Institut für Nutztierwissenschaften der Universität Hohenheim
Fachgebiet Tiergenetik und Züchtung
Prof. Dr. Jörn Bennewitz

Investigations on methodological and strategic aspects of genomic selection in dairy cattle using real and simulated data

DISSERTATION

zur Erlangung des Doktorgrades
der Fakultät Agrarwissenschaften
der Universität Hohenheim

vorgelegt von

LAURA ISABEL PLIESCHKE

M. Sc. (Agr.)
aus Leverkusen,
Nordrhein-Westfalen

2017

Die Dissertation wurde mit dankenswerter Unterstützung der Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. (ASR) angefertigt.

Institut für Nutztierwissenschaften der Universität Hohenheim
Fachgebiet Tiergenetik und Züchtung
Prof. Dr. Jörn Bennewitz

Investigations on methodological and strategic aspects of genomic selection in dairy cattle using real and simulated data

DISSERTATION

vorgelegt von

LAURA ISABEL PLIESCHKE

M. Sc. (Agr.)
aus Leverkusen,
Nordrhein-Westfalen

2017

Die vorliegende Arbeit wurde am 14. März 2017 von der Fakultät Agrarwissenschaften der Universität Hohenheim als „Dissertation zu Erlangung des Grades eines Doktors der Agrarwissenschaften“ angenommen.

Tag der mündlichen Prüfung:	10. Juli 2017
Leiter der Prüfung:	Prof. Dr. Markus Rodehutscord
Berichterstatter, 1. Prüfer:	Prof. Dr. Jörn Bennewitz
Mitberichterstatter, 2. Prüfer:	Prof. Dr. Henner Simianer
3. Prüfer:	Prof. Dr. Hans-Peter Piepho

Die Dissertation wurde mit dankenswerter Unterstützung der Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. (ASR) angefertigt.

Contents

General Introduction 4

Chapter One 9

Chapter Two 24

Chapter Three 36

General Discussion 54

General Summary (English) 71

General Summary (German) 73

Danksagung 75

Eidesstattliche Versicherung 76

Curriculum Vitæ 78

General Introduction

Genomic selection (Meuwissen et al., 2001) is a widely used tool that brought large changes to plant and animal breeding in recent years. Especially for dairy cattle breeds, this methodology has great advantages and therefore has been implemented in the breeding programs of the most important dairy cattle breeds (e.g. Hayes et al., 2009; Pryce and Daetwyler, 2012). In Germany, genomic selection was implemented in 2010 for the Holstein Friesian breed (Reinhardt et al., 2011) and in 2011 for the breeds Fleckvieh and Brown Swiss (Edel et al. 2011; LfL, 2011).

One advantage of genomic selection is its ability to predict the individual Mendelian sampling deviation without the knowledge of own performance or offspring performance (Pryce and Daetwyler, 2012). In the last few years, the standard selection practice for bulls used in artificial insemination (AI) in dairy cattle has changed from the time-consuming conventional procedure of progeny testing to genomic selection, and selection decisions have been made much earlier. Due to this reduction of the generation interval, the annual genetic gain can be increased (or even doubled) by using genomic selection (Schaeffer, 2006). This reduction of the generation interval is directly related to the use of genomically tested young bulls in the population and is therefore linked to the acceptance of these bulls among breeders and producers. However, comparing reliabilities of the genomic method with those from the progeny testing (Powell et al., 2003), it is important to note that although genomic selection provides relatively reliable estimates for many traits, the more reliable estimates for the selection of AI bulls can be achieved with the conventional system of progeny testing.

Among other things the quality of genomic breeding values depends on the size and composition of the so called reference population. Animals belonging to the reference population are both phenotyped and genotyped and are used to estimate the single nucleotide polymorphism (SNP) effects that are needed to predict genomic breeding values of future selection candidates (Goddard and Hayes, 2007; van Grevenhof and van der Werf, 2015). During the course of using genomic selection in the breeding program the composition of the reference population changes: fewer bulls are selected (Buch et al., 2011; Pryce et al., 2008) and the number of sires put in service each year have almost halved compared to the pre-genomic period. As a consequence fewer AI bulls enter the reference population and, in addition, these bulls are already pre-selected, which

was not the case in the conventional system (Schaeffer, 2014). As a consequence of these two aspects it can be expected that without taking appropriate actions the reliabilities of genomic breeding values will eventually deteriorate and predictions will eventually be biased due to the effect of pre-selection.

Despite these problems the new methodology has many advantages and there has been a continuous effort to develop further the scientific understanding of the mechanisms involved.

This study was part of a project called "Zukunftswege" funded by the Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. Within this project there are different work packages defined which deal with the further development of genomic breeding value estimation for the Fleckvieh and Brown Swiss cattle. Various issues should be dealt with, including taking into account the genetic diversity of the two breeds and ensuring the quality of genomic breeding value estimation using different methods.

The presented work thus deals with two different approaches to improve genomic selection and to ensure the reliability and unbiasedness of genomic breeding values. Chapter one covers a theoretical aspect of genomic selection models, specifically the question of how variation between the allele frequencies in subpopulations of the same breed influences genomic predictors and how this relates to the role of genetic groups in genomic BLUP. Chapter two and chapter three deal with the question of how to preserve and improve the quality of genomic prediction by a great enlargement of the reference population using the genotypes and phenotypes of female animals.

Chapter one shows a simple method to decompose the genomic relationship matrix $\mathbf{G}$ into two independent covariance matrices, where $\mathbf{G}^*_{\mathbf{A}}$ describes the covariance that results from systematic differences in allele frequencies between groups at the pedigree base and $\mathbf{G_s}$ describes genomic relationships corrected for these differences. By the use of this decomposition and with the help of F_{st} statistics (Weir and Cockerham, 1984), it is possible to assign genetic distances between subgroups within the same population to either a heterogeneous genetic structure already present at the base of the pedigree and/or to breed divergence during the breeding process. Three models were tested in a forward prediction on six traits using Brown Swiss and dual-purpose Fleckvieh cattle data to examine the relative importance of the genetic heterogeneity in the pedigree base.

The aim of chapter two was to explore the potential of increasing the reliability of breeding values of young selection candidates by genotyping a fixed number of first-crop daughters of each reference bull from one or two generations in a balanced and regular system of genotyping and adding these to the reference population. A basic population scenario that mimics the situation in dual-purpose Fleckvieh cattle with respect to important key parameters was developed using stochastic simulation. Several scenarios were compared with respect to model-derived reliabilities, validation reliabilities and unbiasedness of predicted values for selection candidates. In the

base scenario the reference set consisted of only genotyped bulls. This reference set was then successively extended by including increasing numbers of daughter genotypes and phenotypes. In the most extended design, with 200 daughters per sire genotyped from two generations, SNP effects were estimated from a reference set of 420,000 cows and 4200 bulls.

In chapter three the approach of chapter two was extended to answer some additional questions. First, the results were complemented for a trait with low heritability but all other aspects were as in chapter two. Additionally the subject of so called 'new traits' was covered. This chapter is therefore structured into two main parts: 1) an 'old trait with low heritability' section and 2) a 'new trait with low heritability' section. In the case of 'new traits', phenotyping was assumed to have started only a few generations back and therefore only a limited number of phenotypes on cows were available. The assignment of animals to the reference population in this situation was done in two ways: either genotypes were available on sires of phenotyped cows only, or genotypes were available and used on phenotyped cows themselves.

All investigations were done having the two largest Bavarian cattle breeds in mind: Chapter one was based on data from Fleckvieh and Brown Swiss cattle and chapter two and three with simulated data resembling the genetic composition and population structure of Fleckvieh. The following chapters present methodological and strategic possibilities for the improvement of some aspects of genomic selection. Other aspects presented in the literature include for example technical issues like chip densities or the so called next generation sequencing. These additional issues will be addressed at the end of this work, together with some additional methodological and strategic aspects which are related to the following chapters.

References

Buch LH, Kargo M, Berg P, Lassen J, Sorensen C. The value of cows in the reference populations for genomic selection of new functional traits. Animal. 2012;6:880-6.

Edel C, Schwarzenbacher H, Hamann H, Neuner S, Emmerling R, Götz KU. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. Interbull Bull. 2011;44:152-6.

Goddard ME, Hayes BJ. Genomic selection. J. Anim. Breed. Genet. 2007;124:323-30.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009;92:433–43.

Landesanstalt für Landwirtschaft. Auch für das Deutsche Braunvieh beginnt das Zeitalter der genomischen Selektion. Pressemitteilung. 2011

Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

Powell RL, Norman HD, Sanders AH. Progeny testing and selection intensity for Holstein bulls in different countries. J Dairy Sci. 2003;6:3386-93.

Pryce JE, Daetwyler HD. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science. 2012;52:107-14.

Pryce JE, Hayes BJ, Goddard ME. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. In Proceedings of the 36th ICAR Biennial Session: 16-20 June 2008; Niagara Falls.

Reinhardt F, Liu Z, Seefried F, Rensing S, Reents R. Genomische Selektion: Stand der Implementierung bei Deutschen Holsteins. Züchtungskunde. 2011:83;248-56.

Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 2006;123:218-23.

Schaeffer LR. Is the animal model obsolete in dairy cattle? University of Guelph, personal communication to the Animal Genetics Discussion Group (AGDG). 2014.

Van Grevenhof IEM, van der Werf JHJ. Design of reference populations for genomic selection in crossbreeding programs. Genet Sel Evol. 2015;47:14.

Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of population structure. Evolution. 1984;38:1358-70.

Chapter One

A simple method to separate base population and segregation effects in genomic relationship matrices

Plieschke et al. Genetics Selection Evolution (2015) 47:53
DOI 10.1186/s12711-015-0130-8

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

A simple method to separate base population and segregation effects in genomic relationship matrices

Laura Plieschke[1*], Christian Edel[1], Eduardo CG Pimentel[1], Reiner Emmerling[1], Jörn Bennewitz[2] and Kay-Uwe Götz[1]

Abstract

Background: Genomic selection and estimation of genomic breeding values (GBV) are widely used in cattle and plant breeding. Several studies have attempted to detect population subdivision by investigating the structure of the genomic relationship matrix **G**. However, the question of how these effects influence GBV estimation using genomic best linear unbiased prediction (GBLUP) has received little attention.

Methods: We propose a simple method to decompose **G** into two independent covariance matrices, one describing the covariance that results from systematic differences in allele frequencies between groups at the pedigree base ($\mathbf{G}^*_A$) and the other describing genomic relationships ($\mathbf{G}_S$) corrected for these differences. Using this decomposition and F_{st} statistics, we examined whether observed genetic distances between genotyped subgroups within populations resulted from the heterogeneous genetic structure present at the base of the pedigree and/or from breed divergence. Using this decomposition, we tested three models in a forward prediction validation scenario on six traits using Brown Swiss and dual-purpose Fleckvieh cattle data. Model 0 (M0) used both components and is equivalent to the model using the standard **G**-matrix. Model 1 (M1) used $\mathbf{G}_S$ only and model 2 (M2), an extension of M1, included a fixed genetic group effect. Moreover, we analyzed the matrix of contributions of each base group (**Q**) and estimated the effects and prediction errors of each base group using M0 and M1.

Results: The proposed decomposition of **G** helped to examine the relative importance of the effects of base groups and segregation in a given population. We found significant differences between the effects of base groups for each breed. In forward prediction, differences between models in terms of validation reliability of estimated direct genomic values were small but predictive power was consistently lowest for M1. The relative advantage of M0 or M2 in prediction depended on breed, trait and genetic composition of the validation group. Our approach presents a general analogy with the use of genetic groups in conventional animal models and provides proof that standard GBLUP using **G** yields solutions equivalent to M0, where base groups are considered as correlated random effects within the additive genetic variance assigned to the genetic base.

Background

Genomic selection [1] and estimation of genomic breeding values (GBV) are currently used for many cattle populations. Genomic best linear unbiased prediction (GBLUP) using relationships estimated based on SNPs (single nucleotide polymorphisms) has been established as one of the most prominent methods for practical applications [2]. The question of how and to what extent population subdivision affects the genomic relationship matrix and genomic predictions was not addressed until applications of GBLUP across breeds or in admixed or crossbred populations were proposed e.g. [3–5]. However, several authors have shown that genomic relationship matrices can be used to detect population subdivision and to calculate measures of genetic distances (e.g. F_{st}) [6, 7].

Conventional methods to estimate breeding values consider that animals with unknown parents belong to

* Correspondence: Laura.Plieschke@lfl.bayern.de
[1]Bavarian State Research Center for Agriculture, Institute of Animal Breeding, Prof.-Dürrwaechter-Platz 1, 85586 Poing-Grub, Germany
Full list of author information is available at the end of the article

an arbitrarily defined base population. Members of this base population are assumed to come from a single population with a mean breeding value of 0 and variance σ_a^2. Since this is rarely true in practical applications, many conventional methods to estimate breeding values include genetic groups or phantom parents [8–10] in the model. A more elaborated approach in the context of multi-breed evaluations was proposed by García-Cortés and Toro [11], who partitioned the elements of the covariance matrix of the additive values into a breed-source term and a segregation term.

In spite of the large number of studies that deal with the use of genetic groups in conventional models, only a few have investigated this issue within the framework of genomic models. Makgahlela et al. [12–14] tested models that accounted for breed effects and compared allele frequencies in subgroups of Nordic Red cattle. They showed that a model that included a fixed breed effect [12, 13] increased the reliability of direct genomic values (DGV) by 2 to 3 % [13] for an admixed Nordic Red population. In a follow-up investigation, they found that using breed- or subpopulation-specific allele frequencies to calculate the genomic relationship matrix (**G**) did not result in higher validation reliabilities, although accounting for specific allele frequencies in the calculation of **G** changed the estimated GBV of some individuals considerably [14]. Tsuruta et al. [15] proposed an approach to assign unknown parent groups in one-step GBLUP for US Holstein cattle data. Their approach can be described as an application of the model that fits standard fixed genetic groups within the context of one-step GBLUP. The question of whether and how population subdivision influences the **G**-matrix was not addressed.

A simulation study by Vitezica et al. [16] compared five BLUP methods and investigated the effect of selection and genome-wide evaluation methods (one-step and multi-step) on bias and accuracy of genomic predictions. They examined the problem of unequal genetic levels between genotyped and non-genotyped animals in the one-step GBLUP procedure, where the genomic relationship matrix **G** and the pedigree-based relationship matrix **A** are combined. They proposed a correction of **G** and concluded that one-step estimation with a corrected **G** results in unbiased estimates of GBV, which have a similar inflation rate and a higher accuracy than estimates obtained with other methods. Christensen [17] presented an alternative approach for one-step models. For admixed populations, he suggested that the pedigree-based relationship matrix should be adjusted by assuming a parametric structure for the relationships between animals in the base population and estimating those parameters. He argued that this approach would be easier to extend and simpler than developing an appropriate method of adjusting the matrix of genomic relationships of genotyped animals across breeds.

The effects of population subdivision on the structure of the genomic relationship matrix **G** have also been investigated in contexts other than when it is used to estimate GBV. There are numerous studies on the calculation of F_{st} statistics [6, 18] and principal component analysis (PCA), e.g. [19, 20], and corresponding extensions to the **G**-matrix [16]. These studies show that it is possible to detect population subdivision with **G** in the same manner as with **A**. This means that **G** includes information about population subdivision and that, in some cases, this information includes the genetic distance between potentially discriminable groups in the base population that is defined by the pedigree. Since base animals are rarely genotyped, these distances cannot be estimated directly. A simple and straightforward method to estimate allele frequencies in the base population was proposed by Gengler et al. [21] and is based on a mixed model approach. In this paper, we estimate allele frequencies in the base of different subpopulations that are present in our datasets and propose a method to separate the genomic relationship matrix (**G**) into two independent components: a base group ($\mathbf{G}_A^*$) component and a segregation ($\mathbf{G}_S$) component. Furthermore, we demonstrate that this decomposition leads to basically identical results as ordinary GBLUP. Finally, we examine models that either ignore the effects of base groups or that consider base groups as fixed effects.

Methods

Material

In total, 7965 genotyped Fleckvieh (FV) and 4257 genotyped Brown Swiss (BS) and 143 genotyped Original Braunvieh (OB) bulls were available for this study. BS and OB data were combined (hereafter called BS/OB, $n = 4400$) into a single dataset because these two subpopulations actually originated from a single breed. The term Brown Swiss is used to denote the modern Braunvieh, which resulted from an exchange of genetic material between Europe and North America. An OB animal is genetically characterized as a descendant of the old European Braunvieh population, with no or only minor genetic contributions from the reimported US Brown Swiss population. This labelling of OB animals within the European Braunvieh population is not necessarily applied in a uniform manner and small differences in the definition can occur between countries.

All animals were genotyped with the Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA). After removing SNPs with low call rates (<90 %), minor allele frequencies less than 2 %, or with a deviation from Hardy-Weinberg equilibrium with $P < 10^{-5}$, 37 718 and 41 254

SNPs were retained for the BS/OB and FV datasets, respectively. Available pedigrees for genotyped animals included 7802 and 16 357 records for the BS/OB and FV breeds, respectively. BS/OB base animals were assigned to nine groups (Table 1) according to origin and date of birth. Since the genetic distances between German, Austrian, Italian and Swiss BS base animals born before 1960 were small (results not shown), they were combined into one base group called EU_b. Base FV animals were assigned to 11 groups with nine groups assigned according to origin and date of birth and two groups assigned to the Red Holstein breed (Table 2).

We estimated DGV for three milk traits and three conformation traits from a dataset that was reduced for the last four years of phenotypic data (referred to as the reduced dataset). Daughter yield deviations (DYD) from the German-Austrian system [22] were used for FV bulls and deregressed MACE (multi-trait across country evaluations) proofs from Interbull [23] for BS/OB bulls. Deregression was done using the method proposed by Garrick et al. [24]. Group effects were not accounted for in the deregression. Traits analyzed were milk yield (MY), protein yield (PY), fat yield (FY), stature (STA), feet and legs (FL) and udder conformation (UD). These traits were *a priori* assumed to have a large genetic trend and/or to show considerable differences between base groups. DGV estimated from the reduced dataset were then compared to DYD and deregressed proofs from the corresponding April 2014 evaluations (current dataset) according to the guidelines of the Interbull GEBV test [25, 26]. In short, the validation group included bulls with no information on the offspring's performances in the reduced dataset but corresponding information in the current dataset. Current information was assumed to be sufficient for the test when the effective daughter contribution (EDC) [27] based on offspring performances was equal to at least 20. The remaining bulls from 2010 with an EDC of at least 1 were included into the training set (*Calib*).

Technically, we tested DGV by a weighted regression of current DYD or deregressed proofs of the animals in the validation group on their DGV estimated from the reduced set. The resulting test statistics are the intercept and slope (b) of this regression as measures of bias and the coefficient of determination (R^2) of this regression as a measure of the reliability of the DGV. The R^2 values were corrected for the uncertainty in DYD, as proposed by [28], i.e. they were divided by the average reliability of the DYD of validation bulls.

For presentation of results, we divided the animals of the validation group into different sub-groups. FV validation animals were assigned to two groups: animals from Germany-Austria (*DEA*) and *others*. BS validation animals were also divided into *DEA* and *others*, and OB validation animals were assigned to a third validation group (*OB*). Numbers of animals included in each validation group are in Table 3. The assignments of validation animals to origins used in this investigation for the purpose of illustration were mainly based on ISO country codes [29] and do not necessarily correspond to assignments based on analyses of genetic contributions from base groups.

Decomposition of G

Assume a common scenario in genomic prediction with *n* animals genotyped for *m* biallelic SNPs. Information on genotypes is collected in an *n* x *m* matrix **C**, using numerical coding that denotes the number of copies of the arbitrarily defined reference allele (0, 1, 2). Let $\mathbf{p}_T$ be the vector of estimated allele frequencies at the *m* SNPs, which for each SNP *j* were derived from genotyped animals.

$$\hat{p}_j = \frac{\sum_{i=1}^{n} C_{ij}}{2n} \quad (1)$$

A genomic relationship matrix $\mathbf{G}_T$ can be calculated and used in GBLUP using these "current" allele frequencies as:

$$\mathbf{G}_T = \frac{\mathbf{M}\mathbf{M}'}{\sum_{j=1}^{m} 2\hat{p}_j\left(1-\hat{p}_j\right)}, \quad (2)$$

where **M** is an *n* x *m* matrix of recoded genotypes, for which each row (= animal) *i* of the matrix of numerically coded genotypes **C** is manipulated in the following manner [30]:

$$\mathbf{M}_i = \mathbf{C}_i - 1 - 2(\mathbf{p}_T - 0.5). \quad (3)$$

Conceptually, this manipulation is equivalent to column-wise centering of **C** if current allele frequencies

Table 1 Number of animals per defined base group for the BS/OB population

	EU_b	DE_b	AT_b	CH_b	IT_b	US_{b1}	US_{b2}	OB_{b1}	OB_{b2}
Year	≤1960	>1960	>1960	>1960	>1960	≤1955	>1955	≤1960	>1960
Number	2093	1482	743	1281	413	489	445	458	398

BS = Brown Swiss and *OB* = Original Braunvieh, assignment was done by country and year of birth with the exception of the OB base groups, which were considered across countries: EU_b = European base group (born before 1960), DE_b = German base group (born after 1960), AT_b = Austrian base group (born after 1960), CH_b = Swiss base group (born after 1960), IT_b = Italian base group (born after 1960), US_{b1} = American base group (born before 1955), US_{b2} = American base group (born after 1955), OB_{b1} = Original Braunvieh base group (born before 1960), OB_{b2} = Original Braunvieh base group (born after 1960)

Table 2 Number of animals per defined base group for FV

	DE_{b1}	DE_{b2}	DE_{b3}	DE_{b4}	HOL_{b1}	HOL_{b2}	AT_b	CZ_b	CH_b	FR_b	Div_b
Year	<1960	≥1960 < 1970	≥1970 < 1980	≥1980	<1960	≥1960	All	All	All	All	All
Number	1368	6055	1661	773	528	427	3452	977	183	228	705

FV = Fleckvieh; assignment was done by country and year of birth with the exception of the Red Holstein and the diverse base groups, which were considered across countries: DE_{b1} = German base group (born before 1960), DE_{b2} = German base group (born between 1960 and 1970), DE_{b3} = German base group (born between 1970 and 1980), DE_{b4} = German base group (born after 1980), HOL_{b1} = Red Holstein base group (born before 1960), HOL_{b2} = Red Holstein base group (born after 1960), AT_b = Austrian base group, CZ_b = Czech base group, CH_b = Swiss base group, FR_b = French base group, DIV_b = base groups with animals with other countries of origin

are used and if each marker is in Hardy-Weinberg equilibrium in the genotyped population.

Assume a subdivision of the genotyped population into *g* groups that systematically differ in allele frequencies, as indicated for example by sufficiently high F_{st} values [31, 32]. Define a *g* x *m* matrix **P** of group-specific allele frequencies that are derived by applying Equation (1) within each group. Using these group-specific allele frequencies, the vector of genotypes for each animal can then be centered by applying Equation (3) using the allele frequencies of the group that it is assigned to. Thus, for animal *i* assigned to group *k* with group-specific allele frequencies $\mathbf{p}_k$, the corresponding row in **C** is manipulated as:

$$\mathbf{M}_i^* = \mathbf{C}_i - 1 - 2(\mathbf{p}_k - 0.5).$$

A **G**-matrix corrected for specific allele frequencies for different groups can then be calculated as:

$$\mathbf{G}_S = \frac{\mathbf{M}^*\mathbf{M}^{*'}}{\sum_{j=1}^{m} 2\hat{p}_j\left(1-\hat{p}_j\right)}, \quad (4)$$

Table 3 Number of animals per validation group for the BS/OB and FV populations and the seven traits considered

		Training set	Validation set		
			DEA	*others*	*OB*
BS/OB	MY	3262	416	346	8
	PY	3262	416	346	8
	FY	3262	416	346	8
	STA	3535	464	350	51
	FL	3551	461	345	43
	UD	3550	458	349	43
			DEA	*others*	-
FV	MY	5276	2589	97	
	PY	5276	2581	97	-
	FY	5276	2581	97	
	STA	5956	2264	139	-
	FL	5956	2272	139	
	UD	5956	2272	139	-

BS = Brown Swiss, *OB* = Original Braunvieh and *FV* = Fleckvieh, *MY* = milk yield, *PY* = protein yield, *FY* = fat yield, *STA* = stature, *FL* = feet and legs, *UD* = udder conformation
Validation sets: *DEA* = German and Austrian validation animals; *others* = validation animals with other countries of origin; *OB* = Original Braunvieh validation animals

with the same denominator as in Equation (2), which is equivalent to expressing this part of the covariance relative to the overall covariance. The discarded component of the original covariance structure, which is caused by differences between group allele frequencies and overall frequencies, can be summarized in a matrix $\mathbf{G}_A$. Treating 2**P** as a matrix of average "genotypes" of groups, a matrix $\tilde{M}$ is calculated by manipulating each group's row g as follows:

$$\tilde{\mathbf{M}}_g = (2\mathbf{P})_g - 1 - 2(\mathbf{p}_T - 0.5).$$

Finally, $\mathbf{G}_A$ is calculated as $\tilde{\mathbf{M}}\tilde{\mathbf{M}}'$ divided by the same denominator as in Equations (2) and (4). The *g* x *g* matrix $\mathbf{G}_A$ can be treated and analyzed in the same manner as the standard **G**-matrix. It can be expanded to give an *n* x *n* matrix $\mathbf{G}_A^*$ based on:

$$\mathbf{G}_A^* = \mathbf{Q}\mathbf{G}_A\mathbf{Q}',$$

where **Q** is the matrix of genetic contributions of each base group to each animal, which can be calculated as:

$$\mathbf{Q} = \mathbf{T}\mathbf{Q}^*,$$

where **T** is a lower triangular matrix that results from decomposing **A** into **TDT'**, as described in [33], and $\mathbf{Q}^*$ is an *n x* g design matrix that assigns genotyped animals to groups. Despite this increase in dimensions, $\mathbf{G}_A^*$ still has rank (*g* – 1). Also, note that:

$$\mathbf{G}_T = \mathbf{G}_S + \mathbf{G}_A^*. \quad (5)$$

Although this decomposition is straightforward, its dependency on the current allele frequencies and the grouping of current animals causes some problems due to ambiguous genetic composition and might not be feasible under practical conditions since new genotypes have to be successively integrated into the system. To circumvent this problem, we propose to replace the current allele frequencies with estimates of base allele frequencies using the estimation procedure developed by Gengler et al. [21]. Using a pedigree that relates genotyped animals to a set of arbitrarily defined but usually ungenotyped base animals and calculating the conventional relationship matrix **A**, the vector of overall base allele frequencies is calculated as a generalized least

squares mean by solving the following equation for each marker *j* (column of **C**):

$$p_j^* = 0.5\left[\left(\mathbf{1}'\mathbf{A}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{A}^{-1}\mathbf{c}_j\right]. \quad (6)$$

Similar to conventional estimation of GBV, base animals can be grouped according to known or assumed population subdivisions and/or generations, when additional differentiation due to considerable genetic trend has to be taken into account. To estimate base group-specific allele frequencies, matrix **1** in Equation (6) is replaced by matrix **Q**. Matrices $\mathbf{G_T}$, $\mathbf{G_S}$ and $\mathbf{G}_A^*$ can then be calculated as described above, using estimates for global and group-specific base allele frequencies and again $\mathbf{G_T} = \mathbf{G_S} + \mathbf{G}_A^*$, as described above.

Models

In order to study the influence of different definitions of base group on the quality of prediction, we examined several models. The general model is a standard mixed animal model with:

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

where **y** is a vector of DYD or deregressed proofs of genotyped animals, **b** is the vector of fixed effects, **u** is the vector of random animal effects, incidence matrices **X** and **Z** relate observations to levels of **b** and **u**, respectively, and **e** is the residual effect. Furthermore, it is assumed that $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V_{yy}})$, $\mathbf{e} \sim N(\mathbf{0},\mathbf{V_e})$ and $\mathbf{u} \sim N(\mathbf{0},\mathbf{V_{uu}})$, with $\mathbf{V_{yy}} = \mathbf{V_{uu}} + \mathbf{V_e}$, $\mathbf{V_e}$ is $\mathrm{diag}(\mathbf{1}/\mathbf{w})^*\sigma_e^2$, where **w** is a vector of weights. The models to be compared are defined in the following.

Standard model (model 0, M0): $\mathbf{X} = \mathbf{1}$ and $\mathbf{V_{uu}} = \mathbf{G_T} \times \sigma_u^2$.
Model 1 (M1): $\mathbf{X} = \mathbf{1}$ and $\mathbf{V_{uu}} = \mathbf{G_S} \times \sigma_u^2$.
Model 2 (M2): $\mathbf{X} = [\mathbf{1} \mid \mathbf{Q}]$ and $\mathbf{V_{uu}} = \mathbf{G_S} \times \sigma_u^2$.

Note that M2 is equivalent to a model that fits standard fixed group effects [34]. Although genomic relationships corrected for unequal base allele frequencies ($\mathbf{G_S}$) are used in M2, it can be shown by least-squares theory that the solutions are identical to a model that uses $\mathbf{G_T}$, if the same matrix **Q** is used to estimate the base allele frequencies and to model the fixed group effects (see Appendix 1). Finally, it can be shown that using the standard genomic relationship matrix $\mathbf{G_T}$ in standard GBLUP (standard model, M0) in the presence of base groups that differ in allele frequencies gives solutions equivalent to the use of a more specific model with genetic groups as random effects and equal variances for the base group and the segregation effects (see Appendix 2), as in the following representation:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Q} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G_S}^{-1}\lambda & \mathbf{Z}'\mathbf{Q} \\ \mathbf{Q}'\mathbf{X} & \mathbf{Q}'\mathbf{Z} & \mathbf{Q}'\mathbf{Q} + \mathbf{G_A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Q}'\mathbf{y} \end{bmatrix},$$

where $\lambda = \sigma_u^2/\sigma_e^2$ and the final estimate for the breeding value is $\hat{\mathbf{u}} = \mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{u}}$. We calculated solutions for the standard model using this more specific model, which, in addition, allowed us to derive estimates for group effects and their prediction errors.

Models were tested in forward prediction by means of the test described in the sub-section Material. To better understand the factors that influence the predictive ability of a specific model for different validation datasets, we analyzed the matrix of base group contributions (**Q**) and derived base group estimates, as well as their prediction errors, using M0 and M2. Differences between group effect estimates were calculated and tested by formulating linear hypotheses.

Distance measures

We calculated F_{st} statistics to illustrate the effects of the proposed decomposition of **G**. F_{st} is a standard measure of genetic distance and can be calculated either by pairwise analysis of differences in allele frequencies between known or assumed subpopulations or breeds [18], or by direct calculation from relationship matrices [6] as:

$$F_{st} = \frac{\tilde{f} - \bar{f}}{1 - \bar{f}},$$

where $\tilde{f}$ is the mean coancestry over all subpopulations and $\bar{f}$ is the average coancestry within a given subpopulation. The term $1-\tilde{f}$ is the average diversity (heterozygosity) and depends on the coancestry within the given subpopulation. F_{st} values are primarily used as a tool to visualize substructures within groups of animals [6, 10, 35]. An F_{st} value of 0.05 can be interpreted as a strong indication of a relevant subdivision [31, 32].

Results

F_{st} statistics

To illustrate the effects of the decomposition of the **G**-matrix, we calculated F_{st} values for both components ($\mathbf{G_S}$ and $\mathbf{G}_A^*$) and for the total **G**-matrix for the 4400 BS/OB animals. Results of the F_{st} statistics are in Fig. 1. Comparison of distances calculated from $\mathbf{G}_A^*$ and $\mathbf{G_S}$ shows that population differences were primarily caused by genetic distances in the base population. A substantial genetic distance existed only between the OB group and the two other groups. This distance was present in both $\mathbf{G}_A^*$ and $\mathbf{G_S}$, but was considerably greater in $\mathbf{G}_A^*$. Interestingly, the distances in $\mathbf{G}_A^*$ and $\mathbf{G_S}$ acted additively

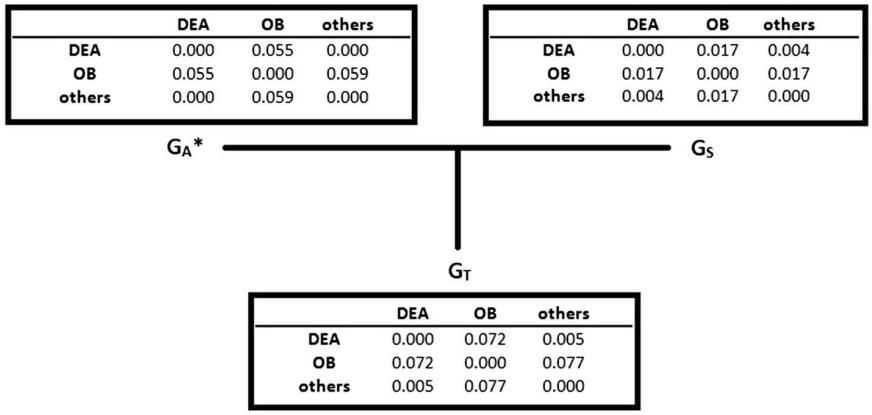

Fig. 1 F_{st} values of the base group term ($\mathbf{G}_\mathbf{A}^*$), the segregation term ($\mathbf{G}_\mathbf{S}$), and total $\mathbf{G}$ ($\mathbf{G}_\mathbf{T}$) for the 4400 BS/OB animals. BS = Brown Swiss and OB = Original Braunvieh

and their sum resulted in the distances calculated from $\mathbf{G}_\mathbf{T}$.

Forward prediction

Results of the forward prediction in terms of the coefficient of determination (R^2), the intercept (a), the slope (b) and corresponding standard errors are in Tables 4 and 5. For both breeds and across all traits, differences between models were small, but M1 consistently resulted in a lower R^2.

Brown Swiss and Original Braunvieh breeds

For the BS/OB data, we found a minimal advantage in terms of the R^2 for model M2 that fitted fixed groups.

Table 4 Results for the coefficient of determination (R^2) from the forward prediction for the BS/OB and FV populations for different models

BS/OB	Trait	M0 ($\mathbf{G}_\mathbf{A}^*$ and $\mathbf{G}_\mathbf{S}$)	M1 ($\mathbf{G}_\mathbf{S}$)	M2 ($\mathbf{G}_\mathbf{S}$ + fixed effects)
R^2	MY	0.416	0.386	0.421
	PY	0.409	0.370	0.417
	FY	0.388	0.349	0.395
	STA	0.499	0.382	0.505
	FL	0.234	0.216	0.220
	UD	0.416	0.394	0.410
FV				
R^2	MY	0.580	0.530	0.557
	PY	0.512	0.463	0.491
	FY	0.548	0.490	0.521
	STA	0.526	0.515	0.516
	FL	0.438	0.425	0.415
	UD	0.406	0.404	0.405

BS = Brown Swiss, *OB* = Original Braunvieh, and *FV* = Fleckvieh, *MY* = milk yield, *PY* = protein yield, *FY* = fat yield, *STA* = stature, *FL* = feet and legs, *UD* = udder conformation

Exceptions were for the traits FL and UD, here the standard random model M0 showed the highest R^2. Across traits, R^2 for M1 was 0.028 to 0.123 lower than that of the best model. Based on results in terms of slope, it should be noted that inflation of genomic predictions was lowest for conformation traits using model M1. For milk traits, the slope was slightly higher and estimates were thus less inflated with the random model M0 than with the fixed model M2.

Fleckvieh breed

Differences in R^2 between M0 and M2 ranged from 0.001 to 0.021. For all six traits, M0 resulted in a higher R^2 than the fixed group model M2. The R^2 achieved with M1 was always lower than that achieved with M0 and M2. Nevertheless, the difference in R^2 between M1 and M0 was only 0.002 for the UD trait. For the other traits, the R^2 that was achieved with M1 was between 0.011 and 0.058 lower than that with M0. Based on slope, model M0 was superior and always led to the lowest inflation of estimates for milk traits. For conformation traits, the fixed model M2 led to the lowest inflation. However, differences between models were relatively small in many cases (between 0.004 and 0.143).

Base group effects

We estimated base group effects based on M0 and M2. Properties of matrix $\mathbf{Q}$ always lead to linear dependencies and no unique solution can be achieved. However, significant differences between group estimates can be derived and tested using linear hypotheses. Results in Tables 6 and 7 are group differences estimated with M2.

Brown Swiss and Original Braunvieh breeds

In the BS/OB dataset, we defined nine different base groups that led to 36 possible contrasts between base

Table 5 Results for the intercept (a), slope (b) and its standard error (s.e.) from the forward prediction for the FV and BS/OB populations for different models

	Trait	M0 ($\mathbf{G}_A^*$ and $\mathbf{G}_S$)		M1 ($\mathbf{G}_S$)		M2 ($\mathbf{G}_S$ + fixed effects)	
BS/OB		a	b (s.e.)	a	b (s.e.)	a	b (s.e.)
	MY	85.551	**0.828** (0.035)	87.672	**0.813** (0.037)	85.091	**0.820** (0.035)
	PY	3.152	**0.768** (0.033)	3.221	**0.748** (0.035)	3.129	**0.765** (0.033)
	FY	3.202	**0.762** (0.035)	3.198	**0.753** (0.037)	3.178	**0.757** (0.034)
	STA	14.934	**0.854** (0.029)	−3.706	**1.020** (0.044)	18.807	**0.817** (0.028)
	FL	1.285	**0.979** (0.061)	−4.480	**1.032** (0.068)	24.889	**0.751** (0.059)
	UD	22.008	**0.786** (0.032)	9.036	**0.904** (0.038)	30.023	**0.711** (0.030)
FV		a	b (s.e.)	a	b (s.e.)	a	b (s.e.)
	MY	62.576	**0.660** (0.019)	76.031	**0.582** (0.018)	76.031	**0.619** (0.018)
	PY	3.213	**0.664** (0.019)	3.914	**0.593** (0.019)	3.914	**0.644** (0.019)
	FY	2.640	**0.734** (0.019)	3.696	**0.650** (0.019)	3.696	**0.729** (0.020)
	STA	0.046	**0.782** (0.024)	0.076	**0.774** (0.024)	0.076	**0.786** (0.025)
	FL	−0.082	**0.900** (0.036)	−0.179	**0.878** (0.036)	−0.179	**1.021** (0.038)
	UD	−0.013	**0.713** (0.033)	−0.031	**0.708** (0.033)	−0.031	**0.736** (0.040)

BS = Brown Swiss, *OB* = Original Braunvieh and *FV* = Fleckvieh; values for the slope are printed in bold and values for the standard error of the slope are shown in brackets. *MY* = milk yield, *PY* = protein yield, *FY* = fat yield, *STA* = stature, *FL* = feet and legs, *UD* = udder conformation

groups. Differences were tested for significance using t-tests. For the PY trait, significant differences were found for the majority of group contrasts and only 5 out of 36 differences were not significant. The largest difference was between the European base group (EU_b) and the German base group (DE_b) (−64.86). Estimates for DE_b were significantly larger than estimates for all other groups. Differences between the EU_b group and the other groups were also large but clearly negative. The smallest difference was between the Swiss base group (CH_b) and the older Original Braunvieh base group (OB_{b1}) (−0.05). The differences between the Austrian (AT_b) and the Italian (IT_b) base groups were relatively small in many cases.

For the STA trait, all group differences were significant, except the difference between the German base group (DE_b) and the younger American base group (US_{b2}). The patterns of differences were quite similar as for PY, although slightly different in magnitude for STA. The largest and smallest differences were also between

Table 6 Differences between base group effects estimated with the fixed model for the BS/OB population for protein yield above the diagonal and stature below the diagonal

	EU_b	DE_b	AT_b	CH_b	IT_b	US_{b1}	US_{b2}	OB_{b1}	OB_{b2}
	≤1960	>1960	>1960	>1960	>1960	≤1955	>1955	≤1960	>1960
EU_b	0	−64.86***	−22.52***	−13.97***	−19.36***	−26.06***	−29.90***	−14.01***	−45.54***
DE_b	25.48***	0	42.35***	50.90***	45.50***	38.80***	34.97***	50.85***	19.32***
AT_b	15.66***	−9.82***	0	8.55***	$3.15^{n.s.}$	$-3.55^{n.s.}$	$-7.38^{n.s.}$	8.50*	−23.03***
CH_b	1.21*	−24.27***	−14.45***	0	−5.40**	−12.10***	−15.93***	$-0.05^{n.s.}$	−31.58***
IT_b	19.63***	−5.85***	3.97***	18.42***	0	−6.70*	−10.53***	5.35*	−26.18***
US_{b1}	11.23***	−14.25***	−4.43***	10.02***	−8.40***	0	$-3.83^{n.s.}$	12.05**	−19.48***
US_{b2}	23.05***	$-2.43^{n.s.}$	7.39***	21.85***	3.42*	11.82***	0	15.88***	−15.65***
OB_{b1}	3.56***	−21.92***	−12.11***	2.35***	−16.08***	−7.67***	−19.50***	0	−31.53***
OB_{b2}	18.05***	−7.43***	2.38***	16.83***	−1.59**	6.82***	−5.01***	14.49***	0

BS = Brown Swiss and *OB* = Original Braunvieh; Protein yield (in kg); Stature (in cm); we calculated the differences row minus column, so negative values indicate superior horizontal groups and positive values indicate superior vertical groups. $^{n.s.}$ = not significant, * = ($p < .05$), ** = ($p < .01$), *** = ($p < .001$). EU_b = European base group (born before 1960), DE_b = German base group (born after 1960), AT_b = Austrian base group (born after 1960), CH_b = Swiss base group (born after 1960), IT_b = Italian base group (born after 1960), US_{b1} = American base group (born before 1955), US_{b2} = American base group (born after 1955), OB_{b1} = Original Braunvieh base group (born before 1960), OB_{b2} = Original Braunvieh base group (born after 1960)

Table 7 Differences between base group effects estimated with the fixed model for the FV population for protein yield above the diagonal and stature below the diagonal

	DE_{b1}	DE_{b2}	DE_{b3}	DE_{b4}	HOL_{b1}	HOL_{b2}	AT_b	CZ_b	CH_b	FR_b	Div_b
	<1960	≥1960 < 1970	≥1970 < 1980	≥1980	<1960	≥1960	All	All	All	All	All
DE_{b1}	0	−16.77***	1.06 $^{n.s.}$	−7.49***	−50.43***	−49.94***	18.21***	−32.21***	10.89***	−28.14***	49.76***
DE_{b2}	−0.29$^{n.s.}$	0	17.83***	9.28**	−33.66***	−33.17***	34.98***	−15.45***	27.66***	−11.37***	66.54***
DE_{b3}	−1.60$^{n.s.}$	−1.31 $^{n.s.}$	0	−8.55***	−51.49***	−51.00***	17.15***	−33.28***	9.82***	−29.20***	48.701***
DE_{b4}	−0.24$^{n.s.}$	0.05 $^{n.s.}$	1.36 $^{n.s.}$	0	−42.94***	−42.45***	25.70***	−24.73***	18.38***	−20.65***	57.25***
HOL_{b1}	5.16***	5.45***	6.76***	5.40***	0	0.49$^{n.s.}$	68.64***	18.21***	61.32***	22.29***	100.19***
HOL_{b2}	−1.49$^{n.s.}$	−1.20$^{n.s.}$	0.11$^{n.s.}$	−1.25$^{n.s.}$	−6.65***	0	68.15***	68.14***	60.83***	21.80***	99.70***
AT_b	−0.14 $^{n.s.}$	0.16$^{n.s.}$	1.46$^{n.s.}$	0.11$^{n.s.}$	−5.30***	1.35$^{n.s.}$	0	−50.43***	−7.32**	−46.35***	31.55***
CZ_b	−3.48***	−3.19$^{n.s.}$	−1.88$^{n.s.}$	−3.24$^{n.s.}$	−8.64***	−1.99$^{n.s.}$	−3.35$^{n.s.}$	0	43.11***	4.08$^{n.s.}$	81.98***
CH_b	−1.79$^{n.s.}$	−1.50$^{n.s.}$	−0.19$^{n.s.}$	−1.55$^{n.s.}$	−6.95***	−0.30$^{n.s.}$	−1.65$^{n.s.}$	1.69 $^{n.s.}$	0	−39.03***	38.88***
FR_b	0.22$^{n.s.}$	0.51$^{n.s.}$	1.82$^{n.s.}$	0.46$^{n.s.}$	−4.95***	1.71$^{n.s.}$	0.35$^{n.s.}$	3.70*	2.01$^{n.s.}$	0	77.91***
Div_b	−3.09***	−2.80$^{n.s.}$	−1.49$^{n.s.}$	−2.85*	−8.25***	−1.60$^{n.s.}$	−2.95*	0.39$^{n.s.}$	−1.30$^{n.s.}$	−3.31**	0

FV = Fleckvieh; Protein yield (in kg); Stature (in cm); we calculated the differences row minus column, so negative values indicate superior horizontal groups and positive values indicate superior vertical groups. $^{n.s.}$ = not significant, * = ($p < .05$), ** = ($p < .01$), *** = ($p < .001$). DE_{b1} = German base group (born before 1960); DE_{b2} = German base group (born between 1960 and 1970), DE_{b3} = German base group (born between 1970 and 1980), DE_{b4} = German base group (born after 1980), HOL_{b1} = Red Holstein base group (born before 1960), HOL_{b2} = Red Holstein base group (born after 1960), AT_b = Austrian base group, CZ_b = Czech base group, CH_b = Swiss base group, FR_b = French base group, DIV_b = base groups with animals with other countries of origin

EU_b and DE_b (25.48) and between the Swiss base group (CH_b) and the European base group (EU_b) (1.21), respectively.

Fleckvieh breed

For the FV breed, almost all group differences were significant for PY. The largest differences were between the older Red Holstein base group (HOL_{b1}) and the Austrian base group (AT_b), between the younger Red Holstein base group (HOL_{b2}) and AT_b and between HOL_{b2} and CZ_b (68.64, 68.15 and 68.14, respectively). The smallest difference was between the two Red Holstein base groups (0.49).

The situation for STA was almost the opposite. Only 16 group differences were significant, while 39 out of 55 differences were not significant. From these 16 significant differences, 10 were between the older Red Holstein base group (HOL_{b1}) and all other base groups.

Base group contributions

Analysis of the matrix of base group contributions (**Q**) revealed several general breed-specific aspects. In addition, it was possible to characterize the validation group, which can help interpretation of other results. Averages and standard deviations of base group contributions for the PY and STA traits are in Tables 8 and 9 for the two breeds.

Table 8 Results of the analysis of the Q-matrix for the BS/OB population

BS/OB		EU_b	DE_b	AT_b	CH_b	IT_b	US_{b1}	US_{b2}	OB_{b1}	OB_{b2}
Year		≤1960	>1960	>1960	>1960	>1960	≤1955	>1955	≤1960	>1960
Calib (3262)	m	0.02	0.02	0.01	0.01	0.01	0.24	0.62	0.03	0.03
	sd	0.04	0.05	0.03	0.03	0.03	0.07	0.12	0.07	0.06
DEA (416)	m	0.02	0.03	0.01	0.00	0.00	0.23	0.62	0.03	0.06
	sd	0.01	0.05	0.07	0.04	0.01	0.04	0.04	0.02	0.04
OB (8)	m	0.25	0.00	0.01	0.05	0.00	0.00	0.00	0.54	0.16
	sd	0.25	0.00	0.02	0.09	0.00	0.00	0.00	0.19	0.15
Others (346)	m	0.01	0.01	0.00	0.01	0.00	0.27	0.67	0.01	0.01
	sd	0.01	0.01	0.01	0.01	0.01	0.03	0.05	0.01	0.02

BS = Brown Swiss and *OB* = Original Braunvieh; averages (m) and standard deviations (sd) of base group contributions are shown. EU_b = European base group, DE_b = German base group (born after 1960); AT_b = Austrian base group (born after 1960), CH_b = Swiss base group (born after 1960), IT_b = Italian base group (born after 1960), US_{b1} = American base group (born before 1955), US_{b2} = American base group (born after 1955), OB_{b1} = Original Braunvieh (born before 1960), OB_{b2} = Original Braunvieh (born after 1960), *Calib* = training set; Validation sets: *DEA* = German and Austrian validation animals, *others* = validation animals with other countries of origin, *OB* = Original Braunvieh validation animals

Table 9 Results of the analysis of the **Q**-matrix for the FV population

FV		DE_{b1}	DE_{b2}	DE_{b3}	DE_{b4}	HOL_{b1}	HOL_{b2}	AT_b	CZ_b	CH_b	FR_b	Div_b
Year		<1960	≥1960 < 1970	≥1970 < 1980	≥1980	<1960	≥1960	All	All	All	All	All
Calib (5273)	m	0.13	0.61	0.04	0.01	0.04	0.03	0.09	0.01	0.04	0.01	0.00
	sd	0.07	0.17	0.04	0.04	0.04	0.05	0.12	0.08	0.04	0.05	0.01
DEA (2581)	m	0.13	0.64	0.05	0.01	0.04	0.02	0.07	0.00	0.04	0.01	0.00
	sd	0.03	0.08	0.02	0.03	0.03	0.02	0.06	0.00	0.02	0.01	0.00
Others (97)	m	0.07	0.36	0.02	0.00	0.09	0.08	0.05	0.25	0.04	0.03	0.02
	sd	0.03	0.14	0.02	0.01	0.05	0.07	0.04	0.13	0.03	0.06	0.02

FV = Fleckvieh; averages (m) and standard deviations (sd) of base group contributions are shown
DE_{b1} = German base group (born before 1960), DE_{b2} = German base group (born between 1960 and 1970), DE_{b3} = German base group (born between 1970 and 1980), DE_{b4} = German base group (born after 1980), HOL_{b1} = Red Holstein base group (born before 1960), HOL_{b2} = Red Holstein base group (born after 1960), AT_b = Austrian base group, CZ_b = Czech base group, CH_b = Swiss base group, FR_b = French base group, DIV_b = base groups with animals with other countries of origin, *Calib* = training set, Validation sets: *DEA* = German and Austrian validation animals, *others* = validation animals with other countries of origin

Brown Swiss and Original Braunvieh

In the BS population, the two American base groups (US_{b1} and US_{b2}) represented between 80 % and 90 % of the overall genetic makeup of the genotyped population (Table 8). No differences in US contributions were detected between the training set (*Calib*) and the validation animals that were assigned to the *DEA* validation set and only a slight increase in US contributions was found in the *others* validation set. The small number of validation animals that was unequivocally assigned to the *OB* group showed a marked difference in this respect, with absolutely no contributions from the US base groups. Standard deviations of contributions for training animals (*Calib*) were also highest for the two US groups. Comparing standard deviations of all contributions between *Calib* and validation groups showed that the validation animals tended to have less variation, again except for the *OB* group.

Fleckvieh

In the FV breed, the second German base group (DE_{b2}) had the largest contribution to all validation groups (Table 9). Average contributions of more than 0.60 of the second German base group to the *Calib* training set and *DEA* validation set were observed and a considerable average contribution of 0.36 to the *others* validation set. The contribution of the Czech group (CZ_b) to the *others* validation set was relatively high (0.25).

As previously, across all base groups, we found similar average contributions to *Calib* and *DEA* and decreasing standard deviations in base group contributions when comparing *Calib* to *DEA*, which indicates an ongoing equalization of contributions.

Discussion

In conventional methods for estimating breeding values, phantom parent groups are used in most practical applications. The reason for this is that the theoretical base population is rarely correctly represented in the available pedigree. The same is of course true for genomic evaluation models. Stratification of the population can be easily determined by F_{st} plots.

Concept and implementation

The decomposition of the standard **G**-matrix that we propose here is primarily an analytical tool. It allows studying the following aspects in some detail: (i) whether and how differences in allele frequencies between base groups contribute to the proportion of genetic variance explained by differences between base groups; and (ii) how the effects estimated for the base groups influence the current population and their genomic predictions. Conceptually, it follows the classical approach for modeling base groups in genetic evaluations and extends it to the GBLUP case. More fundamentally, it theoretically shows that parts of the genetic variation represented by the **G**-matrix can be assigned to systematic differences in allele frequencies between base populations. This implies that standard GBLUP is equivalent to a model that fits random genetic groups, where differences in group means are modeled as part of the natural additive-genetic variance (assumed to be known in the present investigation). Recently, Makgahlela et al. [13] showed that, in the case of the largely admixed Nordic Red population, a model that fits a fixed genetic group has some advantage in terms of the reliability of DGV over the standard GBLUP model. Modeling groups as fixed might be advantageous if true differences between groups are larger than what can be attributed to differences in allele frequencies of genetic markers. This can arise from inconsistent linkage disequilibrium phases between quantitative trait loci (QTL) and markers between subpopulations or breeds, or from

different QTL segregating within groups. Both aspects have been used in the past to explain why across-breed genomic predictions based on 50 k genotypes have low accuracy [36–38].

As in the classical approach for modeling base groups, we assigned base animals to groups and calculated a matrix of genetic contributions **Q** using standard methodology. This matrix **Q** was then used to estimate average allele frequencies using mixed-model methodology, as described by Gengler et al. [21]. As mentioned in the Methods section, estimation of average allele frequencies in base groups is not essential for the proposed decomposition of **G**. However, it provides a convenient way to integrate new animals under practical conditions. Conceptually, it divides the genetic distance between any pair of animals into two parts, i.e. a distance that already exists in the base population and a distance that originates from the history of the breed as documented by the known pedigree. Moreover, estimating allele frequencies in base groups from subsets of genotypes may lead to similar problems as in standard applications of models that fit genetic groups, i.e., if the amount of data to estimate allele frequencies in base groups reliably is not sufficient, it can result in a loss of accuracy and introduction of bias [39]. Then, this tradeoff between defining all possible relevant base groups and estimability needs to be taken into account. A closer examination of the required size and properties for an optimal design of base groups is beyond the scope of this paper.

Group effects were not accounted for when deregressing MACE breeding values for BS/OB animals because (i) group effects or group contributions are usually not reported to Interbull by the participating countries; (ii) Interbull introduces its own group categorizations based on birth year of bull dams for MACE evaluation; and (iii) Interbull does not report group effects or group contributions back to the participating countries. Because of these limitations, we cannot exclude that our results for BS/OB animals may be influenced in one way or the other by the properties of MACE breeding values.

Since we tested different models only in a single forward prediction, the generalization of our results is not straightforward. However, from a practical point of view, the steps that we followed allowed us to better characterize the genetic composition of the validation groups. This in turn might help to decide if a standard GBLUP model is sufficient or whether a different model should be preferred. However, modeling genetic groups in any of the proposed ways is neither intended nor expected to improve the prediction for a standard animal with a pedigree that has many generations and that is sufficiently complete. Predictions for an animal with an incomplete pedigree or a limited number of genotyped ancestors should, however, benefit from the inclusion of group effects in one form or the other.

Models

We compared three models, which treated effects of base groups as random (M0), as fixed (M2), or ignored them completely (M1). Model M1 consistently showed the lowest R^2 values across both breeds and all traits. This was expected, since ignoring part of the genomic information should not result in increased predictive ability. However, it is interesting to note that the segregation term itself results in a relatively good prediction. Using M1, we observed differences in the decrease of the model R^2 between traits, with the UD trait being the least influenced by $\mathbf{G}^*_A$. We cannot exclude that there might be cases where omission of base groups will increase the R^2 of predictions. However, the slopes of the regression of current DYD or deregressed proofs on DGV that we used as a test statistic here gave no indication that omitting $\mathbf{G}^*_A$ without adjusting the genetic variance could lead to less inflated estimates. Recently, Makgahlela et al. [14] compared predictions using a genomic relationship matrix based on average allele frequencies across breeds with predictions using breed-specific allele frequencies in the Nordic Red dairy cattle population. This comparison is conceptually quite close to what we did in the comparison between the reduced model (M1) and the fixed model (M2). The authors found a smaller predictive power and greater inflation of DGV when considering breed-specific allele frequencies. Since using breed-specific allele frequencies without modeling differences in allele frequencies in the base population is equivalent to our reduced model (M1), in this respect, their results are consistent with those presented here.

In terms of predictive power, M2 was better than M0 for all milk traits and one conformation trait for the BS/OB data (Table 5). With the FV data, we saw a clear advantage of M0 for all traits. In a preliminary study [40], we had reported that the OB and current BS populations were separated by a fairly large genetic distance. The validation BS/OB group that we used here included only very few OB animals. The observed genetic distance and the fact that this group of animals is small compared to the overall validation group might explain the small superiority of M2 observed for the BS/OB data. Genetic distances of similar magnitude were not detected in the FV population, for which M0 was clearly the best model. However, the German-Austrian cooperation for genetic evaluations in FV [22] recently fully opened the routine evaluations for the Czech population, which shows some differences in genetic composition compared to the current German-Austrian breeding population (Table 9).

Additional investigations will be necessary to verify if M0 is still superior with an extended base population that will very likely be the result of this extended cooperation.

Genetic contributions and base group effects

Analysis of the matrix of genetic contributions **Q** revealed some interesting features. For example, on the one hand, the analysis of average contributions of genetic groups to current animals revealed that US animals had a strong impact on the current BS population in Europe. On the other hand, a substantial contribution of the "old" European base group (EU_b) to the *OB* validation group was found. Averages and standard deviations of contributions are also an indirect indicator for how accurate base allele frequencies and base group effects could be estimated from the current data. However, since information in **Q** naturally implies some degree of collinearity, this factor has to be taken into account also. Finally, differences in trait means between base groups can only be detected if there is enough variation in base group contributions within the training set (*Calib*). Such variation was observed for both breeds and was considerably smaller for the dominant groups of the validation set. This was expected since, in the last 20 years, much less migration has occurred in both populations, which probably resulted in less admixture in the more recent groups. Although this was not the primary focus of this investigation, it was interesting to note the extremely strong genetic contribution of American Brown Swiss animals to the current BS population. The validation group *OB* was clearly an exception in the sense that a small or even non-existing contribution of American Brown Swiss cattle defines what an OB animal is. In contrast, the strong contribution of the DE_{b2} group to the FV population seems to be an artifact of the completeness of the pedigree used, i.e. most of the pedigrees traced back to this base group.

For both breeds and for the traits analyzed here, it was possible to estimate significant differences between the means of base groups in most cases (Tables 6 and 7). Treating base groups as fixed or random resulted in similar patterns, although they were more pronounced in the case of fixed effects. The observed effects were quite consistent with our expectations and seem to be reasonable when considering the limits that were imposed on estimability and precision by the collinearity and dependencies in **Q** (**Q** has no full column rank). For example, the two Holstein base groups in the FV dataset had a clear advantage for protein yield, which is not surprising since Holstein bulls were introgressed for exactly that reason. In some cases, such as the advantage found for the DE_b group in BS, knowing that the base group definition for DE_b also comprised relatively young base animals was helpful, whereas assignment to American Brown Swiss was more linked to a specific period further back in the history of the breed.

Both the distribution of genetic contributions and precision of base group effects emphasize that when considering genetic grouping in genetic evaluation models, the question of estimability and relevance for the current population should always be included [39]. However, as already noted above, it is not reasonable to believe that the model used has a strong impact on predictive power if the animals used for validation show no differences in their genetic composition with respect to the base groups and if the majority of them have complete pedigrees of sufficient depth.

Additional considerations

This investigation demonstrates that, in many cases, the genomic relationship matrix includes an important component of variation that has no corresponding counterpart in the conventional numerator relationship matrix. However, many practical applications of the estimation of GBV include a step for scaling the genomic relationship matrix to the numerator relationship matrix to set them on the same genetic base (see for example [41]). Based on our results, it seems more suitable to do this scaling based on matrix $\mathbf{G_S}$ only. This component of the **G**-matrix should be free of the effects of systematic differences in allele frequencies between base groups (represented in $\mathbf{G}^*_\mathbf{A}$), which might otherwise exacerbate the derivation of correct scaling factors. This issue was also raised by Makgahlela et al. [14] and might be of special importance for applications of one-step genomic evaluations [16, 17, 42, 43]. Furthermore, it suggests that estimating genetic parameters for genomic evaluations using $\mathbf{G_T}$ might be preferred over a simple transfer of the parameters estimated with the numerator relationship matrix.

Possible extensions of M0, for example with an individual λ for group effects or – in the most general form – using an identity matrix instead of $\mathbf{G_A}$, e.g. [39], as well as an individual λ for group effects were beyond the scope of this paper. In addition, these extensions would require the estimation of a variance component for groups, which would be difficult to do due to the typically small number of degrees of freedom for the variance between group means. Using $\mathbf{G_A}$ but assuming an individual λ for group effects is also somewhat questionable from a conceptual point of view, since it would be necessary to describe the covariance between and within subpopulations based on the same distance between allele frequencies but with different genetic variances.

Conclusions

We showed that the proposed decomposition of the **G**-matrix is helpful to examine the relative importance of base group and segregation effects in a dataset. The commonly

used genomic relationship matrix **G** is equivalent to our model M0, where base groups and segregation terms are considered as random effects with the same genetic variance. Although it is interesting to examine contributions of different founder populations from a scientific point of view, we also conclude that the standard model M0 is preferred in many cases, e.g. if base group effects are small or difficult to estimate, or if the current population is homogenous with balanced base group contributions. However, a fixed model (M2) might be preferred if base group effects are large (i.e. in the range of differences between breeds rather than between subpopulations) or if the genomic evaluation comprises two or more separated populations with only weak genetic links.

Appendix 1

Proof that model 2 (fixed group effects model using $\mathbf{G}_S$ as covariance of individual genetic values) and a corresponding model using $\mathbf{G}_T$ as covariance of individual genetic values will lead to identical solutions for fixed and random effects.

As shown in Appendix 2, the standard model and model 0 are equivalent. Following that, BLUP solutions of a model using $\mathbf{G}_T$ as covariance of breeding values can be equivalently written as:

$$\hat{\mathbf{u}} = \mathbf{G}_S\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}} + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}},$$

where **Q** is a matrix of genetic contributions of random groups to animals with observations as described in Methods and $\tilde{\mathbf{y}}$ is the vector of observations corrected for the GLS-estimates of fixed effects. If the same matrix **Q** is used to model the fixed group effects, as it is generally done, this might be written as:

$$\hat{\mathbf{u}} = \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right) + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right).$$

By omitting the global mean since it cannot be estimated simultaneously and by replacing $\hat{\mathbf{b}}$ by its GLS-estimate, this can be further manipulated to give:

$$\begin{aligned}
\hat{\mathbf{u}} &= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right) + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right)\\
&= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right) + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{y}\text{-}\mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{Q}\hat{\mathbf{b}}\\
&= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right)+\mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{y}\text{-}\mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{Q}\left(\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{Q}\right)^{-1}\mathbf{Q}\mathbf{V}_{yy}^{-1}\mathbf{y}\\
&= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right) + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{y}\text{-}\mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\mathbf{y}\\
&= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{Q}\hat{\mathbf{b}}\right),
\end{aligned}$$

and identical solutions for the random effect **u** are the consequence. It follows that the product of the design matrix **Q** and the contrast for the random group effect (represented by the second term above) is also zero, which is a necessary prerequisite for the resulting estimates, for the fixed genetic groups to be equal in both models also [44]. As a general consequence of the cited publication [44], any extension of **V** in the GLS-estimate of **b** of the form:

$$\mathbf{V}^* = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}',$$

for an arbitrary matrix **U**, where **X** is the same design matrix used to estimate the fixed effect itself, results in GLS-estimates for the fixed effects that are identical to those using **V** alone [44].

Appendix 2

Proof that the standard model is equivalent to the random group model M0.

Let the standard model be:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where **y** is a vector of observations, **b** is a vector of fixed effects, **u** is a vector of random breeding values, **e** is a vector of residuals and **X** and **Z** are known design matrices. For simplification of the presentation **Z** is assumed to be an identity matrix and is omitted. Furthermore, $\mathbf{y} \sim N(\mathbf{X}\mathbf{b}, \mathbf{V}_{yy})$, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_{uu})$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V}_e)$ where:

$$\mathbf{V}_e = \mathbf{I} \times \sigma_e^2 = \mathbf{R},$$

$$\mathbf{V}_{uu} = \tilde{\mathbf{G}}_T \times \sigma_u^2 = \mathbf{G}_T,$$

and

$$\mathbf{V}_{yy} = \mathbf{G}_T + \mathbf{R}.$$

Assume a decomposition of the coefficient matrix $\tilde{\mathbf{G}}_T = \left(\tilde{\mathbf{G}}_S + \tilde{\mathbf{G}}_A\right) \times \sigma_u^2 = \mathbf{G}_S + \mathbf{G}_A^*$ where $\mathbf{G}_A^*$ can be expressed as the product of a matrix of fixed regression coefficients **Q** and a matrix **GA**, that describes the covariance of random slopes, so $\mathbf{G}_A^* = \mathbf{Q}\mathbf{G}_A\mathbf{Q}'$. The BLUP estimates for random breeding values are:

$$\hat{\mathbf{u}} = \mathbf{G}_T\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\mathbf{X}\hat{\mathbf{b}}\right) = \mathbf{G}_T\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}},$$

with $\hat{\mathbf{b}}$ being the generalized least squares estimates of **b**. It follows that:

$$\begin{aligned}
\mathbf{V}_{yy} &= \mathbf{G}_T + \mathbf{R}\\
&= \mathbf{G}_S + \mathbf{G}_A^* + \mathbf{R}\\
&= \mathbf{G}_S + \mathbf{Q}\mathbf{G}_A\mathbf{Q}' + \mathbf{R},
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mathbf{u}} &= \mathbf{G}_T\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}}\\
&= \left(\mathbf{G}_S + \mathbf{G}_A^*\right)\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}}
\end{aligned}$$

$$= (\mathbf{G}_S + \mathbf{Q}\mathbf{G}_A\mathbf{Q}')\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}}$$

$$= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}} + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}}.$$

Let the random group model be:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{Q}\mathbf{g} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of observations, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random genetic values, $\mathbf{g}$ is a vector of random group effects, $\mathbf{e}$ is a vector of residuals and $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices. For simplification of the expressions, $\mathbf{Z}$ is assumed to be an identity matrix and is omitted. $\mathbf{Q}$ is a matrix of genetic contributions of random groups to animals with observations as described in Methods. Furthermore, $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V}_{yy})$, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_{uu})$, $\mathbf{g} \sim N(\mathbf{0}, \mathbf{V}_{gg})$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V}_e)$ where:

$$\mathbf{V}_e = \mathbf{I} \times \sigma_e^2 = \mathbf{R},$$

$$\mathbf{V}_{uu} = \tilde{\mathbf{G}}_S \times \sigma_u^2 = \mathbf{G}_S,$$

$$\mathbf{V}_{gg} = \tilde{\mathbf{G}}_A \times \sigma_u^2 = \mathbf{G}_A,$$

$$\mathbf{V}_{yy} = \mathbf{G}_S + \mathbf{Q}\mathbf{G}_A\mathbf{Q}' + \mathbf{R},$$

$$= \mathbf{G}_S + \mathbf{G}_A^* + \mathbf{R}.$$

This is identical to the phenotypic variance assumed by the standard model if the same $\mathbf{Q}$ is used.

The BLUP solutions for random animal and group effects are:

$$\hat{\mathbf{u}} = \mathbf{G}_S\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\hat{\mathbf{Xb}}\right) = \mathbf{G}_S\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}},$$

and

$$\hat{\mathbf{g}} = \mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\left(\mathbf{y}\text{-}\hat{\mathbf{Xb}}\right) = \mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}}.$$

Let the full estimate for the breeding value (the ranking criterion) be:

$$\hat{\hat{\mathbf{u}}} = \hat{\mathbf{u}} + \mathbf{Q}\hat{\mathbf{g}}$$

$$= \mathbf{G}_S\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}} + \mathbf{Q}\mathbf{G}_A\mathbf{Q}'\mathbf{V}_{yy}^{-1}\tilde{\mathbf{y}},$$

this is identical to the breeding value solution of $\hat{\mathbf{u}}$ of the standard model if $\mathbf{Q}$ is identical in both models.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
LP performed the analysis and drafted the manuscript. LP, CE, RE and KUG designed the study. CE and LP developed methods. CE, ECGP, RE, JB and KUG revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments
We want to thank the contributors of the genotype pool Germany-Austria as well as the Intergenomics consortium for providing the genotypes. We gratefully acknowledge the Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. for their financial support within the research cooperation "Zukunftswege". Furthermore, we wish to thank the editors JCM Dekkers and H Hayes as well as two unknown reviewers for their helpful suggestions to improve the final manuscript.

Author details
[1]Bavarian State Research Center for Agriculture, Institute of Animal Breeding, Prof.-Dürrwaechter-Platz 1, 85586 Poing-Grub, Germany. [2]Institute of Animal Husbandry and Breeding, University Hohenheim, Garbenstraße 17, 70599 Stuttgart, Germany.

Received: 18 December 2014 Accepted: 27 May 2015
Published online: 23 June 2015

References
1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
2. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194:597–607.
3. Ibáñez-Escriche N, Fernando RL, Toosi A, Dekkers JCM. Genomic selection of purebreds for crossbred performance. Genet Sel Evol. 2009;41:12.
4. Harris BL, Johnson DL. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J Dairy Sci. 2010;93:1243–52.
5. Erbe M, Hayes BJ, Matukumalli LK, Goswam S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 2012;95:4114–29.
6. Caballero A, Toro MA. Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Genet. 2002;3:289–99.
7. Álvarez I, Royo LJ, Gutiérrez JP, Fernández I, Arranz JJ, Goyache F. Relationship between genealogical and microsatellite information characterizing losses of genetic variability: Empirical evidence from the rare Xalda sheep breed. Livest Sci. 2008;115:80–8.
8. Thompson R. Sire evaluation. Biometrics. 1979;35:339–53.
9. Quaas RL, Pollack EJ. Modified equations for sire models with groups. J Dairy Sci. 1981;64:1868–72.
10. Westell RA, Quaas RL, Van Vleck LD. Genetic groups in an animal model. J Dairy Sci. 1988;71:1310–8.
11. García-Cortés LA, Toro MA. Multibreed analysis by splitting the breeding values. Genet Sel Evol. 2006;38:601–15.
12. Makgahlela ML, Mäntysaari EA, Strandén I, Koivula M, Sillanpää MJ, Nielsen US, et al. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. Interbull Bull. 2011;44:42–6.
13. Makgahlela ML, Mäntysaari EA, Strandén I, Koivula M, Nielsen US, Sillanpää MJ, et al. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J Anim Breed Genet. 2013;130:10–9.
14. Makgahlela ML, Strandén I, Nielsen US, Sillanpää MJ, Mäntysaari EA. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of multibreed population. J Dairy Sci. 2014;97:1117–27.
15. Tsuruta S, Misztal I, Lourenco DAL, Lawlor TJ. Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holstein. J Dairy Sci. 2014;97:5814–21.
16. Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res. 2011;93:357–66.
17. Christensen OF. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol. 2012;44:37.
18. Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of population structure. Evolution. 1984;38:1358–70.
19. Patterson N, Price AL, Reich D. Population structure and Eigen analysis. PLoS Genet. 2006;2, e190.
20. Zou F, Lee S, Knowles MR, Wright FR. Quantification of population structure using correlated SNPs by shrinkage principal components. Hum Hered. 2010;70:9–22.
21. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: applications to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1:21–8.

22. Edel C, Schwarzenbacher H, Hamann H, Neuner S, Emmerling R, Götz KU. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. Interbull Bull. 2011;44:152–6.
23. Schaeffer LR. Multiple-country comparison of dairy sires. J Dairy Sci. 1994;77:2671–78.
24. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41:55.
25. Mäntysaari E, Liu Z, VanRaden PM. Interbull validation test for genomic evaluations. Interbull Bull. 2010;41:17–22.
26. Interbull CoP. Appendix VIII - Interbull validation test for genomic evaluations – GEBV test. 2013. https://wiki.interbull.org/public/CoPAppendixVIII?action=print&rev=44. Accessed 12 June 2014.
27. Fiske WF, Banos G. Weighting factors of sire daughter information in international genetic evaluations. J Dairy Sci. 2001;84:1759–67.
28. Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol. 2010;42:5.
29. International Organization for Standardization. Codes for the representation of names of countries and their subdivisions – Part 1: Country codes. 3rd ed. Geneva: ISO copyright office; 2013.
30. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
31. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 2005;15:1468–76.
32. Chen C, Durand E, Forbes F, Francois O. Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. Mol Ecol Notes. 2007;7:747–56.
33. Mrode RA. Linear models for the prediction of animal breeding values. 2nd ed. Oxfordshire: CABI Publishing; 2005.
34. Quaas RL. Additive genetic model with groups and relationships. J Dairy Sci. 1988;71:1338–45.
35. Nei M. Analysis of gene diversity in subdivided populations. Proc Nat Acad Sci USA. 1973;70:3321–3.
36. Harris BL, Johnson DL, Spelman RJ. Genomic selection in New Zealand and the implications for national genetic evaluation. In Proceedings of the 36th International Committee for Animal Recording Biennial Session:16–20 June 2008; Niagara Falls. 2009:325–30.
37. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009;92:433–43.
38. de Roos APW, Hayes BJ, Goddard ME. Reliability of genomic predictions across multiple populations. Genetics. 2009;183:1545–53.
39. Phocas F, Laloë D. Should genetic groups be fitted in BLUP evaluation? Practical answer for the French AI beef sire evaluation. Genet Sel Evol. 2004;36:325–45.
40. Plieschke L, Edel C, Pimentel E, Emmerling R, Bennewitz J, Götz KU. Influence of foreign genotypes on genomic breeding values of national candidates in Brown Swiss. In Proceedings of the 10th World Congress of Genetics Applied to Livestock Production: 17–22 August 2014;Vancouver. https://asas.org/docs/default-source/wcgalp-proceedings-oral/078_paper_8984_manuscript_342_0.pdf?sfvrsn=2. Accessed 12 June 2014.
41. Meuwissen THE, Luan T, Woolliams JA. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet. 2011;128:429–39.
42. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656–63.
43. Aguila I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93:743–52.
44. Rao CR. Least-squares theory using an estimated dispersion matrix and its application to measurement of signals. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1967. 1:355–72.

Chapter Two

Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates

Plieschke *et al. Genet Sel Evol (2016) 48:73*
DOI 10.1186/s12711-016-0250-9

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates

Laura Plieschke[1*], Christian Edel[1], Eduardo C. G. Pimentel[1], Reiner Emmerling[1], Jörn Bennewitz[2] and Kay-Uwe Götz[1]

Abstract

Background: Extending the reference set for genomic predictions in dairy cattle by adding large numbers of cows with genotypes and phenotypes has been proposed as a means to increase reliability of selection decisions for candidates.

Methods: In this study, we explored the potential of increasing the reliability of breeding values of young selection candidates by genotyping a fixed number of first-crop daughters of each sire from one or two generations in a balanced and regular system of genotyping. Using stochastic simulation, we developed a basic population scenario that mimics the situation in dual-purpose Fleckvieh cattle with respect to important key parameters. Starting with a reference set consisting of only genotyped bulls, we extended this reference set by including increasing numbers of daughter genotypes and phenotypes. We studied the effects on model-derived reliabilities, validation reliabilities and unbiasedness of predicted values for selection candidates. We also illustrate and discuss the effects of a selected sample and an unbalanced sampling of daughters. Furthermore, we quantified the role of selection with respect to the influence on validation reliabilities and contrasted these to model-derived reliabilities.

Results: In the most extended design, with 200 daughters per sire genotyped from two generations, single nucleotide polymorphism (SNP) effects were estimated from a reference set of 420,000 cows and 4200 bulls. For this design, the validation reliabilities for candidates reached 80 % or more, thereby exceeding the reliabilities that were achieved in traditional progeny-testing designs for a trait with moderate to high heritability. We demonstrate that even a moderate number of 25 genotyped daughters per sire will lead to considerable improvement in the reliability of predicted breeding values for selection candidates. Our results illustrate that the strategy applied to sample females for genotyping has a large impact on the benefits that can be achieved.

Background

Genomic selection and genomic breeding value estimation were implemented in several cattle breeding programs in the last few years. Since the introduction of this methodology, there has been a constant attempt to further improve it and to increase the reliabilities of genomic breeding values. One key factor is the size of the reference set [1, 2]. Nowadays, there are several international organizations that promote the exchange of genotypes on a regular basis to enlarge reference sets and to improve the quality of genomic predictions of the participating countries. In dual-purpose Fleckvieh (FV) cattle, genomic selection was implemented in 2011 and genetic evaluation centers in Germany and Austria cooperate in a joint genetic and genomic evaluation that uses a common genotype pool [3]. Currently, the reference set for FV includes approximately 9000 bulls with phenotypic measures on most traits.

Several studies have reported that sharing genotypes within breeds results in large benefits for the reliability of

*Correspondence: Laura.Plieschke@lfl.bayern.de
[1] Bavarian State Research Center for Agriculture, Institute of Animal Breeding, Prof.-Dürrwaechter-Platz 1, 85586 Poing-Grub, Germany
Full list of author information is available at the end of the article

genomic predictions e.g. [4–6]. However, most opportunities to increase the genotype pool by exchanging genotypes have been exploited and, in most cases, the growth of reference sets within breeds is restricted to the yearly increase in number of genomically preselected young bulls receiving daughter proofs. As a consequence, fewer bulls are progeny-tested than in pre-genomic selection programs [7, 8] and the proportion of old bulls increases over time. Since the reliability of genomic predictions also depends on the degree of relationship between reference and predicted animals [9], this 'aging' of the reference set may lead to decreased reliabilities. As a demonstration of that effect, Cooper et al. [10], for example, excluded subsets of old bulls and found that older bulls in the reference set had only a minimal impact on the reliability of the genomic breeding values of predicted animals. In addition, preselection of young reference bulls may influence the quality of genomic predictions. Schaeffer [11] predicted a situation where considerable bias was introduced on genomic evaluations by strong preselection [12–14] of young bulls based on their genomic breeding values.

Another possibility to increase the size of the reference set is to use information from genotyped and phenotyped females, which can have a beneficial influence on the quality of genomic predictions. Thomasen et al. [15] found that, by adding female genotypes in the reference set, more genetic gain with a lower rate of inbreeding can be achieved compared to a breeding scheme where the reference set grows only from the addition of newly progeny-tested bulls. Pryce et al. [8] showed that by adding 10,000 cows to a reference set of 3000 Holstein bulls, the reliability of genomic predictions of 437 young bulls in the validation set was improved by 4 to 8 %. Calus et al. [16] also combined cows and bulls in a single reference set and found that the highest validation accuracies were achieved with the combined dataset compared to scenarios with a reference set that included only cows or only bulls. Furthermore, since usually cows are not strongly preselected, inclusion of their genotypes and phenotypes may also contribute to reduce the biasing effects of preselection as pointed out by Schaeffer [11]. Last but not least, genotyping cows might be especially important for creating reference sets for so-called new traits or expensive-to-measure traits [7, 17, 18] and, most likely, will be the basis of new and useful management tools for farmers [8].

If female genotypes are to be included in a genomic system, one of the key questions is which cows should be genotyped. Pryce et al. [8], Wiggans et al. [19] and Dassonneville et al. [20] discussed preferential treatment as a potential problem related to the inclusion of bulls' dams into the reference set. Dassonneville et al. [20] found that the inclusion of records on elite cows resulted in overestimation of genomic enhanced breeding values for all animals. Thus, even if genotypes are available for elite cows as a consequence of using genomic predictions for the selection of bulls' dams, in the end, they should not be part of the reference set.

In a preliminary study [21], we performed a deterministic simulation based on nuclear pedigrees extracted from the German-Austrian FV population and showed that there is a benefit from including genotyped cows into the reference set. We quantified the effects of this inclusion on the reliability of genomic breeding values of young selection candidates and found marginal to considerable gains in reliability (between 1 and 40 %) depending on the scenario. However, we were not able to quantify the effects of selection on the results and we could not quantify the cumulative effects at the population level. Therefore, in this study, we examined the following three main effects by means of a stochastic simulation: (1) effects of selection on validation reliability, (2) effects of genotyping randomly selected cows on the accuracy of prediction, and (3) effects of some alternative strategies for sampling the genotyped daughters.

Methods

Simulation

We used the open access software QMSim [22] to run a simulation with five repetitions. Our aim was to simulate a population that resembled the German-Austrian dual-purpose Fleckvieh cattle population for several key characteristics (e.g. linkage disequilibrium (LD) structure, allele frequencies and effective population size).

Simulation of the population

QMSim first simulated a so-called historical population, which consisted of 2000 unrelated animals with a balanced sex ratio. These animals were randomly mated for 2500 generations. To create a sufficiently strong LD structure as observed in FV, a bottleneck was introduced after 2500 generations by reducing the number of breeding animals to 150 for one generation, which corresponds approximately to the effective population size in FV i.e. 160 based on the observed LD structure [23]. This estimate is quite close to that based on pedigree data [24]. After this bottleneck, population size was increased within one generation again to 31,500 animals (30,000 dams and 1500 sires), which represented the founder animals (generation 0) of the so-called 'recent' or pedigreed population. The recent population was propagated for another 10 generations. In each generation of the recent population, 15,000 female and 15,000 male offspring were generated by mating 30,000 dams and 1500 breeding sires. Generations overlapped and in each

generation 30 % of the dams and 70 % of the sires were replaced. These two replacement parameters were quite similar to the situation observed in the real FV population. Breeding animals were selected based on their estimated breeding value (EBV) which was calculated within QMSim with a reliability of 0.6. This was done to mimic a genomic selection program where dams are selected based on a combination of pedigree information and own performance and sires are selected on their genomic breeding value.

Males of generation 5 to 10 were genotyped (Table 1). Sires belonging to generations 5 to 8 (n = 4200) were assigned to the reference set. The remaining animals of generations 9 and 10 were used as validation set for forward prediction. Note that whereas sires in generation 9 (n = 1050) were young bulls that were selected by QMSim based on a genomic breeding value but without daughter performances, the animals of generation 10 (n = 15,000) were unselected candidates. The validation animals were further characterized by the status of their sire i.e. a reference animal or not. Figure 1 gives an overview of the structure of the simulation.

Simulation of the genome

We simulated 30 chromosomes, each 100 cM long. On each chromosome, 1660 single nucleotide polymorphisms (SNPs) and 30 quantitative trait loci (QTL) were evenly distributed (49,800 SNPs and 900 QTL in total). After routine checks [3, 25], nearly 38,000 valid SNPs and approximately 700 QTL that were still segregating in the reference set (both numbers slightly varying between replicates of the simulation) were available. The routine checks were as follows: (1) SNPs that deviated from Hardy–Weinberg equilibrium (HWE) with a p-value less than 10^{-5} and (2) SNPs with a minor allele frequency (MAF) lower than 0.02 were excluded from the dataset. We assumed a sex-linked trait and a single observation for each female with a heritability set to 0.4. The polygenic nature of the trait was ensured by the relatively large number of QTL and their effects were drawn from a uniform distribution (option 'uniform' from QMSim) to prevent the occurrence of a few isolated large QTL effects. With a uniform distribution, the mean of the effects is related to the variance and, thus, the range of the QTL effects is limited. We performed a couple of tests with QMSim and the results confirmed our assumptions (data not shown).

Simulation of the daughter sets

In the main part of the simulation, we generated 200 daughters for each of the reference bulls of generations 7 and 8 (which represented a total of 420,000 additional female genotypes and phenotypes). Due to memory requirements and some limitations of the QMSim software, we did not simulate the daughter genotypes with QMSim directly. Instead, based on the known haplotypes (SNPs and QTL) of the reference bulls of these two generations, we simulated different male gametes by recombination and randomly mated them with gametes of potential dams of the same cohort (excluding sisters, daughters and dams) that was simulated by applying the same strategy. Assuming a Poisson distribution for cross-overs, recombination was simulated by generating on average one random cross-over per Morgan for each chromosome. Using the observed QTL status of each daughter and the known (true) QTL effects from the QMSim simulation, we calculated the true breeding value (TBV) for each daughter.

Phenotypes

We generated yield deviations (YD, [26]) for daughters using the TBV and a random residual. Depending on the design investigated, these daughter phenotypes were used to calculate daughter yield deviations (DYD, [26])

Table 1 Assignment of animals to the reference or validation set in the different scenarios

Generation	Number of individuals			Explanation
	Base scenario	Extended scenarios step 1	Extended scenarios step 2	
5	1050	1050	1050	Reference set
6	1050	1050	1050	
7	1050	1050	1050 + daughters	
8	1050	1050 + daughters	1050 + daughters	
9	1050	1050	1050	Validation set
10a	4516	4516	4516	
10b	10,484	10,484	10,484	

Validation animals were further divided according to the status of the corresponding sire (member of the reference set or not), resulting in three validation groups. Sires of validation animals in generations 9 and 10a were part of the reference set and sires of validation animals in generation 10b were not part of the reference set. First, daughters of the sires of generation 8 were added to the reference set (step 1) and then daughters of the sires of generation 7 were also added (step 2)

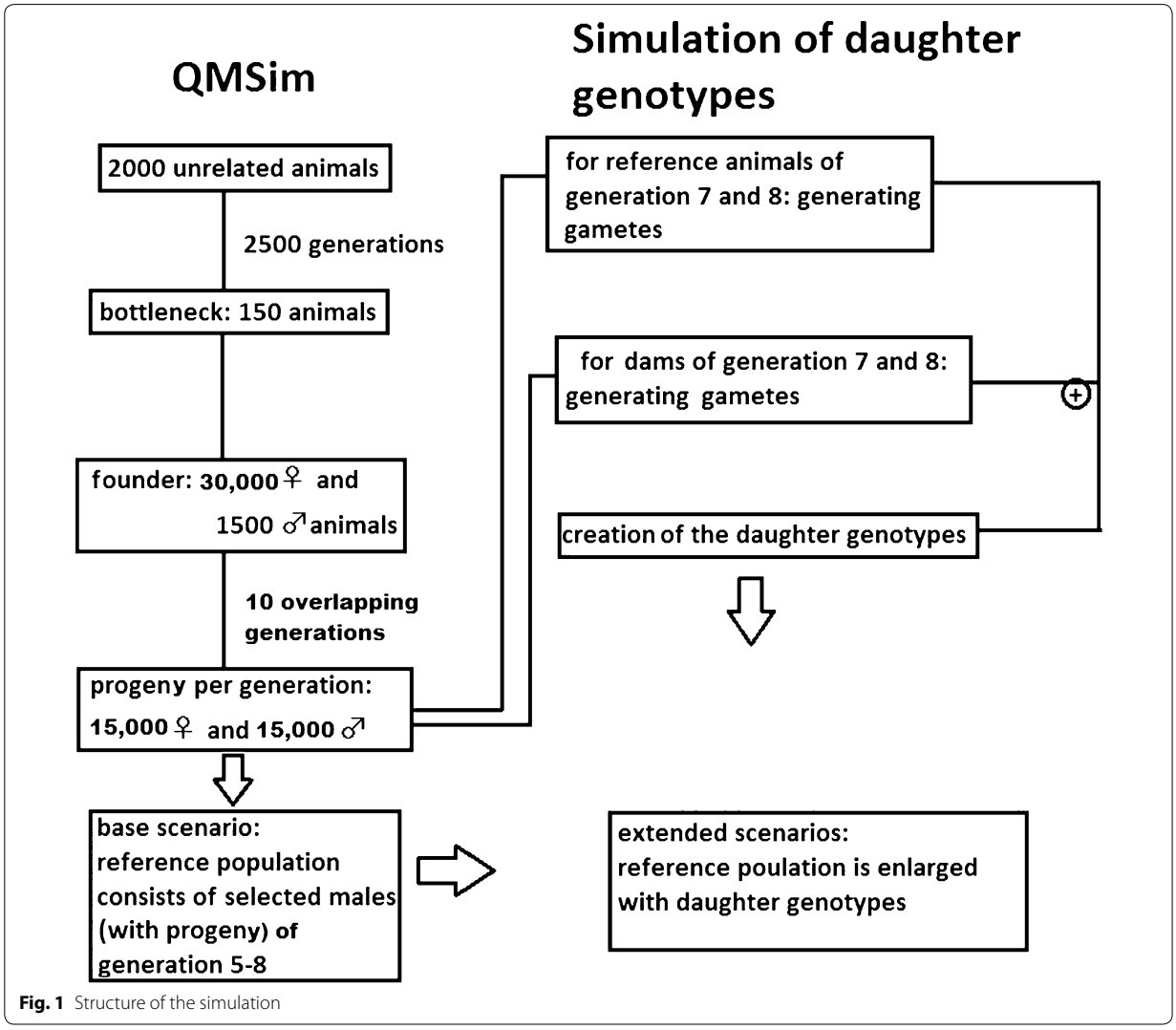

Fig. 1 Structure of the simulation

of the corresponding bull or were directly included in the reference set. In this way, YD of the reference daughters were automatically omitted from the daughter yield deviation (DYD) of the sire and double-counting in the extended scenarios was avoided. To account for different variances of the YD and DYD, phenotypes were weighted with the equivalent number of own performances (EOP, [27]) calculated as

$$\mathrm{EOP} = \lambda \frac{\mathrm{R}^2_{\mathrm{phen}}}{1 - R^2_{\mathrm{phen}}},$$

where $\lambda = \frac{\sigma^2_e}{\sigma^2_a}$ with σ^2_a being the additive genetic variance and σ^2_e the residual variance and $\mathrm{R}^2_{\mathrm{phen}}$ the reliability of the DYD or YD.

Designs

In a more general analysis, we investigated the effects of selection on validation reliability and model-derived reliability parameters. To be able to identify these selection effects, we repeated the basic scenario using the same parameters for QMSim except that we replaced directional selection on EBV by random selection.

In the main part of the simulation, we included large numbers of genotyped cows into the reference set. The general sampling strategy was to genotype a random sample of fixed size of phenotyped daughters of each artificial insemination (AI) bull in defined cohorts. We investigated 10 different scenarios: one base scenario and nine extended scenarios. In the base scenario, the reference

set consisted only of sires of generations 5 to 8. For the extended scenarios, an increasing number of the generated female genotypes and phenotypes were integrated into the reference set. Tables 1 and 2 give an overview of the different scenarios.

To assess how robust the benefits are with respect to our general sampling strategy, we changed the composition of the sample of scenario −/50 (Table 2). Instead of including a random sample of daughters as was done in scenario −/50, we selected the best 50 daughters of each sire for scenario $-/50_s$ (selection was done on YD). In the scenario $-/25_r25_s$, we selected 25 daughters at random and combined them with the 25 best remaining daughters of the corresponding sire. Finally, we also ran one unbalanced scenario ($-/50_{ub}$) with different numbers of daughters per sire to test the effect of moderate unbalancedness but the overall number of genotyped females was kept the same as in scenario −/50. This was done by randomly selecting five daughters for 330 sires, 50 daughters for 621 sires and all 200 daughters for 99 sires. The different numbers of the daughter sets per sire were chosen arbitrarily but we ensured that the total number of genotyped females was maintained and that each sire was represented by at least some daughters. Moreover, random assignment of the different numbers of daughters to the sires was also conducted.

Table 2 Scenarios with corresponding number of animals and composition of the reference set

Scenario	Reference set	
	Number of sires	Number of daughters
Base	4200	0
−/25	4200	26,250
−/50	4200	52,500
−/100	4200	105,000
−/200	4200	210,000
50/50	4200	105,000
100/100	4200	210,000
200/200	4200	420,000
$-/50_s$	4200	52,500
$-/25_r25_s$	4200	52,500
$-/50_{ub}$	4200	52,500

The names of the extended scenarios are derived from the number of daughters per sire which are included in the reference set and the sire's generation. The number before the slash in the scenario's name is the number of daughters per progeny-tested bull of generation 7 (i.e. step 2 of the extended scenarios) and the number after the slash is the number of daughters per progeny-tested bull of generation 8 (i.e. step 1 of the extended scenarios). The $-/50_s$ is a scenario in which the best daughters were selected to be genotyped, $-/25_r25_s$ is a scenario in which 25 random daughters per sire and the 25 best daughters per sire were selected and genotyped and $-/50_{ub}$ is a scenario in which an unbalanced number of daughters for all sires was selected

Genomic prediction

Due to the large number of genotypes, we used a SNP-best linear unbiased prediction (BLUP) model [28] to calculate direct genomic values (DGV) and reliabilities. The model equation is as follows:

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

and the corresponding mixed model equations are:

$$\begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}M} \\ \mathbf{M'R^{-1}X} & \mathbf{M'R^{-1}M} + \mathbf{I}/\sigma_g^2 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{M'R^{-1}y} \end{pmatrix}$$

where

$$\sigma_g^2 = \frac{\sigma_a^2}{\sum_{j=1}^{m}(2p_jq_j)},$$

and $\mathbf{y}$ is the vector of observations (here DYD or YD), $\mathbf{b}$ the vector of fixed effects (in our simulation only an overall mean), $\mathbf{g}$ is the vector of random marker effects, and $\mathbf{e}$ the vector of residual effects. Matrix $\mathbf{X}$ is a design matrix which links the observations to the respective fixed effects and $\mathbf{M}$ is the centered coefficient matrix of marker genotypes and p_j and q_j are base allele frequencies of marker j estimated for generation 0 [29]. Centering was done by subtracting two times the base allele frequency estimate from the corresponding column of $\mathbf{M}$. Matrix $\mathbf{R}$ is a diagonal matrix with σ_e^2/w_i on the diagonal, where w_i is the EOP of the i-th observation and matrix $\mathbf{I}$ is an identity matrix of order *m* (number of markers).

DGV are calculated as:

$$\text{DGV} = \hat{\mathbf{b}} + \mathbf{M}\hat{\mathbf{g}},$$

and the corresponding predicted error variances (pev) are calculated as:

$$\text{pev(DGV)} = \mathbf{M}^*\mathbf{C}_s^{-1}\mathbf{M}^{*\prime},$$

where $\mathbf{M}^*$ is matrix $\mathbf{M}$ extended with a column of ones, and $\mathbf{C}_s^{-1}$ is the inverse of the left hand side of the SNP-BLUP-MME (mixed model equation). The inclusion of the overall mean in the calculation of the pev can be questioned and may lead to slightly higher theoretical reliabilities. We empirically compared results including and omitting the overall mean and found differences that were smaller than the rounding precision of the results. Moreover, because the overall mean is included in each scenario, its impact on the contrasts between scenarios can be ignored.

The reliability of the DGV of the i-th animal can then be calculated as:

$$R_i^2 = 1 - \frac{\text{diag(pev(DGV))}_i}{\text{diag}(\mathbf{G})_i\sigma_a^2},$$

where diag(pev(DGV))_i is the i-th diagonal element of the pev(DGV) and $\text{diag}(\mathbf{G})_i$ the i-th diagonal element of

the genomic relationship matrix (**G**) which is 1 plus the genomic inbreeding coefficient. Matrix **G** is defined as follows:

$$\mathbf{G} = \frac{\mathbf{MM}'}{\sum_{j=1}^{m}(2p_j q_j)}.$$

In addition, we calculated a weighted regression of TBV on DGV for validation animals. We used the model fit of this regression as a measure of validation reliability (ρ^2) and the slope (b) as a measure of the bias that describes the inflation of estimates [30].

To quantify the effect of incomplete LD between SNPs and QTL on the difference between model-derived theoretical reliabilities and validation reliabilities, we included an analysis where we extended the marker genotype coefficient matrix **M** by QTL genotypes. We present the results in the context of the comparison between designs with directional and with random selection (ρ^2_{QTL}).

Results

Comparison of the simulated dataset with the Fleckvieh population

Comparison of the extent of LD between the simulated dataset and the real Fleckvieh dataset [31], revealed a good agreement with slightly higher values of the linkage parameter r^2 [32] for the simulated data at shorter distances. The average distance between a QTL and the nearest SNP in the simulated data was 60 kb. Allele frequencies for the simulated dataset were more evenly distributed than those for the real FV data, for which a slight shift to lower allele frequencies was observed. These results are illustrated in Figure S1 [see Additional file 1: Figure S1] and Figure S2 [see Additional file 2: Figure S2].

Simulation

For ease of interpretation, we separated the presentation of results for generation 9 from those for generation 10, in order to highlight the fact that generation 9 represents a group of individuals that are already pre-selected on an EBV including Mendelian sampling information in the course of the simulation process. This selection does have an effect on validation statistics [30]. In contrast, generation 10 is strictly unselected. Results for generation 10 were further divided according to the status of the sire (member of the reference group or not). A more detailed categorization of the results for these two generations is provided in Tables S1 and S2 [see Additional file 3: Tables S1 and S2]. There was a general tendency for scenarios with the same number of genotyped females (scenario −/100 compared to scenario 50/50 and scenario −/200 compared to scenario 100/100) showing nearly identical results. For the sake of clarity, we do not present results for the redundant scenarios. All the results shown are averages over five repetitions of the simulation. Standard errors of the results presented in the main body of the paper were less than 1.3 % for validation reliabilities (except for one scenario i.e. $-/25_r25_s$ where standard errors were between 3.2 and 4.1 %) and less than 0.02 for regression slopes.

General effects of selection

Table 3 shows model-derived reliabilities (R^2) and validation reliabilities (ρ^2) for a scenario with directional selection and a scenario with random selection. Model-derived reliabilities were slightly higher for the scenario with directional selection than for the scenario with random selection, which indicated that, with directional selection, the pattern of family sizes differs and results in a more informative structure for validation animals. Comparing R^2 with ρ^2 for randomly selected populations, we found slightly lower validation reliabilities when only SNPs were considered. When QTL were included in the SNP panel, validation reliabilities (ρ^2_{QTL}) were slightly higher than R^2. In the scenario with directional selection the validation reliabilities for generation 10 were lower than with random selection (40 to 51 and 33 to 40 %, respectively). When, in addition, the validation sample was selected on information that included Mendelian sampling information as in generation 9, the decrease

Table 3 Model-derived reliabilities (R^2) and validation reliabilities (ρ^2) in the base scenario with directional and random selection

Validation set	Sire status	Number of individuals	Base scenario					
			Random selection			Directional selection		
			R^2	ρ^2	ρ^2_{QTL}	R^2	ρ^2	ρ^2_{QTL}
9	Reference	1050	54	51	59	58	26	32
10a	Reference	4516	54	51	58	58	40	48
10b	Not reference	10,484	48	40	49	48	33	41

Validation animals were divided according to whether their sire was in the reference set or not. For the purpose of illustration (and only here), we included results of analyses in which the segregating QTL were included in the SNP panel used for estimation and prediction (ρ^2_{QTL})

in validation reliabilities was even more pronounced (26 to 51 %). Selection on parent average (PA) in the validation group did not result in inflated predictions (slope estimates ranged from 0.93 to 0.99 for generation 10a).

Effects of genotyped daughters

Table 4 presents validation reliabilities for the three validation groups for the basic scenario and five extended scenarios. Using results for group 10a as a starting point, it can generally be stated that introducing an increasing number of genotyped daughters into the reference set clearly had a positive impact on the validation reliability. Beginning with scenario −/100, validation reliabilities reached values of 70 % and more. If the sire of a validation animal was not a member of the reference set (generation 10b), the overall validation reliability was reduced, but the general trend observed was the same. As expected, the effect of the contribution of a missing sire to the overall reliability decreased as information increased. When the validation group itself was selected (generation 9), the validation reliabilities for all scenarios were lower than for the other validation groups. Again, the impact of this decrease was more pronounced when the number of cows in the reference set was smaller.

Effects of the composition of the daughter samples

Table 5 illustrates some aspects of the composition of the sample of daughters that were chosen for genotyping. Starting with values for R^2, ρ^2 and b for scenario −/50 as a reference point, we found a lower validation reliability and a noticeable increase in inflation of genomic predictions when a selected daughter group was genotyped (scenario $-/50_s$), even if selection was based on the criterion of moderate reliability as in this case. Comparing the base scenario (Table 4) to scenario $-/50_s$ (Table 5), the benefit from adding 52,500 genotyped daughters was small with respect to validation reliability. The negative effect of this preselection can be partially compensated by a combination of directly and randomly selected daughters (scenario $-/25_r25_s$, Table 5), but nevertheless the results were lower than those for a scenario where only 25 randomly selected daughters per sire were included (scenario −/25, see Table 4). A moderately unbalanced scenario (scenario $-/50_{ub}$, Table 5), however, had no detectable effect on reliabilities or regression slopes.

Discussion

In this study, we show that even small groups of daughters per sire can have large beneficial effects on model-derived reliabilities as well as validation reliabilities. A straightforward strategy to achieve these beneficial effects is to genotype a balanced random sample of daughters per sire. With respect to the structure of the validation sample, the results for generation 10 represent the ideal validation sample because it comprises the complete male offspring of the previous generation. In the following discussion, we refer to the results for validation group 10a unless otherwise indicated.

Table 4 Validation reliability (ρ^2) for six different scenarios

Validation set	Sire status	Number of individuals	ρ^2					
			Base	−/25	−/50	−/100	100/100	200/200
9	Reference	1050	26	44	53	62	72	80
10a	Reference	4516	40	56	65	73	80	86
10b	Not reference	10,484	32	51	60	69	77	84

Validation animals were divided according to whether their sire was in the reference set or not

Table 5 Model-derived reliabilities (R^2 were virtually equal across all scenarios), validation reliability (ρ^2) and regression slopes of the –/50 scenario and the three additional scenarios

Scenarios				−/50		$-/50_s$		$-/25_r25_s$ [a]		$-/50_{ub}$	
Validation set	Sire status	Number of individuals	R^2	ρ^2	b	ρ^2	b	ρ^2	b	ρ^2	b
9	Reference	1050	81	53	0.82	35	0.60	40	0.98	53	0.79
10a	Reference	4516	81	65	0.95	42	0.76	48	1.22	65	0.95
10b	Not reference	10,484	76	60	0.92	37	0.70	44	1.14	60	0.91

Validation animals were divided according to whether their sire was in the reference set or not

[a] Higher standard error compared to the other scenarios

Effects of selection

This section is included to illustrate the general effects of selection on validation statistics and to clarify the extent to which the results obtained can be explained by the fact that our population is under selection. The results (Table 3) are in good agreement with expectations and results found by other authors [33–35]. Surprisingly, at first, model-derived theoretical reliabilities were slightly higher for the scenario with directional selection than for the scenario with random selection. However, by analyzing family structures, we found that with directional selection the pattern of family sizes differed, resulting in a more informative structure for validation animals in scenarios with directional selection (results not shown). Model-derived theoretical reliabilities and validation reliabilities show relatively good agreement for the scenario with random selection. The slightly lower values for validation reliabilities are presumably a consequence of the fact that the LD between SNPs and QTL is not perfect and consequently some parts of the additive-genetic variance are not captured by SNPs [36]. However, by simply adding the QTL to the model, we found that validation reliabilities were slightly higher than model-derived theoretical reliabilities. In this case also, the theoretical model is only an approximation of the underlying true model.

The lower values for validation reliabilities under directional selection must be considered as a consequence of selection in the parental generation [33]. When the validation sample itself was selected on a criterion that included Mendelian sampling information, as was the case in generation 9, the decrease in validation reliabilities was even more pronounced. These results are in agreement with previous studies about the effects of selection on theoretical and validation reliabilities [35, 37].

Size and structure of the daughter samples

We tested different scenarios for which increasing numbers of genotyped and phenotyped daughters per sire were included in the reference set. By genotyping 25 daughters per sire from a single generation (corresponding to an overall number of 26,250 genotyped females, Table 2), the validation reliability was considerably improved, from 40 % in the base scenario to 56 % (Table 4, scenario –/25). As the number of daughters increased, the validation reliability showed a nearly linear increase. If we assume that proofs from progeny-testing typically show a validation reliability of about 70 % [38], this threshold is reached in scenario –/100 for validation group 10a and in scenario 100/100 for all other validation groups. With the largest number of genotyped daughters in scenario 200/200 (corresponding to a total of 420,000 genotyped females in the reference set), all validation groups reached reliabilities of 80 % or more. This indicates that large numbers of (unselected) females in the reference set can largely compensate for unfavorable effects such as selection in the parental generation or the effect of a sire for which daughter proofs are not available. As already mentioned, we did not find any relevant differences between scenarios with equal total numbers of females (e.g. scenarios 50/50 and –/100). The similarity between the results of these scenarios is interesting. We expected that a scenario with daughters from two generations such as scenario 50/50 would lead to (slightly) higher validation reliabilities than scenario –/100 because with overlapping generations a larger number of sires would have genotyped daughters in scenario 50/50 and therefore more haplotypes would have been sampled. However, it seems that the existing diversity of haplotypes is already sufficiently covered when genotyping only one generation. In addition, beneficial effects can be reduced by an additional round of meiosis [21]. This implies that a large fraction of the benefits can be already generated in the first generation of a genotyping strategy that considers randomly selected females. Other studies found increases in validation reliabilities when including cows in the reference set but the reported increases were generally much lower e.g. [8, 16, 39]. We see several reasons for such differences. The most obvious one is certainly the larger number of cows that were assumed to be genotyped and phenotyped. Pryce et al. [8] and Koivula et al. [39] added approximately 10,000 genotyped cows to the reference set and Calus et al. [16] only ~1600 first lactation heifers. Other reasons might be related to key parameters such as the reliability of the phenotype [36], effective population size or the LD structure. Moreover, all studies mentioned above used real data that can be differently influenced by selection.

The concept that we propose here is based on genotyping and phenotyping a random sample of (preferably) first-crop daughters of each sire from a generation. We examined how deviations from this design would influence results. Comparison of the results of scenario –/50 (random daughter sample, Table 4) and scenario $-/50_s$ (selected daughter sample, Table 5), showed that with scenario $-/50_s$ the beneficial effect of an additional pool of 52,500 genotypes in the reference set on validation reliability is almost null when compared to the base scenario. Even worse, preselection of daughters caused an increase in inflation as indicated by the low regression slopes (Table 5). One possible explanation is that reference animals that are selected based on their within-family deviation lead to biased family means and also to biased estimates of the deviations from the family mean. Schaeffer [11] argued that the animal model might become obsolete due to the fact that, in the future, only preselected

young bulls will become reference animals. The consequence of this preselection would be that the phenotyped sons of a sire would not represent a random sample of all sons of this sire. Schaeffer [11] expected a relevant increase in inflation as a consequence of this development and given our results this expectation might be at least partly justified. Although not explicitly covered here, it seems likely that the integration of elite cows in the reference set will result in an even stronger bias, because elite cows are not only selected, they frequently receive also preferential treatment so that even their phenotypes are biased. Studies of Wiggans et al. [19] and Dassonneville et al. [20] dealt with the consequences of preferential treatment and provide further evidence of its biasing effects.

The negative result of scenario $-/50_s$ can be only partly removed by a combination of selected and unselected daughters (scenario $-/25_r25_s$; same number of daughters, Table 5). This result indicates that the combination of selected and unselected data cannot yield precise and unbiased estimates. Moreover, the results of scenario $-/25_r25_s$ are lower than those of scenario $-/25$ (Table 4), which indicates that it might be relevant to exclude the genotypes of (pre-)selected daughters from the reference set if this information is available. This kind of monitoring presents an additional challenge especially to single-step genomic BLUP, in which putting a restriction on the reference set is not conceptually intended, an important aspect that was already emphasized by other authors [40].

Another factor with a strong impact on the validation results is the heritability of the trait. In a pilot study [21], we found that for traits with medium to high heritabilities ($h^2 = 0.35$), 100 genotyped daughters per bull increased the marginal reliability [41] by up to 17 % (depending on the scenario) whereas in situations with very low heritabilities ($h^2 = 0.05$), the same number of daughters increased the reliability by up to 4 % only. Our study was limited to a trait with a heritability of 0.4 to investigate several other questions. However, it may be expected that with a lower heritability, less substantial improvements would be found.

In the literature, there are other strategies for genotyping cows. Jiménez-Montero et al. [42] found higher reliabilities when cows selected from both extremes of the distribution of phenotypes were genotyped instead of the best ones or a random sample. We hypothesize that such a strategy would be better suited for traits for which only a few QTL with large effects segregate. Such traits are not common in dairy cattle [43] and therefore we focused our study on a trait with polygenic characteristics, for which no advantage of genotyping extreme animals is expected. Moreover, such a sampling strategy would require trait-specific daughter samples, which is an obstacle for practical implementation. In Calus et al. [16], cow genotypes of entire herds are integrated in the reference set. This strategy could indeed ensure the representativeness of the cow sample if some precautions are taken. We found no disadvantages with moderate unbalancedness in scenario $-/50_{ub}$ in which we ensured that each bull was at least represented by a sample of five daughters. Further investigations on this subject are necessary to clarify which degree of unbalancedness can be tolerated before the accuracy of prediction deteriorates.

In real world breeding programs, it is reasonable to assume that there is a limited interest for the farmers to genotype randomly selected cows and to keep all of them for an unbiased performance recording. Thus, for practical implementation, it would be necessary to find a solution to finance the genotyping costs and to keep track of the cows sampled for the reference set. However, this independent financing solution, once established as a component of the breeding program, might be the only way to ensure a neutral, unselected daughter sample in the long term.

The simple balanced genotyping designs proposed here led to very stable improvements as indicated by the small standard errors of reliabilities and slopes. The only exception was for scenario $-/25_r25_s$, which showed more variation in the results. This indicates that some sampling designs are more robust than others with respect to the improvements that can be achieved.

Conclusions

Extending the reference set by adding a large number of cows with genotypes and phenotypes increases the reliability of breeding values of young selection candidates and may overcome the deterioration of validation reliabilities that are caused by intense preselection of young bulls. We showed the benefits from genotyping a random sample of (first-crop) daughters of all sires from one or two generations. It is possible to obtain reliabilities for selection candidates that are as high as, or even higher than, the reliabilities that have been formerly observed for young progeny-tested bulls. We found that the benefits that can be achieved are sensitive to the strategy used to sample females for genotyping.

Additional files

Additional file 1: Figure S1. LD-structure of the real Fleckvieh population (r2_Fleckvieh data, [30]) and of the simulated population (r2_simulated data) according to distance between SNPs in kb.

Additional file 2: Figure S2. Distribution of the allele frequencies. (A) Simulated data, approximately 38,000 segregating SNPs; (B) Real data on Fleckvieh cattle, approximately 41,000 segregating SNPs.

Additional file 3: Tables S1 and S2. More detailed results on validation reliabilities (ρ^2) (**Table S1**) and on the model-derived theoretical reliabilities (R^2) (**Table S2**) for six scenarios. Description: Validation animals were divided according to the generation of their sire.

Authors' contributions
LP performed the analysis and drafted the manuscript. LP, CE, ECGP, RE and KUG designed the study. LP and CE developed methods. CE, ECGP, RE, JB and KUG revised the manuscript. All authors read and approved the final manuscript.

Author details
[1] Bavarian State Research Center for Agriculture, Institute of Animal Breeding, Prof.-Dürrwaechter-Platz 1, 85586 Poing-Grub, Germany. [2] Institute of Animal Science, University Hohenheim, Garbenstraße 17, 70599 Stuttgart, Germany.

Acknowledgements
We gratefully acknowledge the Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. for their financial support within the research cooperation "Zukunftswege". Furthermore, we wish to thank the editors J van der Werf and H Hayes as well as two unknown reviewers for their helpful suggestions to improve the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 1 March 2016 Accepted: 13 September 2016
Published online: 28 September 2016

References

1. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008;3:e3395.
2. Goddard ME. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009;136:245–57.
3. Edel C, Schwarzenbacher H, Hamann H, Neuner S, Emmerling R, Götz KU. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. Interbull Bull. 2011;44:152–6.
4. Schenkel FS, Sargolzaei M, Kistemaker G, Jansen GB, Sullivan P, Van Doormaal BJ, et al. Reliability of genomic evaluation of Holstein cattle in Canada. Interbull Bull. 2009;39:51–7.
5. Lund MS, de Roos SP, de Vries AG, Druet T, Ducrocq V, Fritz S, et al. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol. 2011;43:43.
6. Su G, Ma P, Nielsen US, Aamand GP, Wiggans G, Guldbrandtsen B, et al. Sharing reference data and including cows in the reference population improve genomic predictions in Danish Jersey. Animal. 2016;10:1067–75.
7. Buch LH, Kargo M, Berg P, Lassen J, Sorensen C. The value of cows in the reference populations for genomic selection of new functional traits. Animal. 2012;6:880–6.
8. Pryce JE, Hayes BJ, Goddard ME. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. In: Proceedings of the 36th ICAR Biennial Session: 16–20 June 2008; Niagara Falls; 2009.
9. Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ. 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. 2012;44:4.
10. Cooper TA, Wiggans GR, VanRaden PM. Short communication: analysis of genomic predictor population for Holstein dairy cattle in the United-States—Effects of sex and age. J Dairy Sci. 2015;98:2785–8.
11. Schaeffer, LR. Is the animal model obsolete in dairy cattle? University of Guelph, personal communication to the Animal Genetics Discussion Group (AGDG). 2014.
12. Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res (Camb). 2011;93:357–66.
13. Patry C, Ducrocq V. Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. Genet Sel Evol. 2011;43:30.
14. Patry C, Ducrocq V. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J Dairy Sci. 2011;94:1011–20.
15. Thomasen JR, Sorensen AC, Lund MS, Guldbrandtsen B. Adding cows to the reference population makes a small dairy population competitive. J Dairy Sci. 2014;97:5822–32.
16. Calus MPL, de Haas Y, Veerkamp RF. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. J Dairy Sci. 2013;96:6703–15.
17. Calus MPL, de Haas Y, Pszczola M, Veerkamp RF. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. Animal. 2013;7:183–91.
18. Egger-Danner C, Cole JB, Pryce JE, Gengler N, Heringstad B, Bradley A, et al. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal. 2015;9:191–200.
19. Wiggans GR, Cooper TA, VanRaden PM, Cole JB. Technical note: adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J Dairy Sci. 2011;94:6188–93.
20. Dassonneville R, Baur A, Fritz S, Boichard D, Ducrocq V. Inclusion of cow records on genomic evaluations and impact on bias due to preferential treatment. Genet Sel Evol. 2012;44:40.
21. Edel C, Pimentel ECG, Plieschke L, Emmerling R, Götz KU. Short communication: the effect of genotyping cows to improve the reliability of genomic predictions for selection candidates. J Dairy Sci. 2016;99:1999–2004.
22. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680–1.
23. Pausch H, Aigner B, Emmerling R, Edel C, Götz KU, Fries R. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel Evol. 2014;45:3.
24. Bundesanstalt Für Landwirtschaft Und Ernährung. Endbericht: Erfassungsprojekt Erhebung von Populationsdaten tiergenetischer Ressourcen in Deutschland: Tierart Rind. 2010 (http://download.ble.de/07BE001.pdf). Accessed 24 Feb 2016.
25. Ertl J, Edel C, Emmerling R, Pausch H, Fries R, Götz KU. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: observations from Fleckvieh cattle. J Dairy Sci. 2012;97:487–96.
26. VanRaden PM, Wiggans GR. Derivation, calculation, and use of national animal model information. J Dairy Sci. 1991;74:2737–46.
27. Edel C, Emmerling R, Götz KU. Optimized aggregation of phenotypes for MA-BLUP avaluation in German Fleckvieh. Interbull Bull. 2009;40:178–83.
28. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
29. Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: applications to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1:21–8.
30. Mäntysaari EA, Liu Z, VanRaden PM. Interbull validation test for genomic evaluations. Interbull Bull. 2010;41:17–22.
31. Ertl J, Edel C, Neuner S, Emmerling R, Götz K-U. Comparative analysis of linkage disequilibrium in Fleckvieh and Brown Swiss cattle. In: Proceedings of the 63rd annual meeting of the European federation of animal science: 27–21 August 2012; Bratislava; 2012.
32. Hill WG. Estimation of effective population size from data on linkage disequilibrium. Genet Res. 1981;38:209–16.
33. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
34. Mäntysaari EA, Koivula M. GEBV validation rest revisited. Interbull Bull. 2012;45:1–5.
35. Gorjanc G, Bijma P, Hickey JM. Reliability of pedigree-based and genomic evaluations in selected populations. Genet Sel Evol. 2015;47:65.
36. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2008;92:433–43.
37. Edel C, Neuner S, Emmerling R, Götz KU. A note on using 'forward prediction' to assess precision and bias of genomic predictions. Interbull Bull. 2012;46:16–9.
38. Powell RL, Norman HD, Sanders AH. Progeny testing and selection intensity for Holstein bulls in different countries. J Dairy Sci. 2003;6:3386–93.
39. Koivula M, Strandén I, Aamand GP, Mäntysaari EA. Effect of cow reference group on validation reliability of genomic evaluation. Animal. 2016;10:1021–6.
40. Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. J Dairy Sci. 2014;97:5833–50.
41. Harris B, Johnson D. Approximate reliability of genetic evaluations under an animal model. J Dairy Sci. 1998;81:2723–8.

42. Jiménez-Montero JA, González-Recio O, Alenda R. Genotyping strategies for genomic selection in small dairy cattle populations. Animal. 2012;6:1216–24.

43. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185:1021–31.

Chapter Three

Genotyping of groups of cows to improve genomic breeding values of low heritability traits and new traits

Submitted for publication

Genotyping of groups of cows to improve genomic breeding values of low heritability traits and new traits

Laura Plieschke[1*], Christian Edel[1], Eduardo CG Pimentel[1], Reiner Emmerling[1], Jörn Bennewitz[2] and Kay-Uwe Götz[1]

[1]Bavarian State Research Center for Agriculture, Institute of Animal Breeding, Prof.-Dürrwaechter-Platz 1, 85586 Poing-Grub, Germany

[2]Institute of Animal Science; University Hohenheim, Garbenstraße 17, 70599 Stuttgart, Germany

[*]Corresponding author

Email addresses:

LP: Laura.Plieschke@lfl.bayern.de

CE: Christian.Edel@lfl.bayern.de

ECGP: Eduardo.Pimentel@lfl.bayern.de

RE: Reiner.Emmerling@lfl.bayern.de

JB: J.Bennewitz@uni-hohenheim.de

KUG: Kay-Uwe.Goetz@lfl.bayern.de

Abstract

Background

Genotyping females and including them into the reference set for genomic predictions in dairy cattle is considered to provide gains in reliabilities of estimated breeding values for young selection candidates.

Methods

By the use of simulation we extended the genomic reference set by including a fixed number of genotyped first-crop daughters for one or two generations of references sires. Moreover, we provided results for the effects of a similar strategy in a situation where for a new trait the recording of phenotypes has recently started. For this case we compared the effect of two different genotyping strategies: First, to phenotype cows but to genotype their sires only, and second, to collect phenotype and genotype on the same cows. We studied the effects on validation reliabilities and unbiasedness of predicted values for selection candidates. We additionally illustrated and discussed the effects of a selected sample and an unbalanced sampling of daughters.

Results

We found, that by extending the reference set with genotyped daughters it is possible to increase validation reliability of genomic breeding values. If the number of phenotypes is limited, as it is in the case of a new trait, it is always better to collect and use genotypes and phenotypes on the same animals instead of using only sire genotypes. We found that the benefits that can be achieved are sensitive to the sampling strategy used when selecting females for genotyping.

Keywords: genomic selection, reference set, genotyping cows, reliability, bias, new traits

Introduction

Genomic selection has changed breeding programs especially in cattle breeding. Bulls can be selected at a younger age with higher reliability. So far, genomic evaluations were primarily implemented for traits with an established performance recording scheme providing phenotypes that were also used in conventional genetic evaluations. For these traits it was possible and straightforward to include all available progeny-tested bulls in the reference set [1]. The changing requirements in the dairy and beef production sector and progress in technology, however, will promote the availability of new phenotypes [2]. It can be assumed that for these new traits only a limited number of phenotypic observations for genetic evaluations will be available in the short to medium term even if a broad performance recording scheme will be established. In other cases new traits will only be recorded on a sample of all cows of the breeding population. For example, only on cows that are milked in automatic milking systems [1] or cows in specific herds. The heritability of many of these new traits with importance for the breeding scheme is often very low [2]. Given these two aspects, it is likely that the reliability of conventional breeding value estimates and the resulting response to selection will be low. Even a genomic reference set consisting of bulls and their daughter yield deviations (DYD) only, will not provide genomic breeding values with high reliability in the foreseeable future due to their low number. It is in relation to this generally unfavorable situation with respect to new traits the additional genotyping of cows providing the phenotypic information has been discussed by some researchers [e.g. 3].

Several studies have shown that the inclusion of females in the reference set of a cattle breed leads to higher reliabilities for young selection animals [e.g. 4-6]. Thomasen et al. [4] found that by additionally including female genotypes in the reference set a higher genetic gain with lower rates of inbreeding can be achieved compared to a breeding scheme where the reference set grows only with the inclusion of newly progeny-tested bulls. Calus et al. [5] also combined cows and bulls in one reference set. In their study the highest validation accuracies were achieved with the combined data set compared to scenarios with only cows or only bulls in the reference set.

This investigation is intended as an extension and follow-up of Plieschke et al. [6] where we have examined the potential to increase reliabilities of breeding values of young selection candidates by genotyping a fixed number of first-crop daughters of each sire of one or two generations in a balanced and regular system of genotyping. In this first investigation we studied a trait of medium to high heritability (0.4). With this short communication we want to add several additional results. First, we want to complement results for a trait with low heritability but keeping all other aspects as in Plieschke et al. [6]. Additionally we want to cover the subject of new traits within the methodological

framework of the approach developed in our first investigation. This paper is therefore structured into two main parts: 1) an 'old trait with low heritability' section and 2) a 'new trait with low heritability' section. In the case of the *old trait* design the whole breeding population of cows is assumed to be phenotyped and the reference set consists of genotyped sires of many generations with a relatively large number of daughters each. In the case of the *new trait* designs, phenotyping has started only recently and therefore only a limited number of phenotypes are available. For genotyping in this situation we investigated two different strategies with genotypes available on sires of phenotyped cows only or genotypes available on the phenotyped cows themselves. We included some additional considerations with respect to the sampling of cows for genotyping and genotyping plus phenotyping, respectively.

Material and Methods

Simulation with QMSim

Most methodological aspects are the same as in Plieschke et al. [6] so we will summarize only the most fundamental aspects. We ran a simulation with four replicates using the open access software QMSim [7]. We first simulated a so-called historical population to create an LD structure sufficiently strong for the situation observed in the Fleckvieh breed. Our founder generation (generation 0) consisted of 30,000 dams and 1500 sires. The pedigreed population was propagated for 10 generations. In every generation of the pedigreed population 15,000 female and 15,000 male offspring were generated by mating 30,000 dams and 1500 breeding sires. Generations were overlapping and in every generation 30% of the dams and 70% of the sires were replaced. We simulated 30 chromosomes each with a length of 100 cM. On each chromosome 1660 markers and 30 QTL were evenly distributed (49,800 markers and 900 QTL in total). After routine checks [8, 9] nearly 38,000 valid markers and approximately 700 QTL still segregating in the reference set (both depending on repetition) were available. We assumed a sex linked trait with polygenic nature and a single observation for every female.

For the sake of brevity in this investigation we only consider the animals of generation 10 as validation sample. This corresponds to validation group 10a in Plieschke et al. [6], and consists of unselected candidates whose sires and/or (half-) sisters are part of the reference set. For these animals genomic breeding values were calculated from several reference sets varying according to the design and scenario investigated. Estimates then were compared to the true breeding values from the simulation.

Simulation of the daughter sets

For each reference bull of generation 7 and 8 we generated 200 additional daughters without using QMSim. We generated gametes by combining known haplotypes (markers and QTL) of these sires and of potential breeding dams from the population. The gametes were randomly mated (excluding sisters, daughters and dams of bulls) to create genotypes of the daughters. When generating the gametes we simulated on average one random crossing-over event per Morgan for each chromosome. True breeding values (TBV) were derived by summing the true QTL effects for all relevant loci of a daughter.

Phenotypes

Phenotypes used were yield deviations [YD, 10] for females and DYDs [10] for sires. YDs for females were generated using the corresponding TBV and a random residual. YDs of all daughters were used to calculate DYDs of the corresponding sire. Females that were themselves part of the reference set were omitted from the calculation of DYDs. Phenotypes were weighted with the equivalent number of own performances calculated as $EOP=\lambda \frac{R^2_{phen}}{1-R^2_{phen}}$, where $\lambda=\frac{\sigma^2_e}{\sigma^2_a}$ and R^2_{phen} is the reliability of the DYD or YD respectively.

Designs

Old trait

The *old trait* design assumed a situation where all cows of the breeding population are phenotyped routinely as in Plieschke et al. [6]. All sires of generations 0 to 8 therefore have DYD and generations 5 to 8 are used as reference set in the base scenario. This scenario was then extended by genotyping daughters of either the last generation or the last two generations of reference bulls and integrating them into the reference set (Table 1). Simulated heritability of the trait was 0.05.

New trait

For the two *new trait* designs we assumed a situation where phenotyping of cows has just begun. Here, phenotypes of daughters were assumed to be available for bulls of generations 8 or generations 7 and 8 only. Two strategies for genotyping were investigated: In design *NTsires* cows are phenotyped but only their sires are genotyped and used in the reference set. In design *NTcows* genotypes of phenotyped cows are available and cows are used directly as the reference set. The number of reference animals therefore differs considerably between these two designs and is summarized in Table 1. Simulated heritability of the trait was 0.05, too.

Special scenarios

We used scenario --/50 but changed the composition of the sample of for the *old trait* and the two *new trait* designs to investigate additional scenarios. In scenario --/50_{sg}, daughters of sires were selected on their phenotype. For the *old trait* design we selected the best 50 out of 200 daughters of each sire of generation 8 for genotyping, using the YD as the selection criterion. Since all 200 daughters were assumed to be phenotyped and contribute to the DYD of their sire this scenario might be labelled as selective genotyping. In the case of the two *new trait* designs a situation of selective genotyping is probably of little relevance. In this case we assumed a situation where the worst 33% of daughters of each sire might not be considered for replacement and therefore will not even reach the stadium of being potential candidates for phenotyping (remaining daughter sample per sire: the best 133 daughters out of the 200 daughters generated for each sire). Then we again sampled randomly 50 daughters to get a comparable sample size to scenario --/50. Such a situation might occur when sampling strategy is intended to be based on an unselected daughter sample, but selection within herds already took place based on a visible phenotype that has not (yet) been recorded. Such a situation might happen with some conformation traits. Maybe even worse might be a situation where genomic breeding values of all cows are available and used for replacement selection within herds. Since only these 50 daughters are assumed to be phenotyped and no other daughters of the sire are phenotyped this situation was labelled selective phenotyping (scenario --/50_{sp}). We also included an unbalanced scenario (--/50_{ub}) with different numbers of daughters per sire genotyped and/or phenotyped but the overall number of females was kept the same as in scenario --/50. This was done by randomly selecting five daughters for 330 sires, 50 daughters for 621 sires and all 200 daughters for 99 sires. Table 2 gives an overview of the two additional scenarios and the number of animals in the reference set.

Genomic prediction

As in Plieschke et al. [6] we calculated direct genomic values (DGVs) and reliabilities using a SNP-BLUP model [11]. The model equation can be described as:

$$\mathbf{y = Xb + Mg + e}$$

and the corresponding mixed model equations are:

$$\begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}M} \\ \mathbf{M'R^{-1}X} & \mathbf{M'R^{-1}M+I}/\sigma_g^2 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{M'R^{-1}y} \end{pmatrix}$$

with

$$\sigma_g^2 = \frac{\sigma_a^2}{\sum_{j=1}^{m}(2p_jq_j)}$$

where **y** is the vector of observations (here DYD or YD), **b** the vector of fixed effects (in our simulation only an overall mean), **g** is the vector of random marker effects, and **e** the vector of residual effects. Matrix **X** is a design matrix which links the observations to the respective fixed effects and **M** is the centered coefficient matrix of marker genotypes and p_j and q_j are the estimated base allele frequencies [12]. Centering was done by subtracting two times the base allele frequency estimate from the corresponding column of **M** [13]. Matrix **R** is a diagonal matrix with σ_e^2/w_i on the diagonal, where w_i is the EOP of the *i*-th observation and matrix **I** is an identity matrix of order *m* (number of markers).

DGVs are calculated as

$$DGV = \hat{\mathbf{b}} + \mathbf{M}\hat{\mathbf{g}}.$$

For the analysis we calculated a weighted regression of TBV on DGV for validation animals. We used the model fit of this regression as a measure of validation reliability (ρ^2) and the regression slope (b) as a measure of bias describing the inflation of estimates [14].

Results

Results shown below are averages of validation reliability (ρ^2) and the regression slope (b) over four repetitions of the simulation. Standard errors of the results presented in the main body of the paper were lower than 1.2 % for validation reliabilities and lower than 0.035 for regression slopes.

Effects of genotyped daughters

Table 3 summarizes validation reliabilities (ρ^2) for 5 different scenarios analyzed. For the *old trait* design in the base scenario without genotyped daughters a validation reliability of 37% was achieved. Extending the reference set with an increasing number of daughters increased the validation reliability by more than 20% (in the most extended scenario with 4200 sires and 210,000 daughters genotyped). Also for the two *new trait* designs an increasing number of phenotyped (and genotyped) daughters lead to increasing ρ^2. The scenario using genotyped cows (*NTcows*) yielded always more reliable DGV than using cow phenotypes only via their genotyped sires**.** Even in the largest scenario with 210,000 females phenotyped and 2100 sires genotyped ρ^2 was only 25%, whereas with the corresponding *NTcows* scenario a validation reliability of more than 50% could be achieved.

Effects of the composition of the daughter samples

Table 4 summarizes validation reliabilities (ρ^2) and regression slopes (b) of three different strategies to genotype and/or phenotype females. Using scenario --/50 as a reference scenario it can be stated that direct selection of the daughter samples (scenario --/50_{sg}) had negative effects on ρ^2 and b in almost all the cases. The strategy of selective genotyping in case of the *old trait* design lead to a validation reliability even lower than the base scenario (ρ^2 = 37%, Table 3) and highly inflated estimates. For the *new trait* design *NTsires* resulted in higher ρ^2 when daughters were selectively phenotyped (scenario --/50_{sp}, Table 4), but at the price of highly deflated estimates. With *NTcows* selective phenotyping lead to lower validation reliability (ρ^2 = 26%) than the reference scenario but still higher than any of the *NTsires* scenarios. However, estimates showed a considerable deflation (b = 1.25).

For the *old trait* and for the new trait design *NTcows* a moderately unbalanced scenario (--/50_{ub}, Table 4), showed no effect on reliabilities or regression slopes whereas for the *new trait* design *NTsires* scenario --/50_{ub} resulted in very low ρ^2 and highly inflated estimates (b = 0.35).

Discussion

Old Trait

General effect of genotyped cows

By adding genotyped daughters to the reference set, it is possible to increase validation reliability for young selection candidates. We found improvements in validation reliability of 5-21% points depending on the scenario. Other studies confirm that it is possible to increase reliability or accuracy for low heritability traits by genotyping females. Jiménez-Montero et al. [15] analyzed accuracies for a low heritability trait with h^2 = 0.1 and a sampling strategy which is comparable to ours and found similar improvements when adding up to 50,000 females to the reference set. Egger-Danner et al. [2] reported comparable results for a trait with a heritability of 0.05.

Our results indicate that such an increase can only be achieved if the sample of cows included in the reference set is not selected based on their phenotype or any other criterion providing information on the individual Mendelian sampling deviation (including genomic breeding values [15]). With selected cows contributing to the reference set, reliabilities might be considerably reduced when compared to a random sample of genotyped cows (scenario --/50) even to a point where the resulting reliability is actually lower than without additional genotypes (base scenario). Moreover, scenarios like --/50_{sg} lead to inflated results (b = 0.36 for --/50_{sg}). Although we were not able to study this in detail, we suppose that the directional selection of daughters for genotyping leads to biased

daughter means and an inconsistency between genotyped and non-genotyped daughters. This might contribute to the unprecise and biased estimates we found. Again observations e.g. by Jiménez-Montero et al. [15] confirm that a sampling strategy that selects only the best cows leads to much lower accuracies than a random sampling strategy. This effect was even more pronounced when sampling was based on the breeding value instead of phenotype [15].

Comparison to higher heritabilities

In Plieschke et al. [6] we investigated the same design and scenarios as described here for *old trait* but with a higher heritability (h^2 = 0.4). We found that for a trait with a heritability of only 0.05 validation reliability for the scenarios investigated is between 3 and 23 % points lower than for a trait with higher heritability. Comparing the results we also found that with an increasing number of daughters, the difference between the two traits also increases indicating that genotyping females will also have a relatively higher effect for high and medium heritable traits. Similar observations were made by Edel et al. [16] in a deterministic approach and a theoretical justification was given by Hayes et al. [17].

New Trait

Comparison of NTsires and NTcows

Since the size of the reference set has a large impact on the reliability of genomic prediction [18] *NTcows* always leads to better results than *NTsires* even though the heritability of an individual cow phenotype is lower than that of a DYD of a sire (5% compared to 39% for scenario --/50). However, the strategy to phenotype females and genotype only their sire is still common for genomic selection programs. Buch et al. [1] also tested two scenarios where they genotyped the phenotyped cows themselves or their sires for a trait with h^2 = 0.05. Their results support our results in general but resulting reliabilities were on a much lower level due to the fact that their reference sets were much smaller in the different scenarios.

Effects of selective phenotyping

Some projects for genotyping females plan to genotype calves from which they then want to sample cows for phenotyping. In such a strategy, it is not possible to prevent some calves from being selected before a phenotype can be recorded. Then the phenotyped cows are assumed to be a random sample, however, they are already pre-selected. This has consequences on validation reliability and, moreover, it leads to highly biased breeding values.

NTsires: We found that scenario --/50_{sp} with selective phenotyping leads to higher validation

reliabilities than scenarios with randomly selected daughters. Results were 12 and 17 % for scenarios --/50 and --/50_{sp}, respectively. Analyzing the variation within daughter samples of sires (results not shown) we found a smaller variation of YDs resulting in lower standard errors for sire means. This is a consequence of the low heritablity of the trait leading to a situation where selection primarily operates in reducing the residual variance of the phenotypes. This results in higher correlations between the mean of the daughter sample and the sire's TBV, but breeding values are overestimated (b = 1.17 and b = 1.45 for scenarios --/50 and --/50_{sp}, respectively). We do not expect to find similar results in practice, because it is unlikely that selection intensities would be the same in all sire families, which would introduce additional variation between sire estimates.

In contrast to the other designs, we found that an unbalanced phenotyping strategy lead to very low validation reliability (only 3 % for scenario --/50_{ub}). Since the size of the reference set with *NTsires* is already very small and DYDs of sires are based on a limited number of lowly heritable daughter phenotypes, reducing the number of daughters for some sires to only 5 virtually eliminated these data points. The effective size of the reference set is therefore decreased and validation reliabilities are reduced.

NTcows: The scenario with selective phenotyping of cow samples had lower validation reliability than the scenario with the randomly selected females. Moreover, the estimated breeding values are highly inflated (b = 1.25 for --/50_{sp} vs. b = 0.97 for --/50). However, the negative effects of selective phenotyping seem to be somewhat weaker than for selective genotyping in the *old trait* design.

General considerations

We only tested one new trait with a low heritability although there are other new traits with a higher heritability like dry matter intake [19] or methane emission [20], for example. The general trends we observed in all our simulations were, however, quite similar no matter what heritability was assumed. Some aspects, like the negative effect of unbalanced daughter samples in *NTsires*, might not be observable with a higher heritable trait.

In this investigation and in Plieschke et al. [6] we calculated the phenotypes to be used in our two-step approach of genomic breeding value estimation based on true breeding values from the simulation plus residuals (YD) and aggregated the DYD of bulls directly based on these YD of daughters. In practice one would also have to cope with biased estimates for the YD of genotyped daughters, which is another argument for random sampling of daughters to be genotyped.

Conclusions

Extending the reference set by adding a large number of cows with genotypes and phenotypes increases the reliability of breeding values of young selection candidates also for low heritability traits. The gain found was much lower than for a trait with higher heritability. In the case of a new trait genotyping of cows seems to be the only realistic option to obtain reasonable reliabilities in due time. We found that the benefits that can be achieved in all cases are sensitive to the sampling strategy used to select females for genotyping.

Acknowledgments

We gratefully acknowledge the Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. for their financial support within the research cooperation ”Zukunftswege“.

Authors’ contributions

LP performed the analysis and drafted the manuscript. LP, CE, RE, JB and KUG designed the study. CE and LP developed methods. CE, ECGP, RE, JB and KUG revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

[1] Buch LH, Kargo M, Berg P, Lassen J, Sorensen C. The value of cows in the reference populations for genomic selection of new functional traits. Animal. 2012;6:880-6.

[2] Egger-Danner C, Cole JB, Pryce JE, Gengler N, Heringstad B, Bradley A, et al. Invited review: Overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal. 2014;9:191-20.

[3] Calus MPL, de Haas Y, Pszczola M, Veerkamp RF. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. Animal. 2013;7:183-91

[4] Thomasen JR, Sorensen AC, Lund MS, Guldbrandtsen B. Adding cows to the reference population makes a small dairy population competitive. J Dairy Sci. 2014;97:5822-32.

[5] Calus MPL, de Haas Y, Veerkamp RF. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. J Dairy Sci. 2013;96:6703-15.

[6] Plieschke L, Edel C, Pimentel ECG, Emmerling R, Bennewitz J, Götz KU. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. Genet Sel Evol 2016;48:73.

[7] Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680-1.

[8] Edel C, Schwarzenbacher H, Hamann H, Neuner S, Emmerling R, Götz KU. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. Interbull Bull. 2011;44:152-6.

[9] Ertl J, Edel C, Emmerling R, Pausch H, Fries R, Götz KU. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: Observations from Fleckvieh cattle. J Dairy Sci. 2012;97:487-96.

[10] VanRaden PM, Wiggans GR. Derivation, calculation, and use of national animal model information. J Dairy Sci. 1991;74:2737-46.

[11] Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

[12] Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: applications to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1:21-8.

[13] VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.

[14] Mäntysaari EA, Liu Z, VanRaden PM. Interbull validation test for genomic evaluations. Interbull Bull. 2010;41:17-22.

[15] Jiménez-Montero JA, González-Recio O, Alenda R. Genotyping strategies for genomic selection in small dairy cattle populations. Animal. 2012;6:1216-24.

[16] Edel C, Pimentel ECG, Plieschke L, Emmerling, R, Götz KU. Short communication: The effect of genotyping cows to improve the reliability of genomic predictions for selection candidates. J Dairy Sci. 2016;99:1999-2004.

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

[18] Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009;92:433–43.

[19] Tetens J, Thaller G, Krattenmacher N. Genetic and genomic dissection of dry matter intake at different lactation stages in primiparous Holstein cows. J Dairy Sci. 2014;97:552-31.

[20] Lassen and Lovendahl. Heritability estimates for enteric methane emissions from Holstein cattle measured using noninvasive methods. J Dairy Sci. 2016;99:1959-67.

Table 1: Scenarios with corresponding number of animals in the reference set.

Scenario	Number of reference animals		
	Old trait	New Trait	
		NTsires	NTcows
Base	4200	/	/
--/25	30,450	1050	26,250
--/50	56,700	1050	52,500
--/100	109,200	1050	105,000
100/100	214,200	2100	210,000

The names of the extended scenarios are derived from the number of phenotyped daughters and the sire's generation. The number before the slash in the scenario's name is the number of daughters per progeny-tested bull of generation 7 and the number after the slash is the number of daughters per progeny-tested bull of generation 8. In the case of *NTsires* the sires of phenotyped daughters are genotyped only. In both cases either *NTsires* or NTcows the column *NTcows* gives also the number of available phenotypes used either as either a DYD (*NTsires*) or directly as a YD (NTcows).

Table 2: Overview of the special scenarios.

Scenario	Sample size	Sampling strategy	Sampling population (daughters/sire)	Label
--/50	50	random	200	reference scenario
--/50$_{sg}$	50	selected*	200	selective genotyp.
--/50$_{ub}$	±50	unbalanced**	200	unbalanced
--/50$_{sp}$	50	random	133***	selective phenotyp.

* the best 50 daughters per sire were selected for genotyping; ** different numbers of daughters per sire were genotyped and/or phenotyped (5, 50 or 200 daughters); *** the best 133 daughters out of the 200 daughters per sire were pre-selected.

Table 3: Validation reliability (ρ^2) for 5 different scenarios for the old trait design and the two new trait designs.

	ρ^2 (%)		
Scenario	**Old Trait**	**NTsires**	**NTcows**
Base	37	-	-
--/25	42	8	20
--/50	46	12	30
--/100	50	18	41
100/100	58	25	54

Table 4: Validation reliability (ρ^2) and regression slope (b) for 3 different scenarios with the same number of daughters but different sampling strategies.

	Old Trait		NTsires		NTcows	
Scenario	**ρ^2 (%)**	**b**	**ρ^2 (%)**	**b**	**ρ^2 (%)**	**b**
--/50	46	1.06	12	1.17	30	0.97
--/$50_{sg/sp}$*	10	0.36	17	1.45	26	1.25
--/50_{ub}	46	1.06	3	0.35	29	0.96

* selective genotyping (sg) in the case of *old trait* and selective phenotyping (sp) in the case of the two *new trait* designs.

General Discussion

In recent years the breeding value estimation system of many large dairy cattle breeds has been extended by the prediction of genomic breeding values. The most common procedure to estimate genomic breeding values in practical applications is GBLUP (VanRaden, 2008). In GBLUP the genomic relationship matrix (**G**) calculated from SNP markers describes the genetic relationships between animals.

All investigations were done having the two largest Bavarian cattle breeds in mind: chapter one was based on data from Fleckvieh and Brown Swiss cattle and chapter two and three with simulated data resembling the genetic composition and population structure of Fleckvieh. The studies presented in the past three chapters have already addressed some methodological and strategic aspects related to genomic selection. These two aspects are discussed in the following section in relation to other related issues. In addition, some technological aspects are addressed which are also related to current developments in genomic selection.

Methodological aspects

The use of GBLUP implies several assumptions, the most obvious one being that the relationship estimated based on markers is a valid estimate of the relationship based on QTL. This might only be true if a trait is assumed to be 'polygenic', meaning that many QTL contribute to trait variance, if the contribution of each QTL is limited and if the linkage disequilibrium (LD) between markers and QTL is sufficiently strong. This means that the allele frequency distribution of marker and QTL are at least comparable. An important point related to methodological aspects is that GBLUP as it is proposed by VanRaden (2008) can be shown to be equivalent to the so called SNP-BLUP approach (Meuwissen et al., 2001) which uses estimates of SNP effects in linear projections (Goddard, 2008; Plieschke et al., 2015).

Different approaches to calculate G

There are several proposals in the literature on how to calculate the genomic relationship matrix **G** but the way proposed by VanRaden (2008, "method one") seems to be the most common approach in practical applications. The formula used is

$$\mathbf{G} = \frac{\mathbf{MM'}}{\sum_{i=1}^{m}(2p_iq_i)}$$

with **M** being constructed as **M** = **K - P**, where **K** is the $(n \times m)$ matrix of gene-contents of the reference allele (2, 1 or 0), **P** is an $(n \times m)$ matrix of which the ith column is $2p_i$, n and m are the number of animals genotyped and the number of markers, respectively and and are known or estimated base allele frequencies. The allele coding proposed by VanRaden (2008) uses assumptions about the allele frequencies in the population and can easily be shown to agree with the quantitative genetics theory that uses $2q$, $(q - p)$ and $-2p$ for genotypes AA, AB and BB respectively (Falconer and Mackay, 1996). If the true but usually unknown base frequencies are used (frequency among animals considered to be the pedigree base) it should at least theoretically lead to a coefficient matrix compatible to the genetic parameter (VanRaden, 2008).

There are other approaches for the calculation of **G** like, for example, the unified additive relationship (UAR) approach of Powell et al. (2010) or "method two" of VanRaden (2008). Method two of VanRaden (2008) is calculated as $\mathbf{G_2} = \mathbf{MDM'}$ where $d_{ii} = \frac{1}{m(2p_iq_i)}$. This method as well as the method of Powell et al. (2010) weights markers by reciprocals of their expected variance instead of summing expectations across loci and then dividing as done in "method one". As a consequence the form of the **G**-matrix is strongly influenced by rare alleles with extreme allele frequencies (Endelman and Jannink, 2012).

The shown approaches calculate a genomic relationship matrix based on IBS (identical-by-state) information. The idea of genomic selection is that each QTL is in sufficiently large LD with the nearby markers and these markers explain a large proportion of the genetic variance. However, according to Habier et al. (2007) genomic breeding values also incorporate information of LD arising from recent family structures. LD generated by family structure can be explained by linkage analysis (LA). This fact implies that genomic selection can also use LA information (Luan et al. 2012). LA information can be used by a genomic identity-by-descent (IBD) matrix, containing identity-by-descent probabilities calculated based on genotypes and the known pedigree. Since an IBD relationship matrix in some sense uses more information than an IBS relationship matrix it seems reasonable to assume that an IBD approach might lead in some cases to higher reliabilities than an IBS approach. However, Luan et al. (2012) found no significant differences between their IBS, IBD and IBD+IBS approach. In a later study (Luan et al., 2014), using simulated data, they found that a relationship matrix based on runs of homozygosity achieved genomic breeding values with higher accuracies than the compared genomic relationship matrices. However, using real data, only small – if any – differences in accuracies were found by Luan et al.

(2014). From the practical point of view, using "method one" of VanRaden (2008) seems to be the fastest and easiest way computing a genomic relationship matrix.

Analyzing the G-matrix

There are several implicit assumptions in the conventional animal model as well in the GBLUP model. An important aspect is that the base population is assumed to be homogenous and is assumed to be in Hardy-Weinberg equilibrium (HWE). It is likely that these assumptions do not hold in many breeding populations. By analyzing **G** instead of the numerator relationship matrix violations of these assumptions can become apparent since they do influence the form of **G**. The question can be raised whether the sometimes unexpected form of **G** is reasonable from a scientific standpoint or not. These differences can be caused by deviations in allele frequencies between founder populations and a disturbance in HWE which can for example lead to an excess of heterozygotes.

A related question of practical importance is for example whether to include international genotypes or genotypes of a so called "equi-breed" into a national genomic evaluation of any country is advisable. In the case of the joined German-Austrian genomic evaluation (Edel et al, 2011) questions of concern were for example: can Original Braunvieh be part of the genomic evaluation for Brown Swiss? Can the Czech population of Fleckvieh cattle be integrated? And, more fundamentally, are the coefficients observed in the **G**-matrix meaningful?

In the study of Plieschke et al. (2014) the question addressed was whether including international genotypes in the German-Austrian genomic evaluation system for Brown Swiss would induce effects on the estimates of the German-Austrian population that are of any relevance. Using principal component analysis (Patterson et al., 2006; Zou et al. 2010) and F_{st} statistics (Weir and Cockerham, 1984; Caballero and Toro, 2002) it was found that there is some degree of genetic separation detectable within the recent genotyped Brown Swiss population. It was found that the inclusion of foreign genotypes in the reference population had a noticeable impact on the breeding values of German-Austrian candidates and results gave an indication that there might be effects beyond a simple numerical enlargement of the reference population as a consequence of population subdivision. Moreover, it was observed that an increase of the reference population did not necessarily lead to an increase in model based reliabilities. This was the case when Original Braunvieh was integrated.

The hypothesis deduced was that **G** includes information related to the genetic distance between potentially discriminable groups in the base population that is defined by the pedigree. Several other studies show that it is possible to detect population subdivision with **G** (Zou et al. 2010; Kadri et al., 2014) analogously to the analyses of the pre-genomic era that were done based on

the numerator relationship. This hypothesis was investigated here in chapter one. By using a simple and straightforward method to estimate allele frequencies in the base population as proposed by Gengler et al. (2007) the genomic relationship matrix (**G**) was separated into two independent components: a base group component $\mathbf{G}^*_{\mathbf{A}}$ and a segregation component $\mathbf{G}_{\mathbf{S}}$. This decomposition allows studying some aspects in more detail: first, whether and how differences in allele frequencies between base groups contribute to the proportion of genetic variance explained by differences between base groups; second, how the effects estimated for the base groups influence the current population and their genomic predictions. The concept extends the classical approach for modeling base groups in genetic evaluations to the GBLUP case. It shows that parts of the genetic variation represented by the **G**-matrix can be assigned to systematic differences in allele frequencies between populations in the base. This implies that standard GBLUP is equivalent to a model that fits correlated random genetic groups, where differences in group means are modeled as part of the natural additive-genetic variance.

Non-linear models

It seems reasonable that for some traits the 'infinitesimal model' assumed in the GBLUP approach does not hold. It can be assumed that the genetic architecture differs between traits (Daetwyler et al., 2010) and that there are some traits that are influenced by some major QTL. For such traits a standard GBLUP approach might not be the best one.

Zhang et al. (2010) showed a possibility to calculate trait specific genomic relationship matrices to take QTL with large effects into account by putting greater weight on loci explaining more of the genetic variance of the trait than other loci. Their approach might in some cases be superior for traits that are influenced by major QTL.

It was also found that Bayesian models using more complex assumptions than GBLUP can lead to higher reliabilities (e.g. Clark et al., 2011; Zeng et al., 2012). However, compared to more complex Bayesian models standard GBLUP and the trait specific GBLUP have some advantages in practical application (Zhang et al., 2010). From a theoretical standpoint it seems to be unreasonable to believe that BLUP models can take full advantage of the LD information (Meuwissen and Goddard, 2010) when using high-density SNP chips. However, the theory of effective chromosome segments demonstrates the interrelation between the limited effective size of our breeding populations, the limited number of segregation segments and the possible limits of resolution and estimability when trying to track phenotypic variation right down to the specific mutation. These arguments might argue in favor of GBLUP and may be used as a justification for why it is still a useful approximation that is used by many countries in routine genomic evaluation besides the simplicity and low computational requirements (Gao et al., 2013).

Multi-step vs single-step

Genomic evaluations for Brown Swiss and Fleckvieh are currently calculated with a GBLUP multi-step approach (VanRaden 2008, Hayes et al., 2009a). A typical multi-step evaluation requires: 1) conventional breeding value estimation to get an estimated breeding value (EBV), 2) calculation of pseudo-observations such as DYDs, 3) estimation of direct genomic values for genotyped animals, and optional 4) combining the direct genomic value with traditional parent averages (PA) or EBV (Hayes et al., 2009a; VanRaden et al., 2009). Those steps are dependent on many parameters and assumptions. If all these steps can be taken together in one step, fewer assumptions have to be made and fewer parameters have to be estimated. The single-step procedure (Legarra et al., 2009 and Christensen and Lund, 2010) is intended to eliminate several assumptions and parameters, and to calculate more accurate genomic evaluations than the multiple-step procedures (Aguilar et al., 2010). In simplified terms, the single step equations are similar to Henderson's mixed model equations from an animal model, but with covariance structure described by an **H**-matrix instead of the numerator relationship matrix **A**. The inverse of the **H**-matrix can be calculated according to Aguilar et al. (2010) as

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

where 2 stands for the group of genotyped animals. One advantage of single-step approaches is that genotyped as well as ungenotyped animals are part of the genomic evaluation. The consequence is that the computing effort is considerably increased. As $\mathbf{A^{-1}}$ can be computed very easily, the additional effort is caused by the $\mathbf{G^{-1}} - \mathbf{A_{22}^{-1}}$ block. In addition, there are some questions that have not yet been fully clarified. For example, the bias which is frequently occurring when using single-step BLUP can only be reduced if the correct scaling factors have been chosen, for which there is no theoretical explanation (Koivula et al., 2015; Pimentel et al., 2016). Currently it is not entirely clear where the bias comes from and how to prevent it. Therefore multi-step GBLUP is still the procedure applied to estimate genomic breeding values in Germany for Fleckvieh and Brown Swiss.

Strategic aspects

Assuming the infinitesimal model holds and the realized relationship matrix **G** to be a sufficiently valid and meaningful representation of the true additive-genetic relationship, Hayes et al. (2009b) showed that given a specific structure of the genome and a defined family structure (e.g. full sibs) it is possible to derive the variance of the realized relationship between family members. Based

on that, it is possible to derive the number of related animals that must be phenotyped and genotyped in order to achieve a desired level of reliability. Hayes et al. (2009b) found that the reliability of genomic breeding values depends to a large extend on the number of family members and the heritability of the trait. In their paper, they used simple family structures to explain their analytical approach. In a deterministic simulation Edel et al. (2016) showed that it is possible to increase reliability of genomic breeding values of young selection candidates when genotyping first-crop daughters of AI bulls. The simulation was based on nuclear pedigrees extracted from the German-Austrian Fleckvieh population. Using this "within-family structure" they found that the information from genotyped females contributes predominantly to the reliability of the Mendelian sampling part of the breeding value estimate of genomic candidates. Limitations of their investigation were that they could not quantify the assumed cumulative effects at the population level arising from the theory of effective chromosome segments.

Genotyping cows for known traits

For chapter two and three a population was stochastically simulated by using the simulation program QMSim. The intention was to simulate a population resembling the current Fleckvieh population with respect to important key characteristics. With this breeding population a reference population for genomic predictions was established, which consisted entirely of bulls with phenotyped daughters and then was extended in a stepwise manner by adding large numbers of daughters with genotypes and phenotypes. Chapter two shows the results of a trait with a heritability of 0.4. Chapter three used the same data structure as used in chapter two but for a trait with heritability of 0.05.

As the number of daughters increased, the validation reliability increased as well. It was found that genotyping females had a higher effect for the trait with high heritability and a lower effect for the trait with low heritability, which is in accordance with Edel et al. (2016) and Hayes et al. (2009b). With decreasing heritability the environmental noise affecting the phenotypic value increases and every daughter contributes only one observation to estimate her individual Mendelian sampling deviation. Therefore, with decreasing heritability more daughters are needed to achieve a desired increase in reliability. If female genotypes are to be included in a genomic system, one of the key questions is which cows should be genotyped. The results of chapter two and three indicate that genotyping a selected daughter sample instead of a random sample decreases the beneficial effect on validation reliability.

Several organizations already include genotyped cows in their reference population. In the United States of America elite females were integrated in genomic breeding value estimation since the introduction of genomic selection (Wiggans et al., 2011). It was found that including those

elite females led to higher bias and therefore cow information was adjusted to be similar to those of bulls. With this adjustment the accuracy of genomic evaluations for Holsteins and Jerseys was increased (Wiggans et al., 2011). In Denmark, Finland and Sweden (DFS) a Nordic "LD-project" started in 2014, in which a low-cost low-density chip was offered to the breeders, who should in turn voluntarily genotype all young animals in their herds (Langdahl, 2014). In 2016 a project called KuhVision started in Germany (DHV, KuhVision). Within this project breeders of German Holstein Friesian have the opportunity to genotype all females at discounted rates, which are dependent on the number of phenotypes.

Genotyping cows for new traits

Genotyping of females has advantages beyond that of increasing the reliability of genomic estimates for traits with established performance recording alone. Genotyping of cows is often mentioned in the context of new or expensive-to-measure traits (Calus et al., 2013; Egger-Danner et al., 2014). In the case of an old trait, the number of phenotypes is expected to be (nearly) unlimited, since a recording system is well established. In the case of a new trait recording of phenotypes would have just started, therefore the number of phenotypes should be limited. For genotyping in this situation there are two different strategies with genotypes available on sires of phenotyped cows only or genotypes available on the phenotyped cows themselves. For the simulated new trait a heritability of 0.05 was assumed in chapter three. For the new trait designs investigated, it was found that it is always better to genotype the phenotyped cows themselves (*NTcows*) instead of using their phenotypic information via their genotyped sires (*NTsires*).

To consider new trait designs might have different reasons (Egger-Danner et al., 2014): new phenotypes will be available as a consequence of technical developments, for example automatic milking or feeding systems that will provide new measurements of fitness and milk parameters or parameters of conformation on a regular basis. Additional aspects that are promoting the necessity for new breeding traits are growing world population, negative effects of climate changes and the need for a higher efficiency when using limited resources. Furthermore, the demand of consumers for the issue of animal welfare is increasing and a decreasing use of antibiotics is desired. However, the motivation of farmers and veterinarians to participate in a non-mandatory monitoring project is usually low. Therefore, recording new traits must provide further benefits to the producer in order to motivate the required extra effort. In Germany there are two examples of such monitoring projects, "GKuh" and "ProGesund". In both projects, veterinarian diagnosis, treatments and observations by the farmer are recorded to get a large database for health related traits. In addition, farmers have the opportunity to use these data to optimize the management of their herd and to make better selection decisions. Having these databases already in place, it

might be reasonable to genotype cows that have been already phenotyped to make better use of these valuable phenotypes.

However, the current breeding value estimation already includes many traits (up to 43 (Egger-Danner et al., 2014)), so it should at least be questioned whether further traits might be useful in the selection decision or whether the clarity and comparability of animals may suffer. Certainly the replacement of indicator traits by directly measured traits has its justification. An example would be the direct estimation of mastitis instead of the somatic cell count. Other traits might be collected in specific herds for research purposes only.

Further advantages when genotyping cows

Genomic selection is working very well for large cattle breeds with a sufficiently large reference population with many progeny-tested bulls. However, the number of progeny-tested bulls is limited for numerically small dairy cattle populations. Su et al. (2015) tested strategies for Danish Jersey to get a larger reference population. In a first step US Jersey bulls were included in the reference population and in a second also a large number of genotyped cows. They found that both including foreign bull genotypes as well as cow genotypes in the reference population greatly increased reliability of genomic prediction in Danish Jersey. Thomasen et al. (2014) also found that integrating females in the reference population of a small breed had several advantages. It was a profitable and fast way to increase reliabilities of genomic predictions. Furthermore, it also increased genetic gain and decreased the rate of inbreeding compared with breeding schemes that only updated reference populations with progeny-tested bulls.

Beside the improvement concerning the reliability of genomic breeding values and the use for new (health) traits, genotyping cows might be especially useful as a management tool for the farmers. Pryce et al. (2009) listed different advantages especially when the whole herd of the farm is genotyped. First, it is easier to identify elite females and the best heifers to become herd replacements. Second, inbreeding can be avoided using genomic assisted mating plans. Third, genetic defects can be managed to a large extent when avoiding matings of cows and bulls that are both identified to carry genetic disease allele. Genotyped females might also be used to estimate non-additive effects. Interactions between genes result in non-additive genetic variation. Dominance is the interaction between genes at the same locus and epistasis is the interaction between genes at different loci. Genotyping an increasing number of females with own phenotypes could lead to a better understanding and use of dominance effects in cattle breeding. It might be possible to further increase reliabilities and also use this information for mating plans. Furthermore, such plans can be used to find a suitable bull for a cow to make a targeted use of overdominance and therefore maximize progeny performance (Wellmann and Bennewitz, 2011).

Wellmann and Bennewitz (2012) found in a simulation study that accounting for dominance effects in GBLUP and Bayesian models led to higher accuracies genomic breeding values than models without dominance effects. Ertl et al. (2014) used real data of 1996 Fleckvieh cows, genotyped with the Illumina HD-chip (Illumina BovineHD BeadChip, Illumina Inc., San Diego, CA; HD-chip), to estimate dominance variance. They found estimates of dominance variance to be from 3.3% to 50.5% of the total genetic variance, depending on the trait. Accuracies did not change when dominance effects were added to the model. However, the data structure in other datasets, for example pig data, might be more useful to get dominance-specific information, because pig data usually contain more full-sibs then dairy cattle data. The dataset of Ertl et al. (2014) contained 3% of full-sibs only. Wellmann et al. (2014) investigated whether accuracy of genotypic values and dominance deviations can be increased by a joint evaluation of bulls and cows. They used the same dataset as in Ertl et al. (2014), increased by 6858 genotyped bulls, and found that using their strategy it was possible to increase the accuracy of estimated genotypic values. Further, they concluded that by genotyping more cows, large scale datasets would become available, which should allow for more accurate prediction of dominance deviations.

Technical aspects

Routine genomic breeding value estimation for dairy cattle is often done based on a 50k-chip (Illumina Bovine SNP50 BeadChips, Illumina Inc., San Diego, CA) containing approximately 50,000 SNP markers. With the introduction of a 777k bovine chip (HD-chip) it was expected that by using this denser chip the accuracy of genomic predictions especially for small breeds could be improved. In several investigations, reliabilities of predicted breeding values were compared when breeding values were predicted either from 50k or from HD genotypes, but only minor, if any, gains in validation reliability were observed (Erbe et al., 2012; Su et al., 2012a, Ertl et al. 2013; VanRaden et al., 2013). It has been suggested that the benefits from HD-chips might be small if most genetic variation is from very small QTL effects (Clark et al., 2011). In addition to the HD-chip and the routinely used 50k-chip, there is also a series of low-density chips (e.g. 3k, 6k or 7k). LD-chips are often used to genotype cows, since they are more cost-effective than chips with higher density (Brøndum et al., 2015; Wiggans et al., 2016). For example, since 2010, genotyping of females with a LD-chip has been implemented at a large scale in Holsteins in the United States (Wiggans et al., 2011; Goa et al., 2015). However, using LD-chips in most cases includes an additional imputation step, in which the low-density genotypes are imputed to the higher density, which is then used in the genomic evaluation. Imputation with a large reference population works quite well in most cases (Pausch et al., 2013; Plieschke et al., 2014). The

quality however depends on population key-parameters like the effective population size or LD (Sargolzaei et al, 2014). This, and the fact that LD-chips are still somewhat costly, might be the reason why the use of LD-chips has not been established in routine genomic evaluation for the predominant South-German breeds.

Technical progress over the last decade has allowed millions of DNA reads to be sequenced at a reasonable cost in a relatively short period of time (Ni et al., 2016). Thus, in the last years there has been another technology getting available and feasible for dairy cattle: whole-genome sequencing (e.g. Meuwissen and Goddard, 2010). Compared to SNP arrays, some advantages of having whole-genome sequence data are expected. For instance, there should be an increase in the ability to predict the genetic value of an individual for complex traits (Meuwissen and Goddard, 2010). Additionally, it would make it possible to identify causal mutations and to increase stability of genomic predictions without updating the reference population (Meuwissen and Goddard, 2010; Pérez-Enciso et al., 2015; van Binsbergen et al., 2016). Goddard (2017) further suggested and envisioned the use of whole-genome sequence data in a SNP-MACE method for combining SNP-effects across countries and even across breeds.

Although the cost of DNA sequencing has decreased in the recent years due to the rapid development of sequencing technology, it is still relatively expensive (Meynert et al., 2014; Ni et al., 2016). Instead of sequencing the whole population it is possible to sequence only some so called key ancestors and impute the remaining genotypes from a lower chip array to the sequence level. Imputation to sequence data might be a cost-effective approach to obtain a large training set of sequenced individuals (van Binsbergen et al., 2015; Pausch et al., 2016). The 1000 bull genomes project (Daetwyler et al., 2014) was set up to build up whole-genome sequence data from various large cattle breeds. The aim of the project was to build a database containing sequence variant genotypes of key ancestors from three different cattle breeds to be able to do genome-wide association studies (GWAS) and genomic prediction based on sequence data and to use this data to identify mutations that influence animal health, welfare and productivity (Daetwyler et al., 2014). This project has rapidly increased the availability of sequence data of important ancestors in these cattle breeds. Although the expectations were great, the results with respect to the reliability of genomic prediction did not show the desired advantage over the other SNP chip arrays (e.g. van Binsbergen et al., 2015; Pérez-Enciso et al., 2015). Moreover, sequence data are very large datasets and therefore difficult to transfer. Additionally, the use of these datasets requires expensive hardware equipment (Pérez-Enciso et al., 2015). Nevertheless, sequencing technology might be helpful to detect SNP variants that cause genetic defects.

Genomic selection is a comprehensive topic that brought large changes to animal and plant

breeding in recent years. Over the past few years, progress has been made. However, it can be expected that much of the ongoing work will not be completed in the short term and further open questions will be addressed in the future.

References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 2010;93:743-52.

Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, Lund MS. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J. Dairy Sci. 2015;98:4107-16.

Caballero A, Toro MA. Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Genet. 2002;3:289-99.

Calus MPL, de Haas Y, Pszczola M, Veerkamp RF. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. Animal. 2013;7:183-91.

Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 2010;42:2

Clark SA, Hickey JM, van der Werf JHJ. Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. 2011;43:18.

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 2010;185:1021-31.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.

DHV. KuhVision. http://www.holstein-dhv.de/seiteninhalte/kuhvision.html. (January 2017)

Edel C, Schwarzenbacher H, Hamann H, Neuner S, Emmerling R, Götz KU. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. Interbull Bull. 2011;44:152-6.

Edel C, Pimentel ECG, Plieschke L, Emmerling, R, Götz KU. Short communication: The effect of genotyping cows to improve the reliability of genomic predictions for selection candidates. J Dairy Sci. 2016;99:1999-2004.

Egger-Danner C, Cole JB, Pryce JE, Gengler N, Heringstad B, Bradley A, et al. Invited review: Overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal. 2014;9:191-20.

Endelman JB, Jannink JL. Shrinkage Estimation of the Realized Relationship Matrix. G3. 2012;2:1405-13.

Erbe M, Hayes BJ, Matukumakki LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 2012;95:4114-29.

Ertl J, Edel C, Emmerling R, Pausch H, Fries R, Götz K-U. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: Observations from Fleckvieh cattle. J. Dairy Sci. 2013;97:487-96.

Ertl J, Legarra A, Vitezica ZG, Varona L, Edel C, Emmerling R, Götz K-U. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet. Sel. Evol. 2014;46:40.

Falconer and Mackay. Introduction to quantitative genetics. Fourth Edition. 1996

Gao H, Su G, Janss L, Zhang Y, Lund MS. model comparison on genomic predictions using high-density markers for different groups of bulls in the nordic Holstein population. J. Dairy Sci. 2013;96:4678-87.

Gengler N, Mayeres P, Szydlowski M. A simple method to approximate gene content in large pedigree populations: applications to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 2007;1:21-8.

GKUH. http://www.gkuh.de/Texte/home.html (December 2016)

Goddard ME. Genomic selection: prediction of accuracy and maximization of long term response. Genetica, 2008;136:245-57.

Goddard ME. Interbull estimation of SNP effects. Interbull Technical Workshop. 6-8 February 2017;Ljubjana.

Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389-97.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009a;92:433-43.

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

Illumina. BovineSNP50 Genotyping BeadChip. www.illumina.com (January 2017).

Illumina. BovineHD Genotyping BeadChip. www.illumina.com (January 2012).

Kadri NK, Guldbrantsen B, Sørensen P, Sahana G. Comparison of Genome-Wide Association Methods in Analyses of Admixed Populations with Complex Familial Relationships. PLos ONE. 2014;9:e88926.

Koivula M, Stranden I, Pösö J, Aamand GP, Mäntysaari EA. Single-step in genomic evaluation using multitrait random regression model and test-day data. J Dairy Sci. 2015;98:2775-84.

Langhdahl C. Status on practical breeding program in VG. 21–22 January, 2014; Copenhagen.

Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 2009:92;4656-63.

Luan T, Woolliams JA, Ødegard J, Dolezal M, Roman-Ponce SI, Bagnato A, Meuwissen THE. The importance of identity-by-state information for the accuracy of genomic selection. Genet. Sel. Evol. 2012;44:28.

Luan T, Yu X, Dolezal M, Bagnato A, Meuwissen THE. Genomic prediction based on runs of homozygosity. Genet. Sel. Evol. 2014;46:64.

Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

Meuwissen T, Goddard M. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics. 2010;185:623-31.

Meynert AM, Ansari M, Fitzpatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 2014;15:247.

Ni G, Strom TM, Pausch H, Reimer C, Preisinger R, Simianer H, Erbe M. Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. BMC Genomics. 2015;16:824.

Patterson N, Price AL, Reich D. Population structure and Eigen analysis. PLoS Genet. 2006;2:e190.

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

Pausch H, MacLeod IM, Fries R, Emmerling R, Bowman PJ, Daetwyler HD, Goddard ME. Evaluation of the accuracy of imputed sequence variants and their utility for causal variant detection in cattle. bioRxiv 085399. 2016. doi: https://doi.org/10.1101/085399

Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 2015;47:43.

Pimentel ECG, Edel C, Emmerling R, Götz K-U. Über die Skalierung der H-Matrix in single-step Analysen. Vortragstagung der DGfZ und GfT 2016 in Hannover.

Plieschke L, Edel C, Bennewitz J, Emmerling R, Götz K-U. Imputation von SNP-Genotypen mit den Programmen FImpute und findhap. Züchtungskunde. 2014;86:81-94.

Plieschke L, Edel C, Pimentel E, Emmerling R, Bennewitz J, Götz KU. Influence of foreign genotypes on genomic breeding values of national candidates in Brown Swiss. Proceedings of the 10th World Congress of Genetics Applied to Livestock Production: 17-22 August 2014; Vancouver.

Plieschke L, Edel C, Pimentel E, Emmerling R, Bennewitz J, Götz KU. Equivalence of genomic breeding values and reliabilities estimated with SNP-BLUP and GBLUP. Proceedings of the 66th EAAP Annual Meeting, 2015; Warsaw.

Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Genet. 2010;11:800-5.

ProGesund. http://www.progesundrind.de/ (December 2016)

Pryce JE, Hayes BJ, Goddard ME. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. In Proceedings of the 36th ICAR Biennial Session. 16-20 June, 2008; Niagara Falls.

Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:278.

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. 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. 2012a;95:4657-65.

Su G, Ma P, Nielsen US, Aamand GP, Wiggans G, Guldbrandtsen B, Lund MS. Sharing reference data and including cows in the referencepopulation improve genomic predictions in Danish Jersey. Animal. 2015;10:1067-75.

Thomasen JR, Sorensen AC, Lund MS, Guldbrandtsen B. Adding cows to the reference population makes a small dairy population competitive. J Dairy Sci. 2014;97:5822-32.

van Binsbergen R, Calus MPL, Bink MCAM, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Selec. Evol. 2015;47:71.

VanRaden PM. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 2008;91:4414-23.

VanRaden PM, Van Tassel CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 2009;92:16–24.

VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, Sonstegard TS, Connor EE, Winters M, can Kaam JBCHM, Valentini A, van Doormaal BJ, Faust MA, Doak GA. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 2013;96:668-78.

Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of population structure. Evolution. 1984;38:1358-70.

Wellmann R, Bennewitz J. Die Berücksichtigung von Dominanzeffekten bei der genomischen Zuchtwertschätzung. Züchtungskunde. 2011;83:361-70.

Wellmann R, Bennewitz J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genet Res (Camb). 2012;94:21-37.

Wellmann R, Ertl J, Emmerling R, Edel C, Götz K-U, Bennewitz J. Joint genomic evaluation of cows and bulls with BayesD for prediction of genotypic values and dominance deviations. Proceedings of the 10th World Congress of Genetics Applied to Livestock Production: 17-22 August 2014;Vancouver.

Wiggans GR, Cooper TA, VanRaden PM, Cole JB. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. J Dairy Sci. 2011;94:6188-93.

Wiggans GR, Cooper TA, VanRaden PM, Van Tassell CP, Bickhart DM, Sonstegard TS. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. J Dairy Sci. 2016;99:4504–11.

Zeng J, Pszczola M, Wolc A, Strabel T, Fernando RL, Garrick DJ, Dekkers JCM. Genomic breeding value prediction and QTL mapping of QTLMAS2011 data using Bayesian and GBLUP methods. BMC Proceedings. 2012;6:S7

Zhang Z, Liu J, Ding X, Bijma P, de Koning D-J, Zhang Q. Best Linear Unbiased Prediction of Genomic Breeding Values Using a Trait-Specific Marker-Derived Relationship Matrix. PLoS ONE. 2010;5:e12648

Zou F, Lee S, Knowles MR, Wright FR. Quantification of population structure using correlated SNPs by shrinkage principal components. Hum Hered. 2010;70:9-22.

General Summary (English)

The aim of this study was to investigate methodological and strategic aspects of genomic selection in Bavarian cattle breeds.

In chapter one a method was developed to separate the genomic relationship matrix into two independent covariance matrices. Here, the base group component describes the covariance that results from systematic differences in allele frequencies between groups at the pedigree base. The remaining segregation component describes the genomic relationship that is corrected for the differences between base populations.

To investigate the proposed decomposition three different models were tested on six traits, where the covariance between animals was described either only by the segregation component or by a combination of the two components. An additional variant examining the effect of a fixed modeling of the group effects was included. In total, 7965 genotyped Fleckvieh and 4257 genotyped Brown Swiss and 143 genotyped Original Braunvieh bulls were available for this study.

The proposed decomposition of the genomic relationship matrix helped to examine the relative importance of the effects of base groups and segregation component in a given population. It was possible to estimate significant differences between the means of base groups in most cases for both breeds and for the traits analyzed. Analysis of the matrix of base group contributions to the populations investigated revealed several general breed-specific aspects. Comparing the three models, it was concluded that the segregation component is not sufficient to describe the covariance completely. However, it also was found that the model applied has no strong impact on predictive power if the animals used for validation show no differences in their genetic composition with respect to the base groups and if the majority of them have complete pedigrees of sufficient depth.

The subject of the chapter two was investigation to systematically increase the reliability of genomic breeding values by integrating cows into the reference population of genomic breeding value estimation. For this purpose a dataset was generated by simulation resembling the German-Austrian dual-purpose Fleckvieh population. The concept investigated is based on genotyping a fixed number of daughters of each AI bull of the last or last two generation of the reference population and, together with their phenotypic performance, to integrate them into the reference

population of the genomic evaluation. Different scenarios with different numbers of daughters per bull were compared. In the base scenario the reference population was made up of 4200 bulls. In the extended scenarios, more and more daughters were gradually integrated in the reference population. The reference population of the most extended scenario contained 4200 bulls and 420,000 cows.

It was found that the inclusion of genotypes and phenotypes of female animals can increase the reliabilities genomic breeding values considerably. Changes in validation reliability of 6-54% for a trait with a heritability of 0.4 depending on scenario were found. As the number of daughters increased, the validation reliability increased as well. It should be noted that the composition of the daughter samples had a very great influence on whether the additional genotyped and phenotyped animals in the reference population can have a positive effect on the reliability of genomic breeding values. If pre-selected daughter samples were genotyped, the mean validation reliability decreased significantly compared to a correspondingly large unselected daughter sample. In addition, a higher bias was observable in these cases.

Chapter three expands the investigations of chapter two by a low-heritability trait, as well as the aspect of so called new traits. The results found in chapter two were confirmed in chapter three for a low-heritability trait. Changes in validation reliability of 5-21% for a heritability of 0.05 depending on scenario were found. The negative effects of pre-selected daughter samples were even more pronounced in chapter three. In the case of an 'old' trait, the number of phenotypes is expected to be (nearly) unlimited, since a recording system is well established. In the case of a new trait recording of phenotypes just started, therefore the number of phenotypes is limited. Two different genotyping strategies were compared for new traits. On the one hand, the sires of the phenotyped cows were genotyped and on the other hand the phenotyped cows were genotyped themselves. It was found in all compared scenarios that it is more sensible to genotype cows themselves instead of the genotyping their sires. However, if usual strategy of phenotyping female animals and genotyping of males is applied, it is at least important to ensure that many daughters are phenotyped in a balanced system. If different numbers of daughters per bull are phenotyped and unbalancedness becomes severe, the average validation reliability decreased significantly.

General Summary (German)

Ziel der vorgelegten Arbeit war es verschiedene Aspekte der genomischen Selektion zu untersuchen.

In Kapitel eins wurde eine Methode entwickelt, die die Verwandtschaftsmatrix in zwei unabhängige Kovarianzmatrizen trennt. Dabei beschreibt die eine Matrix die Basisgruppenkomponente, eine Kovarianzkomponente die auf Allelfrequenzunterschiede zwischen verschiedenen Gruppen der Basispopulation zurückzuführen ist. Die Segregationskomponente beschreibt die genomische Verwandtschaft, welche um diese Unterschiede zwischen verschiedenen Basispopulationen korrigiert wurde.

Die Zerlegung wurde anhand drei verschiedener Modelle für sechs Merkmale untersucht. Dabei wurde die Kovarianz zwischen Tieren entweder nur durch die Segregationskomponente oder über eine Kombination beider Komponenten beschrieben. Eine zusätzliche Variante untersuchte den Effekt einer fixen Modellierung der Gruppeneffekte. Zur Verfügung standen für diese Studie die Genotypen von 7965 Fleckviehbullen sowie von 4257 Braunvieh- und 143 Original Braunviehbullen.

Über den beschriebenen Weg war es möglich, den Umfang der Basisgruppeneffekte und der Segregationskomponente in den beiden untersuchten Populationen sichtbar zu machen. Es konnten signifikante Unterschiede zwischen den mittleren Basisgruppeneffekten in den meisten Fällen für beide Rassen und für die analysierten Merkmale gefunden werden. Die Analyse der Genanteile der diversen Basisgruppen an den untersuchten Populationen machte zudem einige rassetypische Aspekte sichtbar. Im Vergleich der drei Modelle, wurde festgestellt, dass die Segregationskomponente nicht ausreicht die Kovarianz vollständig zu beschreiben. Es wurde allerdings auch festgestellt, dass das verwendete Modell keinen starken Einfluss auf die Vorhersagekraft hat, wenn die zur Validierung verwendeten Tiere in ihrer genetischen Zusammensetzung weitgehend homogen sind und die Mehrheit von ihnen ein vollständiges Pedigree mit ausreichender Tiefe aufweist.

Gegenstand des zweiten Kapitels waren Berechnungen zur systematischen Steigerung der Sicherheiten genomischer Zuchtwerte durch die Aufnahme von Kühen in die Referenzstichprobe der genomischen Zuchtwertschätzung. Hierfür wurde eine Simulationsstudie durchgeführt, durch die die deutsch-österreichische Fleckviehpopulation widergespiegelt werden sollte. Das in der Simulation untersuchte Genotypsierungskonzept beruht darauf, eine fixe Anzahl erstlaktierender

Töchter eines jeden Besamungsbullen aus der jeweils letzten bzw. vorletzten und letzten Generation der Referenzstichprobe zusätzlich zu genotypisieren und zusammen mit ihren phänotypischen Leistungen in die Referenzstichprobe der genomischen Zuchtwertschätzung zu integrieren. Verglichen wurden verschiedene Szenarien mit unterschiedlicher Anzahl an Töchtern je Bulle. Im Basisszenario bestand die Referenzstichprobe aus 4200 Bullen, in den erweiterten Szenarien wurden schrittweise immer mehr Töchter integriert. Wobei die Referenzstichprobe im letzten Szenario 4200 Bullen und 420.000 Kühe umfasste.

Es konnte gezeigt werden, dass durch die Aufnahme von Genotypen und Phänotypen weiblicher Tiere die Sicherheiten genomischer Zuchtwerte erheblich gesteigert werden können. Die gefundenen Zuwächse betrugen dabei zwischen 6 und 54 % bei einem Merkmal mit einer Heritabilität von 0,4, wobei die Zuwächse mit zunehmender Anzahl an Töchter ebenfalls weiter anstiegen. Die Zusammensetzung der Töchtergruppen hatte einen großen Einfluss darauf, ob die zusätzlichen genotypisierten und phänotypisierten Tiere in der Referenzstichprobe einen positiven Effekt auf die Sicherheiten genomischen Zuchtwerte haben können und wie hoch dieser ist. Genotypisierte man ausschließlich vorselektierte Töchtergruppen, sank die mittlere Validierungssicherheit erheblich im Vergleich zu einer entsprechend großen unselektierten Töchtergruppe. Außerdem waren Effekte einer deutlichen Verzerrung der Zuchtwerte beobachtbar.

Kapitel drei erweiterte die Untersuchungen aus Kapitel zwei um ein niedrig-erbliches Merkmal sowie um den besonderen Aspekt der so genannten neuen Merkmale. Die in Kapitel zwei gefundenen Ergebnisse konnten in Kapitel drei auch für ein niedrig-erbliches Merkmal bestätigt werden. Bei einer Heritabilität von 0,05 konnten die Sicherheiten in den verschiedenen Szenarien zwischen 5 und 21 % gesteigert werden. Eine gezielte Auswahl der genotypisierten Töchtergruppen führte auch hier zu negativen Effekten auf die ansonsten erzielbaren Sicherheiten und führte zu einer Verzerrtheit der genomischen Zuchtwerte. Im Falle eines Merkmals mit etablierter Leistungsprüfung kann davon ausgegangen werden, dass die Anzahl der Phänotypen (nahezu) unbegrenzt ist.. Im Falle der neuen Merkmale trifft dies nicht zu. In diesem Zusammenhang wurden zwei verschiedene Strategien der Genotypisierung verglichen. Zum einen wurden die Väter der phänotypisierten Kühe genotypisiert und zum anderen wurden die begrenzte Anzahl phänotypisierter Kühe selber genotypisiert. Es konnte in allen verglichenen Szenarien gezeigt werden, dass es sinnvoller ist, die Kühe selbst zu Genotypisieren statt deren Väter. Sollte dies nicht möglich sein und man nutzt die neuen Phänotypen wie bisher nur über die Väter, ist zumindest darauf zu achten, dass Töchter in balancierter Weise phänotypisiert werden. Bei niedrig erblichen Merkmalen und stark begrenzter Verfügbarkeit von Phänotypen kann Unbalanciertheit deutlich negative Effekte auf die mittlere Validierungssicherheit haben.

Danksagung

An erster Stelle möchte ich mich bei Herrn Prof. Dr. Jörn Bennewitz für die Überlassung des Themas und die Unterstützung und Betreuung dieser Arbeit sehr herzlich bedanken. Für die Übernahme des Koreferates bedanke ich mich bei Herrn Prof. Dr. Simianer und Herrn Prof. Dr. Piepho.

Mein Dank gilt ebenfalls Dr. Christian Edel, welcher immer mein erster Ansprechpartner war. So auch Prof. Dr. Kay-Uwe Götz, Dr. Reiner Emmerling und Dr. Eduardo Pimentel der LfL in Grub, die mich ebenfalls sehr unterstützt haben. Außerdem ein großes Dankeschön an alle weiteren Kollegen und Freunde, die mich während meiner Promotion begleitet haben.

Einen ganz herzlichen Dank meinem Freund und meiner Familie.

Mein Dank gilt ebenfalls der Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. für die finanzielle Unterstützung im Rahmen des Projektes „Zukunftswege“.

Eidesstattliche Versicherung

Eidesstattliche Versicherung

gemäß § 8 Absatz 2 der Promotionsordnung der Universität Hohenheim zum Dr.sc.agr.

1. Bei der eingereichten Dissertation zum Thema
   Investigations on methodological and strategic aspects of genomic selection in dairy cattle using real and simulated data
   handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Ich habe nicht die Hilfe einer kommerziellen Promotionsvermittlung oder -beratung in Anspruch genommen.
4. Die Bedeutung der eidesstattlichen Versicherung und der strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Die Richtigkeit der vorstehenden Erklärung bestätige ich. Ich versichere an Eides Statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift

Eidesstattliche Versicherung

Belehrung

Die Universität Hohenheim verlangt eine Eidesstattliche Versicherung über die Eigenständigkeit der erbrachten wissenschaftlichen Leistungen, um sich glaubhaft zu versichern, dass die Promovendin bzw. der Promovend die wissenschaftlichen Leistungen eigenständig erbracht hat.

Weil der Gesetzgeber der Eidesstattlichen Versicherung eine besondere Bedeutung beimisst und sie erhebliche Folgen haben kann, hat der Gesetzgeber die Abgabe einer falschen eidesstattlichen Versicherung unter Strafe gestellt. Bei vorsätzlicher (also wissentlicher) Abgabe einer falschen Erklärung droht eine Freiheitsstrafe bis zu drei Jahren oder eine Geldstrafe.

Eine fahrlässige Abgabe (also Abgabe, obwohl Sie hätten erkennen müssen, dass die Erklärung nicht den Tatsachen entspricht) kann eine Freiheitsstrafe bis zu einem Jahr oder eine Geldstrafe nach sich ziehen.

Die entsprechenden Strafvorschriften sind in § 156 StGB (falsche Versicherung an Eides Statt) und in § 161 StGB (Fahrlässiger Falscheid, fahrlässige falsche Versicherung an Eides Statt) wiedergegeben.

§ 156 StGB: Falsche Versicherung an Eides Statt

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§ 161 StGB: Fahrlässiger Falscheid, fahrlässige falsche Versicherung an Eides Statt:

Abs. 1: Wenn eine der in den §§ 154 und 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

Abs. 2: Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Absätze 2 und 3 gelten entsprechend.

Ich habe die Belehrung zur Eidesstattlichen Versicherung zur Kenntnis genommen.

Ort und Datum Unterschrift

Curriculum Vitæ

Laura Isabel Plieschke

Personal Profile

Date of Birth	01.01.1988
Place of Birth	Leverkusen
Nationality	German

Education

2013-2016	**PhD Student**, *Bavarian State Research Center for Agriculture, Institute of Animal Breeding*.
2010-2013	**Study of Agricultural Sciences**, *University of Hohenheim*. Master of Science
2007-2010	**Study of Agricultural Sciences**, *University of Hohenheim*. Bachelor of Science
1998-2007	**Secondary Education**, *Marienschule Opladen*. Abitur

Work Experience

2012-2013	**Internship as part of the master thesis**, *Bavarian State Research Center for Agriculture, Institute of Animal Breeding*.
Summer 2010	**Professional internship (8 weeks)**, *Horse farm Gilles in Leverkusen*.
Summer 2009	**Professional internship (8 weeks)**, *Organic farm Höffgen in Solingen*.
Summer 2008	**Professional internship (8 weeks)**, *Fruit farm Flügel in Leichlingen*.