Illumina BovineSNP50 BeadChip (im Folgenden 50K Chip genannt) typisiert waren, wurde unterschiedlich aufgeteilt, so dass sich zwischen 800 bis 2‘200 Tiere in der Trainingsstichprobe und die jeweils restlichen Tiere in der Validierungsstichprobe befanden. Zwei BLUP-Modelle, eines mit einem zufälligen genomischen Effekt und eines mit einem zufälligen polygenen und einem zufälligen genomischen Effekt, wurden zur Vorhersage verwendet. Die höchste Genauigkeit der Vorhersage konnte mit der größten Trainingsstichprobe erreicht werden. Eine große Trainingsstichprobe bei gegebenem limitierten Datenmaterial impliziert aber auch, dass gleichzeitig die Validierungsstichproben klein und damit die Standardfehler der beobachteten Genauigkeiten sehr hoch sind. Falls es das Ziel einer Studie ist, signifikante Unterschiede zwischen Modellen nachzuweisen, ist es besser größere Validierungsstichproben zu verwenden. Eine fünffache Kreuzvalidierung scheint in vielen Fällen ein guter Kompromiss zu sein.

Die Verwandtschaftsstruktur zwischen den Tieren in der Trainings- und der Validierungsstichprobe hat einen großen Effekt auf die Genauigkeit der genomischen Zuchtwertschät-

zung. Momentan sind noch genügend nachkommengeprüfte Bullen in den Trainingsstichproben vorhanden, mit denen die Tiere in der Validierungsstichprobe hoch verwandt sind. Wenn die genomische Selektion konsequent angewendet wird, ist es möglich, dass solche Individuen für die Trainingsstichprobe knapper werden. Deshalb enthält ***Kapitel 3*** eine Studie, die untersucht, wie sich die Verwandtschafts- und Altersstruktur auf die Genauigkeit der genomischen Zuchtwerte von jungen Bullen auswirkt. Ein Datensatz mit 5‘698 Bullen der Rasse Holstein Friesian, die alle mit dem 50K Chip typisiert wurden und zwischen 1981 und 2005 geboren wurden, war die Basis dieser Arbeit. In allen Szenarien wurden die 500 jüngsten Bullen dieses Datensatzes als Validierungsstichprobe verwendet. Verschiedene Trainingsstichproben mit je 1‘500 Individuen wurden ausgewählt, um die genomischen Zuchtwerte der jungen Tiere (Selektionskandidaten) vorherzusagen: eine zufällige Auswahl an Bullen, die ältesten und jüngsten verfügbaren Tiere, Tiere mit Verwandtschaftskoeffizienten kleiner 0.25 oder 0.5 zu allen Selektionskandidaten, oder Tiere, die am stärksten mit den Selektionskandidaten verwandt waren. Verglichen mit dem Szenario mit der zufälligen Auswahl führte eine Verringerung der Verwandtschaft zu einer sichtbaren Abnahme der Genauigkeit der genomischen Vorhersage. Die Genauigkeit für die Szenarien mit den hoch verwandten Tieren bzw. den jüngsten Tieren in der Trainingsstichprobe war hingegen höher. Für die praktische Anwendung bedeutet dies, dass in stark verwandten Gruppen wie Elitebullen der Rasse Holstein Friesian keine weiteren Probleme für die Vorhersage junger Tiere zu erwarten sind, solange Väter, Voll- und Halbgeschwister in der Trainingsstichprobe vorhanden sind. Neue nachkommengeprüfte Bullen sollten deshalb kontinuierlich zur Trainingsstichprobe hinzugefügt werden – sonst wird eine klare Abnahme der Genauigkeit schon nach ein oder zwei Generationen zu sehen sein.

Kapitel 4 beschäftigt sich mit zwei weiteren Faktoren, die die Genauigkeit der genomischen Vorhersage beeinflussen können: Markerdichte und Methodenwahl. Bis jetzt wurden normalerweise 50K SNPs für die genomische Zuchtwertschätzung verwendet, aber seit Kurzem ist auch ein neues hochdichtes SNP-Array mit 777K SNPs verfügbar. Dies lässt die Frage aufkommen, ob die höhere Markerdichte zu einem Anstieg in der Genauigkeit führen kann. Je mehr Marker verfügbar sind, umso größer wird auch die Notwendigkeit, Methoden zu entwickeln, die einen Teil der Marker als nicht informativ (d.h. ohne Effekt auf das untersuchte Merkmal) zulassen. Deshalb wurde eine neue und effiziente Bayes’sche Methode (*BayesR*) entwickelt, die annimmt, dass die SNP Effekte aus einer Reihe von Normalverteilungen stammen, die unterschiedliche Varianzen haben. Die Anzahl der SNPs pro Verteilung wird nicht festgesetzt, sondern mit Hilfe einer Dirichlet-Verteilung modelliert. In ***Kapitel 4*** wird außerdem auf die Frage eingegangen, wie sich die Genauigkeit der Vorhersage im Fall von Trainingsstichproben mit mehreren Rassen bei unterschiedlicher Markerdichte verhält. Bei Milchrinderrassen sind große Trainingsstichproben erforderlich, um robuste Schätzer der

SNP-Effekte zu erhalten, aber gerade bei kleinen Rassen kann es schwierig sein, solch große Trainingsstichproben aufzubauen. Trainingsstichproben, die Tiere mehrerer Rassen enthalten, können deshalb eine Möglichkeit sein, dieses Problem zu umgehen. Mit 50K SNPs war der Erfolg solcher Mehrrassen-Trainingsstichproben gering, was darauf zurückgeführt wurde, dass die Haplotypenstruktur über die Rassen hinweg bei dieser Markerdichte nicht konsistent war. Der hochdichte SNP-Chip könnte hier allerdings Verbesserungen für die Vorhersage über Rassen hinweg bringen. Die Veränderungen in der Genauigkeit der genomischen Zuchtwertschätzung innerhalb einer Rasse und über Rassen hinweg wurden mit Daten von australischen Bullen der Rassen Holstein Friesian und Jersey, die mit dem 50K Chip typisiert und auf 777K SNPs imputet waren, und zwei verschiedenen Methoden (GBLUP, *BayesR*) untersucht. Die Verwendung von imputeten hochdichten Markern führte zu keinem signifikanten Anstieg der Genauigkeit innerhalb einer Rasse und nur zu einer geringen Verbesserung der Genauigkeit in der kleineren Rasse im Mehrrassen-Szenario. *BayesR* lieferte gleichwertige oder in vielen Fällen höhere Genauigkeiten als GBLUP. Eine Eigenschaft von *BayesR* ist außerdem, dass es möglich ist, aus den Ergebnissen Erkenntnisse zur genetischen Architektur des Merkmals zu erhalten, z.B. indem man die durchschnittliche Anzahl an SNPs in den verschiedenen Verteilungen betrachtet.

Die Genauigkeit der genomischen Zuchtwertschätzung kann mit verschiedenen Validierungsprozeduren berechnet werden, sobald reale Daten vorhanden sind. In manchen Situationen kann es jedoch von Vorteil sein, wenn man die erwartete Genauigkeit der Vorhersage im Vorfeld einer Studie abschätzen kann, z.B. um zu wissen, welche Größe die Trainingsstichprobe haben sollte oder wie hoch die Markerdichte sein sollten, um eine bestimmte Genauigkeit zu erreichen. Verschiedene deterministische Formeln zur Abschätzung der erreichbaren Genauigkeit sind in der Literatur verfügbar, die alle auf den mehr oder weniger gleichen Parametern beruhen. Einer dieser Parameter ist die Anzahl unabhängig segregierender Chromosomensegmente (M_e), die normalerweise mit Hilfe von theoretischen Werten wie der effektiven Populationsgröße (N_e) deterministisch bestimmt wird. In ***Kapitel 5*** wird ein Maximum-Likelihood Ansatz beschrieben, der es ermöglicht, M_e basierend auf systematisch angelegten Kreuzvalidierungsexperimenten empirisch zu bestimmen. Darauf aufbauend wurden verschiedene deterministische Funktionen zur Vorhersage der Genauigkeit verglichen und so modifiziert, dass sie am besten zu den vorhandenen Datensätzen passten. Mit 5‘698 Holstein Friesian-Bullen, die mit dem 50K Chip typisiert waren, und 1‘333 Braunvieh-Bullen, die mit dem 50K Chip typisiert und auf 777K SNPs imputet waren, wurden mit GBLUP verschiedene k-fache Kreuzvalidierungen (k=2, 3, …, 10, 15, 20) durchgeführt. So konnte eine genomische Zuchtwertschätzung bei unterschiedlichen Größen der Trainingsstichprobe nachgebildet werden. Weiterhin wurden alle Szenarien mit verschiedenen Subsets der vorhandenen SNPs (10‘000, 20‘000, 30‘000, 42‘551 SNPs für Holstein Friesian, und

jeder, jeder zweite, jeder 4., … jeder 256. SNP für Braunvieh) durchgeführt, um den Einfluss der Markerdichte erfassen zu können. Der Maximum-Likelihood Ansatz wurde angewendet, um M_e für die beiden vorhandenen Datensätze bestmöglich zu schätzen. Die höchste Likelihood wurde erreicht, wenn eine modifizierte Form der deterministischen Formel von Daetwyler *et al.* (2010, Genetics **185:**1021-1031) für die Modellierung der erwarteten Genauigkeit die Grundlage bildete. Die wahrscheinlichsten Werte für M_e, wenn alle vorhandenen Marker genutzt wurden, waren 1‘241 (412) und 1‘046 (197) für die Merkmale Zellzahl und Milchmenge für Holstein Friesian (Braunvieh). Die Werte für M_e für Braunvieh und Holstein Friesian unterschieden sich deutlich, während N_e für beide Populationen (berechnet auf Basis des Pedigrees oder über die Struktur des Kopplungsungleichgewichts) sehr ähnlich war. Die Schätzungen für M_e variierten zwischen verschiedenen Merkmalen innerhalb von Populationen und über Populationen mit ähnlichen Populationsstrukturen hinweg. Dies zeigt, dass M_e wahrscheinlich kein Parameter ist, der sich nur aus N_e und der Länge des Genoms berechnen lässt. Die Modifizierung der Formel von Daetwyler *et al.* (2010) bestand darin, einen Gewichtungsfaktor hinzuzufügen, der berücksichtigt, dass die maximale Genauigkeit bei gegebener Markerdichte auch mit unendlich großer Trainingsstichprobe nicht 1 sein muss. Dies basiert auf der Annahme, dass die vorhandenen SNPs nicht die ganze genetische Varianz wiedergeben können. Auch dieser Gewichtungsfaktor wurde empirisch bestimmt. Die quadrierten Werte, d.h. der Prozentsatz der genetischen Varianz, die erklärt wird, lagen zwischen 76% und 82% für 10‘000 bis 42‘551 SNPs bei Holstein Friesian und zwischen 63% und 75% für 2‘451 bis 627‘306 SNPs bei Braunvieh. Zwischen dem natürlichen Logarithmus der Markerdichte und dem Gewichtungsfaktor bestand ein linearer Zusammenhang bis zu einer populationsspezifischen Grenze hinsichtlich der Markerdichte (~ 20‘000 SNPs bei Braunvieh). Oberhalb dieser Grenze fand sich ein Plateau, was bedeutet, dass das Hinzufügen von weiteren Markern den Anteil der genetischen Varianz, der erklärt wird, nicht mehr verändert.

1st CHAPTER

General Introduction

GENERAL INTRODUCTION

The aim of this thesis is to investigate different factors that influence the accuracy of genomic breeding value prediction. This chapter therefore provides a short history and description of this breeding approach and introduces the relevant methodology.

(Genomic) Breeding value estimation

A comprehensive system for estimating reliable breeding values is one of the key points of an efficient breeding program and a useful selection process. The introduction of best linear unbiased prediction (BLUP) (e.g. Henderson, 1975) set a benchmark in the field of animal breeding. Based on BLUP systems, individual breeding values with maximum achievable reliability can be obtained based on pedigree information across many generations and phenotypic information from the individual itself or from any relatives. Besides the traditional animal model, different models have been developed that are able to handle different breeding programs and/or data structure, e.g. sire models for reducing computational demands when breeding values should be calculated only for sires based on progeny records, multi-trait models for combining correlated traits in one model (e.g. Henderson & Quaas, 1976) in which missing values are not that critical, or random regression models for processing longitudinal and test-day data (e.g. Schaeffer, 2004). At least for the production traits, nowadays all conventional evaluation systems in dairy cattle are based on such conventional BLUP approaches. Procedures like Multiple(-Trait) Across Country Evaluations (MACE; e.g. Schaeffer, 1994) have made it possible to compare conventional breeding values on an international scale as well. With the availability of the first genetic markers in the late 80s and 90s of the 20th century, discussions have started on how this new information could be used to improve selection schemes, i.e. introducing a so called marker-assisted selection (MAS).

Most of the traits studied in livestock breeding have a quantitative genetics background which means that the observed phenotypes are on a continuous scale and the observed genetic variance is caused by more than one gene. All gene loci that contribute to the variation in a specific trait are called quantitative trait loci (QTL). Often it is not known where in the genome they are located and how large their contribution to genetic variance is. Early studies have proposed that the number of loci influencing a specific trait will be small to medium (e.g. Hayes & Goddard, 2001), but nowadays the general opinion is that most traits are probably influenced by hundreds of loci with most of them having a very small effect on the trait (e.g. Reed *et al.,* 2008). Since positions of possible QTL are mostly unknown in advance, genetic markers with known positions can be used as proxies. If QTL and marker are located near to each other, they are often in high linkage disequilibrium which enables a large proportion of

the genetic variance caused by the QTL to be captured by the marker. Genetic markers in those days were normally a small set (few hundred) of microsatellites or restriction fragment length polymorphisms which were thought to be a good basis to find positions of relevant QTL. Different statistical approaches have been developed to map QTL positions based on effects of markers on phenotypes (e.g. Sillanpää & Corander, 2002; Meuwissen & Goddard, 2004). However, effects have often been overestimated (e.g. Utz *et al.*, 2000) and could not be confirmed in an independent data set which made it impossible to include MAS in a regular breeding scheme. The success of MAS has never really been stunning across livestock species with the only exception in dairy cattle being the discovery of DGAT1 (Grisart *et al.*, 2004) and France being the only country that really has consequently implemented MAS within a breeding program (Guillaume *et al.*, 2008) for a longer time.

In 2001, the idea of using dense marker sets to predict total genetic values came up (Meuwissen *et al.*, 2001) which has revolutionized the field of animal breeding in a way and at a speed not shown by many innovations before. The idea behind this approach is that with dense marker maps (thousands or tens of thousands of markers) all QTL affecting a specific trait will be in high linkage disequilibrium with at least one marker or chromosomal segment. This is why it should be possible to capture all or a major part of the genetic variance of a trait with a sufficiently dense marker map. Despite looking for particular QTL with large effects in the previous MAS approach Meuwissen *et al.* (2001) described statistical approaches where effects of many markers spread across the genome or of the respective haplotypes are estimated simultaneously. Without applying any significance threshold, all marker or haplotype effects are summed up afterwards to obtain the total genetic value (which will later be called “genomic breeding value”) of an individual.

The advantages of selection based on genomic breeding values over conventional schemes are clear: Using genomic information directly makes it possible to capture Mendelian sampling effects which is not possible with pedigree-based approaches. This may have a positive effect on the inbreeding rate per generation (Lillehammer *et al.*, 2011) and the accuracy of breeding values. Given that a sufficient number of individuals with phenotypes are available to estimate the marker effects, genomic breeding values can be obtained also for individuals that are not phenotyped, but just genotyped. This means that accurate breeding values for young individuals or even embryos are available and selection (“genomic selection”) is possible based on these genomic breeding values. In the years after this idea had come up, many studies used simulated data sets (e.g. Habier *et al.*, 2007; Solberg *et al.*, 2008; de Roos *et al.*, 2009) to test different prediction approaches and implementation scenarios and ideas on how to integrate genomic selection into existing breeding programs from e.g. an economical point of view were based on deterministic considerations (e.g. Schaeffer, 2006)

since appropriate data have not yet been available to assess the impact of genomic breeding values and genomic selection in real data.

Availability of SNP data

In 2001, it was not clear when appropriate data would be available to predict genomic breeding values with a level of reliability that is necessary for an application under practical circumstances in routine evaluations. It has been favorable for this approach that the full sequence of the bovine genome became available in 2009 (e.g. Liu *et al.*, 2009; Zimin *et al.*, 2009) and that the genotyping technology made a great leap forward in the first years of this century so that a huge amount of genomic marker data have become available up to now.

In genomes of mammals, different kinds of sequence variants exist that can be used as markers – amongst others microsatellites, copy number variations, insertions, deletions and single nucleotide polymorphisms (SNPs). For practical implementation of genomic breeding value prediction, genome-wide markers roughly distributed equally over the genome and available in large quantity are necessary. SNPs fulfill these criteria and are therefore an optimal marker type for genomic prediction approaches. A SNP is a polymorphism that occurs at a single base and is normally biallelic. In mammalian genomes, millions of those SNPs are available (e.g. 2.44 Mio SNPs have been discovered in a single Simmental bull (Eck *et al.*, 2009); 15.8 Mio within 133 Holstein Friesian and Simmental bulls (Hayes *et al.*, 2012)). With new technologies, it is possible to obtain genotypes for an individual at many thousands of SNPs in one step at reasonable costs by using a SNP array within high-throughput genotyping platforms. Two world-wide acting companies, Illumina Inc. (http//www.illumina.com) and Affymetrix Inc. (http://www.affymetrix.com), provide commercial and customized SNP arrays of which Illumina's arrays comprising around 6'000 (Illumina BovineLD BeadChip), 54'000 (Illumina Bovine50 BeadChip, referred to as "50K Chip" in the following) and 777'000 SNPs (Illumina BovineHD BeadChip) are most common in cattle. Up to now, in official genomic evaluation, SNPs of the 50K Chip mostly build the basis of genomic breeding values since most elite bulls have been genotyped with this SNP array.

Imputation of genotypes

Two years ago, the new SNP array, the Illumina BovineHD BeadChip (referred to as "HD chip" in the following), became available which includes around 777.000 SNPs. Scientists awaited this new array eagerly since it was hoped that a more than 10-fold higher marker

density compared to the 50K Chip would increase the accuracy of genomic prediction considerably and would have great advantages for small breeds since multi-breed prediction would be more successful. Findings with the HD Chip in real data will be discussed in ***Chapter 4*** and in the ***General Discussion***.

Many individuals, especially many progeny-tested bulls, had been genotyped with the 50K chip or low-density SNP chips before the HD chip has become available, so it was clear from the beginning that not all of those individuals would be re-genotyped with the HD chip. It was thus necessary to follow another strategy which is called "imputation". Imputation aims at reconstructing genotypes of un-genotyped marker loci in silico. The basic steps of an imputation process are always the same: Assume a sample of individuals genotyped with a marker set A is available and these individuals should be imputed to a larger marker set B. First, another sample of individuals genotyped with marker set B must be available ("reference"). Second, haplotypes are reconstructed ("phased") based on relationship-based linkage and/or populations-wide linkage disequilibrium for individuals genotyped with marker set B and individuals genotyped with marker set A using an appropriate software tool. Based on those haplotypes, alleles at marker loci not included in marker set A but in B can be reconstructed for individuals that have just been genotyped with marker set A.

Different software for reconstructing haplotypes and imputing missing data has been developed especially in the human genetics community (e.g. "fastPHASE" by Scheet & Stephens, 2006; "MaCH" by Li *et al.*, 2010; "ShapeIt" by Delaneau *et al.*, 2012). Many of these programs, however, have limited ability to process hundreds or thousands of samples with tens of thousands of SNPs in an acceptable time frame or they are not able to process data without a reference set with predefined haplotypes. One exception is BEAGLE (Browning & Browning, 2007) which is widely used in the human genetics framework as well as in the field of livestock genetics and provides all necessary features. To overcome the problems previously described, further software has been developed in the last years in the livestock breeding sector, too (e.g. "findhap" by VanRaden *et al.*, 2011; "FImpute" by Sargolzaei *et al.*, 2011; "AlphaImpute" by Hickey *et al.*, 2011).

Apart from the choice of the program the size and the composition of the reference set are the two factors that mainly influence the accuracy of imputation (e.g. Pausch *et al.*, 2013). Larger reference sets and a larger number of near relatives apparently increase the accuracy; however, the more animals have to be genotyped with the higher marker density, the more costs will be incurred. One of the strategies often used is therefore to select key ancestors in a way that the proportion of genes they have contributed to the actual population is maximized (Goddard & Hayes, 2009) and to genotype these ancestors with the HD chip.

Recently, there have been studies available with real data that assessed the accuracy of imputation when imputing up to the HD chip. Erbe *et al.* (2012) found imputation accuracies with “BEAGLE” for Australian Holstein Friesian and Australian Jersey bulls based on around 100 HD genotyped key ancestors of 97.5% and 95.6%, respectively. Brøndum *et al.* (2012) compared correlations between true and imputed genotypes in different Nordic breeds with single breed and multi-breed reference sets and obtained values of around 0.93 (0.95) in Danish Red and 0.97 (0.98) in Finnish Ayrshire with single (multi) breed references of around 200 (556) individuals using “BEAGLE”. A sire in the reference set improved the accuracy and decreased the allele error rate in the imputed offspring. With around 1100 individuals genotyped with HD VanRaden *et al.* (2013) showed that more than 99% of the genotypes could be imputed correctly with “findhap” in Holstein Friesian bulls genotyped with 50K. Pausch *et al.* (2013) investigated different imputation methods in Simmental data and found imputation accuracies of greater than 0.97 with only 100 key ancestors in the reference using a combination of pre-phasing with “BEAGLE” and imputing afterwards with “MiniMac” (Howie *et al.*, 2012).

From these results it can be concluded that imputing genotypes from 50K to HD is feasible and accurate so that imputed genotypes can be used for further studies. In this thesis, imputed high density genotypes will be the basis for genomic prediction within and between dairy cattle breeds in ***Chapter 4***.

Genomic evaluation and selection in dairy cattle

Why do genomic selection schemes have such a striking success especially in dairy cattle? Four parameters determine the genetic gain of a breeding scheme: Genetic standard deviation, selection intensity, accuracy of breeding value estimation and the generation interval. In the following, the genetic standard deviation is assumed to be constant.

With classical progeny testing schemes in dairy cattle the accuracy of breeding values is very high in progeny tested bulls (up to 0.99) but high accuracies can only be obtained when many performance records of daughters become available (normally >80, Pryce & Daetwyler, 2012), i.e. when the bull is already at least 5 years old. For young bulls, a parent average can be calculated but this is too imprecise to build a basis for concrete selection decisions and just a pre-selection, namely which bull becomes a testing bull, is done at that point in time. Therefore, generation intervals on the bulls’ side are quite high and are the limiting factor in classical breeding schemes. A further point is that keeping testing bulls over years up to the point where selection will be made based on progeny records is quite expen-

sive. Genetic gain could thus be increased and costs could be reduced a lot if more accurate breeding values of a bull can be obtained earlier in life. Genomic breeding values can be predicted for young individuals not as accurately as with progeny performance but accurately enough. This will allow two strategies: a more precise pre-selection for testing bulls is possible and/or young bulls can be directly used without waiting for any progeny records. Furthermore, genomic selection could also be applied in the bull dams' path allowing the selection of elite cows taking place earlier in life and being more precise.

Schaeffer (2006) showed with deterministic considerations that in a classical four path breeding scheme genetic gain can be doubled and costs per bull can be reduced dramatically when applying genomic selection consequently in the bulls' and the bull dams' path. In actual studies with stochastic simulations of genomic breeding schemes, restrained values of around 20% (Lillehammer *et al.*, 2011) up to extreme values of over 100% (e.g. de Roos *et al.*, 2011) increase in genetic gain can be found depending on selection intensity and generation interval assumed in the studies.

A regular and official genomic evaluation which is the basis for genomic selection was first introduced for the breed Holstein Friesian in the US and Canada in 2009 and many countries have followed since that time (e.g. Germany in 2010, Australia in 2011). But efforts have also been made to use genomic breeding values for other dairy breeds (e.g. USA for Jersey and Brown Swiss in 2009, Germany for Simmental and Brown Swiss in 2011). Most countries have started with small training sets of a few hundred or a few thousands individuals genotyped with 50K SNPs, but as the number of individuals in the training set is crucial, various cooperation consortia have been established (e.g. EuroGenomics in Holstein Friesian (David *et al.*, 2010), Intergenomics in Brown Swiss (Zumbach *et al.*, 2010), etc.) which helps to improve accuracy of genomic prediction.

Up to now, the procedure to estimate/predict genomic breeding values for bulls in a genomic evaluation is a two-step-method, i.e. first a classical breeding value estimation based on pedigree information and progeny records is performed for all proven bulls. Outcomes of this step are then used as dependent variables in the genomic breeding value estimation. Almost all genomic evaluation systems are based on a best linear unbiased prediction system to predict genomic breeding values whose basic methodology will be described below. Different variables can be used as quasi-phenotypes which all have some advantages and some disadvantages: Estimated breeding values themselves, deregressed proofs (Garrick *et al.*, 2009) or daughter yield deviations (VanRaden & Wiggans, 1991). In this thesis, estimated breeding values will be used in ***Chapter 2***, ***3*** and ***5*** while daughter yield deviations will be used in ***Chapter 4***.

Methods in genomic breeding value prediction

For the following models, m is defined as the number of SNPs, n as the number of all genotyped individuals, and n_o as the number of genotyped individuals with observations.

BLUP framework

The simplest best linear unbiased prediction (BLUP) model in the genomic context is the following:

$$\mathbf{y} = \mathbf{1}'_{n_o}\mu + \boldsymbol{Z}_o\mathbf{b} + \mathbf{e} \quad [1]$$

where $\mathbf{y}$ is a vector of observations (quasi-phenotypes), μ is an overall mean, $\boldsymbol{Z}_o$ is a matrix of genotypes of individuals with observations and is of dimension $n_o\, x\, m$, $\mathbf{b}$ is a vector of random SNP effects and $\mathbf{e}$ is a vector of random residual effects. $\mathbf{b}$ is assumed to be normally distributed with $\mathbf{b} \sim N(0, \mathbf{I}\sigma^2_{SNP})$ and $\mathbf{e}$ is assumed to be normally distributed with $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$. This model is often referred to as *random regression BLUP*, *ridge regression BLUP* or *RR-BLUP*. In this model, only individuals with observations are used for estimating the marker effects directly, but genomic breeding values ($\mathbf{g}$) can be predicted in the next step also for any further genotyped individuals using

$$\hat{\mathbf{g}} = \mathbf{Z}\hat{\mathbf{b}} \text{ with } \mathbf{Z} = \begin{bmatrix} \boldsymbol{Z_o} \\ \boldsymbol{Z_{no}} \end{bmatrix}$$

and $\boldsymbol{Z_{no}}$ is a matrix of genotypes of individuals without observations.

With e.g. 50K SNPs, however, more than 50.000 SNP effects have to be estimated with this model. This may be computationally very extensive, especially if a variance component estimation step is included. Furthermore, animal breeders are often less interested in the SNP effects themselves, but more in genomic breeding values. Different authors (Habier *et al.* 2007; Goddard, 2009; Hayes *et al.*, 2009) have shown that an equivalent model to RR-BLUP exists that leads to the solution for genomic breeding values directly. For this, we first have to define any genomic relationship matrix $\mathbf{G}$ with the form

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{\boldsymbol{t}}$$

where $\mathbf{Z}$ is the same as before and $\boldsymbol{t}$ is a scaling factor. $\mathbf{G}$ will be used to model the covariance matrix between individuals. Now, model

$$\mathbf{y} = \mathbf{1}'_{n_o}\mu + \mathbf{W}\mathbf{g} + \mathbf{e} \quad [2]$$

with $\sigma_g^2 = t \cdot \sigma_{SNP}^2$, $\mathbf{g}$ being a vector of genomic breeding values with $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ and $\mathbf{W}$ being an incidence matrix relating observations to the random effects. Note that $\mathbf{G}$ may contain individuals without observations. This model is often referred to as *genomicBLUP* or *GBLUP*.

The phenotypic variance covariance matrix of $\mathbf{y}$ in [1] is

$$\mathbf{V_1} = \mathbf{Z_o I}\sigma_{SNP}^2\mathbf{Z_o'} + \mathbf{I}\sigma_e^2 = \mathbf{Z_o Z_o'}\sigma_{SNP}^2 + \mathbf{I}\sigma_e^2$$

and the phenotypic variance covariance matrix of $\mathbf{y}$ in [2] is

$$\mathbf{V_2} = \mathbf{WG}\sigma_g^2\mathbf{W'} + \mathbf{I}\sigma_e^2 = \frac{\mathbf{Z_o Z_o'}}{t} \cdot t \cdot \sigma_{SNP}^2 = \mathbf{Z_o Z_o'}\sigma_{SNP}^2 + \mathbf{I}\sigma_e^2 = \mathbf{V_1}$$

Therefore, [1] and [2] will lead to the same solution for fixed effects and genomic breeding values. Equation [2] provides many advantages: It is computationally less demanding since the number of random effects that have to be estimated equals the number of individuals which is usually much smaller than the number of markers. Second, all individuals, those with observations and without observations, can be handled in one step and estimates of genomic breeding values are obtained directly. Furthermore, even when solving [2] first, SNP effects can be calculated afterwards based on the variance components estimates in [2] without the necessity of modeling matrices of dimension markers x markers (e.g. Strandén & Garrick, 2009):

$$\hat{\mathbf{b}} = \frac{\hat{\sigma}_g^2}{t}\mathbf{Z_o}'\left(\mathbf{WG}\sigma_g^2\mathbf{W'} + \mathbf{I}\hat{\sigma}_e^2\right)^{-1}\left(\mathbf{y} - \mathbf{1}'_{n_o}\hat{\mu}\right) = \frac{\hat{\sigma}_g^2}{t}\mathbf{Z_o}'\mathbf{V_2}^{-1}\left(\mathbf{y} - \mathbf{1}'_{n_o}\hat{\mu}\right)$$

For all derivations so far, $\mathbf{1}'_{n_o}\hat{\mu}$ could be replaced with a general $\mathbf{X\beta}$, i.e. any type of fixed effects can be modeled, just as well further random effects could be included in the model. In many practical applications, a random polygenic effect is added to the model (e.g. Liu *et al.*, 2011) assuming that the available markers cannot capture all genetic variance (see Dekkers, 2007). GBLUP models for predicting genomic breeding values will be used in ***Chapters 2***, ***3***, ***4*** and ***5*** of this thesis.

Construction of the genomic relationship matrix

There are different approaches how to build a genomic relationship matrix. The aim is always to use all available marker information to describe the covariance structure between genotyped individuals. While the pedigree based relationship matrix presents expected relationship coefficients between two individuals, any genomic relationship matrix shows the realized values of relationship and is assumed to be more accurate than the one based on pedigree. When using realized values, Mendelian sampling effects are accounted for in the relationship

coefficients while this is not possible when using expectations. This also means that it is possible to distinguish further between individuals e.g. within full-sib groups where all members always have the same expected value, but may differ in the realized relationships. In dairy cattle, the difference between pedigree based and marker based relationships is reflected most notably in increased accuracies of prediction for young individuals when predicting genomic breeding values instead of using the pedigree index (e.g. VanRaden *et al.*, 2009).

One of the first studies that presented a marker based relationship matrix was Hayes & Goddard (2008). They calculated the relationship based on the concept of a similarity index (Eding & Meuwissen, 2001), a method that has not been used very often afterwards. Many further concepts are based on the basic formula $\mathbf{G} = t^{-1}\mathbf{Z}\mathbf{Z}'$ in which the elements in $\mathbf{Z}$ and the scaling factor t differ between approaches. $\mathbf{Z}$ is always a matrix of marker genotypes of all genotyped individuals with individuals in rows and markers in columns. The elements of $\mathbf{Z}$ can directly represent the allele counts (e.g. Habier *et al.,* 2007), namely 0, 1 and 2 for AA, AB and BB, or allele counts that are centered in a way that the heterozygotes are represented by 0, i.e. -1, 0 and 1 for AA, AB, BB. VanRaden (2007) stated that correcting the marker genotypes by the expected mean would lead to unbiased predictions since then the expected value of $\mathbf{u}$ is 0. This is why he proposed to calculate $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{M} - \mathbf{P} = \mathbf{M} - \begin{pmatrix} 2q_1 & 2q_2 & \cdots & 2q_m \\ \vdots & \vdots & \vdots & \vdots \\ 2q_1 & 2q_2 & \cdots & 2q_m \end{pmatrix} \qquad [3]$$

with $\mathbf{M}$ being a matrix of genotypes coded 0, 1, 2 and $\mathbf{P}$ being a matrix where each column vector $\mathbf{p}_i$ contains two times the allele frequency of the i^{th} SNP (q_i). For estimates of the genomic relationship coefficients and for further calculations of genomic breeding values, it does not matter to which of the original alleles the frequency belongs, but it has to be the allele frequency of the allele where the homozygous case on a locus is coded with 2.

In the beginning of genomic breeding value estimation, a common approach for determining t was to use $t = m,$ but this does not take the fact into consideration that marker genotypes at different markers may have different variances. Habier *et al.* (2007) and VanRaden (2007) proposed to build t as

$$t = 2\sum_{i=1}^{m}\left(q_i(1-q_i)\right) \qquad [4]$$

where q_i is the allele frequency at marker locus i. This kind of standardization is based on the fact that $var(\mathbf{z}_i) = 2q_i(1-q_i)$ and makes the pedigree-based relationship and genomic relationship comparable on the same scale (VanRaden, 2008). VanRaden (2008) also argues that minor alleles will get more weight in the genomic breeding values using this centralization process in [3] and the standardization with [4], but this argument does not hold, since it has been shown that the estimated effects do not differ irrespective of the marker coding (Strandén & Christensen, 2011) when using this kind of genomic relationship matrix.

There are other approaches that do not fit to the form $\mathbf{G} = t^{-1}\mathbf{Z}\mathbf{Z}'$, but which standardize each marker separately and then add all marker information together (e.g. VanRaden, 2008; Astle & Balding, 2009; Yang *et al.*, 2011), so that

$$\mathbf{G} = \frac{1}{m}\sum_{i=1}^{m}\frac{(\mathbf{m}_i - \mathbf{p}_i)(\mathbf{m}_i - \mathbf{p}_i)'}{2\big(q_i(1-q_i)\big)} \qquad [5]$$

where $\mathbf{m}_i$ is the i^{th} column in a matrix of genotypes coded 0, 1, 2. Vector $\mathbf{p}_i$ is defined as in [3] and contains two times the allele frequency q_i. In praxis, the differences in estimates of genomic breeding values obtained with a genomic relationship matrix based on [3] or based on [5] are often negligible, however if many low frequency alleles are in the sample, [5] may consider them better.

Goddard *et al.* (2011) noted that especially in the data sets where the marker density is not high, the estimates of the realized values may have high sampling errors and $\mathbf{G}$ may be biased. Goddard *et al.* (2011) therefore suggested using

$$\mathbf{G} = \mathbf{A} + r\left(\frac{1}{m}\sum_{i=1}^{m}\frac{(\mathbf{m}_i - \mathbf{p}_i)(\mathbf{m}_i - \mathbf{p}_i)'}{2\big(q_i(1-q_i)\big)} - \mathbf{A}\right) \qquad [6]$$

where $\mathbf{A}$ is the pedigree based relationship matrix and r is a regression coefficient that can be estimated based on the variance of the non-diagonal values of $\mathbf{G}$ from [5].

The genomic relationship matrix presented in VanRaden (2007) which uses [3] and [4] will be used in all analyses in ***Chapter 2***, ***3*** and ***5***, and the genomic relationship matrix described in [6] will be used in ***Chapter 4*** of this thesis.

Bayesian methods

The second group of approaches proposed for the prediction of genomic breeding values is the wide field of Bayesian methods. This paragraph will just give a short overview over a few well-known approaches, while there are many others around. *BayesA* and *BayesB* have

been proposed in the initial paper on genomic breeding value prediction (Meuwissen *et al.*, 2001) while *BayesCπ* (Habier *et al.*, 2011) followed later.

The general model for *BayesA* and *BayesB* is

$$\mathbf{y} = \mathbf{1}'_{n_o}\mu + \sum_{i=1}^{m} \mathbf{z}_i u_i + \mathbf{e}$$

while for *BayesCπ* it is

$$\mathbf{y} = \mathbf{1}'_{n_o}\mu + \sum_{i=1}^{m} \mathbf{z}_i u_i \delta_i + \mathbf{e}$$

where $\mathbf{y}$ is a vector of observations, μ is an overall mean (which is treated as random in the Bayesian framework), $\mathbf{z}_i$ is a vector of genotypes for all individuals at the i^{th} marker locus, u_i is the random SNP effect of the i^{th} marker and $\mathbf{e}$ is a vector of random residual effects. For *BayesCπ*, δ_i can be 1 (with probability $1-\pi$) or 0 (with probability π) and indicates whether the i^{th} marker is present in the model or not. Since the estimation of the parameters depends on other parameters in the model, all these Bayesian models have to be processed using MCMC algorithms over tens of thousands of iterations with a long burn-in phase. Genomic breeding values can be predicted afterwards as $\hat{\mathbf{g}} = \sum_{i=1}^{m} \mathbf{z}_i \hat{u}_i$ where $\hat{u}_i$ is the estimated SNP effect at the i^{th} SNP averaged over the values obtained for all post-burn-in cycles. The three approaches mentioned above mainly differ (amongst others) in two points, namely in the modeling of the variance of the random SNP effects as well as in the values used for parameter π and their way to determine these values. With *BayesA*, the posterior of the variance of each SNP effect is modeled by a SNP specific inverse χ^2-distribution and π is set to 0 for all markers, i.e. all markers contribute to modeling the genomic breeding values. For *BayesB* the posterior variance of the SNP effect is also SNP specific, but, in contrast to *BayesA,* π is set to a fixed value > 0 that is defined arbitrarily. The parameter π is used for *BayesB* in the following way:

$\sigma^2_{u_i} \sim \chi^2(\nu, S)$ with probability $1-\pi$

$\sigma^2_{u_i} = 0$ with probability π

Values for π that are often used are in the range of 0.9 to 0.99 (e.g. Meuwissen *et al.*, 2001; Habier *et al.*, 2010) which means that most of the SNPs have no effect and only a few contribute to variation in the genomic breeding values. *BayesA* is a special case of *BayesB* with $\pi = 0$ (Gianola *et al.*, 2009).

With *BayesCπ* the variance of the distribution for the SNP effects is also drawn from a scaled inverse χ^2-distribution, but is assumed to be the same for all markers. The parameter π is not set to a specific value, but is modeled with a uniform prior distribution $\pi \sim uniform(0,1)$. If one fixes π to be 0, then this special case of *BayesCπ* will be very similar to *GBLUP* (Habier *et al.*, 2010).

Both *BayesB* and *BayesCπ* have the advantage that they include a parameter $\pi \neq 0$ that allows a situation in which not all markers contribute to the model. The weakness of *BayesB* is the long computing time and the fixation of π, while for *BayesCπ* the assumption that all markers having an effect come from the same distribution may not be realistic – at least for traits where there are a few larger and many small effects. A further Bayesian method takes the advantages of the previous methods and avoids the disadvantages: In *BayesR* (Erbe *et al.*, 2012), SNP effects are assumed to be 0 or to come from different normal distributions that differ in their variance with specific probabilities:

$$u_i \begin{cases} = 0 & with\ probability\ p_1 \\ \sim N(0, \sigma_2^2) & with\ probability\ p_2 \\ \sim N(0, \sigma_3^2) & with\ probability\ p_3 \\ \sim N(0, \sigma_4^2) & with\ probability\ p_4 \end{cases}$$

The entries of vector $\mathbf{p} = (p_1,\ p_2,\ p_3,\ p_4)$ are not fixed but are modeled with a Dirichlet distribution and σ_2^2, σ_3^2 and σ_4^2 are defined as specific proportions of the total genetic variance. This method will be presented in more detail in ***Chapter 4*** of this thesis.

Accuracy of prediction and Cross-validation

Different measures can be used for validating results of genomic breeding value prediction methods. The most common parameters for model assessment in terms of prediction are the correlation between the true and predicted genomic breeding value as proxy for the accuracy of prediction and the slope of the regression of true on predicted breeding values to control the bias. Accuracy of prediction in genomic BLUP models can also be obtained from theoretical considerations in the mixed model framework, but in this thesis accuracy of prediction will always be assessed as the observed correlation from cross-validation studies.

Cross-validation is a technique of model validation that has its origin in the field of psychology. In the early 1930s it was common to use multiple regression approaches to explain behavior of persons or other events. The common procedure was using all available data to search for the multiple regression equation that explained the depending variable best (e.g. expressed by the multiple correlation coefficient). This means that the equation was derived

and evaluated in the same data set which led to a decrease in the accuracy when applying this model to predict the dependent variable in an independent data set. Larson (1931) was one of the first authors who tried to develop a study design for describing the amount of decrease of accuracy when having a limited number of observations available. He split a data set of school boys into two comparable groups and used one group to find the best multiple regression equation that uses test scores of different subjects to predict the score in another subject. Then, he tried to predict values for the second group based on the model trained with the first group and correlated the predicted scores in the second group with the observed ones. This was the basic idea, for what later would be called "cross-validation", namely splitting the data set in groups – one for derivation (training) and one for prediction (validation) – and getting a realistic idea of the prediction ability of the model.

Kurtz (1948) gave the best example why validation in an independent data set is mandatory: The aim of the study was to predict success as life insurance sales managers based on the results of the Rorschach Test, which is a psychological test. A scoring system was developed in a group of 70 sales manager, but it was found to be *"completely useless"* (Kurtz, 1948) when applied to a further group. Mosier (1951) gave the first definition of a cross-validation procedure: *"In cross-validation we have weights based on one sample and we determine their effectiveness on a second sample where both samples are representative of the population to which the weights will be applied for prediction."* From then on different cross-validation strategies have been established and different ways of best splitting the data set have been developed. Without claiming to be complete, the following cross-validation strategies can be listed (assume N to be the total size of the data set; see e.g. Arlot, 2010; Burman, 1989):

- *Leaving-one-out:*

N replicates have to be run in which there is exactly one observation used for validation and $N - 1$ observations are used for training. This strategy is almost unbiased, but computationally very demanding.

- *Double cross-validation:*

This strategy implies splitting the data set in two groups of equal size. The first group is used for training and the second group for validation and then vice versa. Replicates could be realized by repeating the procedure with a different random splitting of the data. Note that fitting of the model is done only with half the data size.

- *Random drawing with/without replicates:*

A specific proportion of observations ($0 < p < 1$) is randomly chosen to be the training set while the remaining observations $\big((1-p)N\big)$ represent the validation set. There is a stratified alternative, namely generating the sets not randomly but based on different criteria. The size of sets can be chosen independently of the number of replicates.

- *k-fold replication:*

The whole data set is divided in k subsets so that there are N/k individuals in each subset. There are k replicates so that each subset acts as the validation set once. Accuracies of prediction are averaged over the k replicates. The number of replicates and the size of training and validation sets thus depend on the chosen factor k. This strategy guarantees that each observation is used for validation exactly one time. A stratified version (e.g. sorted by age) is possible.

Leave-one-out strategies are very popular in other scientific fields like geo-statistics, but in the context of genomic data in livestock the size of the data set is normally too large to run leave-one-out cross-validations. The usual strategies are thus k-fold or random drawing strategies. All these strategies have the aim to describe the prediction ability of a model. The evaluation is normally done with one of the two following parameters: One can measure the accuracy directly by considering the correlation between predicted and true observations in the respective validation set. The second criterion often studied is the error of prediction, e.g. by measuring the mean squared error. If the cross-validation design implies replicates, values can be averaged over folds and/or replicates.

In animal breeding, cross-validation has become very popular with the appearance of genomic breeding value estimation. Normally, phenotypes or conventional breeding values are not available for the individuals for which genomic breeding values should be predicted. However, it is necessary to assess properties of models and to predict the potential accuracy of genomic prediction for those individuals. Thus, cross-validation within the set of genotyped and phenotyped individuals has become a frequently applied tool and different cross-validation strategies have been used in studies with real data sets (e.g. Lee *et al.*, 2008; Luan *et al.*, 2009; Habier *et al.*, 2010). Cross-validation strategies in different forms will be used in the following chapters: Random drawing with replicates in ***Chapter 2***, stratified validation without replicates in ***Chapter 4*** and k-fold replication in ***Chapter 5***.

Objectives of this thesis

The first publication describing genomic breeding value prediction of Meuwissen *et al.* in 2001 has presented first ideas of this new methodology. In the following years, different studies on testing this new approach in simulated and/or first real data sets and different papers on theoretical aspects of the methodology have been published. Apart from others some important factors have emerged that seems to be crucial for the obtained level of prediction accuracy: Meuwissen *et al.* (2001) themselves showed that there are differences in accuracy of prediction caused by the choice of the prediction model. Habier *et al.* (2007) showed that the prediction accuracy can differ between individuals that are related in different degrees to the training set. De Roos *et al.* (2008) demonstrated that a much larger marker density (~300K) than available at that time will be necessary to obtain high prediction accuracies across breeds. Dekkers (2007) described that there will be a maximal achievable accuracy unequal 1 with a specific marker set depending on how much genetic variance can be explained by the given markers. Many more examples could be given. All of these studies make clear that there is a necessity to take a closer look on how accuracy of prediction is determined by various criteria.

The aim of this study was thus to investigate different validation strategies and several factors that may influence the accuracy of genomic prediction in any way:

Chapter 2 shows how different cross-validation strategies influence the correlation between genomic and true breeding value based on a series of cross-validation runs in real dairy cattle data with random assignment of individuals to folds.

Chapter 3 deals with the influence of relationship and age structure between training set and validation set within a large data set of German Holstein Friesian bulls. A validation set of the 500 youngest bulls is predicted with various training sets differing in age and relationship structure to the validation set.

Chapter 4 studies the influence of the underlying marker density and investigates possibilities to process data from different breeds in a combined breeding value estimation. Data sets from Australian Holstein and Australian Jersey genotyped with 50K SNPs and imputed to 777K SNPs are used in purebred and multi-breed validation schemes. Furthermore, a new Bayesian method (*BayesR*) is presented and the influence of the model choice is also studied.

Chapter 5 presents a method to improve deterministic equations that try to predict the expected level of accuracy based on population parameters. Holstein Friesian and Brown Swiss data sets build the basis for cross-validation runs which themselves are the empirical

basis to estimate the number of independently segregating chromosome segments as well as the maximal achievable accuracy with a given marker set. Both estimates are then used to find an optimal deterministic equation.

Chapter 6 includes a general discussion on factors affecting the accuracy of genomic prediction.

REFERENCES

Arlot, S. (2010): A survey of cross-validation procedures for model selection. *Stat. Surv.* **4:**40-79.

Astle, W., and Balding, D. J. (2009): Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat. Sci.* **24:**451-471.

Browning, S. R., and Browning, B. L. (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.

Brøndum, R. F., Ma, P., Lund, M. S., and Su, G. (2012): *Short communication:* Genotype imputation within and across Nordic cattle breeds. *J. Dairy Sci.* **95:**6795-6800.

Burman, P. (1989): A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76:**503-514.

David, X., de Vries, A., Feddersen, E., and Borchersen, S. (2010): International Genomic Cooperation. EuroGenomics significantly improves reliability of Genomic evaluations. *Interbull Bull.* **41:**77-78.

Dekkers, J. C. M. (2007): Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* **124:**331-341.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012): A linear complexity phasing method for thousands of genomes. (2012): *Nat. Meth.* **9:**179–181.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008): Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. *Genetics* **179:**1503-1512.

de Roos, A. P. W., Hayes, B. J., and Goddard, M. E. (2009): Reliability of Genomic Predictions Across Multiple Populations. *Genetics* **183:**1545-1553.

de Roos, A. P. W., Schrooten, C., Veerkamp, R. F., van Arendonk, J. A. M. (2011): Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. *J. Dairy Sci.* **94:**1559–1567.

Eck, S. H., Benet-Pagès, A., Flisikowski, K., Meitinger, T., Fries, R., and Strom, T. M. (2009): Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery. *Gen. Biol.* **10:**R82.

Eding, H., and Meuwissen, T. H. E. (2001): Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *J. Anim. Breed. Genet.* **118:**141-159.

Erbe, M., Hayes, B. J., Matukumalli, L. K., Goswami, S., Bowman, P. J., Reich, C. M., Mason, B. A., and Goddard, M. E. (2012): Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density SNP panels. *J. Dairy Sci.* **95:**4114-4129.

Garrick, D. J., Taylor, J. F., and Fernando, R. L. (2009): Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **41:**55.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009): Additive Genetic Variability and the Bayesian Alphabet. *Genetics* **183:**347-363.

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

Goddard, M. E., and Hayes, B. J. (2009): Genomic selection based on dense genotypes inferred from sparse genotypes. *Proc. Assoc. Adv. Anim. Breed. Genet.* **18:**26-29.

Goddard, M. E., Hayes, B. J., and Meuwissen, T. H. E. (2011): Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* **128:**409-421.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., Mni, M., Simon, P., Frère, J.-M., Coppieters, W., and Georges, M. (2004): Genetic and functional confirmation of the causality of the *DGAT1 K232A* quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA* **101:**2398-2403.

Guillaume, F., Fritz, S., Boichard, D., and Druet, T. (2008): *Short Communication*: Correlations of Marker-Assisted Breeding Values with Progeny-Test Breeding Values for Eight Hundred Ninety-Nine French Holstein Bulls. *J. Dairy Sci.* **91**:2520-2522.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007): The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* **177:**2389-2397.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010): The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* **42:**5.

Habier, D., Fernando, R. L., Kizilkaya, K, and Garrick, D. J. (2011): Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* **12:**186.

Hayes, B. J., and Goddard, M. E. (2001): The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33:**209-229.

Hayes, B. J., and Goddard, M. E. (2008): Technical note: Prediction of breeding values using marker-derived relationship matrices. *J. Anim. Sci.* **86:**2089-2092.

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009): Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. Camb.* **91:**47-60.

Hayes, B., Anderson, C., Daetwyler, H., Fries, R., Guldbrandtsen, B., Lund, M., Boichard, D., Stothard, P., Veerkamp, R., Hulsegge, I., Rocha, D., Van Tassell, C., Coote, D., Goddard, M., and The 1000 Bull Genomes Consortium (2012): Towards genomic prediction from genome sequence data and the 1000 bull genomes project. *Book of Abstracts of ICQG 2012* (Edinburgh), p. 55.

Henderson, C. R. (1975): Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31:**423-447.

Henderson, C. R., and Quaas, R. L. (1976): Multiple Trait Evaluation Using Relatives' Records. *J. Anim. Sci.* **43:**1188-1197.

Hickey, J.M., Cleveland, M., Gorjanc, G., Tier, B., van der Werf, J.H.J., and Kinghorn, B. (2011): An Imputation Strategy which Results in an Alternative Parameterization of the Single Step Genomic Evaluation. *Interbull Bulletin* **44:**38-41.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012): Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44:**955-959.

Kurtz, A. K. (1948): A Research Test of the Rorschach Test. *Pers. Psychol.* **1:**41-51.

Larson, S. C. (1931): The shrinkage of the coefficient of multiple correlation. *J. Educ. Psychol.* **22:**45-55.

Lee, S. H., van der Werf, J. H. J., Hayes, B. J., Goddard, M. E., and Visscher, P. M. (2008): Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. *PLoS Genet.* **4:**e1000231.

Li Y., Willer C. J., Ding J., Scheet P., and Abecasis, G. R. (2010): MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genet. Epidemiol.* **34:**816-834.

Lillehammer, M., Meuwissen, T. H. E., and Sonesson, A. K. (2011): A comparison of dairy cattle breeding designs that use genomic selection. *J. Dairy Sci.* **94:**493–500.

Liu, Y., Qin, X., Song, X.-Z. H., Jiang, H., Shen, Y., Durbin, K. J., Lien, S., Kent, M. P., Sodeland, M., Ren, Y., Zhang, L., Sodergren, E., Havlak, P., Worley, K. C., Weinstock, G. M., and Gibbs, R. A. (2009): *Bos taurus* genome assembly. *BMC Genomics* **10:**180.

Liu, Z., Seefried, F. R., Reinhardt, F., Rensing, S., Thaller, G., and Reents, R. (2011): Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Gen. Sel. Evol.* **43:**19.

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. E. (2009): The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* **183:**1119-1126.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001): Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **157:**1819-1829.

Meuwissen, T. H. E., and Goddard, M. E. (2004): Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol.* **36:**261-279.

Mosier, C. I. (1951): Symposium: The Need and Means of Cross-Validation. I. Problems and designs of cross-validation. *Educ. Psychol. Meas.* **11:**5-11.

Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K.-U., and Fries, R. (2013): Imputation of high-density genotypes in the Fleckvieh cattle population. *Genet. Sel. Evol.* **45:**3.

Pryce, J. E., and Daetwyler, H. D. (2012): Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim. Prod. Sci.* **52:**107-114.

Reed, D. R., Lawler, M. P., and Tordoff, M. G. (2008): Reduced body weight is a common effect of gene knockout in mice. *BMC Genet.* **9:**4.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2011): FImpute – An efficient imputation algorithm for dairy cattle populations. *J. Dairy Sci.* **94**(E-Suppl. 1):421.

Schaeffer, L. R. (1994): Multiple-Country Comparison of Dairy Sires. *J. Dairy Sci.* **77:**2671-2678.

Schaeffer, L. R. (2004): Application of random regression models in animal breeding. *Livest. Prod. Sci.* **86:** 35-45.

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

Scheet, P., and Stephens, M. (2006): A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *Am. J. Hum. Genet.* **78:**629-644.

Sillanpää, M. J., and Corander, J. (2002): Model choice in gene mapping: what and why. *Trends Genet.* **18:**301-307.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008): Genomic selection using different marker types and densities. *J. Anim. Sci.* **86:**2447-2454.

Strandén, I, and Garrick, D. J. (2009): *Technical note:* Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* **92:**2971-2975.

Strandén, I., and Christensen, O. F. (2011): Allele coding in genomic evaluation. *Genet. Sel. Evol.* **43:**25.

Utz, H. F., Melchinger, A. E., and Schön, C. C. (2000): Bias and Sampling Error of the Estimated Proportion of Genotypic Variance Explained by Quantitative Trait Loci Determined From Experimental Data in Maize Using Cross Validation and Validation With Independent Samples. *Genetics* **154:**1839-1849.

VanRaden, P. M. (2007): Genomic Measures of Relationship and Inbreeding. *Interbull Bull.* **37:**33-36.

VanRaden, P. M. (2008): Efficient Methods to Compute Genomic Predictions. *J. Dairy. Sci.* **91:**4414-4423.

VanRaden, P. M., and Wiggans, G. R. (1991): Derivation, Calculation, and Use of National Animal Model Information. *J. Dairy Sci.* **74:** 2737-2746.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009): *Invited review:* Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92:**16-24.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., and Weigel, K. A. (2011): Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* **43:**10.

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., Sonstegard, T. S., Connor, E. E., Winters, M., van Kaam, J. B. C. H. M., Valentini, A., Van Doormaal, B. J., Faust, M. A., and Doak, G. A. (2013): Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* **96:**668-678.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011): GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88:**76-82.

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C. P., Sonstegard, T. S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J. A., and Salzberg, S. L. (2009): A whole-genome assembly of the domestic cow, *Bos taurus. Genome Biol.* **10:**R42.

Zumbach, B., Jorjani, H., and Dürr, J. (2010): Brown Swiss Genomic Evaluation. *Interbull Bull.* **42:**44-51.

2nd CHAPTER

Assessment of Cross-validation Strategies for Genomic Prediction in Cattle

M. Erbe, E. C. G. Pimentel, A. R. Sharifi and H. Simianer

Georg-August-University, Animal Breeding and Genetics Group, 37075 Goettingen, Germany

Published in *Proceedings of the 9th WCGALP* (ISBN 978-3-00-031608-1), No. 0553

http://www.kongressband.de/wcgalp2010/assets/pdf/0553.pdf

INTRODUCTION

The basic idea of cross-validation procedures is to divide a data set into a reference and a validation set, to omit any kind of information of the validation set and to predict this information, e.g. phenotypes, with a model trained exclusively in the reference set. The accuracy of prediction can be used to evaluate the underlying model and to compare alternative models. In the field of genomic selection, cross-validation can be used for assessing the accuracy of genomic breeding values (GEBVs) predicted with a specific model (e.g. Blonk *et al.*, 2010; Luan *et al.*, 2009) and for comparing the quality of different approaches used for estimation of GEBVs (Lund *et al.*, 2009; Goddard and Hayes, 2007). However, the way of subdividing the data set is known to influence the results obtained by cross-validation (Luan *et al.*, 2009; Lee *et al.*, 2008). Therefore, we studied the changes in results when using different numbers of animals for the reference and the validation set and tried to find optimal subdivision strategies for the different objectives of cross-validation.

MATERIAL AND METHODS

Data

We used a sample of 2,294 Holstein bulls, which were genotyped with the Illumina 50K SNP chip. SNPs with a minor allele frequency lower than 5%, with missing position or a call rate lower than 95% were excluded. After filtering, there were 39,557 SNPs remaining for further analyses. Missing genotypes at these SNP positions were imputed using fastPHASE (Scheet and Stephens, 2006). All bulls had pedigree information and breeding values for somatic cell score with an accuracy > 0.87, which were used as quasi-phenotypes for the following analyses.

Methods to predict the GEBVs

We used two best linear unbiased prediction (BLUP) models for the estimation and prediction of the GEBVs. The first model included a random genomic and a random polygenic effect (Model A), while the second one included a random genomic component only (Model B). For model A, we fitted

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{Wg} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of the phenotypes (breeding values for somatic cell score) for all bulls in the reference set, μ is the overall mean, $\mathbf{Z}$ is the incidence matrix for the random polygenic effect, $\mathbf{u}$ is a vector containing a random polygenic effect for each individual, $\mathbf{W}$ is the inci-

dence matrix for the random genomic effect, $\mathbf{g}$ is a vector containing a random genomic effect for each animal and $\mathbf{e}$ is a vector of random residual terms. $\mathbf{u}$ is assumed to follow $N(0, \mathbf{A}\sigma_u^2)$ where $\mathbf{A}$ is the pedigree based relationship matrix. $\mathbf{g}$ is distributed $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$ where $\mathbf{G}$ is a marker based relationship matrix, which was built according to VanRaden (2008) based on all available SNPs ($n = 39{,}557$). In model A, the total breeding value was the sum of the polygenic and the genomic breeding value.

In model B, the polygenic component was omitted, hence the model was

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

Here, information of only one eighth of all available SNPs ($n = 4{,}945$) was used to build $\mathbf{G}$. The prediction accuracy of model A is expected to exceed the one of model B. As can be seen in the model design, we did not estimate an effect for each single SNP, but used the genomic relationship matrix to model a genomic effect for each individual. It was thus possible to estimate variance components in each step in each replicate by using ASReml (Gilmour *et al.*, 2009). With the corresponding variance components, effects were estimated and GEBVs for the bulls in the validation set were predicted.

Cross-validation procedure

The whole data set was divided into a reference and a validation set. Phenotypes of the animals in the validation set were assumed to be unknown. First, a random sample of 100 animals was drawn for the validation set while the remaining 2,194 bulls built the reference set. In the next step, the size of the validation set was increased by 100 by moving 100 randomly chosen individuals from the reference to the validation set. This was done stepwise until 1,500 bulls were in the validation set. For each step, GEBVs were predicted for the bulls in the validation set with the information from the animals in the reference set. The whole procedure was repeated 60 times.

Criteria for comparison

Pearson's correlation coefficient between the realized and the predicted phenotypes for the animals in the validation set was calculated for each step in each replicate. In case the model did not converge during the process of variance component estimation, the correlation for this step in the particular replicate was considered to be NA.

The correlation between realized and predicted phenotypes was also used for testing whether the models differed significantly from each other. Therefore, the correlation coefficients were transformed so that they follow approximately a normal distribution and the difference

between the correlation coefficient was tested against being zero (Sachs and Hedderich, 2009). The test of significantly different correlations was applied in each step in each replicate and the obtained p-values were averaged over the replicates.

RESULTS AND DISCUSSION

Figure 1 shows a Box-Whisker-Plot of the correlations between realized and predicted phenotypes for the animals in the validation set. As expected, Model A was found to be more accurate than Model B. With both methods, the highest correlations could be found when the number of animals in the validation set (n_y) was small. With $n_y = 100$, the median of the correlations obtained was 0.689 and 0.627 with Model A and B, respectively. The median was almost constant with n_y ranging between 100 and 600 and then decreased continuously with both methods. The median of the correlation was 0.577 and 0.536 for $n_y = 1500$ with Models A and B, respectively. Due to the design of cross-validation the smaller the validation set, the larger is the reference set. A larger reference set will lead to a more accurate estimation of the variance components and thus to a better estimation of the effects. Therefore, also a better prediction of the phenotypes for the animals in the validation set is possible and higher correlations are expected with small validation sets.

However, variation over the replicates was also highest in the case of very small validation sets. For example, with $n_y = 100$, the results varied between 0.545 and 0.794 with Model A and ranged from 0.509 to 0.786 with Model B. Lee *et al.* (2008) described similar tendencies concerning the accuracy and the variation of prediction of phenotypes.

Since a higher number of values can be used for calculating the correlation coefficient when the number of animals in the validation set is higher, the correlation coefficient is estimated better with larger values of n_y. Therefore, even if the absolute distance between the models regarding the correlation coefficients seems to be similar with all sizes of the validation set, a significance differentiation between Models A and B was only possible with n_y ranging between 700 and 1,300 (Figure 2).

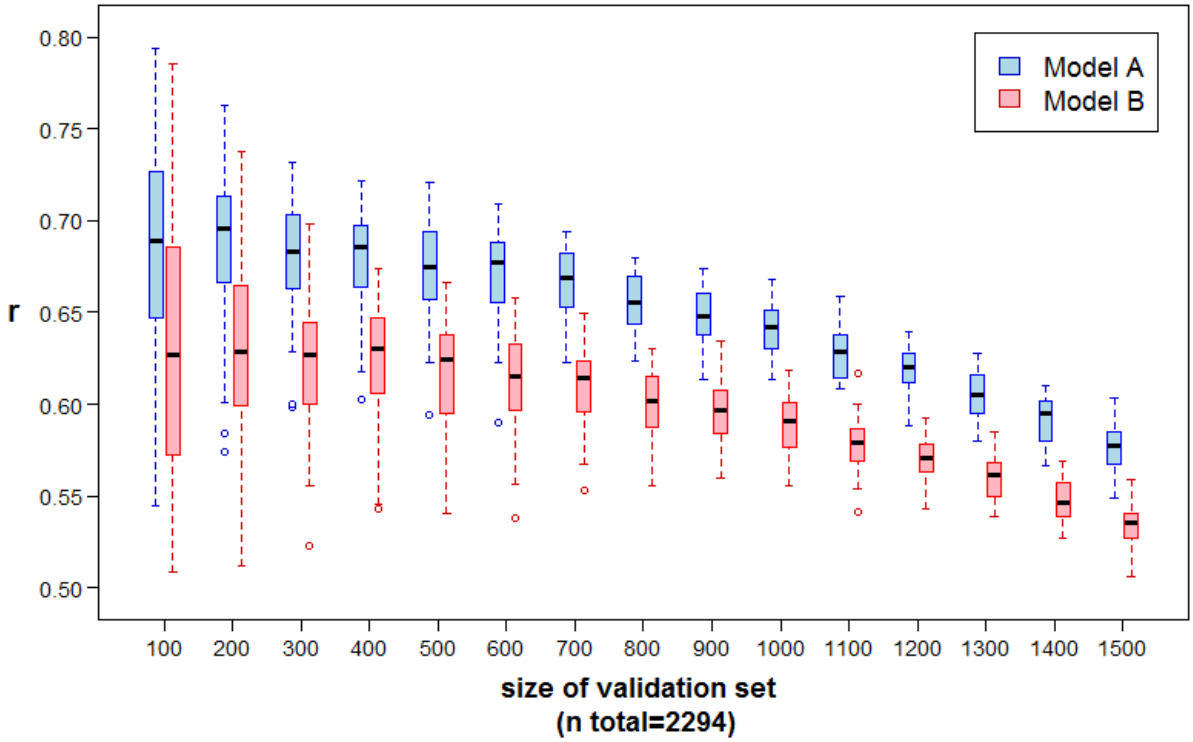


Figure 1: Box-Whisker-Plot of the correlations between realized and predicted phenotypes for the animals in the validation set.

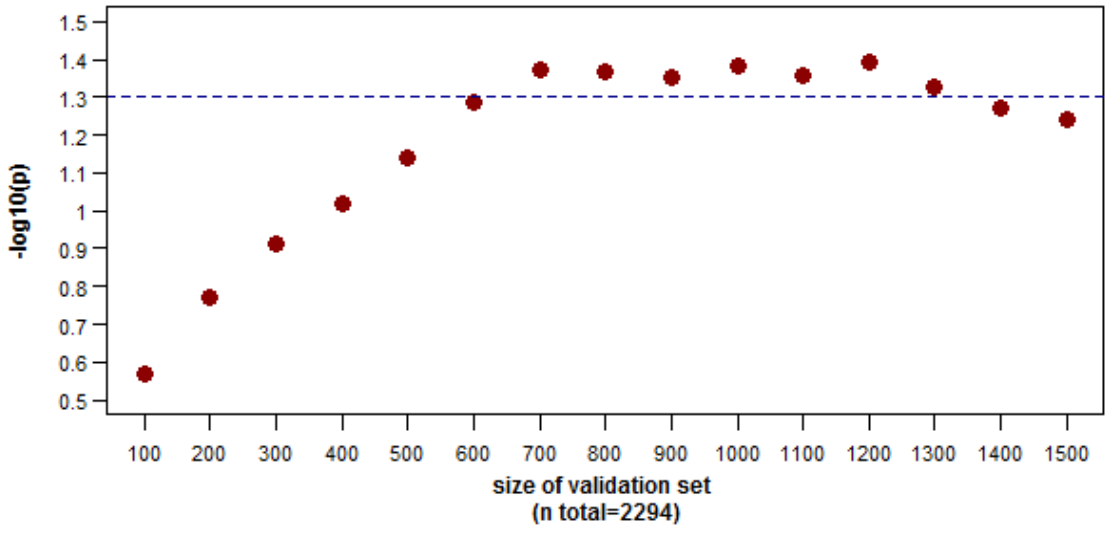


Figure 2*:* -log(p)-values (averaged over the replicates) of the test for a difference in correlation coefficients of Models A and B. The dashed line symbolizes the significance threshold on a 5% error level.

CONCLUSION

The optimal subdivision of a data set depends on the objective of cross-validation. The highest correlations can be obtained with the size of the reference set being large and therefore the validation set being small. A small validation set, however, also leads to a high variation in the obtained correlations and results depend strongly on the sample chosen for the validation set. A five-fold subdivision, using 20 per cent of the data as validation set, seems to be a good compromise. Larger validation sets provide more accurate estimation of correlation coefficients. Hence, if the aim is to differentiate significantly between two models, larger validation sets are recommended.

ACKNOWLEDGMENTS

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” and the project FUGATO-plus GenoTrack. We thank the Vereinigte Informationssyteme Tierhaltung v.W. (VIT), Verden, for providing data.

REFERENCES

Blonk, R.J.W., Komen, H., Kamstra, A. *et al.* (2010). *Genetics*, 184: 213-219.

Gilmour, A.R., Gogel, B.J., Cullis, B.R. *et al.* (2009). ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Goddard, M.E. and Hayes, B.J. (2007). *J. Anim. Breed. Genet.*, 124: 323-330.

Lee, S.H., van der Werf, J.H.J., Hayes, B.J. *et al.* (2008). *PLoS Genetics*, 4(10): e1000231.

Luan, T., Woolliams, J.A., Lien, S. *et al.* (2009). *Genetics*, 183: 1119-1126.

Lund, M.S., Sahana, G., de Koning, D.-J. *et al.* (2009). *BMC Proceedings*, 3(Suppl 1): S1.

Sachs, L. and Hedderich, J. (2009) *Angewandte Statistik*. Springer, Dodrecht, Heidelberg, London, New York.

Scheet, P. and Stephens, M. (2006). *Am. J. Hum. Genet.*, 78: 629-644.

VanRaden, P.M. (2008). *J. Dairy Sci.*, 91: 4414-4423.

.

3rd CHAPTER

Effect of Relationship and Age Structure Between Training and Validation Set on the Accuracy of Genomic Breeding Value Prediction Using Genomic BLUP

M. Erbe[1], F. Seefried[2] and H. Simianer[1]

[1] Georg-August-University, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany

[2] Vereinigte Informationssysteme Tierhaltung w.V. (VIT), Heideweg 1, 27283 Verden/Aller, Germany

Published in *Proc. Assoc. Advmt. Anim. Breed. Genet.* **19:** 287-290.

SUMMARY

A dataset of 5698 Holstein Friesian bulls born between 1981 and 2005 was used to study the influence of different relationship levels between a training set and the set of candidates for whom genomic breeding values (GBV) are to be predicted. Traits studied were milk yield and somatic cell score. Different scenarios were modeled while always the GBV of the 500 youngest bulls of the available data set were predicted. The correlation between true breeding value and GBV was used as evaluation criterion. The prediction of the youngest bulls was best when other bulls of the same age or only slightly older or bulls which were especially highly related to the candidates were used to train the model while there was a decrease of accuracy, especially for GBV in somatic cell score, when the oldest bulls formed the training set. Reducing the maximum relationship between all candidates to the training set to less than 0.5 led to a decrease in accuracy. The decrease was even stronger when the maximum relationship was limited to less than 0.25. It seems that accuracy of prediction of GBV depends clearly on the relationship and age structure between the validation and the training set which is in accordance with some previous studies. Therefore, it is implicitly necessary to continuously fill the training sets used for predicting young bulls with new progeny tested bulls to avoid the reduction of maximum relationship.

INTRODUCTION

In the last years, prediction of genomic breeding values has become a popular tool for predicting reliable breeding values of not yet progeny tested bulls of young age, especially in dairy cattle populations. Different studies (e.g. Lund *et al.* 2009; Habier *et al.* 2010) have shown that accuracy of prediction is clearly influenced by the relationship between bulls in the training and in the validation set. Since the methodology of genomic selection is new, there are still enough progeny tested bulls available which are strongly related to the candidates and can be used to train the models. However, in a few years, if genomic selection will be consequently applied, there may be a lack of such animals. It is thus necessary to further investigate how the relationship and age structure influences the accuracy of genomic breeding values of young bulls.

MATERIALS AND METHODS

Data

We used a sample of 5698 Holstein bulls, which were genotyped with the Illumina 50K Single Nucleotide Polymorphism (SNPs) chip. SNPs with a minor allele frequency lower than 1%, with missing position or a call rate lower than 95% were excluded. After filtering, there were 42,551 SNPs remaining for further analyses. Missing genotypes at these SNP positions were imputed using Beagle 3.2 (Browning and Browning 2007).

The bulls were born between 1981 and 2005. The average of the mean pedigree-based relationship between a random bull and all others was 0.093 while the mean of the maximum relationship was 0.459. 1832 bulls had a genotyped father and 1974 had one or both grandsires genotyped. There were 77.2% of bulls having at least 10 half or full sibs. The average inbreeding coefficient was 0.045. All bulls had pedigree information and breeding values for somatic cell score and milk yield. Average accuracy of the breeding values of the validation bulls was 0.89 and 0.96 for somatic cell score and milk yield, respectively. For bulls in the training sets, it was between 0.92 and 0.96 for somatic cell score and between 0.97 and 0.98 for milk yield in the different scenarios.

Method to predict GBV

Genomic breeding values were predicted using best linear unbiased prediction (BLUP) based on the model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of quasi-phenotypes (breeding values of milk yield or somatic cell score, respectively) for all bulls in the training set, $\mathbf{1}$ is a column vector of ones, μ is the overall mean, $\mathbf{Z}$ is the incidence matrix for the random genomic effect, $\mathbf{u}$ is a vector containing the random genomic effect (i.e. the genomic breeding value) for each animal and $\mathbf{e}$ is a vector of random error terms. $\mathbf{u}$ is assumed to be distributed $\mathbf{u} \sim N(0, \mathbf{G}\sigma_g^2)$ and $\mathbf{e}$ is assumed to follow $N(0, \mathbf{I}\sigma_e^2)$. $\mathbf{G}$ is a genomic relationship matrix which was built based on all SNPs available after quality control following VanRaden (2008). Variance components were estimated once with the complete data set using ASReml 3.0 (Gilmour *et al.* 2009) and were then used for all runs.

Validation strategy

The dataset was used for studying the influence of relationship and age structure on prediction of genomic breeding values (GBV). For this, we ran different scenarios with a constant

set of candidates (validation set) whose GBV were predicted using different training sets to train the model. Since the usual application of genomic prediction in cattle is the prediction of genomic breeding values for young bulls without phenotypes and not yet progeny tested, we used the 500 youngest bulls in our data set (all born in 2005) as the validation set for all scenarios. For each scenario, 2000 bulls fulfilling scenario specific criteria were chosen from the remaining data set. Prediction was then replicated 10 times in each scenario using always a random sample of 1500 out of the 2000 bulls at a time. As a standard for comparison to all other scenarios the training set comprised first of all completely randomly chosen bulls (***random***). For two further scenarios, the 2000 bulls were the oldest ones (***old***) and the youngest ones (***young***) of the remaining data set. To study the changes in accuracy of prediction when the relationship between training and validation set was reduced, we performed three scenarios where the training set contained only animals with a maximum pedigree-based relationship less than 0.25 ($rel_{max} < 0.25$) to all candidates. In the first of these three scenarios, we only controlled the maximum relationship ($<.\mathbf{25}$) while in both the others we also controlled the age structure ($<.\mathbf{25}\boldsymbol{y}$: youngest bulls with $rel_{max} < 0.25$, $<.\mathbf{25}\boldsymbol{o}$: oldest bulls with $rel_{max} < 0.25$). In one further scenario, a maximum relationship of 0.5 was allowed ($<.\mathbf{50}$). The last scenario (***maxrel***) tried to maximize the relationship between training and validation set by including all available near relatives (i.e. sire, grandsires, full and half sibs) of all candidates to the training set and filling the rest with bulls having a relationship of greater than 0.25 to as many candidates as possible.

Criterion for comparison

For the evaluation of the prediction, the correlation ($r_{GBV,TBV}$) between predicted GBV and true breeding value (TBV) was used. For obtaining $r_{GBV,TBV}$, first Pearson's correlation coefficient between the estimated breeding values (used as phenotypes) and the predicted GBV for the animals in the validation set was calculated in each scenario for each replicate. This correlation coefficient was then divided by the mean accuracy of the estimated breeding values of the animals in the respective validation set. To compare the relationship structure between different scenarios, the maximum and mean relationship of each of the 500 youngest bulls to all animals in the particular training set was calculated as well as the average number of animals in the training set to whom each of the candidates was related with a relationship coefficient greater or equal 0.25.

RESULTS AND DISCUSSION

Results for all scenarios and both traits regarding the mean accuracy of prediction and the key data of the relationship structures are given in Table 1.

Boxplots of the accuracy of prediction measured by $r_{GBV,TBV}$ for all scenarios are shown in Figure 1 for milk yield and somatic cell score. For both traits, the prediction was slightly better when random samples of young bulls were used to train the model in comparison to a random sample of bulls regardless of their age. These samples often contain large groups of half sibs of candidates so that the mean and maximum relationship was rather high in comparison to other scenarios. This may explain why prediction was better here.

Table 1: Accuracy of prediction and relationship measurements in different scenarios and both traits (milk yield and somatic cell score).

Scenario	$r_{GBV,TBV} \pm s.e.$ milk yield	$r_{GBV,TBV} \pm s.e.$ somatic cell score	Maximum relationship	Mean relationship	No of animals $rel_{max} \geq 0.25$
random	0.630±0.006	0.667±0.004	0.375	0.098	11
old	0.568±0.006	0.563±0.016	0.395	0.094	3
young	0.649±0.005	0.718±0.007	0.334	0.104	25
$<.\mathbf{50}$	0.543±0.006	0.626±0.006	0.318	0.100	9
$<.\mathbf{25}$	0.489±0.009	0.524±0.009	0.223	0.090	0
$<.\mathbf{25}\boldsymbol{o}$	0.534±0.005	0.454±0.011	0.221	0.090	0
$<.\mathbf{25}\boldsymbol{y}$	0.543±0.007	0.573±0.006	0.221	0.090	0
maxrel	0.685±0.005	0.731±0.003	0.430	0.109	28

Results for correlations between predicted genomic breeding values and true breeding values ($r_{GBV,TBV}$) were averaged over the ten replicates. Relationship criteria were measured between each candidate in the validation set and all animals in the respective training set and then averaged over all 500 candidates and the ten replicates. The last column shows the average number of animals in the training set a candidate is related to with a relationship coefficient greater or equal 0.25.

Including all animals in the training set which were directly related to the candidates (scenario ***maxrel***) led only to a slight increase in accuracy for both traits in comparison to the scenario ***young***. This was expected due to the fact that relationship between all young Holstein Friesian bulls is quite high on average. Therefore, candidates and bulls in the training sets were related to a large extent even if a random sample of young bulls regardless of the relationship structure was used for the training set.

An unambiguous trend of reduced prediction ability was observed when the relationship between training and validation set was limited to a specific maximum value as well as when the age difference between training and validation set became greater. For somatic cell score, the prediction was lowest when using the oldest available bulls with a maximum relationship of less than 0.25 to every candidate, while for milk it was lowest with a random sample with a maximum relationship restricted to less than 0.25 to every candidate.

We even could find a reduction of accuracy when there were only no more sires (and full sibs) of the candidates in the training set (scenario $<.\mathbf{50}$). Lund *et al.* (2009) presented similar tendencies when excluding sires from the training sets for three different traits in a sample of Nordic Holstein bulls. If the maximum relationship was limited to less than 0.25, the reduction in prediction ability was even worse, especially for somatic cell score.

This is in accordance with the work of Habier *et al.* (2010) who showed a continuous decrease of accuracy in different traits when reducing the permitted maximum relationship step by step in a limited sample of Holstein Friesian bulls. A limitation of $rel_{max} < 0.25$ means that no sires, grandsires, half and full sibs were used to train the model. From a practical point of view, this is a scenario which would become relevant after only two generations when the breeders fail to rebuild the training sets with enough new progeny tested bulls.

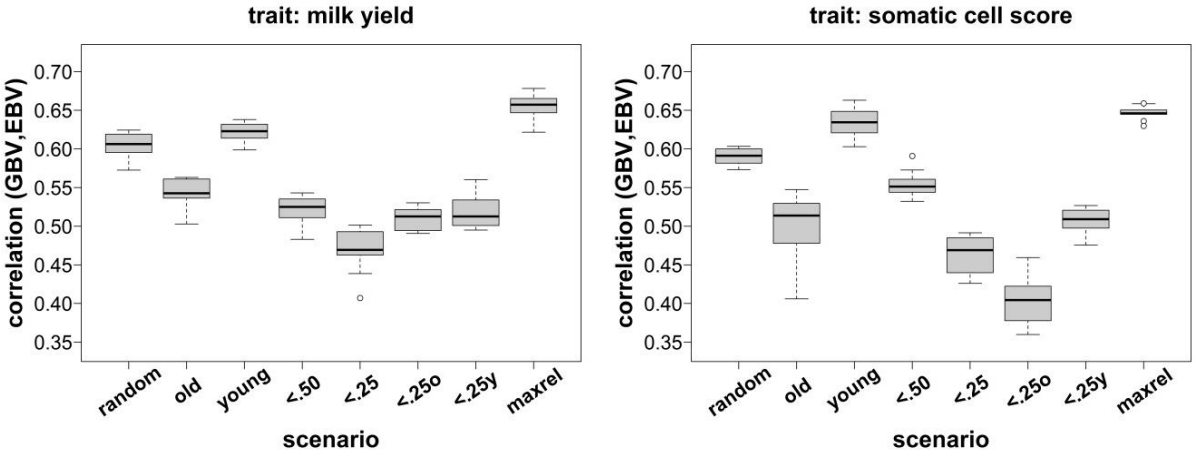


Figure 1: Boxplots of the accuracy of prediction for milk yield and somatic cell score for all scenarios.

CONCLUSIONS

Different training sets were used to train the model and to predict genomic breeding values for the 500 youngest bulls of the available data set. Different levels of relationship and age structure between training and validation set led to differences in accuracy of prediction. Reducing the relationship implicated an apparent decrease of accuracy of prediction. Therefore, in all kinds of validation or cross-validation procedures, relationship and age structure of the sample should be accounted for to ensure fair assessment of the predictive ability.

Concerning practical application of GBV prediction, especially in strongly related samples like progeny tested Holstein Friesian bulls, there seems to be no critical point as long as sires, half or full sibs are included in the training sets. For future prediction, though, a decrease of accuracy is expected when maximum and therefore also mean relationship between the training individuals and the candidates will decrease. If not enough new progeny tested bulls are continuously added to the training set, which may be the case in genomic selection schemes minimising the generation interval (Lillehammer *et al.* 2011), accuracy of prediction will deteriorate in perceivable steps even after only one or two generations.

ACKNOWLEDGMENTS

This research was funded by the German Federal Ministry of Education and Research within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” (Funding ID: 0315526).

REFERENCES

Browning S.R. and Browning B.L. (2007) *Am. J. Hum. Genet.* **81:** 1084.

Gilmour A.R., Gogel B.J., Cullis B.R. and Thompson R. (2009) ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Habier D., Tetens J., Seefried F.-R., Lichtner P. and Thaller G. (2010) *Genet. Sel. Evol.* **42:** 5.

Lillehammer M., Meuwissen T.H.E and Sonesson A.K. (2011) *J. Dairy. Sci.* **94:** 493.

Lund M.S., Su G., Nielsen U.S. and Aamand G.P. (2009) *Interbull Bulletin* **40:** 162.

VanRaden P.M. (2008) *J. Dairy Sci.* **91:** 4414.

4th CHAPTER

Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels

M. Erbe[1]*, B. J. Hayes[234]*, L. K. Matukumalli[5], S. Goswami[6], P. J. Bowman[23], C. M. Reich[23], B. A. Mason[23] and M. E. Goddard[27]

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

[2] Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria 3083, Australia

[3] Dairy Futures Cooperative Research Centre, Victoria 3083, Australia

[4] La Trobe University, Bundoora, Victoria 3086, Australia

[5] Bovine Functional Genomics Laboratory, USDA, Beltsville, MD 20705

[6] Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110

[7] Faculty of Land and Food Resources, University of Melbourne, Parkville, Victoria, 3010, Australia

* These authors contributed equally to this manuscript.

Published in *Journal of Dairy Science* **95:** 4114-4129.

http://dx.doi.org/ 10.3168/jds.2011-5019

ABSTRACT

Achieving accurate genomic estimated breeding values for dairy cattle requires a very large reference population of genotyped and phenotyped individuals. Assembling such reference populations has been achieved for breeds such as Holstein, but is challenging for breeds with fewer individuals. An alternative is to use a multi-breed reference population, such that smaller breeds gain some advantage in accuracy of genomic estimated breeding values (GEBV) from information from larger breeds. However, this requires that marker-quantitative trait loci associations persist across breeds. Here, we assessed the gain in accuracy of GEBV in Jersey cattle as a result of using a combined Holstein and Jersey reference population, with either 39,745 or 624,213 single nucleotide polymorphism (SNP) markers.

The surrogate used for accuracy was the correlation of GEBV with daughter trait deviations in a validation population. Two methods were used to predict breeding values, either a genomic BLUP (*GBLUP_mod*), or a new method, *BayesR*, which used a mixture of normal distributions as the prior for SNP effects, including one distribution that set SNP effects to zero. The *GBLUP_mod* method scaled both the genomic relationship matrix and the additive relationship matrix to a base at the time the breeds diverged, and regressed the genomic relationship matrix to account for sampling errors in estimating relationship coefficients due to a finite number of markers, before combining the 2 matrices. Although these modifications did result in less biased breeding values for Jerseys compared with an unmodified genomic relationship matrix, *BayesR* gave the highest accuracies of GEBV for the 3 traits investigated (milk yield, fat yield, and protein yield), with an average increase in accuracy compared with *GBLUP_mod* across the 3 traits of 0.05 for both Jerseys and Holsteins.

The advantage was limited for either Jerseys or Holsteins in using 624,213 SNP rather than 39,745 SNP (0.01 for Holsteins and 0.03 for Jerseys, averaged across traits). Even this limited and nonsignificant advantage was only observed when *BayesR* was used. An alternative panel, which extracted the SNP in the transcribed part of the bovine genome from the 624,213 SNP panel (to give 58,532 SNP), performed better, with an increase in accuracy of 0.03 for Jerseys across traits. This panel captures much of the increased genomic content of the 624,213 SNP panel, with the advantage of a greatly reduced number of SNP effects to estimate. Taken together, using this panel, a combined breed reference and using *BayesR* rather than *GBLUP_mod* increased the accuracy of GEBV in Jerseys from 0.43 to 0.52, averaged across the 3 traits.

INTRODUCTION

To accurately predict genomic breeding values for selection candidates with no phenotype of their own, a very large reference population of genotyped and phenotyped individuals is required to derive the prediction equation (Goddard, 2009; VanRaden *et al.*, 2009; Brøndum *et al.*, 2011). Although this has been achieved for breeds such as Holstein-Friesian dairy cattle in some countries (e.g., Wiggans *et al.*, 2011), for smaller breeds, assembling such large reference populations is likely to be very challenging (particularly for breeds with limited numbers of progeny-tested sires available for use in the reference population). An alternative is to use a multi-breed reference population, such that the total number of individuals in the reference set is large. For this strategy to actually increase the accuracy of genomic estimated breeding values (GEBV) within a breed requires 1) sufficiently dense markers such that the associations between the marker alleles and the alleles at the QTL affecting the traits are consistent across breed and 2) at least a proportion of the QTL segregating in several of the breeds.

de Roos *et al.* (2008) demonstrated that associations between alleles of pairs of SNP (using 1 SNP as a surrogate for a QTL) were conserved across Holstein, Jersey, and Angus populations, provided that markers were <10 kb apart. They concluded that to find markers that are in linkage disequilibrium with QTL across diverged breeds, such as Holstein, Jersey, and Angus, would require approximately 300,000 markers. The Bovine HapMap Consortium (Gibbs *et al.*, 2009) reached a similar conclusion, demonstrating that among *Bos taurus* breeds, associations between alleles at different SNP were 90% conserved across breed provided the SNP were less than 10 kb apart. In a simulated data set with the same level of linkage disequilibrium both within and across breeds as observed for real Holstein and Jersey populations, de Roos *et al.* (2009) demonstrated that the most accurate genomic predictions were achieved when phenotypes from all populations were combined in 1 reference set, provided the marker density was sufficiently high (equivalent to a marker every 10 kb).

In real data, marker density has been limited to a marker approximately every 60 kb (approximately 50,000 SNP genome wide, termed 50K). In a multi-breed beef cattle population, Kizilkaya *et al.* (2010) demonstrated limited across population predictive ability using these 50K SNP. Hayes *et al.* (2009a) and Pryce *et al.* (2011) both demonstrated very limited or no increase in accuracy of genomic predictions using these SNP with combined Holstein Jersey, and Holstein, Jersey and Fleckvieh dairy cattle reference populations, respectively.

With the recent development of an approximately 777K bovine array [Illumina Bovine high density (HD); Illumina Inc., San Diego, CA], the hypothesis that the accuracy of genomic predictions for some breeds can be improved by using a multi-breed reference population,

provided marker density is sufficiently high, can be tested. One challenge here is that a very large number of animals have been genotyped with 50K, and are unlikely to be regenotyped with the approximately 777K SNP. In this study, we explore imputation of genotypes (e.g., Browning and Browning, 2009; Marchini and Howie, 2010) as an efficient strategy to derive a large reference set with 800K genotypes.

We then explore alternative methods for deriving the SNP prediction equation. A widely used method for genomic prediction is genomic BLUP (GBLUP; e.g., VanRaden, 2008; Goddard, 2009), in which the expected relationship matrix among the animals in the population is replaced with the realized relationship matrix (or genomic relationship matrix) derived from markers. An approach is outlined for calculating the genomic relationship matrix, which takes into account both the inbreeding since the breeds diverged from a common population, and the inbreeding that has occurred since the founders of the pedigree used to derive the expected relationship matrix. This allows the genomic relationship matrix and expected relationship matrix to be combined to maximize the accuracy of prediction. Further, with such dense SNP data, an efficient strategy may be to allow a proportion of SNP to be removed from the prediction model. We outline a new computationally efficient method that allows this.

MATERIALS AND METHODS

Data

The Illumina Bovine SNP50v2.0 and BovineHD chips were used to genotype the animals. The bovine Bead-Chips were processed by following the Infinium protocol from Illumina, and the BeadChips were scanned using the iScan scanner. The raw data was analyzed using GenomeStudio software.

Two genotype data sets were used in this study. The first was heifers and bulls genotyped with the Illumina High-Density Bovine SNP chip (which we will call the 800K panel). The second data set was 2,257 Holstein and 540 Jersey Bulls genotyped with the Illumina Bovine 50K array (which we will call the 50K panel; Matukumalli *et al.*, 2009). For the first genotype data set, 903 Holstein-Friesian heifers from a feed conversion efficiency trial (Pryce *et al.*, 2012), 93 Holstein-Friesian key ancestor bulls, and 93 key ancestor Jersey bulls were genotyped with the Illumina High-Density Bovine SNP chip, which has 777,963 SNP markers. The SNP positions used were from UMD 3.1 (University of Maryland, College Park, MD). Stringent quality control procedures were applied to the data. These included the use of the Illumina GenCall score, which describes the performance of genotyping each SNP in each individual. From previous experience, genotype calls with GenTrain score (GenCall) >0.6 are

high quality; below this value they were excluded. There were 650,934 SNP genotyped at GenCall >0.6. Furthermore, 343 mitochondrial SNP, 1,124 Y chromosome SNP, and 1,735 unmapped SNP were excluded. Some 55 SNP with duplicate map positions were removed so 625,925 SNP remained. Forty-eight individuals with fewer than 90% of SNP genotyped at GenCall <0.6 were removed. Across the remaining samples, 99.6% of SNP were genotyped at GenCall >0.6. Animals with excess heterozygosity (>0.4) were removed, as this is a good indicator of sample contamination. Five animals were identified with heterozygosity above this threshold; however, all of these had already been removed in the step above (i.e., >90% of SNP genotyped). The final stage of filtering was for SNP with very low minor allele frequency (SNP with less than 10 copies of the rare allele in the population were removed). An additional filter was imposed to filter SNP with low imputation accuracy; this is described below.

In the second set of animals (2,797 Holstein and Jersey progeny-tested bulls), genotyped for the 50K panel, quality filters were imposed as described in Hayes *et al.* (2009a). Further, SNP that were not on the 800K panel after quality control in the data set were removed, leaving 39,745 SNP of the 50K panel. Mendelian consistency checks were performed on both 50K and 800K data, and genotypes failing Mendelian consistency checking were set to missing.

Phenotype data for the 2,797 bulls were daughter trait deviations (DTD; e.g., VanRaden and Wiggans, 1991) for milk yield, fat yield, and protein yield, from single-trait models.

Imputation

Imputation of the 50K data set to 800K genotypes was performed with BEAGLE software (Browning and Browning, 2009). Prior to this step, cross-validation was used to assess the accuracy of imputation that could be achieved. The Holstein heifers that were genotyped for the 800K panel were split into 2 subsets at random. In the second split, the genotypes were cut down to the 39,745 SNP on the 50K panel. Imputation was then performed, and the accuracy of imputation was taken as the proportion of genotypes that were correctly imputed. This process was then repeated, but using the second split to impute into the first split. To assess the value of having key ancestors genotyped on the 800K panel for imputation, both runs were repeated with the key ancestors 800K genotypes added.

For the Jerseys, there were only 93 key ancestor bull genotypes for the 800K panel. The accuracy of imputation was assessed using cross-validation again, but dropping 20 bulls at random as the set with 39,745 SNP. This was performed 5 times. It became obvious as a result of imputing 50K to 800K in the Holstein heifer data cross-validations that a small num-

ber of SNP (1,231) were imputed very inaccurately, with accuracy across animals below 80% (Figure 1). Accuracy here is defined as the proportion of genotypes that are correctly imputed. We postulated that these SNP could be mismapped. We attempted to remap the SNP using linkage disequilibrium information. For each of the 1,231 SNP, the R^2 with all the other 624,924 SNP was calculated using genotype frequencies as described by Zaykin *et al.* (2008). The weighted (by distance from the center of the window) average R^2 was calculated in 20 SNP windows across the genome. If the window with the highest average R^2 with the remapped SNP was greater than 1,000 kb different to the position in the original map file, the new position of the SNP being remapped was at the center of the window with the highest weighted average R^2 value. This algorithm is implemented in ldMapper, a program available from the authors.

The imputation was performed again using the proposed new positions of the SNP. This greatly improved the accuracy of imputation for 601 of the SNP; however, 630 of the SNP were still poorly imputed (Figure 1). These were removed from the data set, giving a final data set of 624,213 SNP for the 800K panel. The cross-validations described above were redone to get the final results. The 800K panel genotypes (actually 624,213 SNP) were then imputed into the 50K bull data set.

Finally, as the BEAGLE imputation as implemented here does not use pedigree information, we tested for Mendelian inconsistencies in the post-BEAGLE (imputed) 800K genotypes. We found that a small proportion of SNP genotypes were inconsistent in sire-son comparisons (e.g., opposing homozygotes), amounting to 0.6% of the genotypes.

Transcriptome Panel

To test both the hypothesis that mutations affecting quantitative traits reside in exons, introns, and regulatory regions, and to potentially reduce the computational demand when calculating genomic predictions, we tested another panel of SNP that were in the 624,213 above (800K panel) and also within or near the transcribed part of the genome. The start-stop positions of the transcribed part of the genome were as defined by L. K. Matukumalli (author on the current paper), plus SNP within 1 kb of these stop or start positions. The transcribed part of the genome was identified from a large collection of mRNA transcripts, mapped to the UMD 3.0 bovine assembly (http://www.cbcb.umd.edu/research/bos_taurus_assembly.shtml). This panel (which we will call the transcriptome panel, TRANS) consisted of 58,532 SNP.

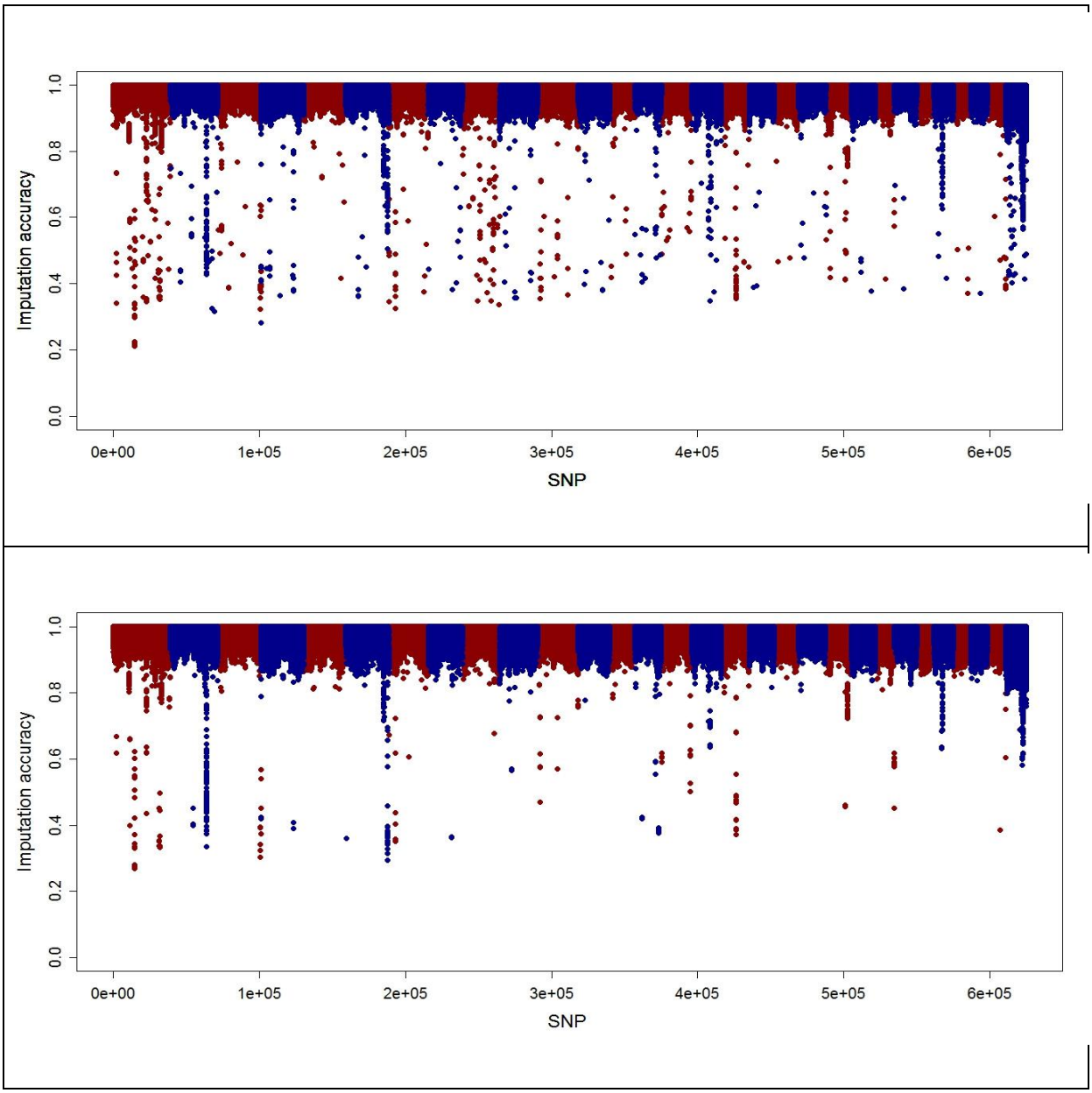


Figure 1: Accuracy of imputation by SNP using BEAGLE software (Browning and Browning, 2009), before and after remapping 1,231 SNPs with < 80% accuracy of imputation in the original data set. Single nucleotide polymorphisms with <80% accuracy of imputation were remapped using linkage disequilibrium (LD), with the new position taken as the position that gave the highest LD in a window of 20 SNP, with all genome positions considered.

Methods for Genomic Prediction

The bulls in each breed were split into reference bulls (those progeny tested before 2007) and validation bulls (those progeny tested in 2007 or later). There were 1,897, 360, 454, 86 Holstein reference, Holstein validation, Jersey reference, and Jersey validation bulls, respectively. Unless otherwise described, the reference set combined both the Holstein and Jersey reference bulls. The surrogate used for accuracy of GEBV was the correlation of GEBV and

DTD in the validation bulls. This surrogate was not corrected for the reliability of the DTD (which averaged 0.8 in the validation sets). The regression of GEBV on DTD was also calculated. For each method, the SNP subsets used were 50K, 800K, and TRANS panels. The methods used to predict GEBV were as follows.

GBLUP

The following model was fitted to the data

$$\mathbf{y} = \mathbf{1}_{\mathrm{n}}\mu + \mathbf{Zg} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{1}_n$ is a vector of 1s, μ is an overall mean, $\mathbf{Z}$ is a design matrix allocating records to breeding values, $\mathbf{g}$ is a vector of genomic breeding values, and $\mathbf{e}$ is a vector of random normal deviates with variance $\mathbf{V}(e) \sim N(0, \sigma_e^2)$, where σ_e^2 is the error variance. The variance of breeding values was $\mathbf{V}(g) = \mathbf{G}\sigma_g^2$, where $\mathbf{G}$ is the genomic relationship matrix derived as in Yang *et al.* (2010), with no consideration of breed, and σ_g^2 is a genetic variance. Then, breeding values for both phenotyped and nonphenotyped individuals can be predicted as

$$[\hat{\mathbf{g}}] = \left[\mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\right]^{-1} [\mathbf{Z}'(\mathbf{y} - 1_n\hat{\mu})],$$

where $\hat{\mathbf{g}}$ is a vector of EBV, $\mathbf{Z}'$ is the transpose of $\mathbf{Z}$, and $\hat{\mu}$ is an estimate of the mean. Variance components were estimated with ASReml software (Gilmour *et al.*, 2002).

GBLUP_mod

Goddard *et al.* (2011) argued that EBV, and particularly the accuracy derived from the coefficient matrix from GBLUP, are biased due to sampling errors in elements of the genomic relationship matrix due to a finite number of markers, where the expectation of $\mathbf{G}$, with $\mathbf{G}$ defined as above, given the estimate of $\mathbf{G}$ ($\widehat{\mathbf{G}}$), $E(\mathbf{G}|\widehat{\mathbf{G}}) \neq \widehat{\mathbf{G}}$. This also means, for example, that information from the expected relationship matrix ($\mathbf{A}$) derived from pedigree and $\mathbf{G}$ cannot be combined to maximize the accuracy of the EBV in a one-step approach (e.g., Misztal *et al.*, 2009). Goddard *et al.* (2011) suggested a new genomic relationship matrix that regressed elements of $\mathbf{G}$ toward $\mathbf{A}$ to account for sampling error in estimating coefficients of $\mathbf{G}$ to create a new matrix $\mathbf{G}^*$:

$$\mathbf{G}^* = [\mathbf{A} + b(\mathbf{G} - \mathbf{A})], \quad [1]$$

where

$$b = V(\mathbf{G})/\left[V(\mathbf{G}) + \frac{1}{m}\right] \quad [2]$$

and $V(\mathbf{G})$ is the variance of the nondiagonal elements of $\mathbf{G}$ obtained with m markers; $V(\mathbf{G})$ can be obtained simply by taking all of the nondiagonal elements of $\mathbf{G}$, where $\mathbf{G}$ is calculated as in Yang *et al.* (2010) and calculating the variance of these elements.

To derive $\mathbf{G}^*$ for a multi-breed population, an appropriate base population relative to which $\mathbf{G}$ and $\mathbf{A}$ are both defined must be chosen. One logical base population in our situation is that immediately before the divergence of Holsteins and Jerseys.

First, a $\mathbf{G}$ matrix can be calculated, which records covariances relative to a base that is a composite breed (c) made up of a proportion of α Holsteins and $(1-\alpha)$ Jerseys.

$$\mathbf{G}_c = \frac{\mathbf{WW}'}{\mathbf{M}}$$

where $\mathbf{W}$ is a centered matrix calculated as $\mathbf{W} = \mathbf{X} - 2\mathbf{p}$, with $\mathbf{p} = \alpha\mathbf{p}_{hol} + (1-\alpha)\mathbf{p}_{jer}$, and $\mathbf{M} = 2\sum_{i=1}^{m} p_i(1-p_i)$. Here, $\mathbf{p}_{hol}$ and $\mathbf{p}_{jer}$ are the average allele frequencies of the 2 allele in Holsteins and Jerseys, respectively; $\mathbf{X}$ is a matrix of animals by SNP, with SNP genotypes coded $0 = 11, 1 = 12,$ or $2 = 22$; $\alpha = \frac{F_{jer}}{F_{jer}+F_{hol}}$ with F_{jer} and F_{hol} defined below; and p_i is the frequency of the 2 allele for the i^{th} SNP. The calculation of $\mathbf{G}_c$ is similar to that described by VanRaden (2008) for a purebred population but with a modification to the allele frequencies to scale $\mathbf{G}$ to the composite base. Our approach is different from that of Harris and Johnson (2010), who also derived $\mathbf{G}$ for a multibreed population. They used the approach of partitioning the diagonals of the matrix into breed fractions to account for different variances among breeds and include segregation variances because of different allele frequencies among breeds. However, their approach will accommodate crossbred animals; ours would need to be extended to do this.

Then, in our approach $\mathbf{G}_c$ is adjusted for the inbreeding that has occurred in both breeds relative to the old base (the base at the divergence of Holsteins and Jerseys):

$$\mathbf{G} = \mathbf{G}_c(1-F) + 2F,$$

where F is the inbreeding relative to an $F1$ base:

$$F = \frac{F_{jer}F_{hol}}{F_{jer} + F_{hol}},$$

$$F_{jer} = 1 - \frac{\sum_{i=1}^{m} 2p_{jer,i}(1 - p_{jer,i})}{\sum_{i=1}^{m}[p_{hol,i}(1 - p_{jer,i}) + p_{jer,i}(1 - p_{hol,i})]},$$

and

$$F_{hol} = 1 - \frac{\sum_{i=1}^{m} 2p_{hol,i}(1 - p_{hol,i})}{\sum_{i=1}^{m}[p_{hol,i}(1 - p_{jer,i}) + p_{jer,i}(1 - p_{hol,i})]}.$$

The pedigree-derived $\mathbf{A}$ must also be converted to the old base (e.g., Powell *et al.* 2010). For the within-Holstein blocks, $\mathbf{A} = \mathbf{A}_{ped}[1 - (F - f_{hol})] + 2(F - f_{hol})$, where f_{hol} is the amount of inbreeding that has occurred since the base of the pedigree within Holsteins; we approximated this as the average of the off-diagonal elements of $\mathbf{A}_{ped}$. The within-Jersey block was constructed in the same way. All elements of the Holstein × Jersey block of $\mathbf{A}$ were 0, as no pedigree links existed between the breeds. Note that in practice, the estimate of f_{hol} and f_{jer} could be an underestimate due to the incompleteness of the pedigree. With an incomplete pedigree the base is less well defined.

Once $\mathbf{G}$ and $\mathbf{A}$ were constructed, the regression of $\mathbf{G}$ toward $\mathbf{A}$ to account for sampling errors in the genomic relationship coefficients (Equation 1) was determined. This was done separately for each breed, and the breed × breed block (e.g., Holstein × Jersey) by calculating the variance of the off-diagonal elements within each of these blocks.

BayesR

The *GBLUP* approaches assume that all markers have a small effect and that these effects are normally distributed (e.g., Habier *et al.*, 2007; Hayes *et al.*, 2009b). Given the large number of markers, a more appropriate prior may be that some of the markers are not in linkage disequilibrium with QTL, so have zero effect, whereas others have a small to moderate effect. This prior was proposed by Meuwissen *et al.* (2001). The challenge of implementing a method that uses such a mixture prior is computational efficiency – for example, in the *BayesB* of Meuwissen *et al.* (2001), sampling of SNP variances from their posterior distributions simultaneously with the SNP effects required a Metropolis Hastings algorithm.

Verbyla *et al.* (2009) described a stochastic search variable selection (*BayesSSVS*) strategy, which maintained the same assumptions about the distributions of SNP effects while maintaining constant dimensionality, which allowed a Gibbs sampling scheme to be used to construct the posterior distributions of the parameters. However, one potential criticism of both *BayesB* and *BayesSSVS* is that the proportion of SNP in each distribution was not sampled

appropriately, such that the means of the posterior distributions of the proportion of SNP with a zero or nonzero effect closely reflected the prior values of these proportions (e.g., "lack of Bayesian learning"; Habier *et al.*, 2011). Here, both to overcome this drawback of *BayesB* and *BayesSSVS*, and for computational efficiency, we propose a new method that assumes that the true SNP effects are derived from a series of normal distributions, the first with zero variance, up to one with a variance of approximately 1% of the genetic variance. The model fitted to the data was

$$\mathbf{y} = \mathbf{1}_n'\mu + \mathbf{Wu} + \mathbf{Zv} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of n DTD for each trait; $\mathbf{W}$ is the $(n\ x\ m)$ design matrix allocating records to the marker effects described above; vector $\mathbf{u}$ is a $(m\ x\ 1)$ vector of SNP effects assumed normally distributed $\left[u_i \sim N\left(0, \sigma_i^2\right)\right]$; $\mathbf{e}$ is a vector of random deviates, where σ_e^2 is the error variance; v_j is the polygenic breeding value of the j[th] animal, $\boldsymbol{V}(v) = \mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is the average relationship matrix; σ_a^2 is the polygenic variance; and $\mathbf{Z}$ is a matrix that allocates records to animals.

The variance of the i[th] SNP effect had 4 possible values:

$$\sigma_1^2 = 0, \sigma_2^2 = 0.0001\sigma_g^2, \sigma_2^2 = 0.001\sigma_g^2, \sigma_2^2 = 0.01\sigma_g^2,$$

where σ_g^2 is the assumed total genetic variance, which was calculated as $\sigma_g^2 = r_{DTD}^2\sigma_{DTD}^2$, with r_{DTD}^2 being the assumed reliability of the DTD, and σ_{DTD}^2 the variance of the DTD. Using these variances results in shrinkage that allows the SNP effects themselves to range from zero effect to moderate effect. The proportions of the SNP in each distribution were $pr1, pr2, pr3,$ and $pr4,$ respectively, in a vector $\mathbf{pr}$.

Bayesian estimation of the parameters was used. The prior distribution of the proportions of SNP in each distribution $\mathbf{pr}$ was the Dirichlet distribution, with $\boldsymbol{\alpha} = \mathbf{1}$ (where $\boldsymbol{\alpha}$ is a 4 × 1 vector of pseudo counts, all with value 1 to give an almost uninformative prior with the numbers of SNP used here). The Dirichlet distribution is a convenient choice of prior, as it is a conjugate before the multinomial distribution, such that the posterior distribution of $\mathbf{pr}$ is $\sim Dir(\boldsymbol{\alpha} + \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector containing the number of SNP in each distribution estimated from the data. To obtain these estimates, we first calculated 4 likelihoods assuming the considered SNP being in 1 of the 4 normal distributions at a time with the respective probability pr_k.

The likelihood that SNP i is in distribution k is

$$LogL(i,k) = -0.5\log|\mathbf{V}| - \frac{0.5\left(\mathbf{y}^{*\prime}\mathbf{y}^{*} - \mathbf{y}^{*}\mathbf{Z}^{*}\hat{\mathbf{u}}\right)}{\sigma_e^2} + \log(pr_k),$$

where $\mathbf{y}^{*}$ is the vector of phenotypes corrected for all marker effects other than marker i, the overall mean, and the polygenic effects ($\hat{\mathbf{u}}$); $\mathbf{Z}^{*}$ is a column vector containing the SNP genotypes of all animals for SNP i ; $\mathbf{V}$ is the variance-covariance structure of a reduced model, including only the effect of the respective SNP and a residual effect; and $\log|\mathbf{V}|$ was calculated as $n\log(\sigma_e^2) + \log\left(\frac{\sigma_i^2\mathbf{Z}^{*\prime}\mathbf{Z}^{*}}{\sigma_e^2} + 1\right)$, where $\mathbf{Z}^{*}$ contains only the information for the current SNP effect.[1]

Then, the probability that SNP i is in distribution k is

$$\frac{1}{\sum_{l=1}^{4}\exp[\mathrm{L}(i,l) - \mathrm{L}(i,k)]}.$$

Based on these probabilities, we selected the normal distribution to sample the SNP effect from using a uniform random variate, using the probabilities of the SNP being in each of the distributions for the current iteration. Over all the SNP, we thus obtained estimates for the elements of $\boldsymbol{\beta}$.

[1] Note that there is an error in the original version of this manuscript. The correct version of this paragraph is:

The likelihood that SNP *i* is in distribution *k* is

$$LogL(i,k) = -0.5\log|\mathbf{V}| - \frac{0.5\left(\mathbf{y}^{*\prime}\mathbf{y}^{*} - \mathbf{y}^{*}\mathbf{Z}^{*}u_k^{*}\right)}{\sigma_e^2} + \log(pr_k),$$

where $\mathbf{y}^{*}$ is the vector of phenotypes corrected for all marker effects other than marker i, the overall mean and the polygenic effects, u_k^{*} is the mean of the posterior distribution of the SNP effect when assumed to be in the k^{th} distribution, $\mathbf{Z}^{*}$ is a column vector containing the SNP genotypes of all animals for SNP i, $\mathbf{V}$ is the variance-covariance-structure of a reduced model including only the effect of the respective SNP and a residual effect and $\log|\mathbf{V}|$ was calculated as $n\log(\sigma_e^2) + \log\left(\frac{\sigma_k^2\mathbf{Z}^{*\prime}\mathbf{Z}^{*}}{\sigma_e^2} + 1\right)$, where $\mathbf{Z}^{*}$ contains only the information for the current SNP effect.

The posterior of $\mathbf{pr}$ cannot be estimated directly, as it is conditional on both the estimates of the SNP effects (to calculate $\mathbf{y}^*$) and estimates of the polygenic effects $\hat{\mathbf{u}}$. A Gibbs sampling scheme was, therefore, used to sample from the posterior distributions of all parameters conditional on the other parameters.

Prior distributions for other parameters were as described by Verbyla *et al.* (2009). The Gibbs sampling scheme was similar to that described by Meuwissen *et al.* (2001) for *BayesA*, but with the addition of a polygenic effect, and with the SNP variances described above. At the end of each iteration, the proportion of SNP in each distribution was sampled from the posterior Dirichlet distribution as described above. We also compared $r(GEBV, DTD)$ from *GBLUP_mod* and *BayesR* to those derived from SNP effects estimated by *BayesA* (Meuwissen *et al.*, 2001).

RESULTS

Accuracy of Imputation

In the Holstein heifer data set, the accuracy of imputation of 50K to 800K was similar across the 2 cross-validations, with an average of 97.4% (Table 1). Adding the key ancestor 800K genotypes improved the accuracy of imputation by 0.5%, despite the limited number of these ancestors. The average accuracy of imputation in the Jersey cross-validations was lower, likely reflecting the much more limited number of animals genotyped for the 800K panel.

Comparison of GBLUP and GBLUP_mod

To check that the proposed modifications to the $\mathbf{G}$ matrix and $\mathbf{A}$ matrix in the *GBLUP_mod* method resulted in relationship matrices expressed relative to the same base population, before $\mathbf{G}^*$ was calculated, we checked the average of the diagonal elements for each breed, and the average off-diagonal elements within and across breeds. These were very close (Table 2). The regressor $\hat{b}$ of $\mathbf{G}$ toward $\mathbf{A}$, which accounts for sampling error in estimating the coefficients of $\mathbf{G}$ is also given for each block. Within a breed, the value of $\hat{b}$ was only slightly less than 1; however, in the across-breed block, the value of $\hat{b}$ was lower at 0.89, reflecting the fact that across-breed genomic relationships are smaller in magnitude, and are estimated with lower precision than within-breed genomic relationships. However, the value of 0.89 is surprisingly high, and may reflect the fact that the Australian dairy herd was upgraded from a largely Jersey base, such that relatively large chromosome segments originating from Jerseys can still be found in cattle classified as Holstein.

Table 1: Accuracy of imputation of 50,000 to 800,000 SNP (50K to 800K) in cross validation of 940 Holstein and 93 Jersey genotypes[1]

	Cross validation	Accuracy of genotype imputation (%)
Holstein		
Heifers only	1	97.4
	2	97.9
	Average	97.7±0.01
Heifers + key ancestors	1	98.0
	2	98.0
	Average	98.0±0.05
Jersey		
	1	96.1
	2	95.0
	3	97.0
	4	95.4
	5	94.2
	Average	95.6±0.05

[1] Cross-validation in the Holstein data set involved splitting the 843 heifers in 2 approximately subsets, and then in silico reducing the numbers of genotypes to the 50K panel. In Jerseys, approximately 20 individuals at each cross-validation were assigned to have their genotypes reduced to the 50K panel.

Table 2: Average of elements of expected and realized relationship matrices ($\mathbf{A}$ and $\mathbf{G}$ respectively), after rescaling to a base which was at the time of divergence of Holsteins and Jerseys[1]

Statistic	Matrix elements	Validation	$\mathbf{A}$	$\mathbf{G}$	$\hat{b}$
Average	Diagonal	Holstein	1.09	1.11	-
Average	Diagonal	Jersey	1.20	1.22	-
Average	Block	Holstein	0.20	0.19	0.96
Average	Block	Jersey	0.42	0.39	0.97
Average	Block	Across breed	0.00	0.01	0.89

[1] The regressor $\hat{b}$ of $\mathbf{G}$ toward $\mathbf{A}$, which accounts for sampling error in estimating the coefficients of $\mathbf{G}$ is given for each block.

Next, we evaluated the effect of using *GBLUP_mod* rather than *GBLUP* on the accuracy and bias of GEBV. For the 800K panel, the accuracy of GEBV [as indicated by the surrogate measure $r(GEBV, DTD)$] from *GBLUP* and *GBLUP_mod* was similar for the Holstein validation data set, but, on average, 0.03 higher for *GBLUP_mod* in the Jersey validation data set

(Table 3). Regressions of DTD on GEBV were closer to 1 in the Jersey validation data sets with *GBLUP_mod* than with *GBLUP* in all traits.

Table 3: Correlations of daughter trait deviations (DTD) and genomic EBV (GEBV) [$r(GEBV, DTD)$] and regressions of DTD on GEBV slopes [$b(DTD, GEBV)$] from *GBLUP* and *GBLUP_mod* methods[1]

		Trait			
Method	Validation	Milk yield	Fat yield	Protein yield	Average
$r(GEBV, DTD)$					
GBLUP	Holstein	0.58	0.58	0.56	0.57
	Jersey	0.33	0.46	0.40	0.40
GBLUP_mod	Holstein	0.58	0.58	0.56	0.57
	Jersey	0.36	0.49	0.44	0.43
$b(DTD, GEBV)$					
GBLUP	Holstein	1.04	1.19	0.94	1.05
	Jersey	0.53	0.86	0.71	0.70
GBLUP_mod	Holstein	1.04	1.16	0.93	1.04
	Jersey	0.69	1.00	0.94	0.88

[1] The *GBLUP_mod* method uses a rescaled genomic relationship matrix, and regresses the $\mathbf{G}$ matrix towards the $\mathbf{A}$ matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers.

Comparison of GBLUP_mod, BayesR, and BayesA for the 800K Panel

Table 4 shows the results for *BayesR*, *BayesA*, and *GBLUP_mod* for the combined reference population and the 800K panel. The *BayesR* method gave higher $r(GEBV, DTD)$for both milk yield and fat yield than *GBLUP_mod*, whereas $r(GEBV, DTD)$ for protein yield was similar. Averaged across the traits, the advantage of *BayesR* over *GBLUP_mod* was 0.05 in $r(GEBV, DTD)$. This advantage was observed in both the Holstein and the Jersey validation data set. The regression of DTD on GEBV (Table 4) was similar for all methods. To compare *BayesR* with a well-known Bayesian method, we also ran *BayesA*. *BayesA* gave similar, but very slightly lower $r(GEBV, DTD)$ for milk yield than *BayesR* and similar results in terms of slope [$b(DTD, GEBV)$].

Table 4: Accuracy of prediction [expressed as $r(GEBV, DTD)$] and slopes [$b(DTD, GEBV)$] of the regression of daughter trait deviations (DTD) on predicted genomic EBV (GEBV) for *GBLUP_mod*, *BayesA,* and *BayesR* with a multi-breed reference population and the 800,000-SNP (800K) panel (the result averaged across traits is also given)

Method[1]	Validation	Milk yield	Fat yield	Protein yield	Average
$r(GEBV, DTD)$					
GBLUP_mod	Holstein	0.58	0.58	0.56	0.57
	Jersey	0.36	0.49	0.44	0.43
BayesA	Holstein	0.61	0.66	0.58	0.62
	Jersey	0.48	0.49	0.46	0.48
BayesR	Holstein	0.62	0.66	0.57	0.62
	Jersey	0.51	0.49	0.46	0.49
$b(DTD, GEBV)$					
GBLUP_mod	Holstein	1.04	1.16	0.93	1.04
	Jersey	0.69	1.00	0.94	0.88
BayesA	Holstein	1.04	1.12	0.94	1.03
	Jersey	0.82	0.91	0.86	0.86
BayesR	Holstein	0.99	1.12	0.91	1.01
	Jersey	0.84	0.92	0.86	0.88

[1] The *GBLUP_mod* method uses a rescaled genomic relationship matrix, and regresses the $\mathbf{G}$ matrix toward the $\boldsymbol{A}$ matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers; *BayesR* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions; and *BayesA* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a Student's *t* distribution. Complete descriptions are given in the text.

Comparison of Different Marker Panels

For genomic predictions within a pure breed, there was no advantage of either the 800K or TRANS panel over the 50K panel when *GBLUP_mod* was used (Table 5). When *BayesR* was used, there was only a very small advantage (and not significant), given the sample size used, in $r(GEBV, DTD)$ of using the 800K or the TRANS panel over the 50K panel in some cases (Table 6). This was of the order of 0.01 averaged across traits for Holsteins, comparing the 800K to the 50K panel, and 0.02 for Jerseys comparing the TRANS panel to the 50K panel (Table 7).

Some improvement for prediction across breeds occurred using only the other breed as the reference when *BayesR* was used with either the 50K or the 800K panel, compared with the *GBLUP_mod* results. For the TRANS panel, the accuracy for predicting Jersey GEBV from a Holstein-only reference looked promising (0.24 average across traits; Table 7). Interestingly, the $r(GEBV, DTD)$ for milk yield was much higher (0.40 and 0.30, respectively) with both

methods when the TRANS panel was used compared with both other panels (Tables 5 and 6).

When a combined reference set was used, *BayesR* clearly outperformed *GBLUP_mod* across all scenarios and traits, especially with prediction of fat yield in Holstein (up to 0.08 higher) and milk yield (0.15 higher) and protein yield (0.05 higher) in Jersey in all panels. The best results for predicting the minor breed (Jerseys) were obtained with a combined reference set, *BayesR* and the TRANS panel [$r(GEBV, DTD)$ = 0.52; Table 7]. This was 0.09 higher than that obtained using *GBLUP_mod*, the combined reference set, and the 800K panel (Table 4).

Table 5: Accuracy of genomic prediction [$r(GEBV, DTD)$] from *GBLUP_mod*[1] using different marker panels and either single-breed or combined reference populations[2]

		Milk yield			Fat yield			Protein yield		
Reference	Validation	50K	800K	TRANS	50K	800K	TRANS	50K	800K	TRANS
Holstein	Holstein	0.61	0.58	0.62	0.58	0.57	0.57	0.57	0.56	0.55
	Jersey	-0.07	-0.01	0.30	-0.24	-0.16	-0.05	-0.31	-0.21	0.05
Jersey	Holstein	0.04	-0.03	0.03	0.16	0.18	0.11	0.14	0.16	0.08
	Jersey	0.38	0.37	0.39	0.49	0.48	0.47	0.43	0.43	0.42
Combined	Holstein	0.60	0.58	0.62	0.58	0.58	0.57	0.57	0.56	0.55
	Jersey	0.35	0.36	0.45	0.47	0.49	0.44	0.40	0.44	0.48

[1] The *GBLUP_mod* method uses a rescaled genomic relationship matrix, and regresses the $\mathbf{G}$ matrix toward the $\mathbf{A}$ matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers.

[2] GEBV = genomic EBV; DTD = daughter trait deviations; 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

Table 6: Accuracy of genomic prediction [$r(GEBV, DTD)$] from *BayesR*[1] using different marker panels and either single-breed or combined reference populations[2]

		Milk yield			Fat yield			Protein yield		
Reference	Validation	50K	800K	TRANS	50K	800K	TRANS	50K	800K	TRANS
Holstein	Holstein	0.62	0.63	0.63	0.64	0.65	0.63	0.55	0.57	0.56
	Jersey	0.27	0.24	0.40	0.12	0.21	0.12	-0.05	0.05	0.21
Jersey	Holstein	0.19	0.03	0.15	0.29	0.29	0.18	0.13	0.10	0.12
	Jersey	0.49	0.48	0.53	0.48	0.46	0.47	0.42	0.41	0.43
Combined	Holstein	0.61	0.62	0.62	0.65	0.66	0.64	0.56	0.57	0.57
	Jersey	0.45	0.51	0.57	0.50	0.49	0.45	0.43	0.46	0.53

[1] *BayesR* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions.

[2] GEBV = genomic EBV; DTD = daughter trait deviations; 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

Distribution of SNP Effects

For *BayesR*, we could calculate the number of SNP in each distribution (explaining 0, 0.01, 0.1, or 1% of the genetic variance). This was achieved by calculating the posterior mean of the sampled proportions of SNP in each of the 4 distributions over all post burn-in iterations, and multiplying them by the total number of SNP. The results show that, on average, only between 7 and 14% (depending on trait) of all SNP contribute to the prediction of genomic breeding value with the 50K panel. Similar absolute numbers of SNP were in distribution 2, 3, and 4 with the 800K panel; that is, the majority of SNP with this panel (over 99%) were estimated to be in the first distribution, which had zero variance (Table 8). When a combined (Holstein and Jersey) reference set was used, for all traits, the number of SNP in the 0.01 distribution was lower than or similar to the purebred Holstein scenario. For distribution 3, the number of SNP was clearly lower than when a single breed reference set was used, whereas it was usually higher for distribution 2. Possible reasons for this are proposed in the discussion. In most cases, the number of SNP in distribution 1 and 2 was clearly lower for fat yield than for both of the other traits with all SNP panels. With the Jersey reference set, more SNP were assumed to explain larger parts of the total variance than with the Holstein reference set. For the TRANS panel, the number of SNP in distribution 1 and 2 could be expected to be higher, as the SNP for this panel were all located in or near transcribed regions. However, we did not observe this trend.

Table 7: Accuracy of genomic prediction [$r(GEBV, DTD)$] from *BayesR*[1] using different marker panels and either single-breed or combined reference populations, averaged across traits

		Panel		
Reference	Validation	50K	800K	TRANS
Holstein	Holstein	0.61	0.62	0.61
	Jersey	0.11	0.17	0.24
Jersey	Holstein	0.20	0.14	0.15
	Jersey	0.46	0.45	0.48
Combined	Holstein	0.61	0.62	0.61
	Jersey	0.46	0.49	0.52

[1] *BayesR* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions.

[2] GEBV = genomic EBV; DTD = daughter trait deviations; 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

Table 8: Average number of SNP in the 4 normal distributions modeled with *BayesR*[1]

	Reference								
	Jersey			Holstein			Combined		
Panel[2]	Milk (kg)	Fat (kg)	Protein (kg)	Milk (kg)	Fat (kg)	Protein (kg)	Milk (kg)	Fat (kg)	Protein (kg)
50K									
1st	35,730	34,201	36,179	34,991	35,917	35,844	34,245	34,558	34,880
2nd	3,677	5,276	3,268	4,612	3,598	3,798	5,410	5,040	4,820
3rd	315	255	287	134	222	93	81	139	36
4th	24	13	10	8	8	10	9	7	8
800K									
1st	620,151	620,026	619,488	620,570	620,544	620,151	620,372	619,526	619,650
2nd	3,727	3,828	4,462	3,390	3,528	3,538	3,579	4,467	4,478
3rd	306	339	254	245	227	122	251	210	77
4th	29	20	9	9	13	9	11	10	8
TRANS									
1st	54,742	54,850	54,242	54,144	55,233	54,953	53,317	54,121	54,272
2nd	3,480	3,210	4,039	4,264	3,064	3,480	5,145	4,257	4,206
3rd	276	455	241	116	225	93	63	143	48
4th	34	17	10	7	11	7	7	11	6

[1] The average number of SNP was calculated as the mean proportion of SNPs in the distribution times the total number of SNP. *BayesR* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions.

[2] 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS =transcriptome panel.

DISCUSSION

In this study, we tested 3 hypotheses: 1) the accuracy of genomic estimated breeding values would be increased using denser marker panels, when the validation animals and reference animals were the same breed, 2) the advantage of using a denser marker panel would be even greater when the validation animals and reference animals were from different breeds, or a combined breed reference set was used, and 3) a method for deriving the prediction equation that could result in a large number of SNP effects being set to zero (e.g., excluded from the prediction model) would result in the greatest advantage from increasing the density of the marker panel.

The support for hypothesis 1) was limited. The $r(GEBV, DTD)$ for the Holstein population did increase when the 800K panel was used rather than the 50K panel, but only by 0.01 averaged across traits, and only when *BayesR* was used. For Jersey (using Jersey reference to predict GEBV in a Jersey validation set), the average $r(GEBV, DTD)$ actually decreased by 0.01 when the 800K panel was used rather than the 50K. In contrast to humans where a very large number of SNP are necessary for accurate genomic predictions due to a large effective population size (e.g., Wray *et al.*, 2007), in modern dairy cattle breeds effective population sizes are sufficiently small that linkage disequilibrium (LD) between SNP and potential QTL is captured even with the 50K panel, and increasing this LD by using a denser panel does not have much effect. Evidence for this is that the proportion of the genetic variance captured by the 50K panel is only slightly lower than that from the 800K panel (Table 9; Haile-Mariam *et al.*, accepted), regardless of which method is used. In sharp contrast to what is observed in human populations, we were able to capture almost 90% of the heritability of our phenotype (DTD) estimated from pedigree with the markers; in human populations this figure is more like 56% for a trait such as human height (Yang *et al.*, 2010). Interestingly, the proportion of variance unexplained with *BayesR* was greatest with fat yield. One explanation for this may be that the largest distribution from which SNP effects are sampled has a variance of 1%, resulting in overshrinking of the effect of DGAT1, such that less variance is explained.

For Jerseys, we must point out that our reference population was small; therefore, any potential advantage in using denser panels may be obscured by the estimation error associated with the greatly increased number of SNP. Further, for Jerseys, the imputation reference set (for imputation of 800K from 50K) comprised only 93 key ancestors, which led to clearly lower imputation accuracies than in Holsteins (Table 1). Inaccurate genotype imputation would have reduced the possible advantages of using the 800K panel (and a multi-breed reference population) for Jerseys.

Table 9: Proportion of genetic variance (estimated from pedigree) unaccounted for by SNP markers, using the Holstein only reference set[1]

		Trait		
Method[2]	Panel	Milk yield	Fat yield	Protein yield
GBLUP_MOD	50K	0.12	0.13	0.17
	800K	0.11	0.12	0.15
BayesR	50K	0.08	0.22	0.12
	800K	0.08	0.18	0.10

[1] For *BayesR*, this was calculated as the estimated polygenic variance from the model divided by the total genetic variance, for *GBLUP_mod*, it was calculated as the variance explained by the modified $\mathbf{G}$ matrix divided by the genetic variance estimated from a model with only a polygenic effect with co(variance) matrix the expected relationship matrix ($\mathbf{A}$).
[2] *GBLUP* = genomic BLUP.

Support for hypothesis 2) was a little more convincing; the average of $r(GEBV, DTD)$ across traits in the Jersey validation set, with Holsteins used as the reference, increased from 0.11 (50K) to 0.17 (800K) when *BayesR* was used (Table 7). With 800K SNP, the persistence of phase among SNP and QTL alleles should be consistent across *B. taurus* breeds (Gibbs *et al.*, 2009). However, this assumes the same QTL are segregating in the different breeds, whereas our results suggest this is only true in a proportion of cases, as discussed below.

There was some support for hypothesis 3). The greatest increase in $r(GEBV, DTD)$ from using the 800K panel rather than the 50K panel were observed when *BayesR* was used rather than *GBLUP_mod* (for example, for prediction of Jersey GEBV from the combined reference population). These results suggest that to take advantage of the increased marker density, methods that either explicitly remove SNP from the model or set their effect to zero (2 ways of achieving the same thing) are necessary.

One possible explanation for our results (especially the limited gains in $r(GEBV, DTD)$ from using 800K compared with 50K) is that we have greatly increased the number of SNP effects to be estimated, without increasing the number of records. Particularly the Jersey population is small, so that the effect of the large increase in the number of estimation errors could erode the accuracy of GEBV. An alternative to using all 800K SNP would be to select a much smaller subset that may be a priori more relevant, thus avoiding the need to estimate a very large number of SNP effects. For our TRANS panel, we selected a subset of SNP from the 800K that was included the transcribed portion of the genome (L. K. Matukumalli, author on the current paper). The TRANS panel worked reasonably well for all traits and led to similar or even better (e.g., in milk yield with *BayesR*) results than with both the other SNP panels.

The average $r(GEBV, DTD)$ for Jerseys was highest using this panel, and accuracies of across breed prediction using the other breed as reference set were quite promising.

Our results for the increase in accuracy for the minor breed (Jerseys) using a combined reference and the 800K panel can be compared with the simulated results from de Roos *et al.* (2009). The simulation those authors used to generate marker associations within and across breeds was based on actual LD within and across similar populations to those considered here. If the divergence time between Holsteins and Jerseys is taken at approximately 300 generations (e.g., de Roos *et al.*, 2008), then their simulation results would suggest that the increase in the accuracy of genomic EBV for Jerseys, as a result of using the 800K panel and combining the reference populations, should have been considerably greater than was observed here. Some of the explanation may be due to too few records to accurately estimate the 800K marker effects, as described above, and imperfect imputation of 800K from 50K, particularly in Jerseys.

However, de Roos *et al.* (2009) also simulated QTL that were segregating in both breeds in most cases. Our results suggest that only some of the QTL segregate across breed. For example, for milk yield, the 9 SNP in Holstein that explained 1% of the genetic variance according to their posterior mean from *BayesR* (Table 8) were tightly clustered in 3 regions, on chromosome 14 (DGAT1), chromosome 5, and chromosome 11. Although the QTL on chromosome 14 and chromosome 5 were detected in Jerseys (as evidenced by clusters of SNP in the fourth distribution of *BayesR*, explaining 1% of the variance, using a Jersey-only reference population), no evidence indicated that the QTL on chromosome 11 was segregating in Jerseys. Further, in Jerseys, QTL were affecting milk yield segregating on chromosomes 23 and 16 (again tracked by SNP with posterior means in the fourth distribution of *BayesR*), and these were not segregating in Holstein. This is a subject for further investigation, but these preliminary results suggest that roughly half the QTL explaining 1% of the genetic variance segregate across Jerseys and Holsteins.

An important question, given our results, is whether further increasing marker density (for example, through whole genome sequencing) will lead to more accurate genomic predictions than from the 50K panel. This question can only be answered once sufficient individual cattle genomes have been sequenced. However, a simulation study (Meuwissen and Goddard, 2010) did show that sequence data, where the actual mutation causing trait variation was included in the data set, led to an increase in the accuracy of GEBV of 3 to 5% over the densest marker panel they simulated. Perhaps even more importantly, the authors demonstrated that in their simulation, prediction equations derived from whole-genome sequence data will lead to a slower decrease in the accuracy of GEBV as the reference population and selection candidates are separated by more generations.

This is in contrast to the accuracies of GEBV from the 50K panel in dairy cattle, which decrease rapidly with genetic distance of the target population from the reference population (Habier *et al.*, 2010). A reduced decay in accuracy may also be achieved with the 800K panel. We do not have the data to test this hypothesis. However, if we divide our validation data set into those bulls that do and do not have a sire in the Holstein reference population, and then compare $r(GEBV, DTD)$ for milk yield for these 2 sets from the 50K and 800K panels, a slightly reduced decay in accuracy for the 800K panel compared with the 50K panel, for bulls with and without a sire (Table 10), was only observed when *BayesR* was used to derive the prediction equation. Results were similar for protein yield; however, for fat yield accuracies were actually higher for the group of validation bulls without sires in the reference. This could have been partially an effect of the DGAT1 mutation – closer inspection showed that the SNP tracking this mutation was at more intermediate frequency in the validation bulls with no sires in the reference, compared with those with sires in the reference. Our results here are only suggestive and would not be significant; more investigation of the effect of increasing marker density, with a greater range of relationship to the reference set, on the rate of decay of prediction accuracy is required.

Another potential advantage of using whole-genome resequencing data in prediction of GEBV may be the potential to capture low-frequency mutations that contribute to genetic variation. Allele frequencies of the SNP on the 50K panel are more or less distributed uniformly (i.e., it is a selection where SNP with very low minor allele frequency are underrepresented; e.g., Matukumalli *et al.*, 2009). This is also true for the 800K data (data not shown). For high and stable LD between SNP and QTL, similar allele frequencies of the loci are necessary. Quantitative trait loci with low minor allele frequencies may thus not be in sufficient LD with a SNP and their variance cannot be captured. This may be one explanation why the difference in proportion of unaccounted genetic variance is small between the 50K and the 800K panel (Table 9). Note that for the 800K panel, animals in the reference set were not genotyped themselves, but imputed. Imputation of SNP with low minor allele frequency is more difficult than for SNP with moderate allele frequencies, which can also result in less accurate estimation of SNP effects and, consequently, missing parts of genetic variance. Whether or not resequencing allows some of these low-frequency variants to be captured will depend on how many animals are sequenced before imputation of sequence data in the reference population.

Table 10: Accuracy [$r(GEBV, DTD)$] for milk yield from *BayesR* and *GBLUP_mod* in the Holstein validation set bulls grouped according to whether or not they had a sire in the Holstein reference population[1]

	Method	
Panel[2]	*BayesR*	*GBLUP_mod*
50K with sire	0.64	0.61
50K without sire	0.55	0.56
800K with sire	0.64	0.60
800K without sire	0.57	0.51

[1] DTD = daughter trait deviations; GEBV = genomic EBV; *GBLUP* = genomic BLUP. *BayesR* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions. The *GBLUP_mod* method uses a rescaled genomic relationship matrix, and regresses the $\mathbf{G}$ matrix toward the $\mathbf{A}$ matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers.

[2] 50K = 50,000-SNP panel; 800K = 800,000-SNP panel.

Regarding the 50K panel, several authors have presented studies analyzing real data sets with different methods for the estimation of the SNP effects. In most studies, accuracies achieved with BLUP approaches were very similar to those achieved with Bayesian methods (e.g., VanRaden *et al.*, 2009). For prediction of a breed from a multi-breed reference set, *BayesR* performed best in our study. As described in previous studies (e.g., Hayes *et al.*, 2010), the superiority of Bayesian approaches is generally greater in traits that are strongly influenced by a few moderate to large genes, which was also observed in our study (compare fat to protein). With *GBLUP_mod*, the variance assumed to be explained is the same for each SNP. Therefore, if more and more markers are used in the model, the expected variance per SNP will be smaller. When modeling traits with 1 or more underlying genes with larger effects, this can be the disadvantage when using *GBLUP_mod* in comparison to a Bayesian method (Meuwissen and Goddard, 2010). This theory would lead to the assumption that prediction with *GBLUP* will be even more disadvantageous when even more SNP are modeled simultaneously. In our study, we saw clearly better results with *BayesR* than with *GBLUP_mod* for the traits fat yield and milk yield, for all marker panels. However, we did not observe that the difference in accuracy between the methods was larger for the 800K panel.

There were generally fewer SNP in the third and fourth posterior distributions from the *BayesR* analysis, those with the largest variance, when a combined-breed reference was used compared with single-breed reference sets (Table 8). This may reflect the fact that many SNP are not in the same phase with QTL across breeds. Then, it could be expected that only the SNP having the same LD structure with the QTL in both breeds would have a

moderate effect when the combined reference is used. Pryce *et al.* (2011) found that a more concentrated set of SNP or even a single SNP captured the effect of DGAT1 in a multi-breed reference population compared with pure-breed reference sets. Following the results of *BayesR*, which showed a decreased number of SNP explaining moderate parts of the variance in the multi-breed reference set for all traits, we also investigated the DGAT1 region and did find a decreased number of SNP capturing the DGAT1 effect when a combined reference set was used (Figure 2). Hayes *et al.* (2009a) concluded that a SNP capturing an effect in a multi-breed reference population must be very close to the potential QTL, as they have to be in high LD across breeds. Assuming that the more concentrated set of SNP with moderate effects implies the SNP are closer located to the QTL, the prediction accuracy will be more persistent over generations than with a purebred reference.

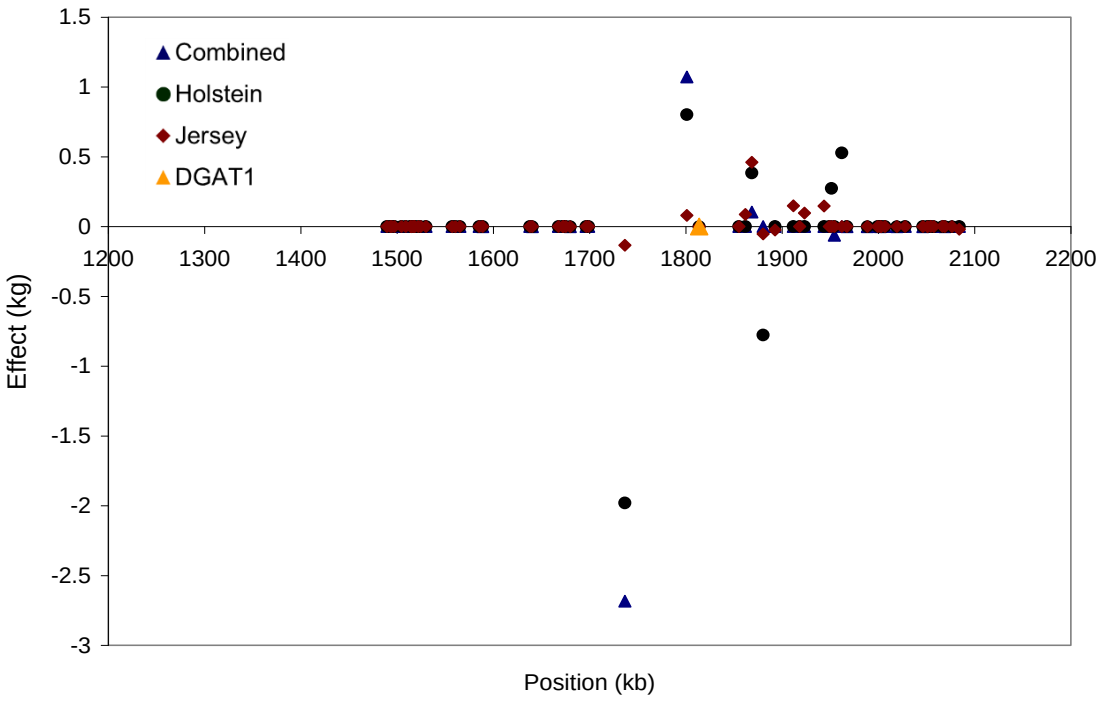


Figure 2: The effect of SNP on fat yield as estimated by a new method (*BayesR*), which used a mixture of normal distributions as the prior for SNP effects, including one distribution that set SNP effects to zero, from different reference populations in the DGAT1 region.

Finally, computer processing times for *BayesR* were reasonable, at 35 h and 20 min for *BayesR* with the multi-breed reference and the 800K panel (Table 11). Using the TRANS panel greatly decreased processing time for all methods, such that this could be applied in national evaluations for dairy cattle. A multi-threaded implementation of the construction of the **G** matrix for a *GBLUP_mod* decreased computing time from several days to 3 min.

Table 11: Processing time (clock time) for multi-breed reference population (2,351 bulls) with 3 SNP panels[1]

	SNP Panel[3]		
Method[2]	50K	800K	TRANS
GBLUP_mod			
Build and invert $\mathbf{G}$	2 min	39 min	3 min
ASREML (1 trait)	20 min	20 min	20 min
BayesA		30 h 55 min	
BayesR	1 h 54 min	35 h 50 min	3h 5 min

[1] Processors were Intel Xeon X5670. For *GBLUP_mod*, multi-threading was used in the construction and inversion of the $\mathbf{G}$ matrix, across 10 threads.

[2] *GBLUP* = genomic BLUP; ASReml = ASReml software (Gilmour *et al.*, 2002). The *GBLUP_mod* method uses a rescaled genomic relationship matrix, and regresses the $\mathbf{G}$ matrix toward the $\mathbf{A}$ matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers; *BayesR* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions; and *BayesA* is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a Student's *t* distribution. Complete descriptions are given in the text.

[3] 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

CONCLUSIONS

In this study, we investigated different marker panels and methods for prediction of genomic breeding values within and across breeds. Two new or modified methods were presented: *GBLUP_mod*, which scales the genomic relationship matrix to an appropriate base and regresses $\mathbf{G}$ toward $\mathbf{A}$ to account for sampling error in estimation of within- and across-breed genomic relationships, and *BayesR*, which assumes that SNP effects follow a mixture of normal distributions, including a distribution with zero variance. Although the *GBLUP_mod* method resulted in less biased breeding values than using an unmodified $\mathbf{G}$ matrix, the *BayesR* method performed best in terms of $r(GEBV, DTD)$ in most studied scenarios, and gave regressions of DTD on GEBV of close to 1. In addition to having the best predictive ability, *BayesR* also presents the possibility of using the results (splitting of SNP into different classes of explained variance) directly for further analyses of, for example, genetic architecture or for SNP selection of less computationally demanding subsets. An additional benefit of the denser marker set of the 800K panel could be seen neither for within- nor for across-breed prediction directly in terms of significant increase of accuracy. However, the 800K panel was the basis for an informative subset of SNP in transcribed parts of the genome, which may be a good alternative to modeling the large number of SNP directly from the 800K panel, balancing the extra genomic information from the 800K with the effect of increased estimation errors from a very large number of SNP in our admittedly small data sets. This

panel (TRANS) in combination with *BayesR* and a combined reference set gave the highest accuracies of prediction in Jerseys, the minor breed in this study.

ACKNOWLEDGMENTS

Parts of the analyses were carried out during a research stay of M. Erbe at the Department of Primary Industries in Victoria, Australia. This research was funded by the German Federal Ministry of Education and Research (Bonn, Germany) within the AgroClustEr “Synbreed–Synergistic plant and animal breeding” (Funding identification: 0315526).

REFERENCES

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. 94:4700–4707.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210–223.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553.

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 179:1503–1512.

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, T. S. Sonstegard, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. C. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Woolliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X.-Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. San Lucas, Q. Fu, M. P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. Barbosa Da Silva, L. P. L. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2009. Genomewide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324:528–532.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2002. ASReml User Guide. Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409–421.

Habier, D., R. L. Fernando, and J. C. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstantinov, and B. J. Hayes. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J. Anim. Breed. Genet. (accepted).

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93:1243–1252.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multibreed dairy cattle populations. Genet. Sel. Evol. 41:51.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6:e1001139.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb.) 91:47–60.

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. 88:544–551.

Marchini, J., and B. Howie. 2010. Genotype imputation for genomewide association studies. Nat. Rev. Genet. 11:499–511.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–631.

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

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655.

Powell, J. E., P. M. Visscher, and M. E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800–805.

Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald, G. C. Waghorn, W. J. Wales, Y. J. Williams, R. J. Spelman, and B. J. Hayes. 2012. Accuracy of genomic predictions of residual feed 14 intake and 250 day bodyweight in 15 growing heifers using 625,000 SNP markers. J. Dairy Sci. 95:2108–2119. http://dx.doi. org/10.3168/jds.2011-4628.

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. 94:2625–2630.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

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.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746.

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. (Camb.) 91:307–311.

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94:3202–3211.

Wray, N. R., M. E. Goddard, and P. M. Visscher. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17:1520–1528.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569.

Zaykin, D. V., A. Pudovkin, and B. S. Weir. 2008. Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180:533–545.

5th CHAPTER

A function accounting for training set size and marker density to model the average accuracy of genomic prediction

M. Erbe[1], B. Gredler[2], F. R. Seefried[2], B. Bapst[2] and H. Simianer[1]

[1] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University, Goettingen, Germany

[2] Qualitas AG, Zug, Switzerland

Accepted in *PLoS ONE*

ABSTRACT

Prediction of genomic breeding values is of major practical relevance in dairy cattle breeding. Deterministic equations have been suggested to predict the accuracy of genomic breeding values in a given design which are based on training set size, reliability of phenotypes and the number of independent chromosome segments (M_e). The aim of our study was to find a general deterministic equation for the average accuracy of genomic breeding values that also accounts for marker density and can be fitted empirically. Two data sets of 5'698 Holstein Friesian bulls genotyped with 50K SNPs and 1'332 Brown Swiss bulls genotyped with 50K SNPs and imputed to ~600K SNPs were available. Different k-fold (k = 2-10, 15, 20) cross-validation scenarios (50 replicates, random assignment) were performed using a genomic BLUP approach. A maximum likelihood approach was used to estimate the parameters of different prediction equations. The highest likelihood was obtained when using a modified form of the deterministic equation of Daetwyler *et al.* (2010), augmented by a weighting factor (w) based on the assumption that the maximum achievable accuracy is $w < 1$. The proportion of genetic variance captured by the complete SNP sets (w^2) was 0.76 to 0.82 for Holstein Friesian and 0.72 to 0.75 for Brown Swiss. When modifying the number of SNPs, w was found to be proportional to the log of the marker density up to a limit which is population and trait specific and was found to be reached with ~20'000 SNPs in the Brown Swiss population studied.

INTRODUCTION

In dairy cattle, prediction of genomic breeding values (GBV) has become a basis for selecting young bulls which are not yet progeny tested. Often, conventional estimated breeding values, daughter yield deviations or deregressed proofs are used as quasi-phenotypes when training genomic prediction models ([1], [2]). The empirical correlation of predicted GBV and the (quasi-)phenotypes used that can be obtained via cross-validation or other empirical validation processes is often used as a measure for the accuracy of prediction (e.g. [2], [3], [4]). However, for selection purposes, we are more interested in the correlation of the predicted GBV and the true breeding value (TBV) which can be approximated from the obtained correlation of GBV and the quasi-phenotype ([5], [6]). In this study, we will refer to the correlation between predicted GBV and TBV ($r_{GBV,TBV}$) as the accuracy of genomic breeding value prediction.

For determining e.g. the required size of the training set or SNP density to achieve a predefined level of accuracy, it would be desirable to be able to assess the expected $r_{GBV,TBV}$ in advance for a GBV prediction study with a given design. Respective deterministic prediction

equations have been suggested ([7], [8], [9], [10]). The approaches have been reported to fit well when applied to a limited number of data points in empirical studies ([10], [11], [12], [13]) and simulated data sets ([9], [10]). In these equations information on the number of animals in the training set, the heritability of the quasi-phenotype used, and the effective number of independently segregating chromosome segments (M_e) are the factors determining the accuracy. Daetwyler et al. [9] showed that the accuracy of the GBV obtained with genomic best linear unbiased prediction (GBLUP) models is independent from the number of underlying QTL. Therefore, this information is not accounted for in the deterministic equations when considering only results from GBLUP approaches. While all approaches referred to so far do not include information on the marker set used, Goddard et al. [10] suggested the number of markers as an additional parameter to account for in the prediction of accuracy.

Derivations of all these deterministic approaches imply that there are no relationship structures between the individuals. Wientjes et al. [13] studied the adaptability of such formulas to different simulation scenarios where selection candidates are related to the reference set in specific manner. They showed that the deterministic equation of [7] as well as the formula of [14] produced similar results for the reliability in comparison with reliabilities obtained with cross-validation also in scenarios where reference and validation individuals were highly related.

The number of independently segregating chromosome segments M_e is a population parameter and is usually estimated based on assumptions of the effective population size (N_e) and the genetic length of the genome in Morgan (L). Different formulas ([8], [10], [15]) on how to determine M_e based on theoretical considerations lead to quite different M_e, which has a major impact on the results of the deterministic prediction of the accuracy. Another possibility is to define the number of independent chromosome segments to be the reciprocal of the variance of the difference of the genomic relationship matrix and the numerator relationship matrix when complex family structures are in the data set ([10], [13]).

By using empirical accuracies obtained via cross-validation in a genomic prediction with real or simulated data, it is possible to determine M_e by rearranging the equation used for predicting of accuracy. With different levels of training set size this may lead to different estimates of M_e (see e.g. [9] with simulated data). Being a population parameter, M_e should have a constant value within one data set independently of the size of the training set used for its estimation, though. Daetwyler [16] proposed using a regression approach for overcoming this problem.

In our study, we suggest determining M_e empirically based on a systematic multi-level cross-validation using a maximum likelihood approach and based on this, we will compare various

deterministic prediction equations. We suggest a modified form of the deterministic prediction equation of [9] with the maximum accuracy that can be obtained with the given marker set as a further parameter, which will be shown to be a function of the natural logarithm of the marker density. All equations will be compared using two dairy cattle data sets of relevant size, and possible practical implications will be discussed.

MATERIAL AND METHODS

Data Sets

To establish and test the methodology, we used a sample of 5'698 Holstein bulls, which were genotyped with the Illumina BovineSNP50 BeadChip. Single nucleotide polymorphisms (SNPs) with a minor allele frequency lower than 1%, with missing or non-autosomal position or a call rate lower than 95% were excluded. After filtering, there were 42'551 SNPs remaining for further analyses. Missing genotypes at these SNP positions were imputed using Beagle 3.2 ([17]). To study the influence of different marker densities, we reduced the number of markers to subsets of 30'000, 20'000, or 10'000, respectively. Markers in the subsets were chosen at random from the complete set.

All bulls used for this study had estimated breeding values based on progeny testing for somatic cell score and milk yield with an accuracy > 0.84, which were used as quasi-phenotypes for the following analyses.

To test the proposed approach in a further data set and with different SNP marker density, we used a set of 1'332 Brown Swiss bulls which was partly genotyped with the Illumina BovineSNP50 BeadChip and partly with the Illumina BovineHD BeadChip with around 777K. For the Brown Swiss bulls genotyped with the Illumina BovineSNP50 BeadChip, genotypes have been imputed to the Illumina BovineHD BeadChip based on a reference set of 727 Brown Swiss cows and 153 bulls using a combination of family and population imputation implemented in the software FImpute ([18]). After quality control, there were 627'306 SNPs available for further analyses. To study different marker densities, the set of markers was also decreased by using each 2^x-th marker where x was 1, 2, …, 8.

Genotype and phenotype data is available from the authors on request.

Cross-validation strategy

Cross-validation was performed in different k-fold scenarios with k = 2, 3, … , 10, 15 or 20. This resulted in different sizes of training sets with different values of k. With a k-fold cross-

validation, k-1 folds are used to predict the remaining fold and this procedure is repeated so that each fold is predicted once. Animals were assigned to the folds randomly. For the evaluation of the GBV prediction, the correlation $r_{GBV,TBV}$ between predicted GBV and TBV was used, which was calculated as $r_{GBV,TBV} = \frac{r_{GBV,EBV}}{r_{EBV,TBV}}$ (e.g. [6]), where $r_{EBV,TBV}$ is the accuracy of the estimated breeding values, which we used as quasi-phenotypes. $r_{GBV,EBV}$ was calculated for each GBV prediction in a fold and then averaged over the k folds within a k-fold scenario. Each k-fold scenario was replicated 50 times, so that there were 50 values of $r_{GBV,EBV}$ available for each k-fold scenario for further analyses.

Genomic BLUP:

Genomic breeding values were predicted using genomic best linear unbiased prediction (GBLUP) based on the model

$$\mathbf{y} = \mathbf{1}'_{n_t}\mu + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of quasi-phenotypes (in our case estimated breeding values of milk yield or somatic cell score, respectively) for all bulls in the training set, $\mathbf{1}'_{n_t}$ is a column vector of ones of length number of animals in the training set (n_t), μ is the overall mean, $\mathbf{Z}$ is the incidence matrix for the random genomic effect, $\mathbf{u}$ is a vector containing the random genomic effect (i.e. the genomic breeding value) for all animals and $\mathbf{e}$ is a vector of random error terms. $\mathbf{u}$ is assumed to be distributed $N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\mathbf{e}$ is assumed to follow $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. $\mathbf{G}$ is the genomic relationship matrix following [14]. Since we wanted to study the effect of different number of markers, we built $\mathbf{G}$ based on different SNP sets. For the basic scenario, we used all SNPs available after quality control (i.e. 42'551 SNPs for the Holstein Friesian data set and 627'306 for the Brown Swiss data set) while for the further scenarios $\mathbf{G}$ was based on a subset of the total available number of SNPs, namely on 30'000, 20'000 and 10'000 SNPs for the Holstein Friesian and 313'653, 156'827, 78'414, 39'207, 19'604, 9'802, 4'901 and 2'451 SNPs for the Brown Swiss data set, respectively. Variance components were estimated once with the respective complete data set in combination with a specific SNP set using ASReml 3.0 ([19]) and were then used for all respective runs using a subset of these data, but the same SNP set.

In the following, we will describe available deterministic equations for prediction of the level of accuracy from the literature and modifications of these formulas we will conduct:

Equation of Daetwyler et al.

Daetwyler et al. [9] presented an equation (D1) to predict the accuracy of a genome-wide genomic breeding value prediction:

$$r_{GBV,TBV\_D1} = \sqrt{\frac{n_t h^2}{n_t h^2 + M_e}} \tag{D1}$$

where n_t is the number of animals in the training set, h^2 is the heritability of the observed trait and M_e is the number of independently segregating chromosome segments. When estimated breeding values (EBV) from a conventional breeding value estimation scheme are used as quasi-phenotypes for genomic prediction, h^2 can be replaced by the reliability of the EBV. This is also true for all further prediction equations that will be described later. Daetwyler et al. [9] suggested using the definition of [8] to calculate M_e, but we will take M_e as a parameter not further determined in our study.

Equation of Goddard et al.

Goddard et al. [10] proposed a new equation for predicting the reliability of genomic prediction which also accounts for the number of markers used. The basic formula in this paper is

$$r^2_{GBV,TBV\_G1} = b\frac{\theta}{\theta + 1}$$

where

$$b = \frac{n_{SNPs}}{n_{SNPs} + M_e}$$

and

$$\theta = \frac{n_t b h^2}{M_e}$$

Goddard et al. [10] proposed a slightly different definition of M_e than [8] but we will not use any of them but keep M_e again as a population parameter to be determined empirically. Using those definitions, the prediction formula for the accuracy can be expressed as

$$r_{GBV,TBV\_G1} = \sqrt{\frac{b^2 n_t h^2}{b n_t h^2 + M_e}} = \sqrt{b}\sqrt{\frac{n_t h^2}{n_t h^2 + \frac{M_e}{b}}} \tag{G1}$$

which is very similar to the one proposed by [9] but with the variable b included to account for the finite number of markers. Note that if $b \to 1$, i.e. for a large number of SNPs and a limited number of M_e, D1 and G1 become identical. Goddard et al. [10] suggested using also a correction factor due to a smaller error variance when using a multiple marker analysis rather than single marker analyses. They refer to [9] and present the optimal prediction equation (G2) for predicting the accuracy as

$$r_{GBV,TBV\_G2} = \sqrt{\left(b\frac{\theta}{\theta+1}\right)\left(1+\frac{\left(b\frac{\theta}{\theta+1}\right)^2 h^2}{2\theta}\right)} \qquad \text{(G2)}$$

Modification of Daetwyler's equation

Assuming a finite M_e D1 will asymptotically approach 1 when $n \to \infty$. Daetwyler [16] stated in the general discussion of his PhD thesis that it may be useful to modify his prediction equation to deal with the fact that the marker density of the Illumina BovineSNP50 BeadChip might not be high enough to capture all genetic variation.

According to [20] the accuracy of the GBV as a predictor of the true breeding value component that is associated with the available marker set is a product of the square root of the proportion of genetic variance associated with the used marker set (w) and the accuracy of genomic breeding values assuming all causal variants are known and considered so that

$$r_{GBV,TBV|M} = w \cdot r_{GBV,TBV}$$

The factor $0 \leq w \leq 1$ can be interpreted as the maximum accuracy that can be obtained with a specific SNP set when the size of the training set is infinite. Using this in model D1 leads to the modified equation (D2) of [9]

$$r_{GBV,TBV\_D2} = w \cdot \sqrt{\frac{n_t h^2}{n_t h^2 + M_e}} \qquad \text{(D2)}$$

Modification of Goddard's equation

Equations G1 and G2 of [10] include also a weighting factor which accounts for the fact that not all genetic variance can be captured if the number of markers is limited. The authors of [10] defined this factor using the number of SNPs and the number of M_e but this may not be the optimal factor. We thus wanted to study the results of prediction when using G2 in a

modified form by setting $\sqrt{b}$ equal to our w and avoiding any further definition of b. This leads to prediction equation G3 defined as

$$r_{GBV,TBV\_G3} = \sqrt{\left(w^2\frac{\tilde{\theta}}{\tilde{\theta}+1}\right)\left(1+\frac{\left(w^2\frac{\tilde{\theta}}{\tilde{\theta}+1}\right)^2 h^2}{2\tilde{\theta}}\right)} \qquad \text{(G3)}$$

with

$$\tilde{\theta} = \frac{n_t w^2 h^2}{M_e}$$

Maximum Likelihood approach

A maximum likelihood approach was used to determine the value of M_e in equations D1, G1 and G2, or the combination of w and M_e in equations D2 and G3 that provide the best fit of the respective model to our cross-validated data over all different training set sizes. We determined the most appropriate estimate of M_e or w and M_e, respectively, by maximizing the Likelihood function

$$L = \prod_{i=1}^{n_{fold}} \prod_{j=1}^{n_{rep}} f(x_{ij})$$

where n_{fold} is the number of different k-fold scenarios, n_{rep} is the number of replicates within one scenario and x_{ij} is the mean accuracy of prediction obtained by cross-validation in the i^{th} scenario in the j^{th} replicate. We assumed that $\mathbf{x}_i$ was approximately normal distributed with

$$\mathbf{x}_i \sim N\left(E(\mathbf{x}_i), \sigma_i^2\right)$$

and observations were independent. $E(x_i)$ was derived from the respective model to predict the accuracy (i.e. D1-D2, G1-G3, respectively) and σ_i^2 was assumed to be the empirical variance in the 50 observed values within the i^{th} scenario. To ensure that the assumption of correlation coefficients being normally distributed random variables is not violated we tested all k-fold results with the 42'551 SNPs in the Holstein Friesian data set with a Shapiro-Wilk test [21].

Most of the parameters used in $E(x_i)$ were determined by the empirical data, namely the heritability, number of animals in the training set and number of markers. Therefore, M_e and

w remain the only parameters in all considered equations to be adjusted. Searching for the maximal likelihood was done using the function "optimize" in R [22] for a one-dimensional search (i.e. for M_e in equations D1, G1, and G2) and the function "optim" in R [22] for a two-dimensional search (i.e. for M_e and w in D2 and G3).

Predicting prediction accuracies

In many applications the prediction accuracy obtained with the data, especially the training set size, at hand is not sufficient. In such cases it would be desirable to be able to determine accurately the required training set size to achieve a pre-defined level of accuracy of genomic prediction. We tried to mimic this exercise to compare the usefulness of a model accounting for the fact that the finite marker set does not account for the full genetic variation (model D2) with that of a model not doing so (model D1). We used subsets of 4'000 Holstein-Friesian bulls to derive the optimal number of M_e (in D1) or M_e and w (in D2) and then predicted the accuracies for a training set in the size of the training set used for the 20-fold cross-validation runs with the whole Holstein Friesian data set (i.e. 5'413 bulls). For this we chose 4'000 bulls randomly out of the whole sample and performed a variance component estimation and all k-fold cross-validation runs (k = 2-10, 15, 20) for the different subsets. Afterwards, we maximized the likelihood as described above. Since there may be a sampling effect when using a random subset of 4'000 bulls, we repeated the whole procedure ten times so that we had predictions for ten different subsets of 4'000 bulls. The range of predicted values for a training set size of 5'413 bulls then was compared with the empirical accuracy from a 20-fold cross-validation with our whole data set, i.e. with a training set size of 5'413 bulls.

RESULTS

The mean and standard errors of the empirical accuracies obtained from the different cross-validation schemes in the Holstein Friesian data are displayed in Figures 1 and 2 for the traits milk yield and somatic cell score. The mean accuracies (± standard errors) ranged from 0.743 ± 0.0005 (0.73 ± 0.0007) with a 2-fold cross-validation and training set size 2'849 to 0.798 ± 0.0002 (0.808 ± 0.0002) with a 20-fold cross-validation and training set size 5'413 for milk yield (somatic cell score).

Our observed accuracies were far away from the bounds of correlation coefficients (-1 and 1) and apparently normally distributed: The results of the Shapiro Wilk test showed that for all k-fold results with 42'551 SNPs in the Holstein Friesian data set the null hypothesis "normally distributed" was not rejected in a single case with $p<0.01$. Therefore, no further transfor-

mation of the data was necessary. Other approaches, like the least squares principle used by [12] to fit model D1 to sequence-based genomic predictions in *Drosophila melanogaster*, can also be used to estimate the model parameters and in our case would lead to very similar results (results not shown).

In the following, we will first describe the results for the estimates of M_e obtained based on the original equations from the literature [9,10] and then based on the modified versions of these equations (i.e. with w added) with different numbers of markers.

Table 1 shows the numbers of M_e obtained by maximizing the likelihood of the empirical accuracies under equations D1, G1 and G2 for both traits. The estimates of M_e were of the same magnitude (~ between 2'000 and 2'800) with all methods while the likelihood obtained with G1 is highest for both traits. Not surprisingly, the estimates were similar for both traits since the empirical accuracies for milk yield and somatic cell score were very similar.

Table 1: Fitted values of the number of independent chromosome segments (M_e) and weighting factors (w) with the Maximum-Likelihood approach and the corresponding natural logarithm of the likelihoods when using the Holstein-Friesian data set.

Method[1]	Trait	M_e fitted	w	% genetic variance captured	Ln(Likelihood)
D1	Milk yield	2783.2	-	-	-3912.5
D2	Milk yield	1045.6	0.875	76.6	2613.1
G1	Milk yield	2282.4	-	-	-1903.9
G2	Milk yield	2821.9	-	-	-4367.6
G3	Milk yield	904.9	0.869	75.5	2611.0
D1	Somatic cell score	2442.3	-	-	495.5
D2	Somatic cell score	1241.0	0.907	82.3	2512.9
G1	Somatic cell score	2036.2	-	-	1272.7
G2	Somatic cell score	2506.0	-	-	340.2
G3	Somatic cell score	1128.4	0.897	80.5	2508.7

[1] D1 uses the formula of Daetwyler et al. (2010) to calculate the expected values of accuracy, G1 and G2 are based on Goddard et al. (2011) without and with the proposed correction factor, respectively. D2 is a modified equation of Daetwyler et al. (2010) while G3 is based on Goddard et al. (2011) with the weighting factor not defined like in the original publication but like in D2.

In Figure 1 and 2, the best curves of predicted accuracy under equations D1, G1 and G2 based on the respective maximum likelihood estimates of M_e in the Holstein Friesian data set

are shown for the traits milk yield and somatic cell score. None of these equations provided a curve of predicted accuracies that matched the empirical data to a sufficient extent. The results obtained under equations D1 and G2 are very similar while G1 provided a slightly better fit in accordance with the superior likelihood value for this model. Nevertheless, all equations led to a downward bias of predicted accuracies for small training set sizes while they showed an upward bias for large training set sizes.

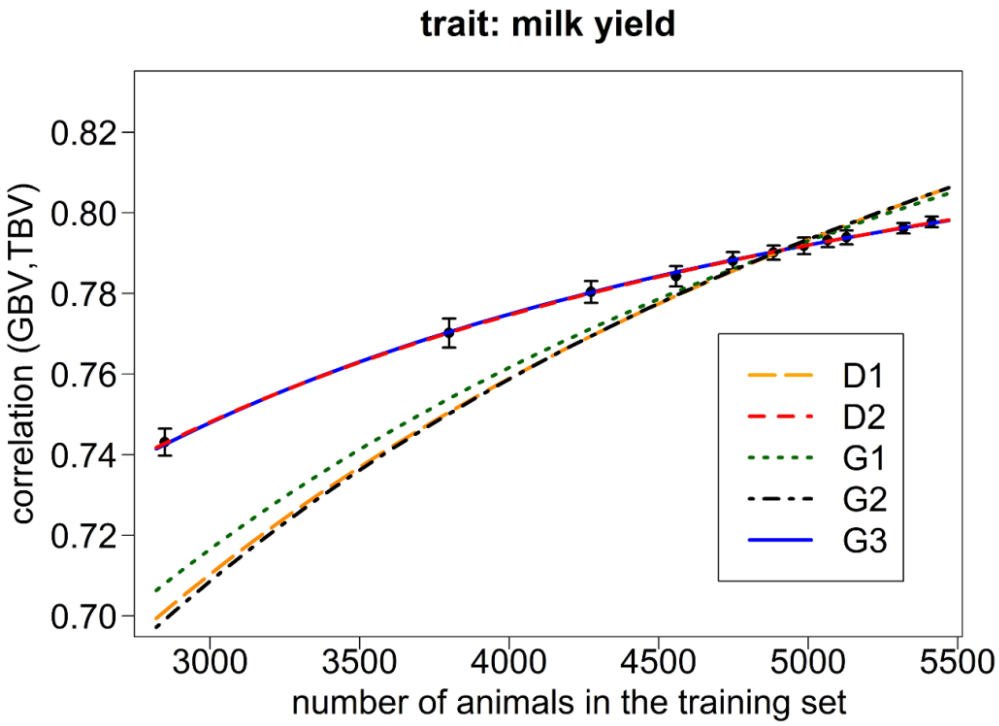


Figure 1: Empirical values and expected values of $r_{GBV,TBV}$ for milk yield in Holstein-Friesian data.

Empirical values of $r_{GBV,TBV}$ and expected values using the number of M_e derived with a Maximum-Likelihood approach for the Holstein-Friesian data set in the original equation of Daetwyler et al. (2010) (D1) as well as in a modified form (D2) and in the equation of Goddard et al. (2011) without (G1) and with (G2) the proposed correction factor, respectively, and with the factor b not further determined (G3). For the empirical values, the mean and the standard deviation over the 50 replicates in each k-fold scenario of the Holstein-Friesian data set are shown.

Maximum likelihood estimates for w and M_e for the Holstein data set with the new equations D2 and G3 used for the calculation of the expectations of the accuracy are also presented in Table 1. The obtained likelihoods were dramatically higher compared to the conventional equations, with D2 slightly outperforming G3 with the present data sets. The estimates of M_e were clearly lower with both equations compared with the original equations and were in the range of ~900 to ~1'240 depending on method and trait. The optimal weighting factor w was

in all cases between 0.87 and 0.91, suggesting that with the given marker set the accuracy of prediction will not approach 1 even if a very large training set is used. According to Dekkers (2007) the squared value of w represents the proportion of genetic variance associated with the markers which in our case would range between 75.5 per cent (milk yield with model G3) and 82.3 per cent (somatic cell score with model D2). This indicates that a large proportion, but not the complete genetic variation in our data set is captured by the SNP set at hand.

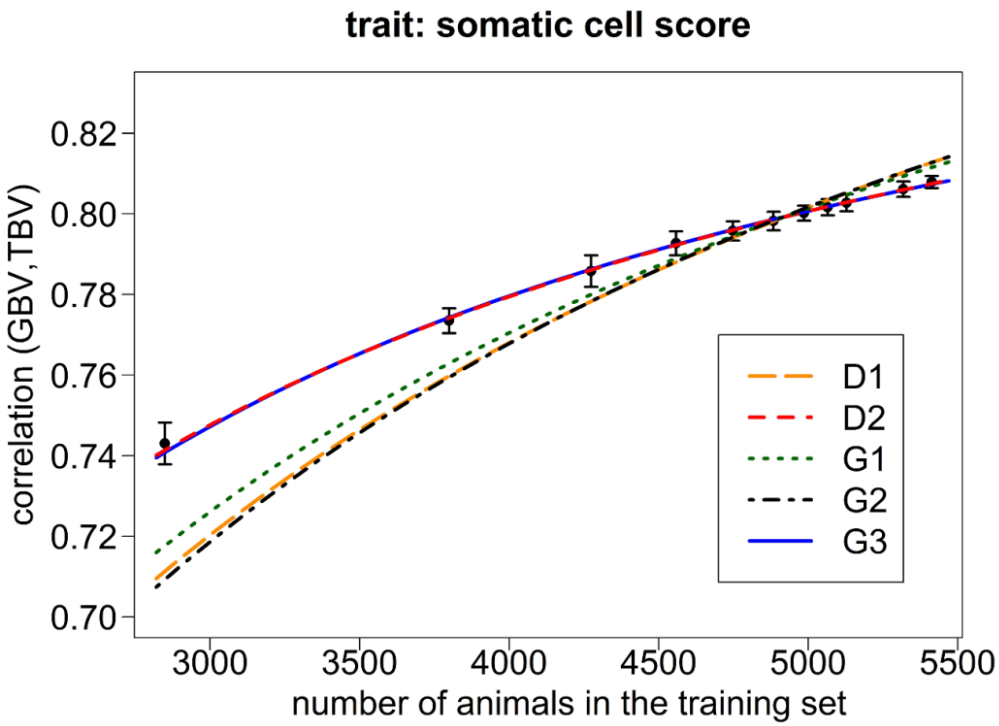


Figure 2: Empirical values and expected values of $r_{GBV,TBV}$ for somatic cell score in Holstein-Friesian data.

Empirical values of $r_{GBV,TBV}$ and expected values using the number of M_e derived with a Maximum-Likelihood approach for the Holstein-Friesian data set in the original equation of Daetwyler et al. (2010) (D1) as well as in a modified form (D2) and in the equation of Goddard et al. (2011) without (G1) and with (G2) the proposed correction factor, respectively, and with the factor b not further determined (G3). For the empirical values, the mean and the standard deviation over the 50 replicates in each k-fold scenario of the Holstein-Friesian data set are shown.

Figures 1 and 2 show prediction curves resulting from the optimal fit of the equations D2 and G3 for the traits milk yield (Fig. 1) and somatic cell score (Fig. 2) within the Holstein Friesian data set. For both traits and with both equations, the predicted accuracies fit the empirical data extremely well and in any case much better than with the conventional equations. By fitting two parameters (M_e and w) the curves could accommodate a different slope of the

empirical accuracy values more flexibly than with the one-parameter equations, which are bound to have their origin in $r_{GBV,TBV} = 0$ and asymptotically have to approach $r_{GBV,TBV} \to 1$.

Since we observed that only a specific fraction of the genetic variance was captured by the available SNP set we were interested in studying the effect of different SNP densities on the shape of the curve of expected accuracies and the respective parameters. Results of the maximum likelihood estimation using equations D2 and G3 with different marker set sizes in the Holstein Friesian data set are given in Table 2.

Table 2: Fitted values of the number of independent chromosome segments (M_e) and weighting factors (w) with the Maximum-Likelihood approach and the corresponding natural logarithm of the likelihoods for different methods and different SNP sets when using the Holstein-Friesian data set.

Method[1]	Trait	No. of SNPs	M_e fitted	w	% genetic variance captured	Ln(Likelih.)
D2	Milk yield	10000	992.3	0.844	71.2	2576.4
D2	Milk yield	20000	1043.9	0.863	74.5	2600.0
D2	Milk yield	30000	1068.6	0.868	75.3	2594.4
D2	Milk yield	42551	1045.6	0.875	76.6	2613.1
G3	Milk yield	10000	791.6	0.838	70.2	2574.2
G3	Milk yield	20000	874.1	0.856	73.3	2597.2
G3	Milk yield	30000	904.1	0.861	74.1	2491.9
G3	Milk yield	42551	904.9	0.868	75.3	2611.0
D2	Somatic Cell Score	10000	1201.3	0.868	75.3	2457.8
D2	Somatic Cell Score	20000	1240.1	0.895	80.1	2496.0
D2	Somatic Cell Score	30000	1250.8	0.904	81.7	2512.3
D2	Somatic Cell Score	42551	1241.0	0.907	82.3	2512.9
G3	Somatic Cell Score	10000	993.5	0.861	74.1	2456.0
G3	Somatic Cell Score	20000	1093.3	0.885	78.3	2491.9
G3	Somatic Cell Score	30000	1127.0	0.894	80.0	2508.1
G3	Somatic Cell Score	42551	1128.4	0.897	80.4	2508.7

[1] D2 is a modified equation of Daetwyler et al. (2010) while G3 is based on Goddard et al. (2011) with the weighting factor not defined like in the original publication but like in D2.

We observed a decreasing trend in the weighting factor w when reducing the number of SNPs but the extent of the decrease was limited, so that even with 10'000 SNPs a high per-

centage of the genetic variance (71.2% for milk yield and 75.3% for somatic cell score, both with model D2) is captured and not much is gained by applying a more than four-fold SNP density. For the optimal number of M_e the trend was not that clear. It was also not expected that the number of M_e changes systematically in one direction since the same animals were used for all analyses. The likelihoods were in the same range for all reduced SNP sets compared to the full SNP set for both methods.

Based on our previous results, we next tried to describe the relationship between the estimates of w obtained and the underlying marker density.

We hypothesize that the maximum accuracy that can be obtained, w, is a function of the natural logarithm of the SNP density. Using the Holstein Friesian data, we found that a function

$$w = a + z\frac{1}{\ln(\#SNPs/L)} \qquad \text{Eq. [1]}$$

where $\#SNPs/L$ is the number of SNPs per Morgan, fitted our empirical data reasonably well (Figure 3). With an intercept of $a = 1.001\ (1.071)$ and a regression coefficient of $z = -0.914\ (-1.173)$ for milk yield (somatic cell score), the coefficient of determination of the fitted regression line was 0.990 (0.971), and the regression coefficients were significant $(p < 0.05)$ for both traits. Note that we had only four data points available, but nevertheless they showed a very clear trend. An intercept of approximately 1 could suggest that with an increasing SNP density (i.e. decreasing values of the reciprocal of the natural logarithm of the SNP density) the accuracy of genomic prediction asymptotically approaches 1. This result also suggested that it will be necessary to use multi-folds of a given marker density to obtain a substantial increase of the prediction accuracy.

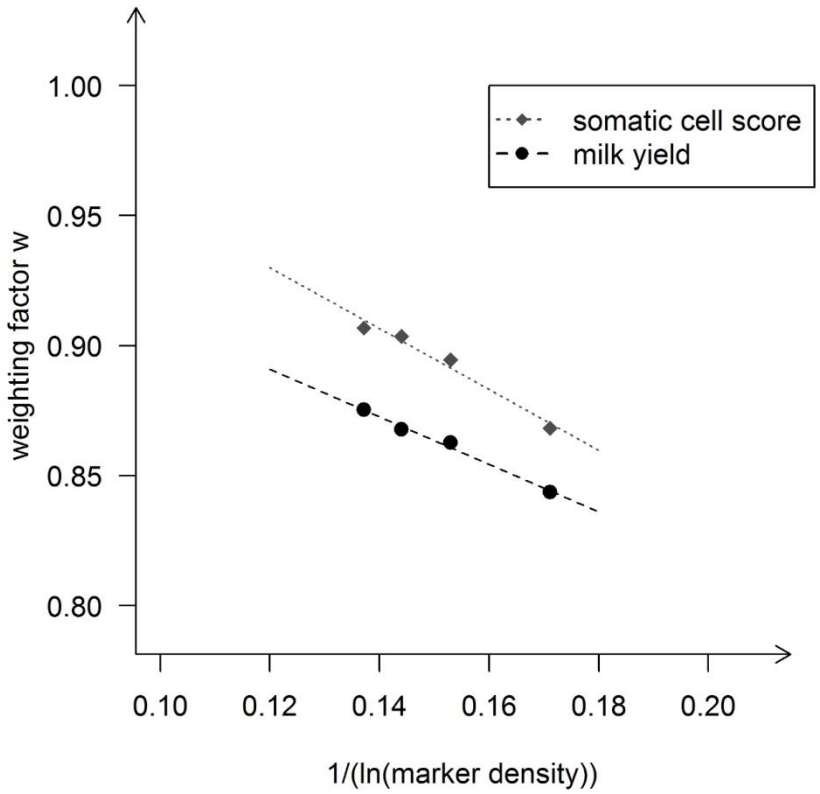


Figure 3: Regression of weighting factor w on the reciprocal of the logarithm of the marker density in Holstein-Friesian.

Regression of the weighting factor w on the reciprocal of the natural logarithm of the marker density for the traits milk yield and somatic cell score in the Holstein-Friesian data set. The marker density was defined as the number of markers used divided by the length of the used parts of the genome in Morgan. The dots mark the values derived with the Maximum likelihood approach using the modified equation of Daetwyler et al. (2010) (D2) to describe the expected value of accuracy and the empirical data sets.

As we had cross-validation results based on different marker densities available, we were also interested in finding a global function for estimating M_e and a weighting factor including all available empirical results. Eq. [1] made it possible to find a global M_e and a factor z depending on the marker density using our suggested maximum likelihood approach. We used D2 for the expected value with $w = \left(1 - z\frac{1}{\ln(\#SNPs/L)}\right)$ and found the highest likelihood with $M_e = 1'151.55$ and $z = 0.853$. A comparison between predicted and empirical values is shown in Figure 4. It can clearly be seen that the empirical values deviate only slightly from the predicted values. Deviations are largest for small training set sizes and/or low marker densities.

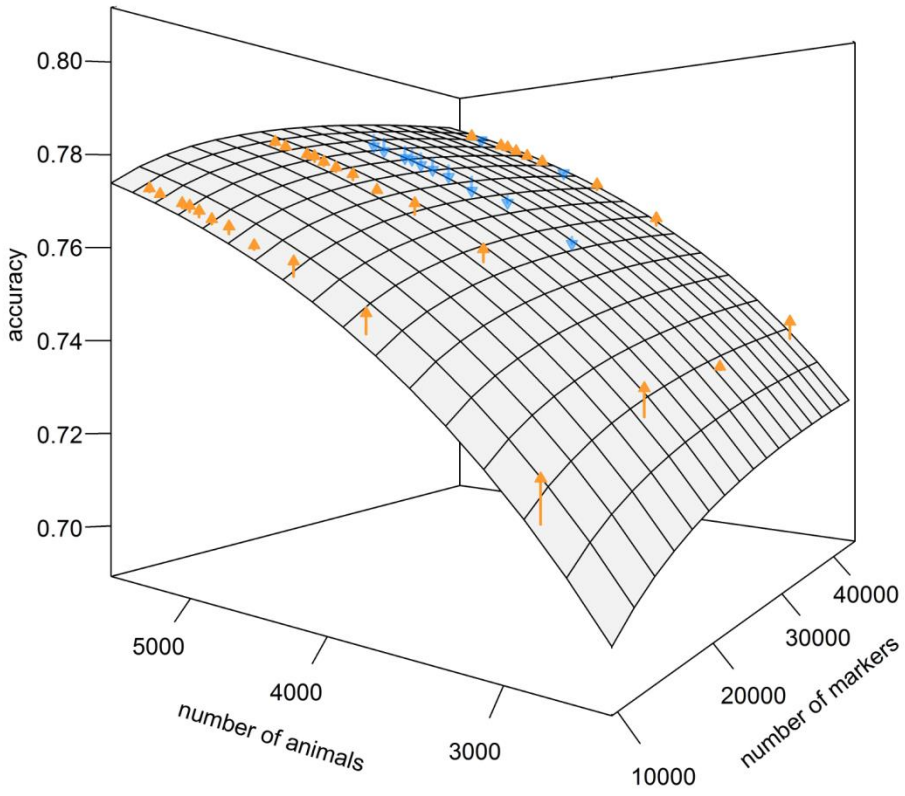


Figure 4: Predicted and empirical values of $r_{GBV,TBV}$ (grid) for different #SNPs and different #animals for Holstein-Friesian.

Predicted values of $r_{GBV,TBV}$ (grid) for different numbers of markers and different number of animals in the training set when using the modified equation of Daetwyler et al. (2010) (D2), an M_e of 1'151.55 and a weighting factor of $(1-0.853/\ln(\#SNPS/L))$. Empirical results obtained with cross-validation experiments with Holstein-Friesian data are symbolized by arrows. Orange arrows represent values that were higher than predicted while blue arrows indicate that empirical values were lower than the predicted ones.

To check the results in an independent data set, we applied the maximum likelihood approach based on D2 also on the Brown Swiss data set. Empirical values from the 2- to 20-fold cross-validation when using the full SNP set are shown in Figure 5. Mean accuracies (± standard errors) ranged from 0.757 ± 0.0013 (0.659 ± 0.0015) with a 2-fold cross-validation and training set size 667 to 0.802 ± 0.0006 (0.730 ± 0.0007) with a 20-fold cross-validation and training set size 1266 for milk yield (somatic cell score).

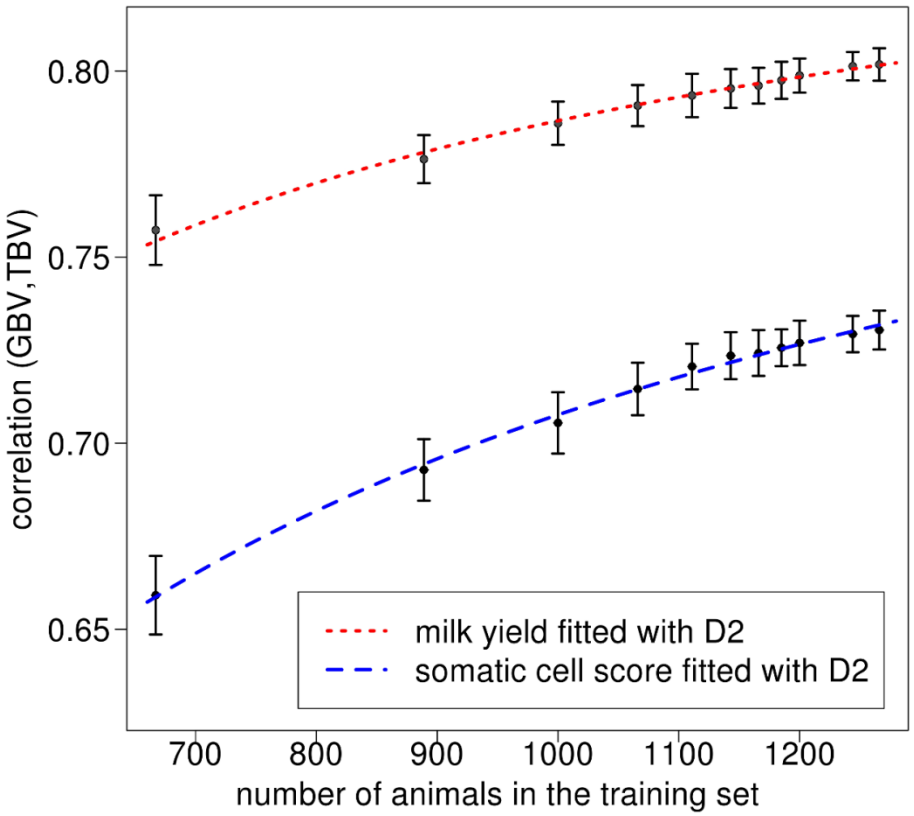


Figure 5: Empirical values and expected values of $r_{GBV,TBV}$ for milk yield and somatic cell score in Brown-Swiss.

Empirical values of $r_{GBV,TBV}$ and expected values using the number of M_e for the Brown Swiss data set derived with a Maximum-Likelihood approach in the modified equation of Daetwyler et al. (2010) (D2). For the empirical values of milk yield and somatic cell score in the Brown Swiss data set, the mean and the standard deviation over the 50 replicates in each k-fold scenario are shown.

Results of the estimation of the number of M_e and w with different SNP sets can be seen in Table 3. Estimates for the number of M_e ranged from 148 to 214 for milk yield and from 277 to 419 for somatic cell score and were thus clearly lower in Brown Swiss than in Holstein Friesian. Estimates of the number of M_e were smaller with milk yield than with somatic cell score as was also observed in the Holstein Friesian data set. The weighting factor w kept constant in both traits (~0.87 for milk yield, ~ 0.85 for somatic cell sore) when decreasing the number of markers up to a point of around 19'000 SNP from where on it decreased considerably. This indicates that the percentage of genetic variance captured with a given SNP set did not increase further when using more than 19'000 SNPs in this data set. Figure 5 shows the prediction curves with the optimized number of M_e and an optimized w as well as D2 for modeling the expected accuracy for both traits and the full SNP set of 627'306 SNPs. As already seen with the Holstein Friesian data, D2 with optimized values for the number of M_e and w fitted the shape of the curve of empirical values very well.

Table 3: Fitted values of the number of independent chromosome segments (M_e) and weighting factors (w) with the Maximum-Likelihood approach and the corresponding natural logarithm of the likelihoods for method D2 and different SNP sets when using the Brown Swiss data set.

Trait	No. of SNPs	M_e fitted	w	% genetic variance captured	Ln(Likelih.)
Milk yield	2451	148.2	0.791	62.6	2111.2
Milk yield	4901	157.2	0.821	67.4	2108.2
Milk yield	9802	192.2	0.849	72.1	2078.3
Milk yield	19604	213.7	0.868	75.3	2075.8
Milk yield	39207	202.2	0.868	75.3	2085.4
Milk yield	78414	199.4	0.868	75.3	2090.9
Milk yield	156827	197.3	0.868	75.3	2095.2
Milk yield	313653	196.5	0.867	75.2	2094.0
Milk yield	627306	196.7	0.866	75.0	2092.2
Somatic Cell Score	2451	277.2	0.735	54.0	1904.7
Somatic Cell Score	4901	354.2	0.792	62.7	1910.0
Somatic Cell Score	9802	378.4	0.824	67.9	1971.9
Somatic Cell Score	19604	418.9	0.850	72.3	1979.7
Somatic Cell Score	39207	405.0	0.845	71.4	1978.6
Somatic Cell Score	78414	411.6	0.849	72.1	1983.0
Somatic Cell Score	156827	414.2	0.850	72.3	1981.4
Somatic Cell Score	313653	412.4	0.850	72.3	1982.0
Somatic Cell Score	627306	412.4	0.851	72.4	1983.9

We also tested the relationship between the weighting factor w and the marker density for the Brown Swiss data set (same approach like in the Holstein Friesian data set). The results are shown in Figure 6. There seems to be a linear relationship up to a number of markers of around 20'000 SNPs (~0.16 when expressed as $\frac{1}{\ln(\#SNPs/L)}$). A linear regression model $w = a + z(\ln(\#SNPs/L))^{-1}$ with $a = 1.03$ and $z = -1.08$ would lead to a coefficient of determination R^2 of 0.998 for milk yield, for example. However, with any further increase of the marker density (i.e. smaller values on the x-axis), the weighting factor did not increase anymore but stayed on a constant level w_{max} (e.g. w_{max}~0.87 for milk yield). This pattern with a linear relationship first and constant values beyond a certain marker density was observed in both traits.

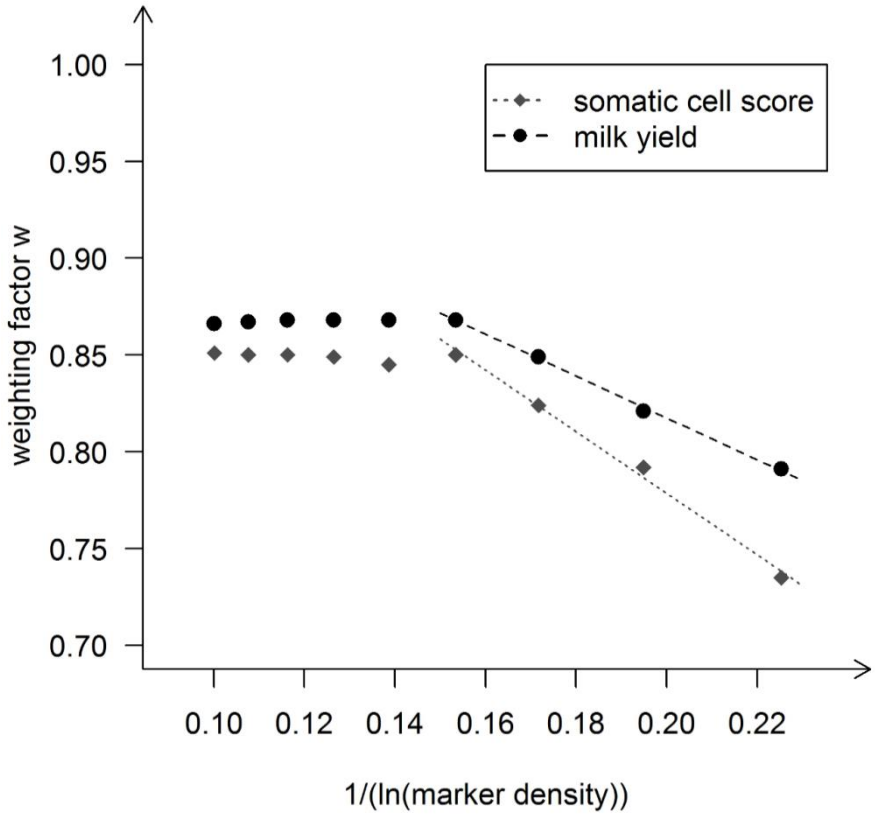


Figure 6: Regression of weighting factor w on the reciprocal of the logarithm of the marker density in Brown-Swiss.

Regression of the weighting factor w on the reciprocal of the natural logarithm of the marker density for the traits milk yield and somatic cell score in the Brown Swiss data set. The marker density was defined as the number of markers used divided by length of the used parts of the genome in Morgan. The dots mark the values derived with the Maximum likelihood approach using the modified equation of Daetwyler et al. (2010) (D2) to describe the expected value of accuracy and the empirical data sets.

Next we studied if our approach can be used to extrapolate the accuracy of prediction beyond the data set used to determine the model parameters. For this, the maximum likelihood approach was applied to ten data sets of 4'000 Holstein Friesian bulls which were the basis of cross-validation runs as described above. Figure 7A displays the resulting prediction curves obtained with model D1 for the trait somatic cell score, for milk yield the picture was very similar (results not shown). The curves varied over the data sets but were reasonably consistent in the level of accuracy and its slope over the different sizes of training sets. The fitted number of M_e ranged between 2'000 and 2'356. When extrapolated to the training set size 5'413 (the one resulting from a 20-fold CV in the full data set), the expected prediction accuracy (averaged over the 10 replicates) was 0.828 ± 0.007. The empirical accuracy obtained from the full data set (0.808±0.002) was clearly outside the range of predicted values obtained with the ten replicates. This suggests that model D1 (and similarly G1 and G2, re-

sults not shown) systematically overestimate the expected prediction accuracy when used for extrapolation beyond the training set size at hand.

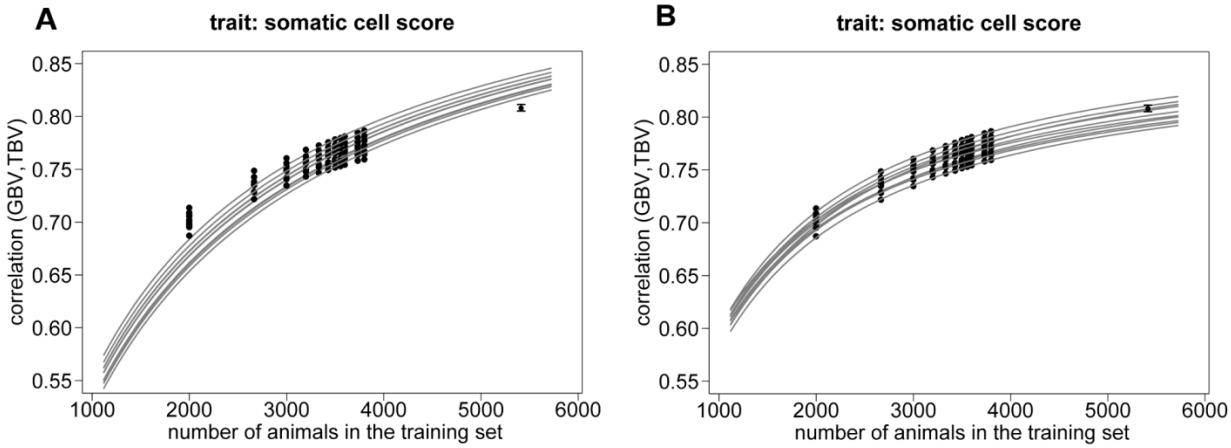


Figure 7A and B: Predicted and empirical values of $r_{GBV,TBV}$ when extrapolating the accuracy.

Empirical values of $r_{GBV,TBV}$ (black dots) of the ten replicates with different k-fold scenarios using 4'000 individuals and of the 20-fold runs of the fifty replicates using 5'698 Holstein-Friesian animals in total. Expected values (grey lines) use the number of M_e derived with a Maximum-Likelihood approach in the original equation of Daetwyler et al. (2010) (D1, Figure 7A) and in the modified equation of Daetwyler et al. (2010) (D2, Figure 7B).

In Figure 7B, D2 was used in the maximum likelihood approach to determine the optimal parameter for the prediction of accuracy based on the cross-validation runs with the ten different data sets of 4'000 bulls for the trait somatic cell score. The optimal weighting factor w ranged between 0.874 and 0.906 for the different data sets while the optimal number of M_e was between 979 and 1'195. These numbers were of the same magnitude as the optimal values we found when using cross-validation results with the total data set of 5'698 bulls. Using the proposed weighting factor made it possible to reflect the increase of accuracy when enlarging the number of animals in the training set. The empirical accuracies (0.808±0.002) obtained with a training set of 5'413 bulls was clearly within the range of accuracies (between 0.793 and 0.815) we would predict when using the parameters optimized for the ten data sets of 4'000 bulls and deviated only slightly from the average predicted value 0.803 ±0.007.

DISCUSSION

The aim of our study was to use empirical data to find a deterministic prediction equation for the accuracy of genomic breeding values that accounts for factors like sample size of the training set and marker density used that fits our data best. We used a maximum likelihood approach to validate different equations to predict the accuracy of GBV. We showed that the likelihood of our approach was best when the estimates of M_e were obtained based on an expected value of the accuracy that also included a weighting factor reflecting the marker density used.

There are different possible reasons why the accuracy of genomic prediction with a specific SNP set may not reach one even if the number of training animals is infinite. First of all, only a fraction of the variance generated by QTL will be tagged by SNPs, i.e. the marker density is too low. Furthermore arrays like the Illumina BovineSNP50 BeadChip were designed such that the allele frequencies of the markers are more or less uniformly distributed ([23]) which leads to an underrepresentation of markers with very low minor allele frequencies. Since similar allele frequencies between marker and QTL are mandatory for obtaining high LD values and capturing the variance of the QTL, QTL with low minor allele frequency may not be represented adequately by the markers on a common SNP chip.

The weighting factor w can be interpreted as the maximum accuracy that can be achieved with the specific marker set in the population at hand assuming an infinite training set size. In our case, we found w to be in a range of ~ 0.875 to 0.9 while the accuracies we could obtain with ~ 5'700 bulls in our Holstein Friesian data set empirically were around 0.8. This means that most of the possibly achievable accuracy is already obtained when having ~ 5'000 bulls in the training set. Genomic heritability (i.e. heritability in the GBLUP model) may be another good indicator of how much genetic variance is captured by the SNPs. Estimates of genomic heritability in our data sets (results not shown) were higher than the estimated squared w (representing the proportion of genetic variance captured by the SNPs), but behaved completely similar in trend (e.g. no increase in genomic heritability in Brown Swiss with additional markers from a number of ~20'000 markers on) compared to w^2.

Having the estimates of M_e and w at hand, one could think about changes in accuracy when enlarging the training set size. Using model D2 with $w = 0.875$ and $M_e = 1'046$ (values obtained for milk yield) we would need a training set size of 54'515 (10'246) animals to reach 99% (95%) of the possibly achievable accuracy with the given SNP density. Duplication of the number of animals in the training set from 5'000 to 10'000 would lead to a mean increase of accuracy of ~ 0.04 from 0.79 to 0.83, while going from 10'000 to 20'000 animals would only lead to an increase of ~ 0.02. Note that these considerations assume that a further ran-

dom set of bulls (i.e. no specific groups like close relatives etc.) is used to enlarge the training set. In general, increasing the number of animals in the training set therefore will not add enough beyond a certain point when set in relation to the additional costs that incur for genotyping and phenotyping the required animals. Reliable knowledge about this case of diminishing returns is crucial when implementing or optimizing genomic selection programs.

Daetwyler [16] used a regression approach to estimate the maximum genetic variance captured by a SNP, which is the squared value of our weighting factor w. He observed four data points within US Holstein data sets for different training set sizes. However, he did not use different k-fold cross validation but validated his theory by augmenting the training set with new animals, including cows, to achieve larger training set sizes. The maximum genetic variance that is captured by the SNP set depends also on the population studied. Adding cows thus may bias the results since a higher genetic variance is expected in the cow population compared to the highly selected group of progeny tested bulls.

The maximum genetic variance found in his study was $q^2_{max} = 0.8 \pm 0.053$ in US Holsteins for Net Merit with the 50k SNP Chip which equates to a w of ~0.89. This is very close to our estimate in a European Holstein data set both for milk yield and somatic cell score. The weighting factor w in principle should be trait specific, but if conventional estimated breeding values (or equivalently de-regressed proofs or daughter yield deviations) are used as quasi-phenotypes for genomic prediction differences between traits should not be large as long as estimated breeding values are sufficiently accurate and homogeneous. Daetwyler [16] also suggested estimating M_e from model D1 based on results from real data [16] and simulated data [9]. For this, they rearranged D1 multiplied by the square root of q^2_{max} so that the number of M_e could be obtained directly. Their results for US Holsteins were in a range of about 900 to 1300 for the number of M_e which is in the same range as the results we obtained with our Holstein data.

All numbers of M_e we derived in Holstein Friesian with D2 or G3 were similar or somewhat smaller than expected compared to the deterministic approach of [10] $(M_{e\_Goddard} = 1'259)$ and clearly smaller than expected compared to the approach of [15] $(M_{e\_Hayes} = 5'800)$ when assuming N_e being 100 and the length of the autosomal genome being 29 Morgan. For Brown Swiss, the approach of [10] would clearly overestimate M_e in comparison to what we found in the empirical data (M_e from 148 to 412). Hayes et al. [11] showed that expected accuracies were very close to empirical results from US Holstein Friesian cattle when using his definition of M_e and an effective population size of 100, a length of the genome of 30 Morgan, and the original equation of [9]. For our data, however, the predicted accuracy using the as-

sumptions of [15] would severely underestimate the accuracies observed in the cross-validation study (results not shown).

Goddard et al. [10] suggested the factor $b = \frac{\#SNPs}{\#SNPs+M_e}$ to estimate the proportion of genetic variance that can be explained by the markers, i.e. our factor w squared. With M_e and $\#SNPs$ in the realistic range reflecting current applications in dairy cattle b will approach 1 very fast. For example, when $M_e = 1'000$ and $\#SNPs = 54'000$, b would be 0.982 and therefore the square root of b (i.e. w) would be > 0.99, which is clearly higher than found in experiments with real data ([2], [24], [25]) including this study.

We found a clear linear relationship between the reciprocal of the logarithm of the marker density and the maximal achievable accuracy (w) of the form $w = a + z(\ln(\#SNPs/L))^{-1}$ where $a \sim 1$ and z is a trait-specific regression coefficient. Such a linear relationship has also been found by [26] in simulated data. Since the relationship is linear to the log of the marker density, it is not surprising that the factor w which can represent the maximal achievable accuracy did not differ much between our runs with different number of SNPs in the Holstein data set. We could not study what will happen with further increasing the marker density in Holstein Friesian, since we did not have access to a sufficiently large set of individuals with high density marker genotypes.

Current results have shown that the accuracy of genomic breeding value prediction within breed did not increase significantly when using imputed 777k SNP marker data rather than 50k SNP data [24]. It seemed that also the proportion of genetic variance captured by the markers was only slightly higher. In our Brown Swiss data set, all bulls had 777k SNP genotypes and we actually saw a stagnation of the percentage of genetic variance explained when the number of markers was greater than ~ 20'000. This means that even with an infinite size of the training set the accuracy of prediction will not be better even if we use 30 times more markers. In Holstein Friesian, w still increased up to ~40'000 markers roughly linearly with the logarithm of the marker density. It thus can only be assumed that the plateau has just not been reached for Holstein Friesian with the observed marker density, which remains to be verified once sufficiently large samples with high density genotypes are available for the Holstein Friesian breed.

The highest possible marker density is achieved when using whole genome sequence data in genomic prediction. In a data set of 157 inbred lines genotyped for ~ 2.5 million SNPs, [12] found that the prediction equation D1 of [9] adapted for the special genetic model of *Drosophila melanogaster* was a good predictor for the accuracy of sequenced-based genomic breeding value estimation looking at different sizes of reference sets. Since the fit of the

original equation of [9] to the empirical accuracies was excellent, it can be concluded that this massive SNP density (~ 1 Mio SNPs/Morgan) recovers the complete genetic variability (i.e. $w^2 \sim 1$) but in contrast to our study the small size of the reference set is the limiting factor in that case.

The results for the estimates of M_e were very different in the two studied breeds. This was surprising because both are modern dairy breeds and rather similar results would have been expected. We thus assessed different characteristics of the two populations (Holstein Friesian and Brown Swiss) to identify potential causes for the difference in the pattern of observed accuracy functions. First, we calculated the effective population size N_e based on pedigree information and found values that were very similar for both breeds ($N_e \sim 75$, obtained with POPREP [27], based on [28]). Based on linkage disequilibrium (using markers available in both sets and formulas of [29] and [30]), estimates for N_e in 6 to 9 generations back was ~133 in Holstein Friesian and ~125 in Brown Swiss. Both analyses suggest that there is no difference between the two breeds regarding N_e. Furthermore, we studied properties of the genomic relationship matrix, namely the eigenvectors and eigenvalues of $\mathbf{G}$ which reflect the degree of population substructure in the sample. To avoid a bias due to the number of SNPs used, we compared the genomic relationship matrix constructed with 42'551 SNPs for Holstein Friesian and 39'207 SNPs for Brown Swiss. The first and the second eigenvectors explained 14.36% (13.32%) and 6.29% (9.96%) of the variance in the Holstein Friesian (Brown Swiss) data set. The first 10 eigenvectors explained around 50% of the variance in both data sets. The differences between the structures of the eigenvectors in the covariance matrix therefore also seem to be negligible.

These results indicate that further parameters have to be found that can determine the proportion of genetic variance explained and the SNP density at which the plateau is reached. They also illustrate that calculating an expected value of M_e just based on the length of the genome and the effective population size may not be sufficient, since empirical values for M_e differ between traits within populations and even between populations with similar N_e and the same length of genome. Furthermore, the results may also indicate that interpretability of population parameters (like e.g. M_e) in such formulas can be limited when they are derived with the suggested goodness-of-fit-approach.

We further showed that model D2 allowed a realistic extrapolation of prediction accuracies with increasing training set sizes, while model D1 systematically overestimated the accuracy for a training set of 5'413 Holstein Friesian bulls when the model parameters M_e and w were derived with a subset of 4'000 bulls. The overestimation was not dramatic for this example, but 5'413 is not that much bigger than 4'000. However, if the difference between the number

of individuals used for fitting the curve and the size of the reference set for which the accuracy is to be predicted increases, the upward bias will accumulate. Especially, as it is expected that number of animals in the training sets will increase in real studies up to ten thousands of training animals, it is critical that a prediction equation is able to fit the slope of the increasing accuracy correctly.

Equations to predict the accuracy of genomic breeding values are often derived for the simple case of a random set of animals that are not related (e.g. [8]) or show an 'average' relationship. In real cattle data, animals are often highly related and stem from specific selected groups, e.g. progeny-tested sires. A general equation, though, should be designed primarily as an indicator for a random animal out of a whole population (e.g. modern dairy cattle). Parameters like the number of M_e and w can be chosen in a way that they describe the underlying population and trait as good as possible, but it is not the goal to obtain exact predictions of accuracies for specific animals in the prediction set. As shown by many studies (e.g. [31]) the relationship between candidates and the training set, which also can be seen as a kind of population stratification, influences the accuracy in a non-random manner. Goddard et al. [10] showed how relationship structures can be used to estimate e.g. the parameter b but this works just in the case where animals have already been genotyped. Another idea on how to determine the maximum achievable accuracy has been recently proposed by de los Campos et al. (2013) [32]. They suggested an approach for the case of imperfect linkage disequilibrium between markers and QTL which is not depending on assumptions like unrelated individuals or parameters like M_e. Further approaches still need to be developed for the "before data collection" case.

CONCLUSION

We suggest a comprehensive model for the average accuracy of genomic breeding values and demonstrate how the model parameters can be estimated using a systematic cross-validation based on empirical data. Integrating all results, we suggest the model

$$r_{GBV,TBV} = min(a + z(\ln(\#SNPs/L))^{-1}; w_{max}) \sqrt{\frac{n_t h^2}{n_t h^2 + M_e}}$$

with the four parameters a, z, w_{max} and M_e, that can be empirically determined via systematic cross-validations as described in this study.

The suggested modification of the original equation of [9] led to a substantially improved fit of the predicted accuracies obtained with cross-validated data and showed its good prediction

ability in the extrapolation to larger training sets. The maximum likelihood approach used for obtaining an estimate of the number of independent chromosome segments led to largely consistent values across different SNP sets. We also propose a function linking the maximally achievable accuracy of genomic prediction to the marker density, suggesting strongly diminishing returns when increasing the sizes of the SNP arrays, which confirms results obtained with different SNP densities in practical applications with dairy cattle.

ACKNOWLEDGEMENT

Genotypic information and estimated breeding values for Holstein Friesian bulls were kindly provided by Vereinigte Informationssysteme Tierhaltung w.V. (VIT), Verden, Germany.

The authors thank Braunvieh Schweiz, the genotype pool Germany-Austria, Associazione Nazionale Allevatori Bovini della Razza Bruna and Beltsville Agricultural Research Centre for provision of genotypes.

Estimated breeding values for Brown Swiss bulls were kindly provided by Braunvieh Schweiz, Zug, Switzerland.

FUNDING

This research was funded by the German Federal Ministry of Education and Research within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” (Funding ID: 0315528C) in association with the Deutsche Forschungsgemeinschaft (DFG) research training group “Scaling problems in statistics” (RTG1644).

The authors gratefully acknowledge co-funding from the European Commission, under the Seventh Framework Program for Research and Technological Development, for the Collaborative Project LowInputBreeds (Grant agreement No 222623). However, the views expressed by the authors do not necessarily reflect the views of the European Commission, nor do they in any way anticipate the Commission’s future policy in this area. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors further acknowledge support by the Open Access Publication Funds of the Göttingen University.

REFERENCES

1. VanRaden PM, Sullivan PG (2010) International genomic evaluation methods for dairy cattle. Gen Sel Evol 42: 7. Available: http://www.gsejournal.org/content/42/1/7. Accessed 30 June 2013.

2. Liu Z, Seefried FR, Reinhardt F, Rensing S, Thaller G, et al. (2011) Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Gen Sel Evol 43: 19.

3. Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, et al. (2009) The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics 183: 1119-1126.

4. Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Gen Sel Evol 41: 51.

5. Legarra A, Robert-Granié C, Manfredi E, Elsen J-M (2008) Performance of Genomic Selection in Mice. Genetics 180: 611-618.

6. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) *Invited review*: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92: 433-443.

7. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3: e3395. Available: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0003395. Accessed 30 June 2013.

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

9. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185: 1021-1031.

10. Goddard ME, Hayes BJ, Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128: 409-421.

11. Hayes BJ, Daetwyler HD, Bowman P, Moser G, Tier B, et al. (2009) Accuracy of genomic selection: Comparing theory and results. Proc Assoc Advmt Anim Breed Genet 18: 34-37.

12. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, et al. (2012) Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster*. PLoS Genet 8: e1002685. Available: http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1002685. Accessed 30 June 2013.

13. Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics 193: 621-631.

14. VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. J Dairy Sci 91: 4414-4423.

15. Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res Camb 91: 47-60.

16. Daetwyler HD (2009) *Genome-Wide Evaluation of Populations.* Ph.D. Thesis: Animal Breeding and Genomics Centre, Wageningen University, Wageningen, NL (2009). ISBN: 978-90-8585-528-6.

17. Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet 81: 1084-1097.

18. Sargolzaei M, Chesnais JP, Schenkel FS (2011) FImpute – An efficient imputation algorithm for dairy cattle populations. J Dairy Sci 94(E-Suppl. 1): 421(333).

19. Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) *ASReml User Guide Release 3.0.* VSN International Ltd, Hemel Hempstead, UK.

20. Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124: 331-341.

21. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52: 591-611.

22. R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/.

23. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF et al. (2009) Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4: e5350. Available: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0005350. Accessed 30 June 2013.

24. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, et al. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95: 4114-4129.

25. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME et al. (2013) Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci 96: 668-678.

26. Meuwissen THE (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41: 35.

27. Groeneveld E, v.d. Westhuizen B, Maiwashe A, Voordewind F, Ferraz JBS (2009) POPREP: a generic report for population management. Genet Mol Res 8: 1158-1178.

28. Pérez-Enciso M (1995) Use of the uncertain relationship matrix to compute effective population size. J Anim Breed Genet 112: 327-332.

29. Sved JA (1971) Linkage Disequilibrium and Homozygosity of Chromosome Segments in Finite Populations. Theor Pop Biol 2: 125–141.

30. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003) Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Res 13: 635-643.

31. Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Gen Sel Evol 42: 5. Available: http://www.gsejournal.org/content/42/1/5. Accessed 30 June 2013.

32. de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet 9: e1003608. Available: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003608. Accessed 03 October 2013.

6th CHAPTER

General discussion

GENERAL DISCUSSION

Preface

This thesis studied different factors that influence the accuracy of genomic prediction in real dairy cattle data sets. The following factors were investigated further in realistic study designs in ***Chapters 2***, ***3*** and ***4***:

- training and validation set size,
- relationship and age structure between training and validation set,
- density of the marker panel,
- composition of the training set and
- model choice.

Chapter 5 attempted to represent the accuracy of genomic prediction as a formula based on population specific parameters.

Important questions arising from the results of the previous chapters will be discussed in the following.

How valuable is cross-validation accuracy and how does it correspond to other parameters assessing the performance of a model?

In all studies in this thesis, cross-validation accuracies have been used which were measured as the correlation between the predicted and the true breeding values (or the quasi-phenotypes). Cross-validation accuracies model an average accuracy over the studied individuals but they cannot show individual levels of accuracy. For general trends how different models work the mean accuracy is usually a good measure.

In mixed model equation (MME) theory, it is possible to obtain the accuracy (r_i) for a specific individual i based on the prediction error variance (PEV) so that

$$r_i = \sqrt{1 - \frac{PEV_i}{\sigma_g^2}} \quad [1]$$

where σ_g^2 is the genetic variance.

To compare the results that would be obtained based on individual accuracies with cross-validation results, individual accuracies were calculated for all scenarios and replicates from ***Chapter 3*** and were averaged within replicates and then over scenarios. Afterwards, mean accuracies per scenario obtained with cross-validation and with individual accuracies were compared. Figure 1 shows that cross-validation accuracies are slightly lower but the trend over scenarios is the same for both measures. Correlations between mean accuracies by cross-validation and by MME calculations over scenarios are 0.925 for somatic cell score and 0.927 for milk yield. This shows that for a comparison of different scenarios, cross-validation accuracies are a good measure for evaluation.

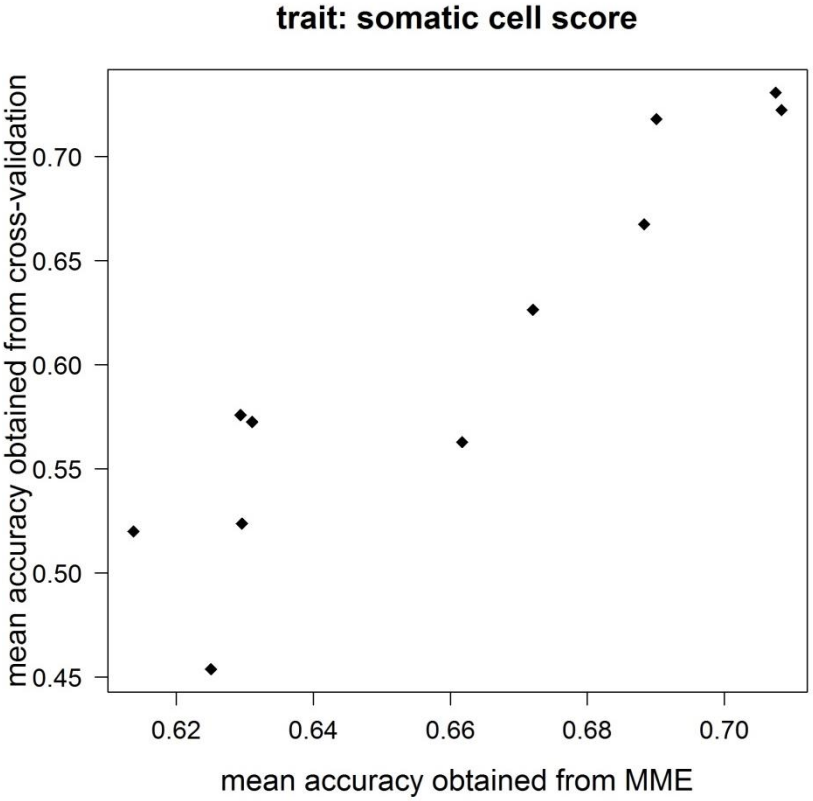


Figure 1: Accuracies based on [1] averaged over individuals and cross-validation accuracies from different scenarios of ***Chapter 3***.

From cross-validation experiments, it is also possible to calculate an empirical prediction error for each individual i as

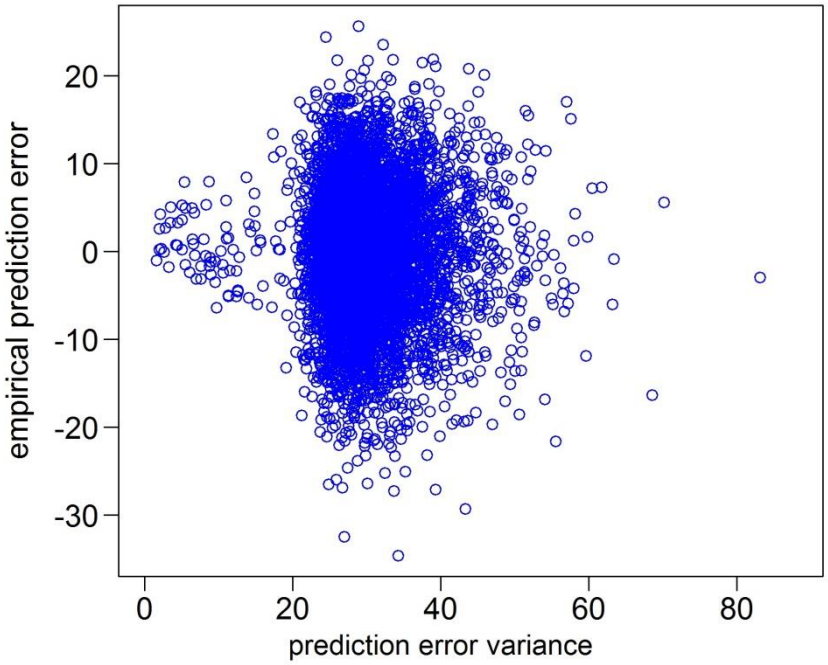
$$Err_i = (y_i - \hat{\mu} - \hat{u}_i) \quad [2]$$

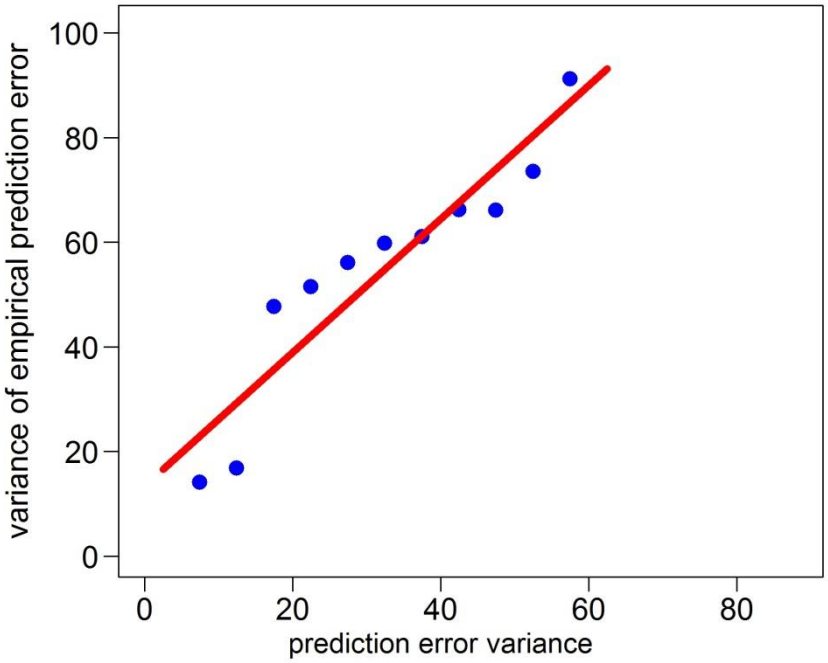
where $\hat{\mu}$ is the estimated overall mean, $\hat{u}_i$ is the predicted genomic breeding value of individual i and y_i is the quasi-phenotype that was masked to obtain $\hat{u}_i$ in the genomic prediction model.

To further study the relation between all these parameters a leave-one-out cross-validation with a data set of 5'698 Holstein Friesian bulls (which were also the basis for the studies in ***Chapter 3*** and ***5***) for the trait somatic cell score was performed. Leave-one-out means that there are 5'698 runs of genomic prediction in which each individual is the validation individual exactly one time and all other individuals are used for training. The following statistics were calculated for the validation individual in each run: prediction error variance, empirical prediction error between masked phenotype and predicted one as in [2] and accuracy of prediction based on the prediction error variance as in [1]. The empirical prediction error could be compared over all runs since $\hat{\mu}$ changed only very slightly between runs.

The empirical accuracy was also considered by calculating the correlation between phenotypes and predicted breeding values for all individuals divided by the square root of the genomic heritability estimated with the whole data set ($h_g^2 = 0.803$). The empirical accuracy was 0.849 while the mean of the individual accuracies based on [1] was 0.849 and was in the range between 0.487 and 0.993. Accounting for inbreeding, namely by calculating $r_i = \sqrt{1 - PEV_i/\left(g_{ii}\sigma_g^2\right)}$ with g_{ii} being the diagonal element of the genomic relationship matrix for individual i did not change the accuracy much (0.849 [0.607; 0.992]).

There was no correlation between the predicted genomic breeding values and the empirical prediction error (-0.002) while there was a highly positive correlation between the quasi-phenotype and the squared empirical prediction error (0.647). Figure 2a shows the relationship between the prediction error variance and the empirical prediction error, whose correlation was 0.04, for all individuals. For Figure 2b, bins of size 5 units of PEV were built and the variance of the empirical prediction error (VEPE) within these bins was calculated. Even when PEV is defined within an individual and VEPE over random individuals, both are expected to show the same tendency, namely the higher the theoretical PEV, the higher should be the empirical variation in deviations of the predicted values from the observed ones. Only results from bins with at least 10 observations are shown in Figure 2b. For low PEV, VEPE remains very low at first, and then increases continuously. Applying a linear regression to those bins ended in a highly significant regression coefficient ($\hat{b} = 1.31$ with a p-value of 0.0000335) and a coefficient of determination of 86.5%. Since numbers of observations per bin were different, the regression may be slightly biased but there is a clear positive trend. Relationships between all those parameters should be studied further to find optimal parameter to describe the assessment of models.





Figures 2a and b: Relationship between the prediction error variances of the validation individuals from 5'698 leave-one-out cross-validation runs and the empirical prediction errors (**Figure 2a**) and between prediction error variance and the variance of the empirical prediction error calculated in bins of 5 units width (**Figure 2b**).

In the past year, different authors (Bijma, 2012; Edel *et al.*, 2012) have considered the question on how selection influences the accuracy measures of genomic prediction. Edel *et al.* (2012) stated that the accuracy obtained with cross-validation in a forward-prediction scheme (i.e. prediction of the youngest) with underlying selection will underestimate the true prediction accuracy. Bijma (2012) argued the other way round, namely that accuracies based on PEV and formulas like [1] are not valuable in populations under selection since they overestimate the actual accuracy. Smaller values obtained with cross-validation than with accuracies from MME were observed in the evaluation of results of ***Chapter 3*** (see Figure 1 in this discussion), while this effect was not present in the leave-one-out cross-validation experiment. Apart from different data sizes and designs (random drawing with replicates vs. leave-one-out) the difference between these two runs was that the first one was stratified (forward-prediction) while the latter considered prediction of all bulls in the set. This may have influenced the results in this direction. Further research on these ideas will be necessary.

Daetwyler *et al.* (2013) suggest the slope of the regression of a linear regression of observed on predicted values should also always be reported. Some authors have done this before (e.g. Su *et al.*, 2012) and it was used as a measure of performance of specific models. Slope of the regression may be important if different sources are merged afterwards to obtain a genomic enhanced breeding value, e.g. by combining pedigree based breeding values and direct genomic breeding values. As long as the information level stays on the level of direct genomic breeding values, the ranking of the bulls is the most important factor and this is not influenced by the slope of the regression. Thus the accuracy of prediction measured as the correlation between observed and predicted values is still the more important measure. At best this measure is obtained in a study design that allows for (stratified) replications since single point estimators are always hard to handle and to interpret. A good overview on further parameters regarding the assessment of a model worth to add to a manuscript is found in Daetwyler *et al.* (2013).

What is the impact of relationship structure on accuracy of genomic prediction?

In ***Chapter 3*** it was discussed that the level of relationship and the age structure between validation individuals and the respective training set clearly influences the level of accuracy. In this study, the validation set was kept constant to assure that the prediction scenario is the one of practical relevance, namely the prediction of the youngest individuals. Not very many studies are available which consider the impact of different relationship levels in real dairy cattle data sets. Habier *et al.* (2010) studied the influence of the maximal relationship to the accuracy of prediction in training and validation sets that were not structured by age. They

also found a decrease in accuracy with lower levels of maximal relationship and they showed that the decrease was stronger when the total number of individuals was smaller.

Clark *et al.* (2012) used data from ~1'750 Merino sheep to model different levels of relationship in a scenario of unrelated individuals, i.e. individuals for whom pedigree information showed no relationship. While in this case no prediction is possible with pedigree BLUP, with genomic BLUP they could still see accuracies of 0.18 (0.28) for live weight at ultra-sound scanning (ultrasound scanned eye muscle depth). These values show another advantage of genomic breeding value prediction, namely that there is at least a good chance to also obtain usable breeding values for individuals where no pedigree information is known. This may be not as important in dairy cattle with deep and comprehensive pedigrees as in many other species. However, this study also shows that unrelatedness within a species of limited effective population size is not comparable to an unrelatedness e.g. in humans with large N_e, otherwise values of clearly higher than 0 would probably not be obtained with this size of training set of unrelated individuals.

What is often shown in recent studies (e.g. Clark *et al.*, 2012) is the correlation between specific relationship parameters and the individual accuracies based on PEV to explain differences in the level of accuracies. In most cases, the "top-ten-relationship" (average of 10 highest relationship coefficients) between validation individual and training set is found to be a very good predictor for the MME accuracy. This is correct, but at the same time it is also not surprising: For an individual without own performance, the PEV is smaller the more information from close relatives is available, i.e. the larger the number of high covariance values with other individuals in the training set. Just a few really high values of covariances count more than many small ones, which makes the "top-ten-relationship" a good parameter for illustration.

Pszczola *et al.* (2012) simulated a dairy cattle population and three traits of different heritabilities. They took two types of relationship in account: The relationship structure between validation and training set (by letting the validation individuals coming from the same generation or from one or two generations further, leading to the same tendencies as discussed before) and the relationship structure within the training set, i.e. how related the individuals are that are used to estimate the SNP effects. They showed a clear influence of the relationship within the training set, with very diverse training sets showing a low mean accuracy having the edge over training sets consisting of highly related individuals.

Wientjes *et al.* (2013) performed a comprehensive study based on a real cattle training set with simulated validation individuals in a way that the effects of different factors on the accuracy of prediction could be studied: For simulating candidates only allele frequency, linkage

disequilibrium structure, haplotype structure and/or family structure from real data were included, respectively. The authors then checked how the simulated validation individuals could be predicted based on marker estimates obtained with the training set of real data. Values of accuracy substantially higher than 0 could only be achieved when whole haploid chromosomes (segregating in the real population) were used to model the validation individuals. Sharing only allele frequencies, LD pattern or small haplotype segments was not enough to predict values for the validation individuals with reasonable accuracy.

What is the ideal training population?

No clear answer to this question exists since this depends on the underlying population, the purpose individuals are genotyped for, and the breeding scheme in the respective population, but some general notes can be given:

Even if probably not the initial idea of genomic breeding value prediction, in the last years it has emerged that relationship between individuals in the training and the validation set is one of the key points for the level of accuracy that can be achieved. It may be that some genomic breeding value approaches are less sensitive to the level of relationship (see e.g. Habier *et al.* (2010) for a discussion), but nevertheless in practical applications mainly simple linear genomic BLUP models are used and these are sensitive to different relationship levels. This is not at all a disadvantage and this is also not at all something that has to be suppressed, because a thoughtful composition of the reference set can compensate this to a high extent.

In ***Chapter 3*** it was discussed that the elimination of all close relatives of the candidates from the training set caused a considerable decrease in accuracy of prediction. This means that after one or two generations of not supplementing the training set with younger individuals the level of accuracy will erode. Updating the training population can be done by adding former selection candidates that have obtained progeny records in the meantime. It may become more difficult after a few generations of consequently applying genomic selection when bulls will be used directly as young individuals and the regular testing scheme will not be maintained anymore. International collaborations to share geno- and phenotypes can play a big role to increase the number of individuals for the training set.

As long as there are still testing bulls, it will be good to also add all those bulls when obtaining progeny records since this may avoid any bias coming up from just adding the elite of the elite individuals. Even when the area of testing bulls may be over this must not mean that phenotyping of individuals becomes less important. Efforts have to be made to ensure that enough progeny records will be available for many individuals so that still a high number of

individuals are available to be added to the training sets. Many females have also been genotyped up to now and could be added to the training sets as well.

Cows that are genotyped are normally elite cows and not a random sample of the cow population. This may lead to a potential bias (e.g. Dassonneville *et al.*, 2012; Pryce & Daetwlyer, 2012) in genomic breeding values when adding only elite cows to the training set, but this is a problem that could be solved. In contrast to bulls with many progeny records, cows can only provide own performances as phenotypes (for adjustments see e.g. Wiggans *et al.*, 2012b), but especially in new approaches like the single step methodology (e.g. Legarra & Ducrocq, 2012) this should be relatively easy to handle. Cows in the training sets can become much more important in genomic selection schemes in small breeds where only a small number of bulls with progeny records can be added per year or as soon as new traits just phenotyped in cows will become relevant.

Individuals within the training set should represent as much of the variation of segregating haplotypes in the population as possible. Thus a wide range of different families should build the training population. Adding females from the production population to the data set may also help to ensure this. Increasing the training set will always help to increase accuracies and can be crucial in small breeds. In large breeds with small effective population size like Holstein Friesian, increasing the training set will not really provide much higher accuracies than already obtained with the actual sizes. Therefore, it may be much more important to keep it on the same size, but up to date.

Is it a general trend that there are no significant changes in accuracy of prediction between 50K and HD SNP Chip?

The study presented in ***Chapter 4*** was one of the first that investigated the benefit of high density (HD) SNP data in different genomic prediction scenarios (purebred vs. multi-breed, BLUP methods vs. Bayesian methods). There was no significant increase in accuracy when using HD SNPs for within-breed prediction and only a slight increase for the minor breed Jersey when using a multi-breed training set. With the new method *BayesR* equal or in many cases better results could be produced than with GBLUP, but there was no strong tendency that Bayesian methods can handle a larger number of markers much better than BLUP methods that weight each marker equally. These results did not match any of the expectations. However, for Jersey there were only 540 bulls available and with larger data sets different results might have been obtained. Further possible reasons for the results with this data set were discussed extensively in ***Chapter 4***.

A few other studies have now been published which all compare accuracy of genomic prediction with 50K and imputed HD genotypes: Su *et al.* (2012) used data sets of ~ 4'500 Nordic Holstein and 4'400 Red Dairy Cattle (RDC) to compare reliabilities of genomic breeding values based on 50K and HD SNP chips. Averaged over three traits (protein yield, fertility and udder health), reliabilities were only 0.5% (0.7%) higher for Holstein (RDC) when using GBLUP and imputed HD SNPs instead of 50K. A Bayesian mixture model produced slightly higher reliabilities than GBLUP, but the advantage was not higher with the HD panel. In Australian Holstein data accuracy of prediction was at most 0.03 higher for HD data than for 50K regardless of whether most probable genotypes or allele dosage from the imputation process were used (Khatkar *et al.*, 2012). Around 10'700 Holstein Friesian bulls and 5'000 Holstein Friesian cows built the training set in the study of VanRaden *et al.* (2013). Averaged over 28 traits, the observed gain in reliability with the HD SNP Chip was only 0.4% with a non-linear model while within the HD SNP scenario the gain was 0.8% with a non-linear model compared to a linear model.

Pryce *et al.* (2012) studied residual feed intake and 250-day body weight in ~1'800 Holstein heifers in Australia and New Zealand within a cross-validation scheme where there were always Australian and New Zealand heifers in the training set while prediction was done either for a subset of Australian or New Zealand heifers. In all cases there was no increase in accuracy in both traits compared to a purebred scenario. Predicting Australian heifers with New Zealand ones did not work at all with any SNP density, while there was an advantage in prediction accuracy for the Bayesian methods with HD when predicting New Zealand heifers with Australian ones. With Holstein Friesian and Jersey to predict Holstein Friesian-Jersey crossbreds no increase in accuracy was found with imputed HD data in New Zealand (Harris *et al.*, 2011) while there were small improvements when using one breed to predict the other.

Some other studies (e.g. Solberg *et al.*, 2011) also show the same tendencies. Even with much larger training sets than in ***Chapter 4*** no clear benefit of the HD data can be seen. The same is true for different methods – none of the different ones used in the studies mentioned above showed a real benefit in regard to the accuracy of prediction with HD data. Since only production traits (milk, fat, protein) which are all known to be influenced in moderate to large parts by DGAT1 were studied in ***Chapter 4***, it was not clear what will happen to other traits like fertility or conformation traits. Further traits have been investigated e.g. in VanRaden *et al.* (2011) or Su *et al.* (2012), but again also for non-production traits there was no obvious improvement in accuracy. Data from dairy cattle breeds with small effective population size will show strong linkage disequilibrium structures which may just be strong enough with the 50K Chip and having more SNPs cannot capture more genetic variance (see results in ***Chapter 4*** and ***5*** and e.g. variance components estimated with 50K and 777K in Su *et al.*,

2012). This may be different in breeds with different genetic background, e.g. in beef cattle breeds or in other farm animal species.

All these studies completely confirm the results from ***Chapter 4***. In addition, one can draw a cautious conclusion: In modern dairy cattle breeds, using data from the HD SNP chip will not affect the accuracy of genomic prediction to a great extent regardless of studied trait, the training set size and the applied model.

Up to now, in most studies with dairy cattle data the sample always consisted of imputed high density data to large parts, i.e. individuals used in the model were not themselves genotyped at high density, but all SNP positions only available at the high density SNP chip had been imputed before. Imputed genotypes are very accurate (e.g. Brøndum *et al.*, 2012) but nevertheless they may not reflect the truth. Assuming that around 700'000 SNP positions per individual have to be imputed from 50K up to the HD SNP Chip, a rate of 97% correctly imputed genotypes still means that on average around 21'000 positions within an individual are not correctly imputed. Segelke *et al.* (2012) have nicely shown how the mean allelic error rate may fluctuate between different chromosomes when imputing from low density SNP chips to 50K. In ***Chapter 4***, it was also shown how possibly misplaced SNPs can influence the accuracy of imputation from 50K to HD at specific areas of the genome. These results imply that imputation accuracy is not uniform across the genome. If the wrongly imputed positions were randomly distributed across the genome, this would probably not pose a big problem. If these wrong genotypes are clustered at specific positions, this may influence the results of genomic prediction. Furthermore, imputation accuracy may not only differ between regions in the genome, but also between individuals that are related to the reference panel in different degrees or have their parents included or not to the reference panel (see e.g. Brøndum *et al.*, 2012; Wiggans *et al.*, 2012a).

Does the imputation process reduce the potential benefit of HD data?

Up to now there is no study available with real data comparing accuracies obtained with real HD genotypes and imputed HD genotypes. VanRaden *et al.* (2011) used simulated data based on a real pedigree structure and found an improvement in reliability of genomic breeding values in young bulls of 1.6% when all individuals were directly genotyped on 500'000 SNPs compared to a scenario where all individuals were genotyped with 50'000 SNPs. When having only a part of the individuals genotyped with 500'000 (~3'800 and ~1'400, respectively) and the remaining ones imputed with findhap, the gain of reliability was only 1.2% (0.9%). Some further studies have investigated the differences between scenarios with im-

puted and real genotypes in the context of SNPs from low density chips (~3'000 or 7'000 SNPs) and the 50K chip. Khatkar *et al.* (2012) found no tendency for a decrease in accuracy of prediction when using imputed 50K data instead of real 50K data with imputation done up from 7K, while accuracies were slightly lower when imputation was from 3K. Similar conclusions can be drawn from Segelke *et al.* (2012) who found that correlations of direct genomic values between evaluations with real 50K data and imputed data from 6K SNPs were higher than between 50K data and imputed data from 3K SNPs in a German Holstein data set. Furthermore, they observed that loss in reliability of genomic breeding values was greater with data imputed from 3K (e.g. with Beagle 2% averaged over traits) than from 6K (0.8%) but the absolute values of loss were small in both cases. They also showed that there is a large effect of the software tool used for imputing, with Beagle being the superior one.

From all these results it can be concluded that there may be some, but no striking loss in accuracy of prediction when imputed genotypes are used instead of observed genotypes. Even if the loss is negligible on average over traits and individuals, one should be aware that this may not be the case for specific individuals with an unfavorable information structure e.g. because they do not have close relatives in the reference set. While there are no studies based on large sets of observed HD genotypes, there is no reason to believe that the lack of superiority of HD-based studies is due to the fact that HD genotypes are only imputed.

The crucial point is rather the diminishing return from adding more and more markers in highly related populations with a distinctive linkage disequilibrium structure. Results in ***Chapter 5*** show that in specific breeds (here: Brown Swiss) this threshold can be as low as around 20'000 markers. Therefore, in the field of genomic breeding values prediction efforts to find specific SNP subsets (e.g. based on a biological background or also with new knowledge from sequencing data) with reasonable size and pooling of available genomic data together to increase the number of individuals used to estimate the SNP effects may have a longer lasting success than the increase of marker density.

How can we limit computational demands with high density data?

One of the major problems arising with larger marker density is that calculations become more and more computationally demanding, especially when using Bayesian methods, but also in GBLUP approaches, e.g. for the creation of the genomic relationship matrix or when calculating SNP effects. Therefore, efforts should be made to reduce the number of SNPs without losing information quality.

First of all, one should reflect if denser markers are really necessary for the purpose the data is used for or if the actual marker density is sufficient. Based on the results so far, for most of the genomic prediction scenarios there was no or very little advantage (see paragraph above) in using imputed HD genotypes within breed for prediction in modern dairy cattle breeds. Given the fact that training sets are refreshed regularly and marker effects are estimated every few months the argument that marker effects from HD genotypes may be more stable over generations does not really count either. The situation is different for the across-breed prediction where some advantages of high-density genotypes have already been found for small breeds and may become greater when imputation accuracy can be increased because of the availability of more HD reference individuals. The situation may also differ in other less related breeds or other species in which 50K genotypes are just not dense enough given the underlying linkage disequilibrium structure.

If high density genotypes are considered useful in a study, the next step should be to check which markers contain redundant information. The simplest strategy is to check for SNPs that are in complete linkage disequilibrium (LD) with a neighboring SNP and delete one of those SNPs (e.g. Su *et al.*, 2012). The search for markers in complete LD could also be applied to all marker combinations on a chromosome without the restriction that they have to be adjacent or LD has to be complete but above a certain threshold. VanRaden *et al.* (2011) checked all pairwise combinations of a marker with the subsequent 349 markers with the threshold being the correlation between the genotypes greater than 0.95 to 0.99 depending on the underlying minor allele frequency. To decide which marker should be removed within a specific pair or groups of markers in high LD, markers with specific properties (e.g. used for parental verification) were preferred. Harris *et al.* (2011) deleted one of a pair of markers within an interval of 250 SNPs when the squared correlation was higher than 0.99. With both strategies, VanRaden *et al.* (2011) and Harris *et al.* (2011) could reduce the high density marker set to a subset of less than 350’000 SNPs, which is less than the half of the actual marker number of the HD chip.

Apart from reduction strategies solely based on the observed LD structure other approaches are imaginable, e.g. trait-specific subsets may reduce computational time and help to improve prediction in specific traits. Furthermore, more biological aspects and background could be used to find a subset that includes all SNPs that are in relevant position within the genome structure. In ***Chapter 4***, a subset of all markers located in the transcribed part of the genome was selected from all high density SNP markers. In all scenarios, the accuracies of genomic prediction with this SNP set were equal or even better than with the full data set with the subset always providing better results when genomic prediction was across breeds

(0.24 (0.52) with the subset and 0.17(0.49) with all HD SNPs when predicting Jersey using a Holstein only (combined) reference set in average over the three studied traits).

All of these techniques keep the advantages of the larger marker set but can reduce the computation time dramatically.

Can higher marker density provide other advantages?

In the first days of the genomic selection idea, estimation of marker effects was seen a step that has to be done once and estimates can then be used for some time – a few years later we know better. As shown in ***Chapter 3*** and in many other studies (e.g. Habier *et al.*, 2010), relationship between individuals in the training set and in the validation set clearly influences the accuracy of prediction. When marker effects are estimated now and are to be applied to predict young individuals in 10 years then at least two generations will be in between the proven individuals now and the selection candidates in 10 years. Apart from a reduction of the overall relatedness of individuals also linkage disequilibrium structures between markers and QTL will change or break down which narrow the predictive ability of the estimated marker effects. Accuracy of genomic prediction would at any point be so low that it would not be worthwhile to apply it.

At the moment, marker effects in official genomic evaluations are re-estimated several times a year to avoid this decrease in accuracy of prediction. Candidates which are selected based on genomic breeding values at the moment (G1) normally have sires or very near relatives (G0) in the training set. These G0 individuals have been selected on the conventional way, have thus very reliable conventional breeding values and can be used in the training set at the time bulls of G1 are to be selected. If we think about selection one generation further in time, young bulls of the next generation (G2) will have sires from G1 which themselves have been largely selected based on their genomic breeding value. Because at the time of selection of G2 individuals there are hardly any daughter records of G1 sires available, conventional breeding values of these sires are not reliable. The G1 bulls are therefore not available for the training set and G2 individuals have to be predicted based on marker effects estimated with individuals two generations back. It may be that in such cases marker effects from high density panels are favorable because markers closer to QTL are available.

Using imputed HD data, VanRaden *et al.* (2013) observed the largest effects in many traits on markers from the original 50K set (i.e. not imputed ones) which could be due to an information loss because of the imputation strategy. The region around DGAT1 was screened further for marker effects obtained with 50K and imputed HD data in ***Chapter 4***. In this case,

markers being closer to the QTL position took over the larger effects which might help to sustain marker effects as good predictors over a longer period of time. In ***Chapter 4***, it was also investigated if the accuracy changed more slightly between "sire in reference" or "sire not in reference" scenarios with the HD data and the 50K data but no consistent results could be obtained over traits. However, this was also not the perfect data set to discuss these ideas intensively. More research should focus on this area in future.

All ideas discussed above were always based on the fact that high density markers should be used for genotype based methods in genomic prediction. Of course, there are many other different research approaches that can benefit a lot from higher marker densities. First of all, there is the wide field of classical single marker genome-wide association mapping where the region of the causative mutations may be narrowed down much more with more markers available. Furthermore, all haplotype-based analyses may also profit since haplotypes can be reconstructed more accurately with smaller marker intervals. These could be haplotype-based approaches in the genomic prediction context as well as in genome-wide association mapping or in the field of selection signature methods.

What is the potential of using sequence data in genomic prediction?

Apart from high density data from the common high-throughput genotype technology, sequence data of many individuals have become available in the last years with over 130 bulls within the 1000 bull genomes project (e.g. Hayes *et al.*, 2012; http://www.1000bullgenomes.com). Around 17.4 million variants in the sequence were found within those bulls (Hayes *et al.* 2012), 15.8 million of which are SNPs. Since re-sequencing of individuals is still very costly, only selected individuals will be re-sequenced directly. Again, imputation will play a big role for obtaining genotypes on all relevant sequence sites also for a large sample of genotyped animals. Fries *et al.* (2012) showed that it is possible to impute from 50 over HD genotypes up to sequence data with an accuracy of over 90%. This is a good range but still not perfect; probably optimized software tools along with more sequenced individuals will lead to an improvement in the next years.

There are a few differences between SNPs or other variants obtained from sequence data compared to data from genotype platforms. This concerns the minor allele frequency distribution which is normally U-shaped when regarding all possible SNPs but is uniform by design for both the 50K and the HD SNP chip selection of SNPs (Matukumalli *et al.*, 2009; VanRaden *et al.*, 2013). Very high linkage disequilibrium can just appear when SNP and causative mutation have roughly the same allele frequency, i.e. detecting rare variants with

common SNP chips is almost impossible. Using SNPs directly from whole-genome sequence data, a rare variant should also be detectable because it is within the set itself or at least a SNP with very similar allele frequency and very close to it can capture its variance to a high percentage. Accuracy of prediction should thus increase because more genetic variance can be captured. However, many individuals have to be sequenced first to really find such rare variants and this may be the crux.

It is very unlikely that causative mutations are amongst the SNP set on commercial SNP arrays not only because they are not known but also because patents may prevent this. When observing all variants in whole genome sequence data with sufficient coverage the causal mutation should be included. For genomic prediction, this means that one is not depending on specific linkage disequilibrium patterns in the population and the causative mutation can reflect the caused genetic variance directly. Including such causative markers in the genomic prediction equations should lead to more stable and reliable marker effect estimates. This would be favorable for prediction in less related samples or when using estimated marker effects over a longer time or even over generations. It should also be favorable for multi-breed sets or for across breed prediction: At the moment the problem arising here is that haplotype phases are not consistent over the range of the given marker densities. Based on simulations de Roos *et al.* (2008) proposed that over 300'000 markers have to be available to assure consistent phases in dairy cattle breeds. However, in real data even with over 600'000 markers from the HD chip accuracy in prediction across breeds was not significantly better than with 50'000 markers (see ***Chapter 4*** or e.g. Harris *et al.*, 2011). Apart from the even higher marker density, sequence data would provide the advantage of having the causative mutation included what renders the existence of consistent phases unnecessary and should provide an improvement for the accuracy of genomic prediction. Admittedly, this idea assumes that many of the causative mutations explaining genetic variance are the same and explain roughly the same proportion of variance in the studied breeds (see ***Chapter 4*** for a discussion of this point).

Methods that allow variable selection (e.g. Bayesian methods) may have a great advantage when modeling so many variables with only a few loci being really causative from the scientific point of view, but implementation may become a great challenge in terms of computer memory and calculation time.

Sequence data provide more variants than just SNPs. Copy number variants or insertions and deletions may also be a source of genomic variation and should not be forgotten. Their proportion of the total genetic variation has to be quantified and if they are of relevant size methods have to be developed to include those variants in genomic prediction processes as well.

Even if it is not a particularly realistic design the simulation of Meuwissen & Goddard (2010) is the only one that deals with SNP data from whole genome sequence. The authors' conclusions are that genomic breeding values will be more consistent over generations after a marker estimation time point when using SNPs from sequence data. Furthermore, they found that when the causative mutations were included in the SNP set accuracy of prediction was 2.5-3.7% higher depending on the number of QTL modeled.

Based on all these theoretical assumptions, having all genetic variation available must influence the accuracy of genomic prediction in various ways. The next years will show if these hypotheses can hold in real dairy cattle data.

How are genetic architecture and accuracy of prediction linked?

In dairy cattle, different traits are relevant in the breeding schemes. It is known that these traits have not the same genetic background. For most of them, it is assumed that they are influenced by many small genes. Some QTL have been detected (e.g. Khatkar *et al.*, 2004; Muncie *et al.*, 2006; Pausch *et al.* 2011) for various traits, but most of them can explain only a small part of the genetic variance. The only exception is DGAT1 (Grisart *et al.*, 2004) on Chromosome 14 that is still segregating in many breeds and determines both performances in fat percentage and milk yield to a considerable extent.

Hayes *et al.* (2010) studied three traits that are assumed to have different genetic background: proportion of black coat color (which is assumed to be determined by only a few loci), fat percentage (determined by the DGAT1 effect but probably many other small QTL) and overall type, for which no major genes are known. They showed that a model allowing different variances per SNP (*BayesA* in this case) had advantages for genomic prediction in the two traits that were affected by (a) major gene(s). This is a trend that is generally observed (e.g. Daetwyler *et al.*, 2010) especially in many simulation studies where the number of QTL modeled is often very small and the effects of QTL are often very large compared to what was found in real data. It is undisputable that methods allowing different variances have advantages in these major gene cases, but in real life – except with traits influenced by DGAT1 – these methods seldom produce significantly higher prediction accuracies, but computational demands are much larger.

Hayes *et al.* (2010) also ranked SNPs by their absolute effect on the respective trait and used only subsets (x% highest ones) to predict genomic breeding values. For coat color, 95% of the accuracy obtained with the full SNP set could be reached with very few SNPs. For overall type, 2'000 SNPs led to 90% of the potential accuracy with the full SNP set.

These results show that different genetic background require different numbers of SNPs for accurate prediction. Hayes *et al.* (2010) selected SNPs based on effects with *BayesA* and afterwards rerun *BayesA* with the SNP subset to figure out how many SNPs are necessary for specific traits. For such studies, *BayesR* (described in ***Chapter 4***) may provide additional information on the genetic architecture and the number of SNPs that will be necessary per trait to keep a specific accuracy level.

Respecting the different genetic architecture, one can hypothesize that lowering the number of SNPs randomly may have the opposite effect regarding the accuracy of prediction than what Hayes *et al.* (2010) described. For traits for which only a few SNPs with high effects capture a large proportion of the variance, missing a specific effect because of randomly narrowing down the marker density will have a strong impact. This should be in contrast to a trait where many loci influence the performance and are distributed over the whole genome.

To test this hypothesis, an additional study using a data set of 5'024 Holstein Friesian bulls with genotypes on 42'551 SNPs was conducted. Two different traits were chosen: fat percentage and somatic cell score. Genomic prediction using GBLUP was performed with different SNP densities, namely one SNP per 68, 136, 273 and 545 kb, respectively (i.e. thinning factor 1, 2, 4, 8 in Figure 3). Genomic prediction accuracy measured as the correlation between genomic breeding values and conventional breeding values was assessed in a five-fold cross-validation in three replicates with random assignment to the folds. Figure 3 shows that the accuracy drops in both traits, but the factor of decrease is clearly larger with fat percentage (strongly influenced by DGAT1) than with somatic cell score (no major genes) what may emphasize the hypothesis. Moser *et al.* (2010) showed genomic prediction results for the traits milk traits and overall type. They used a trait-specific subset (SNPs with highest effects on the respective trait) and subsets with evenly spaced SNPs. Down to a number of 3'000 SNPs they did not see a great difference in prediction accuracy when using evenly spaced or trait specific SNPs. Below this threshold, the trait specific subset outperformed the evenly spaced SNP set, suggesting that at this low marker density evenly spaced SNPs capture too less genetic variance.

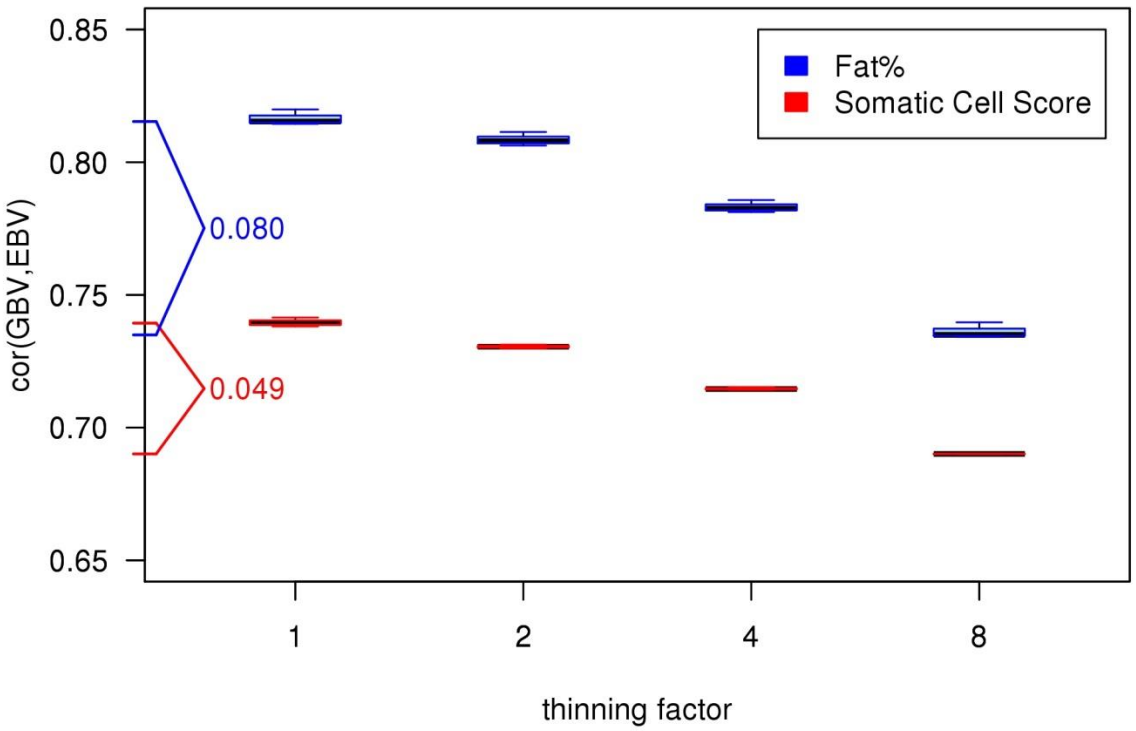


Figure 3: Correlation between predicted genomic breeding values and conventional breeding values for different SNP sets for the traits somatic cell score and fat percentage. For thinning factor 1, 42'551 SNPs were used while for thinning factor 2, 4, 8 a half, a quarter and an eighth of 42'551 SNPs were used.

For traits with different genetic architecture, it may be that the threshold from which on prediction accuracy decreases with evenly spaced SNPs varies in its absolute value and in the marker density it occurs. For Brown Swiss, the values were very similar for milk yield and somatic cell score in the study in ***Chapter 5***, but one has to recall that DGAT1 is almost fixed in Brown Swiss (e.g. Kaupe *et al.*, 2004).

In general, it can be directly seen from the results of ***Chapter 5*** that the number of SNPs needed to obtain this threshold depends not only on the length of the genome and the effective population size. Both values were very similar in the studied populations, but the threshold was reached with Brown Swiss with less than 30'000 SNPs while there was no threshold observable with Holstein Friesian up to over 40'000 SNPs. Apart from different other factors the different total sizes of the data sets and the choice of the studied traits may also have an impact on the obtained thresholds. To further determine these factors it will thus be worthwhile to repeat all k-fold cross-validation scenarios with data sets of the same size and same total number of SNPs for different traits once the data basis has become available.

Main conclusions from this thesis

In this thesis, different influence factors on the level of accuracy of genomic prediction in dairy cattle were studied. The main conclusions from the previous chapters can thus be summarized as:

1. The choice of the validation schemes clearly influences the obtained level of accuracy and should be taken into account when comparing results from different studies.

2. Relationships between individuals clearly influence the accuracy of genomic prediction and (cross-)validation schemes have to be adapted to mimic this situation correctly.

3. The training set size influences the accuracy to a high degree, but the necessary number of individuals needed for a pre-defined level of accuracy is different in different breeds and for different traits within a breed.

4. Methods used for genomic prediction produce very similar results at least with the marker densities used so far. However, specific Bayesian methods may have advantages by providing more information about the genetic architecture.

5. The high density panel has not fulfilled the expectations in the genomic prediction context.

6. The accuracy of prediction is linear to the natural logarithm of the marker density up to a population specific threshold. Increasing the marker density beyond this threshold will not lead to higher accuracy of genomic prediction.

7. Describing the effective number of independently segregating segments in the genome just based on the effective population size and the genome length is critical since it seems to vary between traits within breeds and between breeds with similar effective population size.

REFERENCES

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. *J. Anim. Breed. Genet.* **129:**345-358.

Brøndum, R. F., Ma, P., Lund, M. S., and Su, G. (2012): *Short communication:* Genotype imputation within and across Nordic cattle breeds. *J. Dairy Sci.* **95:**6795-6800.

Clark, S. A., Hickey, J. M., Daetwyler, H. D., and van der Werf, J. H. J. (2012): The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* **44:**4.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010): The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics* **185:**1021-1031.

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013): Genomic Prediction in Animal and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* **193:**347-365.

Dassonneville, R., Baur, A., Fritz, S., Boichard, D., and Ducrocq, V. (2012): Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. *Genet. Sel. Evol.* **44:**40.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008): Linkage Disequilibrium and Persistence of Phase in Holstein- Friesian, Jersey and Angus Cattle. *Genetics* **179:**1503–1512.

Edel, C., Neuner, S., Emmerling, R., and Götz, K. U. (2012): A Note on using ‚Forward Prediction' to Assess Precision and Bias of Genomic Predictions. *Interbull Bull.* **46:**16-19.

Fries, R., Pausch, H., Jansen, S., Aigner, B., and Wysocki, M. (2012): Assessment of the genomic variation in a cattle population by low-coverage re-sequencing. *Book of Abstracts of the 63rd EAAP* (Bratislava), p. 353.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., Mni, M., Simon, P., Frère, J.-M., Coppieters, W., and Georges, M. (2004): Genetic and functional confirmation of the causality of the *DGAT1 K232A* quantitative trait nucleotide in affecting milk yield and composition. *Proc. Natl. Acad. Sci. USA* **101:**2398-2403.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010): The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* **42:**5.

Harris, B. L., Creagh, F. E., Winkelman, A. M., and Johnson, D. L. (2011): Experiences with the Illumina High Density Bovine BeadChip. *Interbull Bull.* **44:**3-7.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., and Goddard, M. E. (2010): Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. *PLoS Genet.* **6:**e1001139.

Hayes, B., Anderson, C., Daetwyler, H., Fries, R., Guldbrandtsen, B., Lund, M., Boichard, D., Stothard, P., Veerkamp, R., Hulsegge, I., Rocha, D., Van Tassell, C., Coote, D., Goddard, M., and The 1000 Bull Genomes Consortium (2012): Towards genomic prediction from genome sequence data and the 1000 bull genomes project. *Book of Abstracts of ICQG 2012* (Edinburgh), p. 55.

Kaupe, B., Winter, A., Fries, R., and Erhardt, G. (2004): *DGAT1* polymorphism in *Bos indicus* and *Bos taurus* cattle breeds. *J. Dairy Res.* **71:**182-187.

Khatkar, M. S., Thomson, P. C., Tammen, I., and Raadsma, H. W. (2004): Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genet. Sel. Evol.* **36:**163-190.

Khatkar, M. S., Moser, G., Hayes, B. J., and Raadsma, H. W. (2012): Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. *BMC Genomics* **13:**538.

Legarra, A., and Ducrocq, V. (2012): Computational strategies for national integration of phenotypic, genomic and pedigree data in a single-step best linear unbiased prediction. *J. Dairy Sci.* **95:**4629-4645.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P., Sonstegard, T. S., and Van Tassell, C. P. (2009): Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS ONE* **4:**e5350.

Meuwissen, T., and Goddard, M. (2010): Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics* **185:**623-631.

Moser, G., Khatkar, M. S., Hayes, B. J., and Raadsma, H. W. (2010): Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genet. Sel. Evol.* **42:**37.

Muncie, S. A., Cassady, J. P., and Ashwell, M. S. (2006): Refinement of quantitative trait loci on bovine chromosome 18 affecting health and reproduction in US Holsteins. *Anim. Genet.* **37:**273-275.

Pausch, H., Flisikowski, K., Jung, S., Emmerling, R., Edel, C., Götz, K.-U., and Fries, R. (2011): Genome-Wide Association Study Identifies Two Major Loci Affecting Calving Ease and Growth-Related Traits in Cattle. *Genetics* **187:**289-297.

Pryce, J. E., and Daetwyler, H. D. (2012): Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim. Prod. Sci.* **52:**107-114.

Pryce, J. E., Arias, J., Bowman, P. J., Davis, S. R., Macdonald, K. A., Waghorn, G. C., Wales, W. J., Williams, Y. J., Spelman, R. J., and Hayes, B. J. (2012): Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. *J. Dairy Sci.* **95:**2108-2119.

Pszczola, M., Strabel, T., Mulder, H. A., and Calus, M. P. L. (2012): Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* **95:**389-400.

Segelke, D., Chen, J., Liu, Z., Reinhardt, F., Thaller, G., and Reents, R. (2012): Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. *J. Dairy Sci.* **95:**5403-5411.

Solberg, T. R., Heringstad, B., Svendsen, M., Grove, H., and Meuwissen, T. H. E. (2011): Genomic Predictions for Production- and Functional Traits in Norwegian Red from BLUP Analyses of Imputed 54K and 777K SNP Data. *Interbull Bull.* **44:**240-243.

Su, G., Brøndum, R. F., Ma, P., Guldbrandtsen, B., Aamand, G. P., and Lund, M. S. (2012): Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *J. Dairy Sci.* **95:**4657-4665.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., and Weigel, K. A. (2011): Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* **43:**10.

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., Sonstegard, T. S., Connor, E. E., Winters, M., van Kaam, J. B. C. H. M., Valentini, A., Van Doormaal, B. J., Faust, M. A., and Doak, G. A. (2013): Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* **96:**668-678.

Wientjes, Y. C. J., Veerkamp, R. F., and Calus, M. P. L. (2013): The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics* **193:**621-631

Wiggans, G. R., Cooper, T. A., VanRaden, P. M., Olson, K. M., and Tooker, M. E. (2012a): Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *J. Dairy Sci.* **95:**1552-1558.

Wiggans, G. R., VanRaden, P. M., and Cooper, T. A. (2012b): *Technical note:* Adjustment of all cow evaluations for yield traits to be comparable with bull evaluations. *J. Dairy Sci.* **95:**3444-3447.

Acknowledgements

I would like to thank

Prof. Dr. Henner Simianer for acting as my main supervisor and giving me the opportunity to work on various fascinating topics. Thank you for the great support and motivation during the last years.

Prof. Dr. Georg Thaller for taking over the co-reference of this thesis.

Prof. Dr. Ben Hayes for giving me the opportunity to visit his group for half a year. Thank you for sharing your ideas with me and enabling me to work on very actual topics.

Ben, **Mike**, **Jennie**, **Hans**, **Kath**, **Phil** and the whole team of the Computational Biology Group at the VABC in Melbourne for a very warm welcome, for fruitful discussions and cooperation and for inviting me to various social activities.

Rasmus, **Emma**, **Michèle**, **Didier** and **Nicolas** for sharing parts of the time in Bell City with me and for nice chats and discussions during our daily trips to the office.

Anne, **Reza**, **Florence** and **Eduardo** for their friendship, for many chats, scientific discussions and leisure activities. You all have supported me in various ways.

my parents, **Helmut** and **Felix** for their love, their belief in me, their understanding and their never-ending support.

Curriculum Vitae

Name:	**Malena Erbe**
Date of Birth:	21.02.1985
Place of Birth:	Roth
Nationality:	German

Education:

Dec 2009 – current	**PhD student**, Georg-August-University Göttingen, Germany
Oct 2007 – April 2009	**M. Sc. "Animal Sciences"**, Georg-August-University Göttingen, Germany *Thesis: "Power and robustness of three whole genome association mapping approaches in selected populations"*
Oct 2004 – Sept 2007	**B. Sc. "Land Management"**, Technische Universität München, Germany *Thesis: "Analyse eines metabolisch relevanten Parameters auf der Basis von routinemäßig erhobenen Milchprüfdaten"*
Sept 1995 – June 2004	Secondary school, Hilpoltstein Degree: Allgemeine Hochschulreife
Sept 1991 – July 1995	Primary school, Hilpoltstein

Work History:

Dec 2009 – current	**Research assistant** Department of Animal Sciences, Animal Breeding and Genetics Group, Georg-August-University Göttingen, Germany
Jun 2009 – Nov 2009	**Research assistant** Chair of Animal Breeding, Technische Universität München, Germany
Dec 2008 – Apr 2009	**Student assistant** Department of Crop Sciences, Plant Breeding Unit, Georg-August-University Göttingen, Germany
Apr 2008 – Jun 2008	**Student assistant** Department of Animal Sciences, Animal Breeding and Genetics Group, Georg-August-University Göttingen, Germany

Research stay abroad:

March 2011 – Aug 2011	**Computational Biology group** Department of Primary Industries (Victoria), Melbourne, Australia (Leader: Associate Professor Dr. Ben Hayes)

Further publications

Albrecht, T., Wimmer, V., Auinger, H.-J., **Erbe, M.**, Knaak, C., Ouzunova, M., Simianer, H., and Schön, C.-C. (2011): Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* **123:**339-350.

Chen, J., Ytournel, F., **Erbe, M.**, Pimentel, E. C. G., and Simianer, H. (2010): Use of Mendelian Sampling Terms in Genomic Models. *Book of Abstracts of the 9th WCGALP (Leipzig)*, p. 272.

Erbe, M., Hayes, B. J., Bowman, P. J., Simianer H., and Goddard, M.E. (2012): Genomic prediction within and between dairy cattle breeds with an imputed high density marker panel. *Book of abstracts of the 63rd EAAP (Bratislava)*, p. 128.

Erbe, M., Reinhardt, F., and Simianer, H. (2011): Empirical determination of the number of independent chromosome segments based on cross-validated data. *Book of abstracts of the 62nd EAAP (Stavanger),* p. 115.

Erbe, M., Ytournel, F., Pimentel, E. C. G., Sharifi, A. R., and Simianer, H. (2011): Power and robustness of three whole genome association mapping approaches in selected populations. *J. Anim. Breed. Genet.* **128:**3-14.

Ober, U., **Erbe, M.**, Long, N., Porcu, E., Schlather, M., and Simianer, H. (2011): Predicting Genetic Values: A Kernel-Based Best Linear Unbiased Prediction with Genomic Data. *Genetics* **188:**695-708.

Pimentel, E. C. G., **Erbe, M.**, König, S., and Simianer, H. (2011): Genome partitioning of genetic variation for milk production and composition traits in Holstein cattle. *Front. Gene.* **2:**19.

Wiebelitz, J., **Erbe, M.**, and Simianer, H. (2012). Genauigkeit der genomischen Zuchtwertschätzung in unterteilten Populationen. *Tagungsband DGfZ-/GfT-Gemeinschaftstagung*: A14.

Ytournel, F., Teyssèdre, S., Roldan, D., **Erbe, M.**, Simianer, H., Boichard, D., Gilbert, H., Druet, T., and Legarra, A. (2012): LDSO: a program to simulate pedigrees and molecular information under various evolutionary forces. *J. Anim. Breed. Genet.* **129:**417-421.