From the Institute of Animal Breeding and Genetics
Justus-Liebig-University Gießen

Exploring the potential of machine learning methods and selection signature analyses for the estimation of genomic breeding values, the estimation of SNP effects and the identification of possible candidate genes in dairy cattle

Dissertation
to obtain the doctoral degree (Dr. agr.)
in the Faculty of Agricultural Science, Nutritional Science
and Environmental Management of
Justus-Liebig-University Gießen, Germany

presented by
Saeid Naderi Darbaghshahi
born in Esfahan, Iran

Gießen, March 2018

1st Referee: Prof. Dr. Sven König
Institute of Animal Breeding and Genetics
Justus-Liebig-University Gießen, Germany

2nd Referee: Prof. Dr. Nicolas Gengler
Animal Science Unit, Numerical Genetics, Genomics and Modeling
Agriculture, Bio-engineering and Chemistry Department
University of Liège - Gembloux Agro-Bio Tech, Belgium

Table of Contents

LIST OF TABLES

2nd Chapter

3rd Chapter

4th Chapter

LIST OF FIGURES

2nd Chapter

3rd Chapter

4th Chapter

5th Chapter

The objective of this thesis was to study a variety of factors that affect the accuracy of genomic predictions applying random forest methodology (RF), genomic BLUP (GBLUP) and single step genomic BLUP (ssGBLUP) method with strong focus on training set design. In the following, selection signature through variation in linkage disequilibrium (LD) within and between dual-purpose black and white (DSN) and Holstein populations was identified.

In chapter 2 a stochastic simulation was applied for genomic predictions of binary disease traits based on cow training set. Composition of training and testing sets were modified in different allocating schemes. In addition, different scenarios were studied according to the quantitative-genetic background of the trait, the genetic architecture as well as low and high density of SNP chip panel. The highest genomic prediction accuracies were achieved when disease incidences within training sets was close to the population disease incidence of 0.20. Decreasing the traits heritability and QTL reduction were associated with decreasing genomic prediction accuracies.

In chapter 3, different disease traits from 6,744 cows with genotypes from 58 large-scale contract herds was used to study the impact of training set composition, the impact of response variable as well as the impact of RF, GBLUP and ssGBLUP methodology on genomic prediction accuracies. Using de-regressed proofs (DRP) as response variables, accuracies were larger compared to pre-corrected phenotypes (PCP) for both methods GBLUP and RF. A further increase in genomic prediction accuracies was realized via ssGBLUP method compared to corresponding scenarios with RF or GBLUB. In addition, RF identified significant SNP close to potential positional candidate gene, i.e., *GAS1*, *GPAT3*, and *CYP2R1* for clinical mastitis, *SPINK5* and *SLC26A2* for laminitis, and *FGF12* for infertility.

Genetic variation between the Holstein and the DSN population as well as between sub-populations was inferred by using XP-EHH method in chapter 4. The analysis was performed on 2,076 genotyped Holstein cows and 261 genotyped DSN cows. The most outstanding XP-EHH score that revealed the regions under recent selection was on chromosome 6 and on chromosome 12 for DSN and on chromosome 20 for Holstein population. Annotation of selection signature regions revealed various genes associated with production traits such as *CLU* and *WARS2*. Furthermore, several hub genes associated with dermatitis digitalis resistance was detected including *FARS2, ACTR8* and *CRY1*.

Keywords: genomic predictions, random forest, disease traits, selection signature.

Ziel dieser Arbeit war die Analyse einer Vielzahl von Faktoren, welche die Genauigkeit der genomischen Zuchtwertschätzung beeinflussen. Hierzu wurden, mit besonderer Fokussierung auf die Konzeption der Referenzstichprobe, Random Forest (RF), genomic BLUP (GBLUP) und single step genomic BLUP (ssGBLUP) Verfahren angewendet. Nachfolgend wurden Selektionssignaturen mithilfe von Variationen im Kopplungsungleichgewicht (LD) innerhalb und zwischen Populationen des Schwarzbunten Niederungs- und Holsteinrinds identifiziert. Im zweiten Kapitel wurde eine stochastische Simulation appliziert, um genomische Schätzungen binärer Krankheitsmerkmale basierend auf der Referenzstichprobe durchzuführen. Die Struktur der Referenz- und Teststichprobe wurde in verschiedenen Zuweisungsschemata modifiziert. Darüber hinaus fand die Anwendung divergenter Szenarien hinsichtlich des quantitativ-genetischen Hintergrundes, der genetischen Architektur sowie der Dichte des SNP Chip statt. Die höchste genomische Schätzgenauigkeit wurde bei Annäherung der Krankheitsinzidenz innerhalb der Referenzstichprobe an die Inzidenz der Population von 0,20 erreicht. Die Reduktion der Merkmalsheritabilität und QTL ging mit einer Verringerung der genomischen Schätzgenauigkeit einher. In Kapitel drei wurde der Effekt der Struktur der Referenzstichprobe, der abhängigen Variablen wie auch der RF, GBLUP und ssGBLUP Methode auf die genomische Schätzgenauigkeit, unter Einbeziehung diverser Krankheitsmerkmale von 6744 genotypisierten Kühen aus 58 Testherden, analysiert. Die Verwendung deregressierter Zuchtwerte (DRP) als abhängige Variable im GBLUP sowie RF Verfahren, führte zu einer Verbesserung der Genauigkeiten im Vergleich zur Nutzung vorkorrigierter Phänotypen (PCP). Ein weiterer Anstieg der Genauigkeit wurde durch Anwendung der ssGBLUP Methode erzielt. Mithilfe der RF Methode, konnten zudem signifikante SNP in der Nähe möglicher Kandidatengene wie *GAS1*, *GPAT3*, *CYP2R1* für Mastitis, *SPINK5*, *SLC26A2* für Laminitis und *FGF12* für Unfruchtbarkeit identifiziert werden. Genetische Variation zwischen der Holstein und DSN Population wie auch Subpopulationen wurde mithilfe der XP-EHH Methode, unter Einbeziehung von 2076 genotypisierten Holstein und 261 DSN Kühen, in Kapitel vier dargestellt. Der auffälligste XP-EHH Score, jüngste Selektionsregionen darstellend, wurde auf Chromosom 6 und 12 für DSN und auf Chromosom 20 für die Holsteinpopulation detektiert. Die Annotation der Selektionssignaturregionen eruierte diverse, mit Produktionsmerkmalen assoziierte Gene wie *CLU* und *WARS2*, wie auch einige, mit Dermatitis Digitalis Resistenz assoziierte Hub Gene, einschließlich *FARS2, ACTR8* und *CRY1*.

Schlüsselwörter: genomische Schätzungen, Random Forest, Krankheitsmerkmale, Selektionssignaturen

1st Chapter

General Introduction

From conventional pedigree-based selection towards genomic selection

In livestock, most of economically important discrete and continuous traits have a complex and quantitative expression that is influenced by genetic and environmental components. The sum of these components is a simple and robust model for the inheritance of quantitative traits, that the genetic component of offspring traits follows a normal distribution around the average of the parents (Barton et al., 2016).

The greatest achievements in livestock selection during past decades relied on quantitative genetic theory and infinitesimal model. The infinitesimal model assumes that genetic differences among individuals are related to contributions of an infinite number of loci with their small effects on the trait (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Under the infinitesimal model, selection is a sort of blind process where genotypes are progressively modified without any real knowledge about gene number, location, effect and frequencies of the favourable alleles. Therefore, it is assumed that these quantitative trait loci (QTL) are homogeneously distributed across the genome (Montero, 2013).

Conventional pedigree-based selection combines only phenotypic data and probabilities that genes are identical by descent (IBD). In this type of selection, a numerator relationship matrix (NRM) is used to describe the additive variance-covariance relationship between all individual pairs in a population (Henderson, 1975). This approach has been very effective to capture the parent average component of the estimated breeding value (EBV), but cannot capture variation in the proportion of genome shared by pairs of relatives due to Mendelian sampling (Daetwyler et al., 2007). Therefore, using information on variation in DNA sequence between animals should lead to the prediction of more accurate breeding values.

The sequencing of the genome is a new tool that provides the genomic information of each individual. These modern sequencing techniques allow genotyping of thousands of variation sources throughout the genome. Recently, high-density single nucleotide polymorphisms (SNP) data is available for many farm animal species (Elsik et al., 2009). It is expected that some of those variations will be close to QTLs of interest. Therefore, SNPs are used as markers under the assumption that they will be inherited jointly to QTLs due to the existing linkage disequilibrium (LD) in the genome; and selection based on this genomic information was named genomic selection (GS). The GS has been proposed as an alternative to conventional pedigree-based selection (Meuwissen et al., 2001), where the NRM could be replaced by a genomic relationship matrix (GRM) to estimate the genomic breeding value (GEBV) of each individual. In GRM, the

relationships reflect the actual proportion of marker alleles shared by identity by state (IBS), as a deviation from the expected proportion of IBS alleles shared in the population (Luan et al., 2014). Therefore, using GS, it could be possible to capture the Mendelian sampling component in the absence of recorded phenotypes (Beaulieu et al., 2014).

Effects of genomic selection on rate of genetic gain

Four main factors affect the rate of genetic gain in a population undergoing artificial selection, and can be expressed in mathematical terms by the following equation (Falconer, 1989):

$$\Delta G = \frac{ir\sigma_a}{L}$$

where, ΔG is rate of genetic gain per year; i is the selection intensity; r or accuracy of selection is the correlation between the true and estimated breeding values of animals; σ_a is the additive genetic standard deviation of the trait of interest, and L is the generation interval. In this respect, the benefits of genomic selection over conventional pedigree-based selection were first reported for dairy cattle (Meuwissen et al., 2001). Meuwissen et al. (2001) mentioned that selection based on GEBV could significantly increase the rate of genetic gain in dairy cattle. With considering the genetic gain equation, the main benefits of genomic selection is due to the reduction in the generation interval and the increase in accuracies of the estimated breeding values of young bulls and bull dams (Schaeffer, 2006; König et al., 2009). Most of current designed animal breeding schemes show that the accuracy of selection is already high (e.g. in progeny testing). However, genomic selection will be especially useful for traits where the conventional pedigree-based selection has several undesirable characteristics that can limit genetic progress. For example, when there are insufficient phenotypic records for the individual itself and its progeny, or when the trait has low heritability or traits that are measured late in life or it needs the slaughtering of animals, or for disease resistance traits that require expensive recording and/or risky challenge testing, and for sex-limited traits (Goddard and Hayes, 2007). Therefore, Goddard (2009) stressed that the main advantage of genomic selection is the increment of the selection accuracy at an early age of the animal when the own phenotype and pedigree is not available. Based on these findings and because of organizational advantages (e.g., high degree of artificial insemination providing many very well proven sires as reference animals), genomic selection was rapidly implemented in dairy cattle breeding programs (Loberg and Dürr, 2009)

where early selection of sires is crucial.

Factors that affect accuracy of genomic prediction

LD, co-segregation and pedigree relationships are sources of genetic information that contribute to accuracy of genomic prediction (Habier et al., 2007). The accuracy of genomic prediction will be high when high LD exists between QTL and SNPs (Zhong et al., 2009; Yin et al., 2014). Additionally, the accuracy due to LD is more likely to persist across generations and breeds than the accuracy due to relationships (Meuwissen et al., 2001; Habier et al., 2007).

Co-segregation and pedigree relationships that can be captured by the genomic model have large contributions when predicting close relatives but have small contributions when predicting individuals that are distantly related to the reference population (Sun, 2014; Habier et al., 2010; Wientjes et al., 2013). Researchers have presented that LD between QTL and SNPs in livestock populations is low, and prediction accuracy mainly comes from co-segregation and pedigree relationships that are implicitly captured by SNP genotypes (Habier et al., 2010; Saatchi et al., 2011; Wientjes et al., 2013).

Marker density is another important factor that affects accuracy of genomic prediction (Habier et al., 2009; Meuwissen, 2009). Due to the fact that markers and QTL can be in LD, more markers capture a higher proportion of genetic variance of the trait (Goddard, 2009). As the all genetic variance is explained by the markers which are scattered in the whole genome, higher marker density could increase the accuracy between adjacent markers and indeed the accuracy of genomic prediction (Meuwissen, 2009; Yin et al., 2014). On the other hand, the ability of a genomic model to capture sources of genetic information depends on the effective number of SNPs (Habier et al., 2013).

Researchers (Hayes et al., 2009; Moser et al., 2012) have shown that the heritability, as a factor underlying the genetic architecture of trait, has a strong relationship with the accuracy of genomic prediction. Most of simulation and real data studies have shown that accuracies increase with increasing heritability (Muir, 2007; Yin et al., 2014).

The variable on which the effects of markers are regressed is an important factor affecting genomic prediction accuracy. The most reliable response variable is the true breeding value (Hayes et al., 2009), that can be only accessed in simulation studies. Daughter yield deviations (Gao et al., 2013), de-regressed proofs (DRP), or phenotypes (Fernanda et al., 2011) are possible dependent variables to be uses with real data.

Selection of animals to genotype and size of the training population are important to increase the accuracy of genomic predictions (VanRaden et al., 2009). As the contribution of sire path is higher than the dam path for the overall genetic improvement for a trait, in most countries, only sires have been included in the reference population (Loberg and Dürr, 2009). Because of the large amount of information from daughters, sire genomic information provides high accuracy of genomic prediction. However, accuracy of genomic prediction can be improved using female genotypes especially because the economically important traits are directly measured in female population. In addition, increasing attention has been directed at recording health traits, and female reference populations for GS of disease traits could be feasible (Ducrocq and Santus, 2011) and could play an important role regarding prediction accuracies. Therefore, it is necessary to figure out an optimum and practically useful training population including a reasonable ratio of genotyped cows especially for cow training sets and novel health traits.

The statistical method is another important factor that could affect the accuracy of genomic prediction (Moser et al., 2009; Su et al., 2014). Therefore, in the next section, properties of some important methods and differences between them are described.

Methods of genomic prediction

Some of the methods that have been proposed to analyze genomic data and genomic prediction are discussed below.

- **Least Squares**

The least square (LS) method, one of the earliest methods, is used to predict the expected value of observations leading to the equation:

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{g}}=\mathbf{X}'\mathbf{y} \text{ and } \hat{\mathbf{g}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where, **X** is an incidence matrix associating observations done of individuals and the regressions for markers; **y**, is a vector of the observations; $\hat{\boldsymbol{g}}$, is a vector of regression coefficients (effects) for markers. It is not possible to use LS method to fit all the effects at the same time to estimate the solutions for a very large number of parameters (p) from a smaller number of phenotypic observations (n). Because there will not be enough degrees of freedom, i.e., p>>n. However, least squares can be used to test all the genomic variables included, by analyzing them one by one for their statistical significance. The major problem that arises from the use of least squares is that a bias occurs by setting the effect of the non-significant genes to zero (Meuwissen et al.,

2001); a small effect does not necessarily mean lack of effect. Therefore, this method overestimates some variables that have statistically significant effects and underestimates variables that are non-significant but with effects.

- **Genomic best linear unbiased prediction**

The genomic best linear unbiased prediction (GBLUP) method is very close to conventional pedigree-based BLUP (Henderson, 1975). In the GBLUP, Goddard (2009) showed that NRM of BLUP is replaced by a GRM. In this method, the markers informations are fitted as random effect and their effects are assumed to be normally distributed with a uniform variance for all markers (Meuwissen et al., 2001). This method does not suffer from large p small n problem since the amount of unknown effects is usually the same as in conventional pedigree-based BLUP (González-Recio et al., 2008). When a large number of animals are phenotyped, but only a sub-sample is genotyped, the single-step genomic BLUP (ssGBLUP) methodology was proposed (e.g., Aguilar et al., 2011; VanRaden, 2012). Application of ssGBLUP allows the estimation of GEBV using pedigree, phenotypic and genomic information simultaneously by blending the NRM and GRM into an H- matrix (Legarra et al., 2014).

- **Bayesian**

An alternative approach to genomic prediction which deals with the shortage in degrees of freedom is based on Bayesian Theorem. There are a number of methods (Bayes A, Bayes B, Bayes C, Bayes LASSO etc) developed under the Bayes Theorem with different underlying assumptions; however, all of them are based on the same framework as follow:

$$P(x|y) \propto P(y|x)\, P(x)$$

where, the probability P(x|y) is called the posterior probability; P(y|x) is a pseudo-likelihood used by frequentists and P(x) is called the prior probability that is derived from the observed data. The difference between GBLUP and Bayesian methods is that in Bayesian methods, variances of the allelic effect are assumed individually.

The different Bayesian methods are distinguished by the assumptions made concerning the distribution of SNP effects with variable variances (Meuwissen et al., 2001). Accordingly, Bayes A assumes a normal prior distribution on the SNPs effects and also in this method it is assumed that the variance of SNPs effects had a scaled inverted Chi-square distribution allowing

some SNPs to have larger effects than they do under an assumption of normality. Bayes B is also described by Meuwissen et al. (2001). In this method, some SNPs (with a probability of π) have no effect on the trait, and another proportion of SNPs (with a probability of 1-π) have an effect drawn from a *t* distribution. Therefore, Bayes B can be reduced to Bayes A by having π=0 (Gianola et al., 2009). Bayes C was proposed (Kizilkaya et al., 2010) to overcome the statistical problems associated with the Bayes B, as the estimation of the probability π or the mixture distribution, which in Bayes C is applied on the SNPs effects instead of the variances (Montero, 2013). The Bayes LASSO has been proposed by Park and Casella (2008) to implement in genomic selection. This approach considers a Laplace (double exponential) prior distribution on the markers effects. The Bayes LASSO performs larger shrinkage on the marker coefficients estimates through zero than methods such as Bayes A (Montero, 2013).

- **Machine learning**

Many machine learning algorithms have emerged recently as a way to optimize predictive ability in a set of data without necessity of adjusting a specific pattern of inheritance (Long et al., 2007). Some of these algorithms are discussed in this chapter.

Reproducing Kernel Hilbert Spaces Regression (RKHS) was proposed by Gianola et al., (2006) as an alternative to SNPs regressions. This method is more attractive for multiple and complex interactions that may exist in the biological system. The results obtained using RKHS method are not worse than the Bayesian methods and in many cases out-performed them in predictive ability (Long et al.,2010).

Boosting is a machine learning ensemble method that has shown competitive behavior in prediction studies in multiple domains (Friedman, 2001). This method combines different predictors in a sequential manner with some shrinkage effect on each (Friedman, 2001). Thereby, it can handle interactions, automatically select variables, missing data and numerous correlated and irrelevant variables. It can construct variable importance in exactly the same way as random forest (RF) (Ogutu et al., 2011). For boosting algorithm, similar or better predictive abilities have been shown in comparison with Bayes A or GBLUP methods, when it has been applied to genome-wide prediction in chicken, swine and dairy cattle (González-Recio et al., 2010; González-Recio and Forni, 2011).

Support Vector Machine (SVP) methods are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and

regression analysis (Montero, 2013). It was also shown to be a particular case of RKHS (Moser et al., 2009).

Neural Networks method has also been proposed to be used in genomic studies. Gianola et al. (2011) concluded that Neural Networks may be useful for predicting complex traits using high-dimensional genomic information.

RF algorithm is another approach that was found to be appropriate for handling genetic markers effects estimation (Breiman, 2001). RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Within each tree, the best splitting predictor (SNP) at each node is chosen from a random set of all predictors. For prediction, votes from each single tree are averaged (Long et al., 2009). In animal genetics, RF has been applied to genome-wide association studies to identify SNP associated with phenotypes and to map QTL on the genome (Minozzi et al., 2014). It considers all markers with their possible interactions, environmental factors and even interactions between markers and environments, which constitutes a major advantage in the study of complex traits as diseases (Sun, 2010). Compared with other methods for binary traits (e.g., classification and regression trees, logistic regressions), RF performed better for a large sample size combined with a low percentage of missing data (García-Magariños et al., 2009). Despite the reported advantages of RF in certain situations (i.e. for complex diseases) it has been seldom uses in genome-assisted evaluations. Generally, RF algorithms and boosting are a suitable alternative to other methods used for genomic evaluations at the expense of a lower interpretability of results (González-Recio et al., 2010) and are the most appealing alternatives to analyze complex discrete traits using dense genomic markers information (González-Recio and Forni, 2011).

Genomic selection for disease resistance using cow training set

Claw disorders, clinical mastitis and infertility are the disease categories with significant incidences in dairy cattle population. These diseases are economically important and impose a large cost on dairy cattle production systems (Hogeveen et al., 2011). Although these traits are characterized by having a relatively low heritability, they have a genetic variation that enables genetic selection (Egger-Danner et al., 2015). Improvement in animal health through genetic selection is advantageous. Because genetic gain is cumulative, and small improvements that build up over time will provide ongoing savings for the long-term development of dairy

populations. In this respect, the main advantage of GS is that candidates can be evaluated and selected without their own or their relatives phenotypic records known (Boichard et al., 2016). On the other hand, because of relatively low heritability of disease traits, GS could be useful to increase genetic gain via increased accuracy of prediction. Hence, using genomic data, selection of candidates can be applied earlier with high accuracy of prediction, and subsequently, it could lead to increase disease resistance in dairy cattle. Furthermore, for novel health traits without organized progeny testing in the past, sires have a limited number of daughter records and in consequent low reliability for sire EBV. Therefore, setting up a training set just based on bulls leads to inaccurate genomic prediction. Accuracy of genomic predictions can be increased by including genotyped cows in training sets (Pryce et al., 2010), or, as a further alternative, by building a large cow training set which are only based on cow phenotypes (Pimentel et al., 2013).

From genome wide association study (GWAS) to identification of selection signatures

Complex traits, such as disease traits, are influenced by alleles segregating at multiple loci. Usually two different approaches: i) quantitative trait loci (QTL) mapping and ii) association mapping will be applied to identify the genetic variants, controlling fitness or productivity. In both approaches the associations between the phenotype of interest with genetic variants will be identified to detect chromosome segment as well as the responsible genes for its variation. Identification of QTL for complex traits is difficult and feasible only with large sample (Xu and Garland, 2017). By applying GWAS across the whole genome, non-random association between genomic markers and the trait of interest will be detected taking the advantage of the historical recombination in the population (Hunter et al., 2013). The power of GWAS to identify a true association between a genomic marker and a trait largely depends on the portion of phenotypic variance explained by the genomic marker (Pardo-Diaz et al., 2015). The efficiency of those methods will be reduced by decreasing the variation in a population for a particular trait. Maximizing the genetic variance within a sample is feasible by using large sample sizes (Korte and Farlow, 2013). However, the result of GWAS in some studies using extremely large sample sizes showed that a single trait could be controlled by many minor effect loci that could explain a small proportion of the heritability of the trait and in consequent, the identification of causal variants will be limited (Rockman, 2011). The GWAS limitation is highly relevant to the analysis of the adaptive immune traits genetics with polygenic inheritance (Rockman, 2011;

Turchin et al., 2012). In contrast to GWAS that evaluate the association between phenotype and genotype, identification of selection signature is based on evolutionary parameters and population genetic using only genomic information (Zhao et al., 2015). Selection signatures are defined as the footprints at specific regions of the genome containing a beneficial mutation and natural or artificial selection causes specific changes in the structural patterns of DNA (i.e. haplotypes) at these regions (Qanbari and Simianer, 2014). Identification of those changes among the loci that are directly affected by selection as well as the linked neutral loci is the principle of selection signature studies. Furthermore, selection causes not only the increase of allele frequencies of beneficial mutations, but also reduces local variability due to the hitchhiking theory (i. e. selective sweeps) (Smith and Haigh, 1974). Depend on assessing signatures of selection through single site differentiation and variation in linkage disequilibrium between populations (inter-populations statistics), different statistical method will be applied to identify selection signature. In the following an inter-populations statistic based on single site differentiation, Wright fixation index (F_{ST}), and one inter-populations statistic based on variation in linkage disequilibrium which were used in this thesis, are explained in detail:

F_{ST} is one of the most popular methods based on single site differentiation for detection of selection signature and reflect genetic differentiation between populations (Wright, 1949). F_{ST} is defined based on the following equation:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

Where H_T and H_S are respectively expected heterozygosities of total population and expected heterozygosities across subpopulations. The F_{ST} ranges from 0 (no differentiation between populations) to 1 (totally distinct populations). Positive selection associate with increase in F_{ST} when the populations under negative or balancing selection show low F_{ST} (Barreiro et al., 2008). The advantage of F_{ST} over LD based methods is that it is SNP-specific and can theoretically reveal the actual genetic variants under selection (Gholami, 2014).

Cross-population extended haplotype homozygosity (XP-EHH)

XP-EHH is long-range haplotype based method introduced by Sabeti et al. (2007), for identification of recent positive selection signatures. XP-EHH is based on EHH values, and evaluates the LD decay across the genome. In this method by enlarging the region of interest to 1 Mb centred on the given core SNP, the decay of LD will be measured. For a bi-allelic SNP

with alleles A and a, EHH is defined as follows (Sabeti et al., 2002):

$$EHH=\frac{\sum_{i=1}^{h_x}\binom{n_i}{2}}{\binom{n_a}{2}\binom{n_A}{2}}$$

where n_A and n_a are the number of haplotypes with alleles A and a, respectively, n_i is the count of the i^{th} haplotype within a sub-population, and h_x represents the number of distinct haplotypes in a genomic region up to a distance x from the core locus.

In order to calculate the XP-EHH for populations 1 and 2, all SNPs located 1Mb away from a given core SNP should be considered in both directions from the core SNP. Afterwards, the EHH will be integrated within these bounds (for the entire interval from the core SNP up to a distance x from the core locus) with respect to the genetic distance of core SNP and x for both populations. Then, unstandardized XP-EHH will be calculated based on the following equation (for more details see Sabeti et al., 2007):

$$XP\text{-}EHH=\log\left(\frac{\int D\ EHH_{pop1}(x)dx}{\int D\ EHH_{pop2}(x)dx}\right)$$

Objectives of the thesis

In this study, cow training set for disease traits depicts a sub population of commercial herds from eastern part of Germany and it is the basis of GEBV for German Holstein cows. Also, genotyping of female calves and heifers in west-Germany have been started by breeders whereas those family farms are not included in the training set. On the other hand, Gernand et al. (2012) reported significant differences for mean of disease incidence between large scale farms of east-Germany and small scale herds of west-Germany. Such differences between cows from contract herds in East Germany and the remaining German Holstein population could be a crucial point when genomic predictions evaluate for disease traits of a genotyped female calf or heifer, and it needs to be verified in advance. Also, assigning a fixed budget for genotyping, the first question that arises is how to choose the best animals to genotype for setting up a training set in order to maximize the accuracy of GEBV. Therefore, mimic cow training and testing sets based on incidence of diseased cows in the training set were designed in the chapter 2, to study:

- effect of composition of the training set (based on incidence of diseased cows in the training set)

- impact of genomic architecture of traits (number of QTL, heritability, the LD structure and density of the marker panel)
- impact of model choice (compare the RF methodology and GBLUP method)
- potential of random forest to detect locations of the most important SNP and how the locations overlapped with true QTL.

In the following chapter (chapter 3) a large dataset of commercial contracted herds and randomly selected cows with phenotypes for health traits was genotyped to study:

- potential of RF to estimate GEBV and compare with GBLUP and ssGBLUP for disease traits of claw disorder, clinical mastitis and female infertility
- effect of cow disease incidences in the training set on accuracies of GBV
- impact of using either DRP or PCP as response variable on accuracies of GBV
- identification of SNP with large effect by using RF and compare with a classical GWAS approach
- annotated genes in close distances of significant SNPs

In chapter 4 detection of selection signature was studied to:

- manifest adaptive genetic variation between the DSN population and the German Holstein
- demonstrate adaptive genetic variation between population strata according to disease incidences and geographical characteristics
- infer biological pathways of the annotated genes (the genes that overlapped with selection signatures).

REFERENCES

Aguilar, I.; Misztal, I.; Tsuruta, S.; Wiggans, G. R.; Lawlor, T. J. (2011). Multiple trait genomic evaluation of conception rate in Holsteins. J. Dairy Science, 94, 2621–2624.

Barreiro, L.B.; Laval, G.; Quach, H.; Patin, E.; Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat. Genet. 40, 340–345.

Barton, N. H.; Etheridge, A. M.; Ve′ber, A. (2016). The infinitesimal model. bioRxiv, 039768.

Beaulieu, J.; Doerksen, T. K.; MacKay, J.; Rainville, A.; Bousquet, J. (2014). Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC genomics 15, 1048-1064.

Boichard, D.; Ducrocq, V.; Croiseau, P.; Fritz, S. (2016). Genomic selection in domestic animals: Principles, applications and perspectives. Comptes Rendus Biologies. 339, 274–277.

Breiman, L. (2001). Random forests. Machine Learning. 45, 5-22.

Daetwyler, H. D.; Villanueva, B.; Bijma, P.;, Woolliams, J.A. (2007). Inbreeding in genome-wide selection. J Anim. Breed. Genet. 124, 369–376.

Ducrocq, V.; Santus, E. (2011). Moving away from progeny test schemes: consequences on conventional (inter)national evaluations. Interbull Bulletin 43 (http://www.interbull.org/images/stories/Ducrocq_copy.pdf).

Egger-Danner, C.; Cole, J. B.; Pryce, J. E.; Gengler, N.; Heringstad, B.; Bradley, A.; Stock, K. F. (2015). Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal, 9, 191-207.

Elsik, C. G.; Tellam, R. L.; Worley, K. C. (2009). The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324, 522–528.

Falconer, D. S. (1989). Introduction to Quantitative Genetics. 3rd ed. Longman Scientific and Technical, New York, NY.

Falconer, D. S; Mackay, T. F. C. (1996). Introduction to Quantitative Genetics, 4th edition. Prentice Hall, London.

Fernanda, V. B.; Neto J. B.; Sargolzaei, M.; Cobuci, J. A.; Schenkel, F. S. (2011). Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. BMC genetics, 12, 80.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat.

29,1189–1232

Gao, H.; Christensen, O. F.; Madsen, P.; Nielsen, U. S.; Zhang, Y.; Lund, M. S.; Su, G. (2012) Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet. Sel. Evol. GSE, 6, 44.

García-Magariños, M.; López-de-Ullibarri, I.i; Cao, R.; Salas, A. (2009). Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Annals of human genetics. 73, 360–369.

Gernand, E.; Rehbein, P.; König von Borstel, U.; König, S. (2012). Incidences of and genetic parameters for mastitis, claw disorders, and common health traits recorded in dairy cattle contract herds. J. dairy science 95, 2144–2156.

Gholami, M. (2014). Selection signature detection in a diverse set of chicken breeds. Dissertation Faculty of Agricultural Sciences, Georg-August-University Göttingen, Germany.

Gianola, D.; Fernando, R. L.; Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173, 1761–1776.

Gianola, D.; los Campos, G. de; Hill, W. G.; Manfredi, E.; Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. Genetics 183, 347–363.

Gianola, D.; Okut, H.; Weigel, K. A.; Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks. A case study with Jersey cows and wheat. BMC genetics 12, 87.

Goddard, M. E.; Hayes, B. J. (2007). Genomic selection. J. Anim. Breed. Genet., 124, 323-330.

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

González-Recio, O.; Forni, S. (2011). Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet. Sel. Evol: GSE, 43, 7.

González-Recio, O.; Gianola, D.; Long, N.; Weigel, K. A.; Rosa, G. J. M.; Avendaño, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics. 178, 2305–2313.

González-Recio, O.; Weigel, K. A.; Gianola, D.; Naya, H., Rosa, G. J. M. (2010). L2-Boosting algorithm applied to high-dimensional problems in genomic selection. Genet. Res. (Camb). 92, 227–37.

Habier, D.; Feranndo, R. L.; Dekkers, J. C. M. (2007). The impact of genetic relationship

information on genomic-assisted breeding values. Genetics, 177, 2389-2397.

Habier, D.; Fernando, R. L.; Garrick, D. J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics, 194, 597-607.

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

Habier, D.; Fernando, R. L.; Dekkers, J. C. M. (2009). Genomic selection using low-density marker panels. Genetics, 182, 343–353.

Hayes, B. J.; Bowman, P. J.; Chamberlain, A. J.; Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci., 92, 433-443

Henderson, C. R. (1975). Rapid method for computing the inverse of a relationship matrix. J. Dairy Sci., 58, 1727–1730.

Hogeveen, H.; Huijps, K.; Lam, T. J. (2011). Economic aspects of mastitis: new developments. New Zealand Veterinary Journal, 59, 16-23.

Hunter, B.; Wright, K. M.; Bomblies, K. (2013). Short read sequenc ing in studiesof natural variation and adaptation. Current Opinion in Plant Biology, 16, 85–91

Kizilkaya, K.; Fernando, F. L; Garrick, D. J. (2010). Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim Sci., 88, 544–551.

König, S.; Simianer, H.; Willam, A. (2009). Economic evaluation of genomic breeding programs. J. Dairy Science., 92, 382–391.

Korte, A.; Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS. A review. Plant methods 9, 29.

Legarra, A.; Chistensen, O. F.; Aguilar I.; Misztal I. (2014). Single step, a general approach for genomic selection. Livestock Production Science, 166, 54–65.

Loberg, A.; Dürr, J. W. (2009). Interbull survey on the use of genomic information. Interbull Bull, 39, 3–14.

Long, N.; Gianola, D.; Rosa, G. J. M.; Weigel, K. A.; Avendaño, S. (2007). Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet., 124, 377–389.

Long, N.; Gianola, D.; Rosa, G. J. M.; Weigel, K. A.; Avendaño, S. (2009). Comparison of classification methods for detecting associations between SNPs and chick mortality.

Genetics Selection Evolution, GSE, 41, 18.

Long, N.; Gianola, D.; Rosa, G. J. M.; Weigel, K. A.; Kranis, A.; González-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics research, 92, 209–225.

Luan, T.; Yu, X.; Dolezal, M.; Bagnato, A.; Meuwissen, T. H.E. (2014). Genomic prediction based on runs of homozygosity. Genetics Selection Evolution, GSE, 46, 64.

Lynch, M.; Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland.

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

Meuwissen, T. H. E. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution, GSE, 41, 35.

Minozzi, G.; Pedretti, A.; Biffani, S.; Nicolazzi, E. L.; Stella, A. (2014). Genome wide association analysis of the 16th QTL- MAS Workshop dataset using the Random Forest machine learning approach. BMC proceedings 8(Suppl 5), S4.

Montero, J. A. J. (2013). Genomic selection in small dairy cattle populations. Ph.D. Thesis. Polytechnic University of Valencia. 242.

Moser, G.; Khatkar, M. S.; Hayes, B. J.; Raadsma, H. W. (2012). Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genetics Selection Evolution, GSE, 42, 37.

Moser, G.; Tier, B.; Crump, R.; Khatkar, M.; Raadsma, H. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution, GSE 41, 56.

Muir, W. M. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet., 124, 342-355.

Ogutu, J. O.; Piepho, H. P.; Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. BMC proceedings, 5 Suppl 3, S11.

Pardo-Diaz, C.; Salazar, C.; Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. Methods in ecology and evolution 6, 445–464.

Park, T.; Casella, G. (2008). The Bayesian Lasso. J. American Statistical Association 103, 681–686.

Pimentel, E. C.; Wensch-Dorendorf, M.; König, S.; Swalve H. H.. (2013). Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genetics selection evolution, GSE, 45.

Pryce, J. E.; Goddard, M. E.; Raadsma, H. W., Hayes, B. J. (2010). Deterministic models of breeding scheme designs that incorporate genomic selection. J. Dairy Science 93, 5455–5466.

Qanbari, S.; Simianer, H. (2014). Mapping signatures of positive selection in the genome of livestock. Livest. Sci. 166, 133–143.

Rockman, M.V. (2011). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution, 66, 1–17.

Saatchi, M.; McClure, M. C.; McKay, S. D.; Rolf, M. M.; Kim, J. W.; Decker, J. E.; Taxis, T. M.; Chapple, R. H.; Ramey, H. R.; Northcutt, S. L.; Bauck, S.; Woodward, B.; Dekkers, J. C. M.; Fernando, R. L.; Schnabel, R. D.; Garrick, D. J. Taylor, J. F. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics selection evolution, GSE, 43.

Sabeti, P. C.; Reich, D. E.; Higgins, J. M.; Levine, H. Z. P.; Richter, D. J.; Schaffner, S. F. et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837.

Sabeti, P. C.; Varilly, P.; Fry, B.; Lohmueller, J.; Hostetter, E.; Cotsapas, C. et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913–918.

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

Smith, J. M.; Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23.

Su, G.; Christensen, O. F.; Janss, L.; Lund, M. S. (2014). Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Science, 97, 6547–6559

Sun, X. (2014). Genomic prediction using linkage disequilibrium and co-segregation. Iowa State University Ames, Iowa. PhD Thesis. 14273.

Sun, Y. V. (2010) "Multigenic modeling of complex disease by random forests."Advances in Genetics, 72, 73–99.

Turchin, M.C.; Chiang, C.W.K.; Palmer, C.D.; Sankararaman, S.; Reich, D.; Hirschhorn, J.N.

(2012). Evidenc e of wides pr ead selection on standing varia-tion in Europe at height-associated SNPs. Nature Genetics, 44,1015–1019

VanRaden, P. M.; Van Tassell, C. P.; Wiggans, G. R.; Sontegard, T. S. G.; Schnabel, R. D.; Taylor, J. F.; Schenkel, F. S. (2009). Reliability of genomic predictions for North American Holstein bulls. J. Dairy Science 92, 16–24.

VanRaden P. M. (2012). Avoiding bias from genomic pre-selection in converting daughter information across countries. Interbull Bulletin, 45, 1–5.

Wientjes, Y. C. J.; Veerkamp, R. F.; Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics, 193, 621-631.

Wright, S., (1949). The Genetical Structure of Populations. Ann. Eugen. 15, 323–354.

Xu, S.; Garland, T. (2017) A Mixed Model Approach to Genome-Wide Association Studies for Selection Signatures, with Application to Mice Bred for Voluntary Exercise Behavior. Genetics 207, 785–799.

Yin, T.; Pimentel, E.C.G.; König van Borstel, U.; König, S. (2014). Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature × humidity-dependent covariate. J. Dairy Science, 97, 2444–2454.

Zhao, F.; McParland, S.; Kearney, F.; Du, L.; Berry, D. P. (2015). Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genetics selection evolution, GSE 47, 49.

Zhong, S.; Dekkers, J. C. M.; Fernando, R. L.; Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. Genetics 182, 355–364

2nd Chapter

Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups

Naderi, S., Yin, T., König, S.

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

Published in Journal of Dairy Science:
http://www.journalofdairyscience.org/article/S0022-0302(16)30382-4/fulltext

ABSTRACT

A simulation study was conducted to investigate the performance of random forest (RF) and genomic BLUP (GBLUP) methodology for genomic predictions of binary disease traits based on cow calibration groups. Training and testing sets were modified in different scenarios according to disease incidences, the quantitative-genetic background of the trait (h^2 = 0.30 and h^2 = 0.10), and the genomic architecture (725 QTL and 290 QTL, population with a high and a low level of linkage disequilibrium (LD)). For all scenarios, 10,005 SNP (depicting a low density 10K SNP chip) and 50,025 SNP (depicting a 50K SNP chip) were evenly spaced along 29 chromosomes. Training and testing sets included in total 20,000 cows (4000 sick, 16,000 healthy, disease incidence 20%) from the last two generations. Initially, 4000 sick cows were assigned to the testing set, and the remaining 16,000 healthy cows represented the training set. In the ongoing allocation schemes, the no. of sick cows in the training set increased stepwise by moving 10% of sick animals from the testing set to the training set, and vice versa. Accordingly, the size of the training and testing sets were kept constant. Evaluation criteria for both applications GBLUP and RF were the correlations between genomic breeding values (GEBV) and true breeding values (TBV) (= prediction accuracy), and the area under the receiving operating characteristic curve (AUROC). Prediction accuracies and AUROC increased for both methods and all scenarios with increasing percentages of sick cows allocated to the training set. Highest prediction accuracies were observed for disease incidences within training sets reflecting the population disease incidence of 0.20. For such an allocation scheme, largest prediction accuracies of 0.53 for RF and of 0.51 for GBLUP, and largest AUROC of 0.66 for RF and of 0.64 for GBLUP, were achieved using 50,025 SNP, a heritability of 0.30, and 725 QTL. Heritability decrease from 0.30 to 0.10, and QTL reduction from 725 to 290, were associated with decreasing prediction accuracies and decreasing AUROC for all scenarios. This decrease was more pronounced for RF. Also the increase of LD had stronger impact on RF compared to GBLUP results. The highest prediction accuracy from the low LD scenario was 0.30 from RF and 0.36 from GBLUP, and increased to 0.39 for both methods in the high LD population. RF successfully identified important SNP in close map distance to QTL explaining a high proportion of the phenotypic trait variation.

Key words: Disease traits, random forest methodology, accuracy of genomic predictions

INTRODUCTION

Classically, large sire calibration groups combined with a two-step procedure are used for the estimation of genetic values based on SNP marker effects in dairy and dual-purpose cattle breeding programs (e.g. Edel et al., 2011). Such procedures utilize highly accurate conventional estimated breeding values (**EBV**) for sires, implying consideration of the traditional flow of traits from official recording systems. Modern dairy cattle breeding programs are aiming at including novel functional traits that reflect animal health, behavior and product quality (e.g. König et al., 2013).

From an animal breeding perspective, substantial cow health improvements and increasing monetary genetic gain imply the direct inclusion of health traits into overall breeding goals or selection indices (König et al., 2009; Egger-Danner et al., 2012). However, for novel traits, and without the organized and widespread progeny testing schemes of the past, sires have a limited no. of daughter records. Only a few daughter records per sire imply sire EBV with low reliability, further, also inaccurate genomic predictions from two-step procedures. Reliabilities of genomic predictions can be improved by including genotyped females in the training set (Pryce et al., 2010; Mc Hugh et al., 2011), or, alternatively, by setting up large cow calibration groups which are solely based on cow phenotypes (Pimentel et al., 2013). In a study by Buch et al. (2012), accuracies of direct genomic breeding values (**GEBV**) were higher when using cow phenotypes instead of sire EBV in training sets. Higher no. of genotyped females allow more accurate selection of bull dams and cow dams, with an associated increase of genetic gain in the whole population (Thomasen et al., 2014).

Cow calibration groups, i.e., combining phenotypes for novel traits with high density genetic markers, offer new perspectives towards breeding for improved disease resistance. Nevertheless, disease related traits are generally categorical, influenced by multiple genes, deviate from the Mendelian inheritance, and show obvious gene by gene as well as gene by environment interactions (Hernandez and Blazer, 2006), which all pose statistical challenges for GEBV estimation. Classically, as introduced in the key paper by Meuwissen et al. (2001), mixed model equations (**MME**) were solved by applying genomic BLUP (**GBLUP**) models or Bayesian methodology. Especially for novel traits, current cow calibration group studies are characterized by a small number of genotyped animals (n), genotyped with dense SNP panels (m) (Kramer et al., 2014). This implies model over-parameterization (m larger than n), suggesting evaluation of alternative methodology, such as random forest (**RF**) applications.

In animal genetics, RF was applied to genome wide association studies (**GWAS**) to identify SNP associated with phenotypes, and in order to map quantitative trait loci (**QTL**) on the genome (Minozzi et al., 2014). In comparison to other GWAS methods for binary traits (e.g., classification and regression trees (**CART**), logistic regressions), RF performed better for a large sample size combined with a low percentage of missing data (Garcia-Magarinos et al., 2009). Also Li et al. (2014) and Nguyen et al. (2015) explored the potential of RF for GWAS. In genomic predictions, González-Recio and Forni (2011) showed that RF performed better than Bayesian regressions in detecting resistant and susceptible animals from genetic markers. Ogutu et al. (2011) applied RF to a calibration group of 2326 genotyped and phenotyped individuals. Correlations between predicted GEBV and true breeding values (**TVB**) were in a moderate range (0.39 to 0.54).

A variety of factors and parameters, e.g. the heritability and genetic architecture of the trait, the linkage disequilibrium (**LD**) structure of the population, and the design of the training set, affect the accuracy of genomic predictions. Especially the size of the training set and the strength of genetic relationships between training and testing samples, contributed to different prediction accuracies from GBLUP and Bayesian applications (Albrecht et al., 2011; Pszczola et al., 2012; de Los Campos et al., 2013). So far, the efficiency of RF for genomic predictions using calibration groups, and how to stratify datasets according to genomic architectures and trait characteristics, is not yet clarified. However, knowledge in this regard is imperative, because cow calibration groups, especially for novel health traits, usually represent only a small sub-population from selected herds. For example, a calibration group of ~ 20,000 cows from only ~ 50 herds located in the eastern part of Germany is the basis for genomic selection for health traits in the German Holstein population (Yin and König, 2016). The registered German Holstein population includes in total 1.72 Million cows from 18,700 herds (ADR, 2014). Substantial differences for mean disease incidences were reported when comparing small-scale farms from the western part with large-scale herds in the eastern part of Germany (Gernand et al., 2012). König et al. (2005) identified genotype by environment interactions for protein yield when stratifying data according to the herd locations either “East” or “West” Germany. In this context, we hypothesize impact of cow calibration group characteristics (here: cows from large-scale contract herds in East Germany) when predicting the disease probability of a genotyped female calf or heifer from a different sub-population.

We applied stochastic simulations to mimic different designs of cow training and testing sets,

being the basis for ongoing genomic predictions for disease traits. Specifically, the objectives of the present study were: i) to compare accuracies of genomic predictions using RF and GBLUP for a simulated binary trait by altering the following parameters: incidence of diseased cows in the training population, heritability, no. of QTL, marker density, and the LD structure of the genotyped population; and ii) to compare RF estimates for effects and locations of the most important SNP with true QTL.

MATERIALS AND METHODS

Simulation of populations

The QMSim software (Sargolzaei and Schenkel, 2009) was used to generate phenotypes, genotypes, and TBV of a dairy cattle population by applying stochastic simulations. At first, over a time span of 1000 generations, a historical population including 20,000 females and 400 males was created to achieve a desired LD level. In order to generate a population with a higher level of LD, a "bottleneck" was simulated. For this purpose, we initiated the same simulation process, but after 1000 generations, the population size decreased during 80 generations to 500 individuals. Afterwards, the population size increased over a time period of 140 generations, resulting again in 20,000 females and 400 males. In the second simulation step in the recent population, all animals from the last generation of the historical population served as founders. The recent population was expanded by simulating additional eight generations using a random mating design. Each mating produced one offspring with an equal probability of being either male or female. Replacement rates were 50% for males and 20% for females, and selection for both genders based on EBV.

Table 1: The evaluated scenarios with respect to the no. of markers and QTL, the heritability of the trait, and the level of linkage disequilibrium (LD)

	Scenario						
Parameter	S_I	S_II	S_III	S_IV	S_V	S_VI	S_VII
h^2	0.3	0.3	0.1	0.3	0.3	0.1	0.1
No. of QTL	725	290	290	725	290	290	290
No. of SNP	10K	10K	10K	50K	50K	50K	50K
Level of LD	Low	Low	Low	Low	Low	Low	High

In the recent population, seven different scenarios (Table 1) were simulated in order to reflect variations with regard to heritability, no. of QTL, no. of SNP and LD. Biallelic SNP markers were evenly spaced along 29 chromosomes, each 80 cM long. Simulations of either 345 or 1725 biallelic markers per chromosome depicted applications with 10,005 SNP (10K chip) and 50,025 SNP (50K chip), respectively. For both marker densities 10K and 50K, two different no. of QTL (either 10 or 25 QTL on each chromosome) affected the trait of interest. QTL effects were sampled from a gamma distribution with a shape parameter of 0.4. The gamma distribution assumes many QTL with small effects, and correspondingly, only a few QTL with large effects. The mutation rate was 2.5×10^{-5} for both markers and QTL per locus and per generation, as used in previous simulations (e.g., Yin et al., 2014). The whole amount of additive-genetic variance was attributed to the QTL, implying no further polygenic effects. Simulations considered a low ($h2 = 0.10$) and a moderate heritability trait ($h2 = 0.30$). The whole set of parameters as used for the simulations is summarized in Table 2. Cows of the last generation were ranked in descending order according to the continuous trait. The value from rank no. 4000 was defined as threshold, implying 4000 animals with code 1 = diseased, and 16,000 animals with code 0 = healthy. Such ranking and transformation into a binary outcome reflected a selection strategy on a continuous trait (e.g. protein yield) in the past generations, and a pronounced antagonistic relationship with a disease trait.

The LD measurement applied in this study was r2. Average r2 in the low LD scenario (S_VI) was 0.224 for distances of 0.05 cM, while the corresponding average r2 in the high LD scenario (S_VII) was 0.425. The PLINK program (Purcell et al., 2007) was used to calculate the amount of LD between the most important QTL (10 QTL on the whole genome explaining the highest amount of genetic variance in relation to the phenotypic variance of the trait), and the most important SNP at the same chromosome identified by applying RF methodology.

Table 2: Parameters of the simulation process

Parameters	Low linkage disequilibrium	High linkage disequilibrium
Historical population		
No. of generations (population size) in phase 1	1000 (20,400)	1000 (20,400)
No. of generations (population size) in phase 2	-	1080 (500)
No. of generation (population size) in phase 3	-	1220 (20,400)
Recent population		
No. of founder sires (dams)	400 (20,000)	
No. of generations	8	
No. of offspring per dam	1	
Mating system	Random	
Replacement ratio for males (females)	0.5 (0.2)	
Criteria for selection / culling	EBV / age	
Sex probability for offspring	0.5	
Genome		
No. of chromosomes	29	
Total length of chromosomes (cM)	2,320	
Marker distribution	Evenly spaced	
No. of QTL alleles	Random (2, 3, or 4)	
Effects of QTL alleles	Gamma (0.4)	
Marker and QTL mutation rate	$2.5\ 10^{-5}$	
Position of marker and QTL	Random	
No. of QTL	290, 725	290
No. of markers	10,005, 50,025	50,025
Heritability of the trait	0.30, 0.10	0.10

Prediction of genomic breeding values and of SNP effects

Genomic breeding values. For the estimation of GEBV, GBLUP and RF were applied to the simulated binary data. For GBLUP, the AI-REML algorithm (DMU software package, Madsen and Jensen, 2010) was used, which allows the specification of a generalized linear mixed model with a logit link function for binary data. The statistical model was:

$$\text{logit}\,(\pi_r) = \log[\pi_r / (1-\pi_r)] = \varphi + \gamma_r$$

π_q = probability of occurrence for the disease of cow *r*

φ = overall mean effect

γ_r = random animal effect

The random animal effect was included by considering the genomic relationships among animals based on SNP marker data. The genomic relationship matrix (**G**-matrix) was constructed according to the method proposed by VanRaden (2008), and applying the **G**-matrix software (Su and Madsen, 2013). Markers with minor allele frequency lower than 0.05 were discarded. A small value (0.01) was added to the diagonal of the **G**-matrix to circumvent problems with matrix singularity.

For RF analyses, the java package RanFoG (González-Recio and Forni, 2011) was applied. In the RF analysis (see Breiman, 2001, for details), thousands of classification trees were constructed by bootstrapping (Efron and Tibshirani, 1993) of the data in the training set. For the construction of each tree, F used on average about two thirds of the observations and a random subset *p* of the *m* SNP (p ~ 2/3*m). Cows which were not included in the bootstrapped sample were defined as "out of bag (**OOB**)", being the testing set for each tree. At each node, data were split in two branches based on the genotype at SNP *j* by minimizing a loss function for classification. Repetition of this procedure implied a large no. of trees (i.e. random forest), until the convergence criterion was achieved. The convergence criterion used classification errors of OOB samples. In the present study, 2000 and 5000 trees were constructed for the 10K and the 50K SNP chip, respectively. Random sampling (bootstrapping) of the data contributed to the formation of de-correlated trees. Each tree reflected the most frequent outcome of the disease for a given combination of SNP genotypes. The average of the predicted value of each tree was the probability for being susceptible to the disease.

Assessing the prediction performance

When allocating cows to training and testing sets, we considered 20,000 cows from the last two recent generations. All cows had genotypes, but phenotypes of the cows in the testing set were assumed to be unknown. The training and testing sets were designed for different percentages of sick and healthy cows (Figure 1). In total, 4000 cows were considered to be sick, and 16,000 cows were healthy, reflecting a disease incidence of 0.20 in the population. Initially, the 4000 sick cows were allocated to the testing set, and all 16,000 cows in the training set were healthy. In ongoing animal allocation schemes, the no. of sick cows in the training set increased by moving 400 sick animals from the testing to the training set. Keeping the size of both sets constant, vice versa 400 healthy cows moved from the training to the testing set. This was done stepwise in increments of 400 cows, until all of the 4000 sick cows moved to the training set. Correspondingly, in the final step, all 4000 cows in the testing set were healthy. For all allocation schemes within each scenario, we performed 10 replicates.

The correlation coefficient between TBV and predicted GEBV from cows in the testing set was the evaluation criterion for the prediction accuracy of GEBV. A further evaluation criterion for phenotype predictions was the area under the receiving operating characteristic curve (**AUROC)**. AUROC is an information criterion to assess the efficiency of classification models by comparing true positive and false positive classifications, which was often used to compare models in clinical studies. As specified by González-Recio et al. (2014), the genetic susceptibility to disease was predicted based on SNP genotypes. Accordingly, animals were classified as either susceptible or non-susceptible, depending on an arbitrary GEBV threshold (t). Genomic predictions in relation to the true genetic susceptibility could either be true positive, false positive, true negative or false negative, classified as positive (**FPR**= (predicted positives) over (negatives); **TPR** = (predicted positives) over (positives)). The receiver operating characteristic (**ROC**) curve depicted TPR in relation to FPR for all thresholds in [0, 1] that can be used to classify animals into sick and healthy. Own R code was used to plot the ROC curve, and to calculate AUROC (higher values for AUROC indicate accurate disease predictions).

Importance of single SNP. The importance of single SNP was calculated based on OOB data, i.e., comparing the difference in prediction accuracy from the original OOB sample of a given tree in relation to another OOB sample, where the genotype of *j-th* SNP has been permuted. The average decrease in accuracy for the *j-th* permuted SNP over all bootstrapped trees determined

the importance of each SNP. The relative importance of a single SNP in the range from 0 to 1 was calculated by defining a SNP ratio, which was a specific single SNP value in relation to the highest SNP value (most important SNP: value of 1). Values of all SNP were saved in the variable importance file of RanFoG.

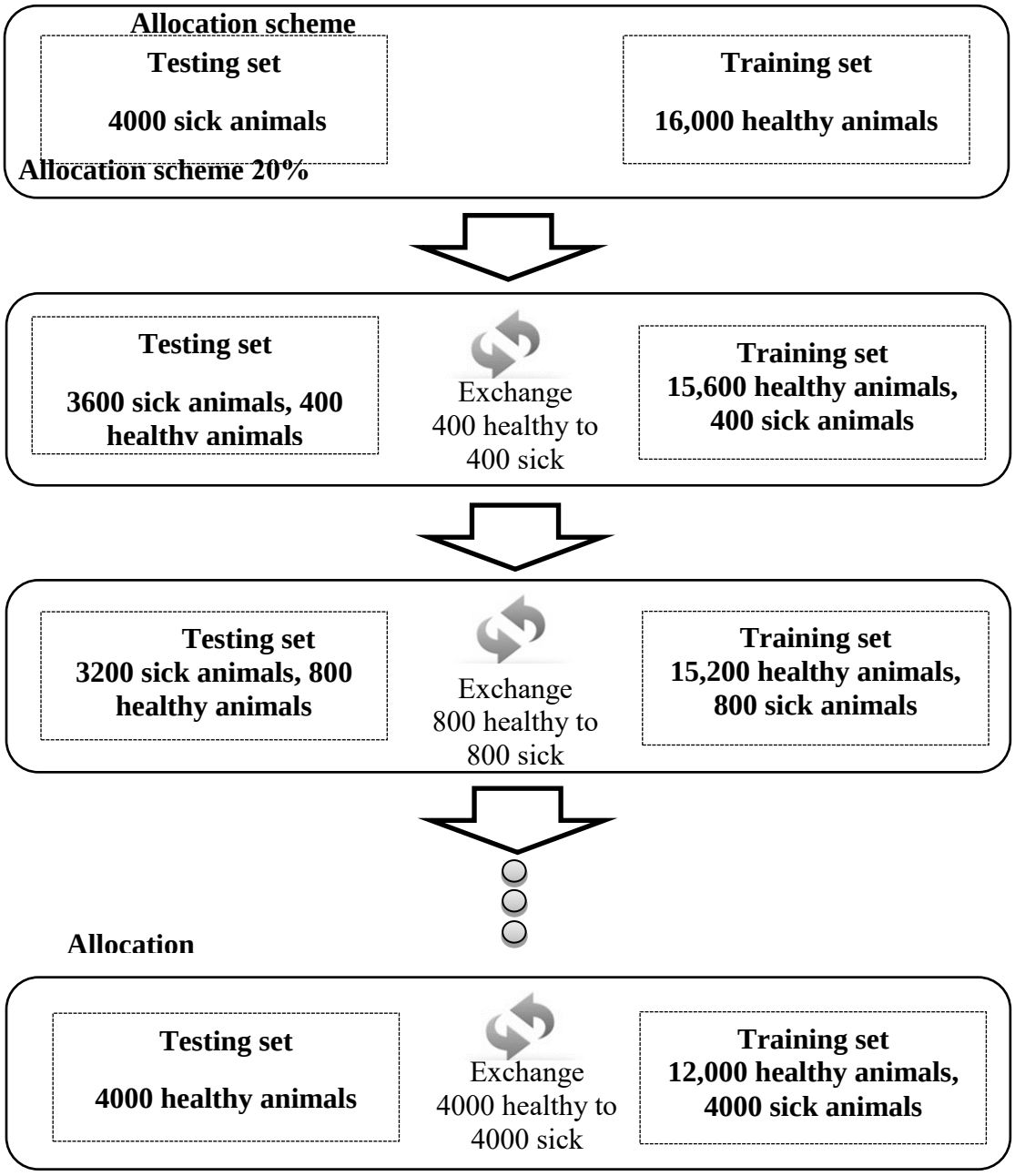

Figure 1: Strategy for the creation of training and testing sets

RESULTS AND DISCUSSION

Accuracy of predictions with low (10K) and higher density (50K) SNP chips

General aspects. For the 10K SNP panel, accuracies of cow GEBV in the testing set estimated with RF and GBLUP are listed in Table 3 (scenarios S_I, S_II and S_III). For all percentages of sick cows in the training set, prediction accuracies from GBLUP always outperformed those from corresponding

RF applications. As indicated by smaller SD, results from RF across the ten replicates were more homogenous compared to GBLUP applications. When increasing the no. of sick cows allocated to training sets from 10% (400 cows = 2.5% of all cows in the training set are sick) to 60% (2400 cows = 15% of all cows in the training set are sick), the prediction accuracy increased correspondingly. For instance, with regard to scenario S_I, the prediction accuracy increased from 0.25 to 0.47 for GBLUP, and from 0.14 to 0.32 for RF. Hence, highest accuracies were observed for percentages of sick cows in training sets being close to the population disease incidence (20%). The prediction accuracy remained quite constant when allocating a higher no. of sick cows to the training set. AUROC values for these scenarios S_I, S_II and S_III are summarized in Table 4. Also the AUROC evaluation criterion reflected superiority of GBLUP compared to RF predictions. AUROC from GBLUP ranged between 0.56 and 0.66, and was slightly smaller for RF. AUROC only slightly increased for a higher no. of sick animals allocated to training sets.

For 10K SNP chip panels and in comparison to RF, GBLUP provided more accurate GEBV, and differentiated more accurately between sick and healthy individuals. González-Recio and Forni (2011) simulated a moderate heritability trait (h2 = 0.25). They randomly selected 2500 animals to establish a training set, in order to evaluate the efficiency of Bayesian regressions and machine-learning algorithms for a binary trait. They reported an accuracy of 0.36 when applying RF, and accuracies ranging from 0.37 to 0.41 for boosting algorithms. Ogutu et al. (2011) applied three machine learning methods and ridge regression BLUP to a calibration group of 2326 animals genotyped for 10K. For a simulated quantitative trait with heritabilities of 0.39 for females and 0.52 for males, the correlation between predicted GEBV and TBV from RF applications was 0.48, and 0.60 for ridge regression.

For the 50K SNP chip and for both methods RF and GBLUP, correlations between GEBV and TBV for individuals without phenotypes from scenarios S_IV, S_V, S_VI and SV_II are listed in Table 5. For all scenarios, prediction accuracies increased with increasing percentages of sick

animals allocated to the training set until a 50% threshold was achieved, and remained stable thereafter. For scenario S_IV and GBLUP applications, even a continuous increase was observed, with highest accuracies for the 100% sick animal allocation scheme. Specifically, for GBLUP and scenario S_IV, prediction accuracies increased from 0.24 (10% of sick animals allocated to the training set) to 0.50 (100% of sick animals allocated to the training set = 25% of sick animals within the training set). With regard to scenario S_IV and RF methodology, accuracies increased from 0.20 to 0.52, and the highest accuracy (0.53) was identified for 3200 sick animals allocated to the training set (= 80% of sick cows allocated to the training set). Such a high percentage of sick animals allocated to the training set reflected similar disease incidences for both groups training set and whole population. Also AUROC evaluation supported the better performance of RF compared to GBLUP for scenarios with the 50K SNP chip (Table 6). For both prediction methods RF and GLUP, AUROC was lowest for low percentages of sick animals allocated to the training set. For example for scenario S_IV, and for only 10% of sick animals allocated to training sets, AUROC values were 0.58 and 0.57 when using RF and GBLUP, respectively.

With regard to the 50K SNP chip and scenario S_IV, RF performed better than GBLUP for all allocation schemes, apart from the 10%, 30% and 90% cases. The increased marker density contributed to RF prediction improvements. An increase of SNP markers shortened the distances between markers and functional mutations. In consequence, SNP close to a QTL were sampled with sufficient frequency, implying that the signal of the QTL is captured by distinct SNP in close map distance. The higher prediction accuracies with larger marker density were in agreement with previous studies (Makowsky et al., 2011; Vazquez et al., 2012). However, VanRaden et al. (2011) identified only minor increase in prediction accuracies when using SNP chips larger than 50K.

Table 3: Correlation between true and predicted genomic breeding values for S_I (10K SNP, h^2 =0.30 and 725 QTL), S_II (10K SNP, h^2 =0.30 and 290 QTL) and S_III (10K SNP, h^2 =0.10 and 290 QTL) from RF and GBLUP applications (the values in parenthesis show the SD from ten replicates).

		Percentage of sick animals allocated to the training set / (percentage of sick animals within the training set)[1]								
Method	Scenario	10% (2.5%)	20% (5%)	30% (7.5%)	40% (10%)	50% (12.5%)	60% (15%)	70% (17.5%)	80% (20%)	90% (22.5%)
GBLUP	S_I	0.25 (0.05)	0.31 (0.06)	0.36 (0.04)	0.42 (0.03)	0.45 (0.03)	0.47 (0.02)	0.46 (0.03)	0.47 (0.03)	0.47 (0.02)
	S_II	0.24 (0.06)	0.33 (0.04)	0.35 (0.001)	0.42 (0.03)	0.45 (0.02)	0.46 (0.03)	0.47 (0.03)	0.48 (0.03)	0.48 (0.03)
	S_III	0.16 (0.05)	0.24 (0.03)	0.26 (0.03)	0.31 (0.05)	0.31 (0.03)	0.32 (0.03)	0.33 (0.04)	0.35 (0.03)	0.34 (0.03)
RF	S_I	0.14 (0.010)	0.19 (0.01)	0.25 (0.03)	0.29 (0.02)	0.31 (0.02)	0.32 (0.03)	0.35 (0.02)	0.34 (0.01)	0.34 (0.02)
	S_II	0.13 (0.02)	0.20 (0.03)	0.25 (0.04)	0.29 (0.02)	0.29 (0.02)	0.32 (0.03)	0.34 (0.03)	0.35 (0.03)	0.34 (0.02)
	S_III	0.07 (0.02)	0.11 (0.01)	0.14 (0.03)	0.16 (0.02)	0.19 (0.01)	0.21 (0.02)	0.20 (0.01)	0.23 (0.01)	0.23 (0.03)

[1]Size of training and testing set was 16,000 and 4000, respectively for all scenarios.

Table 4: The area under the receiving operating characteristic curve (AUROC) for S_I (10K SNP, h^2 =0.30 and 725 QTL), S_II (10K SNP, h^2 =0.30 and 290 QTL) and S_III (10K SNP, h^2 =0.10 and 290 QTL) from RF and GBLUP applications (the values in parenthesis show the SD from ten replicates).

		Percentage of sick animals allocated to the training set / (percentage of sick animals within the training set)[1]								
Method	Scenario	10% (2.5%)	20% (5%)	30% (7.5%)	40% (10%)	50% (12.5%)	60% (15%)	70% (17.5%)	80% (20%)	90% (22.5%)
GBLUP	S_I	0.62 (0.026)	0.64 (0.029)	0.64 (0.015)	0.65 (0.021)	0.66 (0.017)	0.64 (0.01)	0.64 (0.012)	0.65 (0.017)	0.63 (0.016)
	S_II	0.56 (0.011)	0.59 (0.011)	0.60 (0.014)	0.60 (0.014)	0.62 (0.008)	0.62 (0.008)	0.61 (0.018)	0.62 (0.018)	0.63 (0.025)
	S_III	0.50 (0.004)	0.52 (0.010)	0.52 (0.010)	0.53 (0.014)	0.54 (0.01)	0.54 (0.02)	0.55 (0.012)	0.56 (0.018)	0.53 (0.009)
RF	S_I	0.56 (0.014)	0.57 (0.008)	0.58 (0.007)	0.59 (0.01)	0.59 (0.008)	0.60 (0.013)	0.60 (0.016)	0.60 (0.01)	0.60 (0.014)
	S_II	0.55 (0.008)	0.57 (0.008)	0.57 (0.01)	0.59 (0.013)	0.59 (0.016)	0.59 (0.014)	0.59 (0.014)	0.61 (0.014)	0.60 (0.020)
	S_III	0.51 (0.012)	0.53 (0.005)	0.53 (0.005)	0.53 (0.012)	0.53 (0.012)	0.54 (0.012)	0.54 (0.014)	0.54 (0.012)	0.53 (0.008)

[1]Size of training and testing set was 16,000 and 4000, respectively for all scenarios.

Table 5: Correlation between true and predicted genomic breeding values for S_IV (50K SNP, h^2 =0.30 and 725 QTL), S_V (50K SNP, h^2 =0.30 and 290 QTL), S_VI (50K SNP, h^2 =0.10 and 290 QTL) and S_VII (50K SNP, h^2 =0.10, 290 QTL and high LD) from RF and GBLUP applications (the values in parenthesis show the SD from ten replicates).

		Percentage of sick animals allocated to the training set / (percentage of sick animals within the training set)[1]									
Method	Scenario	10% (2.5%)	20% (5%)	30% (7.5%)	40% (10%)	50% (12.5%)	60% (15%)	70% (17.5%)	80% (20%)	90% (22.5%)	100% (25%)
GBLUP	S_IV	0.24 (0.09)	0.36 (0.07)	0.40 (0.08)	0.43 (0.06)	0.43 (0.08)	0.46 (0.07)	0.49 (0.06)	0.50 (0.07)	0.51 (0.06)	0.50 (0.05)
	S_V	0.23 (0.09)	0.36 (0.08)	0.38 (0.06)	0.41 (0.05)	0.44 (0.05)	0.45 (0.05)	0.47 (0.07)	0.49 (0.07)	0.48 (0.06)	0.44 (0.06)
	S_VI	0.23 (0.05)	0.22 (0.06)	0.27 (0.10)	0.28 (0.06)	0.32 (0.05)	0.31 (0.07)	0.33 (0.04)	0.32 (0.09)	0.33 (0.08)	0.36 (0.05)
	S_VII	0.19 (0.08)	0.26 (0.08)	0.28 (0.08)	0.33 (0.08)	0.34 (0.08)	0.36 (0.07)	0.38 (0.07)	0.38 (0.07)	0.37 (0.06)	0.39 (0.07)
RF	S_IV	0.20 (0.05)	0.36 (0.04)	0.39 (0.06)	0.46 (0.05)	0.47 (0.04)	0.52 (0.05)	0.51 (0.05)	0.53 (0.05)	0.50 (0.05)	0.52 (0.03)
	S_V	0.15 (0.04)	0.28 (0.07)	0.30 (0.02)	0.36 (0.04)	0.39 (0.01)	0.44 (0.02)	0.44 (0.03)	0.44 (0.03)	0.48 (0.02)	0.44 (0.01)
	S_VI	0.13 (0.004)	0.18 (0.02)	0.17 (0.05)	0.17 (0.01)	0.20 (0.07)	0.23 (0.03)	0.26 (0.05)	0.30 (0.04)	0.28 (0.06)	0.27 (0.03)
	S_VII	0.17 (0.02)	0.20 (0.04)	0.20 (0.02)	0.24 (0.02)	0.26 (0.04)	0.33 (0.01)	0.32 (0.02)	0.39 (0.02)	0.34 (0.01)	0.35 (0.01)

[1]Size of training and testing set was 16,000 and 4000, respectively for all scenarios.

Table 6: The area under the receiving operating characteristic curve (AUROC) for S_IV (50K SNP, h^2 =0.30 and 725 QTL), S_V (50K SNP, h^2 =0.30 and 290 QTL), S_VI (50K SNP, h^2 = 0.10 and 290 QTL) and S_VII (50K SNP, h^2 = 0.10, 290 QTL and high LD) from RF and GBLUP applications (the values in parenthesis show the SD from ten replicates).

		Percentage of sick animals allocated to the training set / (percentage of sick animals within the training set)[1]								
Method	Scenario	10% (2.5%)	20% (5%)	30% (7.5%)	40% (10%)	50% (12.5%)	60% (15%)	70% (17.5%)	80% (20%)	90% (22.5%)
GBLUP	S_IV	0.57 (0.06)	0.59 (0.036)	0.60 (0.041)	0.59 (0.031)	0.60 (0.037)	0.61 (0.042)	0.63 (0.031)	0.63 (0.040)	0.64 (0.046)
	S_V	0.53 (0.06)	0.58 (0.05)	0.59 (0.05)	0.60 (0.02)	0.60 (0.02)	0.60 (0.02)	0.62 (0.02)	0.61 (0.04)	0.62 (0.03)
	S_VI	0.54 (0.04)	0.52 (0.05)	0.53 (0.03)	0.53 (0.02)	0.53 (0.02)	0.53 (0.02)	0.54 (0.02)	0.55 (0.04)	0.53 (0.04)
	S_VII	0.56 (0.019)	0.56 (0.020)	0.57 (0.023)	0.57 (0.025)	0.57 (0.010)	0.58 (0.019)	0.58 (0.019)	0.58 (0.014)	0.57 (0.007)
RF	S_IV	0.58 (0.01)	0.64 (0.02)	0.64 (0.02)	0.66 (0.02)	0.66 (0.035)	0.63 (0.041)	0.64 (0.021)	0.66 (0.012)	0.64 (0.009)
	S_V	0.47 (0.02)	0.58 (0.02)	0.57 (0.01)	0.61 (0.01)	0.62 (0.02)	0.62 (0.02)	0.62 (0.01)	0.63 (0.02)	0.65 (0.03)
	S_VI	0.57 (0.05)	0.54 (0.01)	0.54 (0.01)	0.56 (0.02)	0.56 (0.01)	0.56 (0.01)	0.57 (0.01)	0.57 (0.01)	0.59 (0.004)
	S_VII	0.56 (0.026)	0.55 (0.016)	0.56 (0.026)	0.56 (0.018)	0.56 (0.011)	0.59 (0.011)	0.57 (0.018)	0.58 (0.024)	0.58 (0.011)

[1]Size of training and testing set was 16,000 and 4000, respectively for all scenarios.

When using the 50K SNP to analyze binary data, RF ranked individuals more accurately than GBLUP, and also precisely distinguished between healthy and affected individuals in most of the allocating schemes. For example, AUROC 0.66 (S_IV, method RF, 80% of sick animals allocated to the training set) indicated for 66% of susceptible individuals larger genomic predictions than for non-susceptible individuals. Likewise, an AUROC of value 0.50 implied that a classification in either being sick or healthy is not better than a random guess: 50% of susceptible individuals with larger predictions than non-susceptible individuals, and vice versa. AUROC values were in the range of previous studies (González-Recio and Forni, 2011; Vazquez et al., 2012), reporting AUROC between 0.58 and 0.70. Larger AUROC in the range from 0.62 to 0.97 were calculated by Nguyen et al. (2015) when using high density 400K SNP chips. Hence, marker density and the size of training sets were identified as the most important parameters effecting AUROC.

Interestingly and presumably attributed to the large-scale cow calibration group, accuracies of GEBV in the present study were higher compared to previous simulation evaluations. Contradictorily, values for AUROC were slightly smaller. Using RF, prediction was based on a random sub-sample of SNP. Therefore, especially for low-density SNP panels, it might be the case that SNP close to a QTL were not sufficiently sampled, implying that the QTL signal was captured by distant SNP.

A wide range of computation times was observed across scenarios. Computation time was generally higher for RF. For example for scenario S_I, the computation time for GBLUP was 46 hours per replicate, but 68 hours per replicate for RF. Ghafouri-Kesbi et al. (2015) reported that GBLUP was faster than RF, and also Neves et al. (2012) confirmed the lower computation time for a GBLUP model compared to RF. Therefore, computation time might be the crucial limitation factor when intending to apply RF to a large dataset of genotyped cows.

Impact of the number of QTL. For the 10K SNP panel, accuracies of GEBV from RF and GBLUP applications for scenarios with identical heritability, but differing no. of QTL, i.e., S_I (725 QTL) versus S_II (290 QTL), were very similar (Table 3). For the low density 10K SNP panel, with either a low or larger no. of QTL (S_I versus S_II), prediction accuracies from GBLUP were higher compared to RF. Also minor impact on AUROC was identified when decreasing the no. of QTL (Table 4). The effect of QTL reduction on AUROC was more pronounced for GBLUP, but AUROC still indicated superiority of GBLUP over RF for a reduced no. of QTL.

For the 50K SNP panel, accuracies of GEBV from RF and GBLUP applications for scenarios

S_IV (725 QTL) and S_V (290 QTL), both considering a trait with heritability 0.30, are shown in Table 5. Especially for RF, a pronounced impact of the QTL no. on prediction accuracies was identified: accuracies strongly decreased for all animal allocation schemes when decreasing the no. of QTL from 725 to 290. Regarding GBLUP, only a slight reduction in prediction accuracies was identified. For most of the cow allocation schemes (apart from the "10%, 30% and 90% cases") and 725 simulated QTL, RF performed better than GBLUP, but opposite results were found when simulating 290 QTL. Calculated AUROC values for these scenarios S_IV and S_V are summarized in Table 6. An evaluation of AUROC exhibits better performances of RF for both QTL scenarios. Also for the lower no. of 290 QTL and despite lower correlations between TBV and GEBV, the RF method separated healthy and sick individuals more precisely than GBLUP (comparison of results for scenario S_V in Tables 6 and 7).

Table 7: The calculated r^2 between each of ten most effective QTL and the most important SNP located in the same chromosome for scenarios S_I (10K SNP chip, $h^2 = 0.3$ and QTL = 725) and S_IV (50K SNP chip, $h^2 = 0.3$ and QTL = 725)

S_I	QTL	Q_141	Q_200	Q_261	Q_300	Q_303	Q_334	Q_402	Q_453	Q_555	Q_574
	Chr.	6	8	11	12	12	14	17	18	22	23
	r^2	0.006	0.004	0.001	0.002	0.001	0.002	0.02	0.001	0.004	0.001
S_IV	QTL	Q_78	Q_86	Q_217	Q_293	Q_342	Q_370	Q_414	Q_525	Q_562	Q_570
	Chr.	3	3	9	12	14	15	18	22	23	24
	r^2	0.28	0.11	0.01	0.06	0.15	0.23	0.01	0.11	0.06	0.01

For a large no. of QTL, Ghafouri-Kesbi et al. (2015) reported higher GBLUP accuracies compared to accuracies from RF and gradient boosting. For their scenario with a low no. of QTL and allele effects drawn from a gamma distribution, gradient boosting was the best performing method. For all methods and contrarily to the results from our present study, Ghafouri-Kesbi et al. (2015) reported generally increasing prediction accuracies with a decreasing no. of QTL. Reasons for these opposite results might be due to the different no. of simulated chromosomes, and different effective population sizes. Ghafouri-Kesbi et al. (2015) spaced 10,000 SNP along five chromosomes, and considered an effective population size of Ne = 100. According to Daetwyler et al. (2010a), prediction accuracy depends on the no. of independent chromosome

segments (**Me**), while Me itself depends on the effective population size and the length of the genome (Goddard, 2009). González-Recio and Forni (2011) identified higher accuracies of GEBV with Bayesian regression applications compared to boosting algorithms and RF methods, when the no. of QTL increased. In the present study, RF showed the smallest accuracy, but the highest AUROC for a scenario assuming a large no. of QTL. In this context, González-Recio and Forni (2011) discussed and compared both evaluation criteria AUROC and accuracy of GEBV. Irrespective of lower accuracies when applying RF to a trait characterized by a high no. of QTL, RF technique was more accurate to discern between healthy and affected individuals. Generally higher sensitivity of RF on QTL alterations than GBLUP (quite stable results for both QTL scenarios) has the following explanation: GBLUP assumed the same variance for each independent chromosome segment regardless of the effect of the segment, while RF based on a sampling technique for predictors (SNP). Therefore, by using high density marker panels combined with a large no. of QTL, SNP in close distance to a QTL were sufficiently frequently sampled. Nevertheless, the fact that most of the important dairy traits are affected by many genes with small effects supports the assumptions made for GBLUP applications (Hayes et al., 2009). GBLUP results from the present study are in agreement with Daetwyler et al. (2010b), who reported independency of prediction accuracies on the no. of QTL.

Impact of heritability. Studying the effect of heritability on prediction accuracies and on AUROC for the 10K SNP chip, and the same no. of 290 QTL, implies a comparison of scenario S_II with scenario S_III (prediction accuracies: Table 3, AUROC: Table 4). A heritability decrease was associated with a pronounced decrease in prediction accuracies for all allocation schemes, and for both methods GBLUP and RF. For both methods GBLUP and RF, and for h^2 = 0.10, highest prediction accuracies were observed for an extremely high percentage of sick animals allocated to the training set (80%, 90%, or 100%). For h^2 = 0.30, such high accuracies were also identified for intermediate percentages of sick animals allocated to training sets, i.e. for the 50% or 60% threshold. An obvious decrease was identified for AUROC when lowering the heritability from 0.30 to 0.10 (Table 4). This decrease was more pronounced for GBLUP compared to RF applications. Similar AUROC were identified from both methods, when h^2 was 0.10 and when allocating more than 40% of sick animals to training sets. Hence, both methods GBLUP and RF revealed the same potential to distinguish between sick and healthy animals for low heritability binary traits according to AUROC, but rediction accuracies were greater for GBLUP than for RF.

The effect of heritabilities (h^2 = 0.30 and h^2 = 0.10) on accuracies of GEBV is depicted for 290

QTL and 50,025 SNP (Table 5, comparison of scenarios S_V and S_VI). As expected, heritability decrease was associated with decreased accuracies for both methods GBLUP and RF, and for all cow allocation schemes. Accuracy decrease was more obvious for RF than for GBLUP applications. Calculated AUROC for both methods GBLUP and RF, within both scenarios S_V and S_VI, are given in Table 6. Also here, apart from an extremely low percentage of sick animals in the training set (10% allocation scheme), the comparison of scenarios S_V and S_VI displayed higher AUROC for the higher heritability scenario S_V. In the low heritability scenario S_VI, AUROC values from RF were throughout higher than from GBLUP applications. For a moderate heritability of 0.30 (scenario S_V), GBLUP partly performed better, especially for low percentages of sick animals in the training set.

The detrimental impact of decreasing heritabilities on GEBV accuracies using GBLUP was proved in several previous studies (e.g., Daetwyler et al., 2010b; Zhang et al., 2010). Daetwyler et al. (2013) found increasing impact of heritabilities on prediction accuracies with an increasing no. of QTL, and they deterministically assed the prediction accuracy $= \sqrt{N_p h^2 / (N_p h^2 + M_e)}$, where N_p is the no. of animals in the training. Guo et al. (2014) identified associations between prediction accuracies and genomic heritabilities of training and testing sets. The increasing genomic heritability was due to higher genetic variations in training and testing sets, contributing to accurate predictions of marker effects.

Impact of LD structure. Scenario S_VII was simulated to evaluate the effect of LD on accuracies of GEBV and on AUROC. Remaining parameters, i.e. a heritability of 0.10 and 290 QTL, were identical with scenario S_VI. Average r^2 in the low LD scenario (S_VI) was 0.224 at distances of 0.05 cM, while the corresponding average r^2 in the high LD scenario (S_VII) was 0.425. With higher LD, accuracies of GEBV increased for all animal allocation schemes and for both methods GBLUP and RF (Table 5). Only for the extreme "10% allocation scheme" and GBLUP, the prediction accuracy was slightly higher for the low LD population (0.23 versus 0.19). Generally, gain in accuracies for the high LD scenario was more pronounced for RF compared to GBLUP applications. With regard to scenario S_VII, highest accuracies were identified for 80% and 100% of sick animals allocated to the training set when applying RF and GBLUP, respectively. A higher level of LD between SNP and QTL implies that more markers capture a higher proportion of the genetic variance of the trait (Goddard, 2009), being a prerequisite for an efficient performance of RF. Results from GBLUP were in agreement with simulations by Yin et al.(2014), who also reported increasing accuracies of GEBV with increasing LD, and with increasing marker density. Meng et al. (2009) simulated SNP with

different effects on a given disease. They identified strong impact of LD on the RF performance when simulating SNP with large genetic effects. Also Daetwyler et al. (2010b) studied the impact of genomic architecture of traits, and they reported strong LD influence on accuracies of GEBV. Results from the present study revealed a stronger impact of LD on prediction accuracies when applying RF compared to GBLUP.

Association mapping

Results from predictive models (e.g. from "genomic predictions") can be used for association mapping (e.g. Biffani et al., 2015), in general, and this holds specifically also for RF applications. The relative importance of the 10,005 SNP generated in scenario S_I by RF is shown in Figure 2a for one specific replicate. This figure 2a depicts the ten QTL explaining the highest amount of genetic variance in relation to the phenotypic variance of the trait. Generally, small distances between the most important SNP with one of the ten top QTL, were identified. For example, the second effective QTL is located on chromosome 6 at 19.64 cM, while the most important SNP identified by RF is located at 22.67 cM at the same chromosome. The short distance between most effective QTL and most important SNP indicates that RF can identify major QTL with high accuracy. Efficiency of RF to identify the most effective QTL was further improved by increasing the no. of SNP from 10K to 50K in scenario S_IV (Figure 2b). The physical distances between the most important SNP identified by RF and the ten most effective QTL decreased, and four important QTL directly overlapped with ten important SNP. For example, one of the most important QTL is located on chromosome 3 at 61.84 cM, while the most important SNP identified by RF on the same chromosome is located at 63.47 cM.

Levels of LD between ten most important SNP identified using RF in scenarios S_I and S_IV, and the most effective QTL located on the same chromosome, are listed in Table 7. For example, r^2 between QTL_78 and the most important SNP on chromosome 3 was 0.28 for the 50K SNP chip, while the highest r^2 for the 10K SNP panel was 0.02 between QTL_402 and the most important SNP at chromosome 17. Minozzi et al. (2014) simulated three different traits, and they also applied RF methodology to identify SNP associated with QTL. They reported that RF was successful in identifying the main QTL. However, in agreement with the present study, RF failed to detect significant associations for small QTL effects. We simulated our data based on purely additive effects. Magarinos et al. (2009) demonstrated that RF was one of the most powerful methods in detecting true association between marker and QTL when considering interactions among SNP.

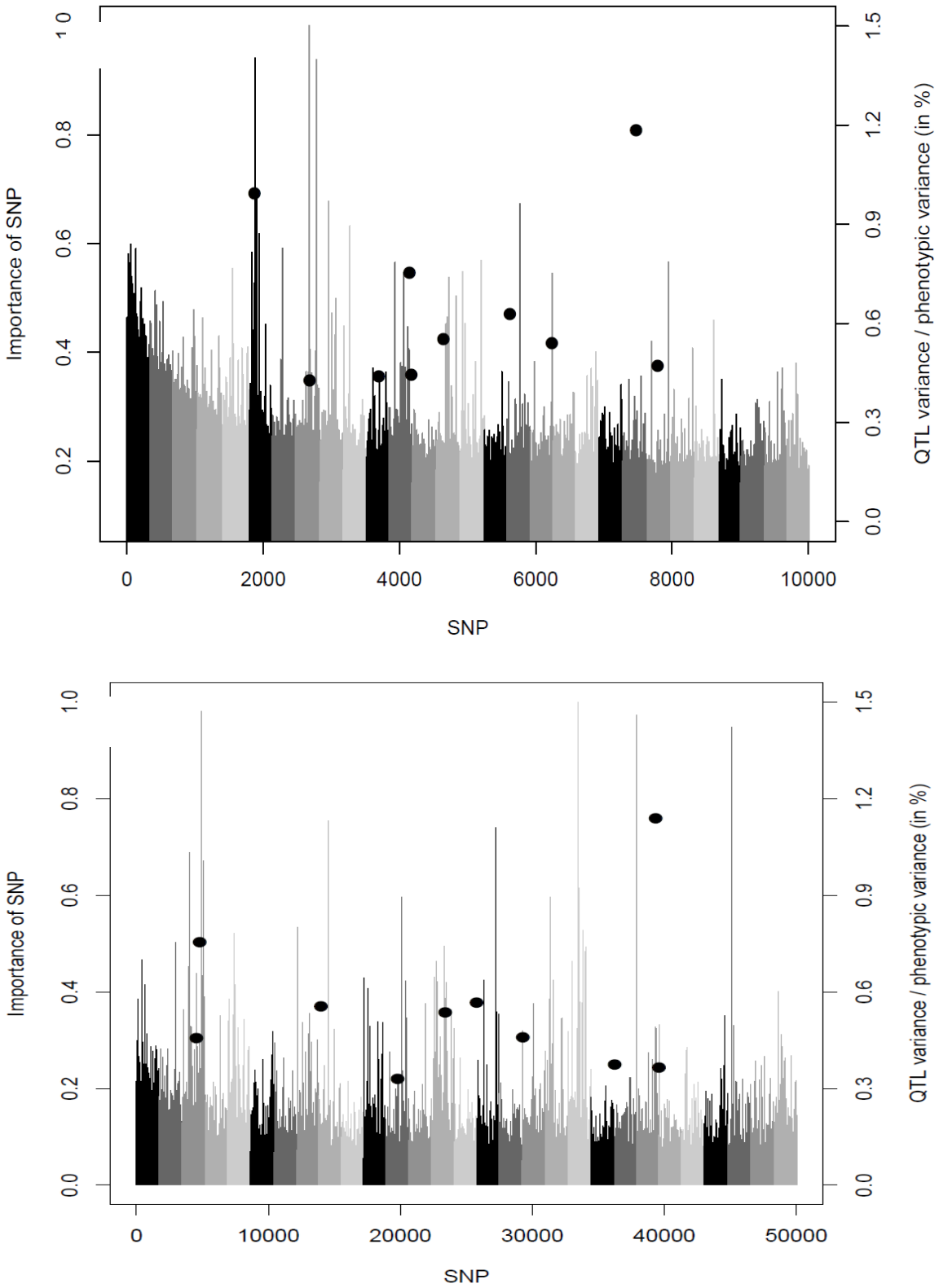

Figure 2: The relative importance of each 10K (**a**) and 50K (**b**) SNP, and positions and percentage of phenotypic variance related to ten top QTL (black circles) along 29 chromosomes.

In previous studies, RF was also successfully applied to real datasets, in order to identify

specific genes. Using RF, Li et al. (2014) confirmed the already known sheep pigmentation gene. Specifically, they identified two markers which are strongly associated with sheep coat pigmentation, and they concluded that RF is a powerful new approach for the exploration of genome wide associations. Also Nguyen et al. (2015) applied RF, and they found 25 important SNP in close distance to gene locations for Parkinson disease.

CONCLUSIONS

For both methods RF and GBLUP, the composition of training populations is one the most important factors affecting prediction accuracies. Optimal compositions of training populations imply disease incidences similar to the population disease incidence, allowing the highest prediction accuracies. The impact of genetic architecture (no. of QTL, level of LD) and of heritabilities on accuracies of GEBV is more pronounced for RF compared to GBLUP applications. The RF method is more precise than GBLUP to differentiate between healthy and sick animals (higher AUROC) for scenarios with larger marker density. Generally, prediction accuracies are higher when using the GBLUP methodology. Only for the scenario combining the highest heritability, the dense marker panel, and the largest no. of QTL, RF performs better than GBLUP. For the low density 10K SNP panel, AUROC values are quite similar for GBLUP and RF applications. With regard to whole-genome screenings, RF identifies important SNP in close distance to a QTL or a candidate gene. One limitation for the application of RF to large cow calibration groups is the demanding computation time.

Acknowledgement

The authors gratefully acknowledge funding from the BMBF and FBF, for the collaborative project "KMU-innovativ-10: Kuh-L–cow calibration groups for the implementation of selection strategies based on high-density genotyping in dairy cattle". We thank the two anonymous reviewers for their detailed and specific suggestions in order to improve our manuscript.

REFERENCES

ADR, 2014. Annual statistics of the German Cattle Breeders Federation. ISSN 1439-8745.

Albrecht, T., V. Wimmer, H. J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer, and C. C. Schön. 2011. Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 2:339–350.

Biffani, S., C. Dimauro, N. Macciotta, A. Rossoni, A. Stella, and F. Biscarini. 2015. Predicting haplotype carriers from SNP genotypes in Baus taurus through linear discriminant analysis. Genet. Sel. Evol.

47:4

Breiman, L. 2001. Random Forests. Machine Learning 45:5–32.

Buch, L. H., M. Kargo, P. Berg, J. Lassen, and A. C. Sørensen. 2012. The value of cows in reference populations for genomic selection of new functional traits. Animal 6:880–886.

Daetwyler, H. D., J. M. Hickey, J. M. Henshall, S. Dominik, B. Gredler, van der Werf, J. H. J., and B. J. Hayes. 2010a. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim. Prod. Sci. 50:1004.

Daetwyler, H. D., M. P. L., Calus, R. Pong-Wong, G., de Los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010b. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031.

de Los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345.

Edel, C., H. Schwarzenbacher, H. Hamann, S. Neuner, R. Emmerling, and K.U. Götz. 2011. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. Interbull Bulletin 44:152-156.

Efron, B., *and R.J.* Tibshirani, R. J. *1993.* An introduction to the bootstrap. *Monographs on Statistics and Applied Probability 57.* New York*:* Chapman & Hall/CRC.

Egger-Danner, C., A. Willam, C. Fuerst, H. Schwarzenbacher, and B. Fuerst-Waltl. 2012. Hot topic: Effect of breeding strategies using genomic information on fitness and health. J. Dairy Sci. 95:4600–4609.

Garcia-Magarinos, M. G., L. U. Inaki, R. Cao, and A. Salas. 2009. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Ann. Hum. Genet. 73:360–369

Gernand, E., P. Rehbein, von Borstel, U. U., and S. König. 2012. Incidences of and genetic parameters for mastitis, claw disorders, and common health traits recorded in dairy cattle contract herds. J. Dairy Sci. 95:2144–2156.

Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar, and A. Nejati-Javaremi. 2015. Predictive ability of random forest, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. Anim. Prod. Sci. Accepted.

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

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet. Sel. Evol. 43:7.

González-Recio, O., Rosa, G. J. M., and D. Gianola. 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest. Sci. 166:217–231.

Guo, Z., D. M. Tucker, C. J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang, and G. Gay. 2014. The impact of population structure on genomic prediction in stratified populations. Theor. Appl. Genet. 127:749–762.

Hayes, B. J., P.J. Bowman, A. J., Chamberlin, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433-443.

Hernandez, L. M., and D. G. Blazer. 2006. Genes, behavior, and the social environment: Moving beyond the nature/nurture debate. National Academies Press, Washington, DC.

König, S., G. Dietl, I. Raeder, and H. H. Swalve. 2005. Genetic relationships for dairy performance between large-scale and small-scale farm conditions. J. Dairy Sci. 88:4087–4096.

König, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci. 92:382–391.

König, S., K. Brügemann, and Pimentel, E. C. G. 2013. Züchterische Strategien für Tier- und Klimaschutz: Was ist möglich und was brauchen wir? Züchtungskunde 85:22–33.

Kramer, M., M. Erbe, F. R. Seefried, B. Gredler, B. Bapst, A. Bieber, and H. Simianer. 2014. Accuracy of direct genomic values for functional traits in Brown Swiss cattle. J. Dairy Sci. 97:1774–1781

Li, Y., J. Kijas, M. Henshall, S. Lehnert, R. McCulloch, and A. Reverter. 2014. Genomewide association for a dominant pigmentation gene in sheep. 10th World Congress of Genetics Applied to Livestock Production 130: 468-475.

Madsen, P., and J. Jensen. 2010. A user's guide to DMU: A package for analysing multivariate mixed models. Version 6, release 5.0. University of Aarhus, Tjele, Denmark.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, Duarte, C. W., D. B. Allison, and de Los Campos, G. 2011. Beyond missing heritability: prediction of complex traits. PLoS genetics 7:e1002051.

Mc Hugh, N., Meuwissen, T H E, A. R. Cromie, and A. K. Sonesson. 2011. Use of female information in dairy cattle genomic breeding programs. J. Dairy Sci. 94:4109–4118.

Meng, Y. A., Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta. 2009. Performance of random forest when SNPs are in linkage disequilibrium. BMC bioinformatics 10:78.

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

Minozzi, G., A. Pedretti, S. Biffani, E. L. Nicolazzi, and A. Stella. 2014. Genome wide association analysis of the 16th QTL- MAS Workshop dataset using the Random Forest machine learning approach. BMC Proc 8(Suppl 5): S4.

Neves, H. H. R., R. Carvalheiro, and S. A. Queiroz. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics 13:1-17.

Nguyen, T. T., J. Z. Huang, Q. Wu, T. Nguyen, and M. Junjie. 2015. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics 16 (Suppl 2):S5.

Ogutu, J. O., H. P. Piepho, and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc 5 Suppl 3:S11.

Pimentel, E. C., M. Wensch-Dorendorf, S. König, and H. H. Swalve. 2013. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genet. Sel. Evol. 45:12

Pryce, J. E., M. E. Goddard, H. W. Raadsma, and B. J. Hayes. 2010. Deterministic models of breeding

scheme designs that incorporate genomic selection. J. Dairy Sci. 93:5455–5466.

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

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, Ferreira, M. A R, D. Bender, J. Maller, P. Sklar, de Bakker, P. I. W., M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25:680–681.

Su, G., and P. Madsen. 2013. User's Guide for Gmatrix version 2, a program for computing Genomic relationship matrix. Available at: http://www.dmu.agrsci.dk/Gmatrix/Doc/.

Thomasen, J. R., A. C. Sørensen, M. S. Lund, and B. Guldbrandtsen. 2014. Adding cows to the reference population makes a small dairy population competitive. J. Dairy Sci. 97:5822–5832.

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

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43:1-11.

Vazquez, A. I., de Los Campos, G., Y. C. Klimentidis, G. J. M. Rosa, D. Gianola, N. Yi, and D. B. Allison. 2012. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. Genetics 192:1493–1502.

Yin, T., and S. König. 2016. Genomics for phenotype prediction and management purposes. Animal Frontiers, doi:10.2527/af2012

Yin, T., E. C. G. Pimentel, V. König, U. Borstel, and S. König. 2014. Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature × humidity-dependent covariate. J. Dairy Sci. 97:2444–2454.

Zhang, Z., J. Liu, X. Ding, P. Bijma, D. J. de Koning, and Q. Zhang. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PloS ONE 5: e12648.

3rd Chapter

Genomic breeding values, SNP effects and gene identification for disease traits in cow training sets

Naderi, S., Bohlouli, M., Yin, T., König, S.

Institute of Animal Breeding and Genetics, University of Gießen, Germany

Published in Journal of Animal genetics:
https://onlinelibrary.wiley.com/doi/full/10.1111/age.12661

SUMMARY

Holstein Friesian cow training sets were created according to disease incidences. The different datasets were used to investigate the impact of random forest (RF) and genomic BLUP (GBLUP) methodology on genomic prediction accuracies. In addition, for further verifications of some specific scenarios, single step genomic BLUP (ssGBLUP) was applied. Disease traits included the overall trait categories i) claw disorders, ii) clinical mastitis and iii) infertility from 80,741 first lactation Holstein cows kept in 58 large-scale herds. A subset of 6,744 cows was genotyped (50K SNP panel). Response variables in all scenarios were de-regressed proofs (DRP) and pre-corrected phenotypes (PCP). Initially, all sick cows were allocated to the testing set, and healthy cows represented the training set. In ongoing cow allocation schemes, the number of sick cows in the training set stepwise increased, by moving ten percent of sick cows from the testing to the training set in each step. The size of training and testing sets was kept constant, by replacing the same number of cows in the testing set with healthy cows from the training set (random selection of cows). For both methods RF and GBLUP, prediction accuracies were larger for DRP compared to PCP. For PCP as response variable, largest prediction accuracies were observed when disease incidences in training sets reflected the disease incidence in the whole population. A further increase in prediction accuracies for some selected cow allocation schemes, i.e., larger prediction accuracies compared to corresponding scenarios with RF or GBLUB, was achieved via ssGBLUP applications. Correlations between GWAS SNP effects with RF importance criteria for single SNP were in a moderate range from 0.42 to 0.57, when considering SNP from all chromosomes or from specific chromosome segments. RF identified significant SNP close to potential positional candidate genes, i.e., *GAS1*, *GPAT3*, and *CYP2R1* for clinical mastitis, *SPINK5* and *SLC26A2* for laminitis, and *FGF12* for endometritis.

Keywords: genomic predictions, random forest, genomic BLUP, disease traits

INTRODUCTION

Claw disorders, clinical mastitis and infertility are the disease categories with largest incidences in the German Holstein cow population (Gernand *et al.* 2012). Those diseases have strong influence on animal welfare (Oltenacu & Broom, 2010), and on farm economy (Hernandez *et al*. 2002; Hogeveen *et al*. 2011). However, improving disease resistance via progeny testing programs, or in

the genomic era via sire training sets, is difficult due to a small number of daughter records per sire for novel functional traits. Since only a small number of sires have daughter records for diseases traits, genomic breeding values (**GBV**) of sires have low accuracies, implying imprecise GBV validation based on sire training sets. As an alternative, Lourenco *et al.* (2015 b) suggested including genotyped and phenotyped females into a sire calibration set. Further consequence is the implementation of a training set only representing females for the estimation of GBV for low heritability disease traits (Ducrocq & Santus 2011; Naderi *et al.* 2016). For novel cow traits, usually a large number of cows are phenotyped, but only a sub-sample is genotyped. For such data, single-step genomic BLUP (**ssGBLUP**) methodology was proposed and applied (e.g., Aguilar *et al*. 2011; VanRaden, 2012). Application of ssGBLUP allows the estimation of GBV using pedigree, phenotypic and genomic information simultaneously (Legarra *et al*. 2014). Misztal *et al*. (2013) indicated superiority of ssGBLUP over multi-step methods (superiority in terms of more accurate GBV) when both phenotypic and genomic information sources are jointly available.

In analogy to the selection of sires for sire training sets, the choice of cows representing the cow training set might have substantial impact on prediction accuracies. The size of the training set and the genetic relationships between training and testing sets were the most important parameters affecting prediction accuracies for GBLUP and Bayesian applications (Habier *et al.* 2007; Habier *et al.* 2010; Albrecht *et al.* 2011; Pszczola *et al.* 2012; de Los Campos *et al.* 2013). Accuracy of genomic predictions also depends on the definition of the statistical model (Moser *et al.* 2009; Su *et al.* 2014), as well as the variable on which the effects of SNP markers do regress. The most reliable dependent variable is the true breeding value (Hayes *et al.* 2009), which only can be generated in simulation studies. For real data, possible response variables are de-regressed proofs **(DRP)**, daughter yield deviations (Gao *et al.* 2013), or pre-corrected phenotypes (**PCP**) (Fernanda *et al.* 2011; Naderi *et al.* 2016). Ostersen *et al.* (2011) reported that the response variable DRP lead to 18 to 39% larger GBV reliabilities compared to estimated breeding values **(EBV**). In contrast, Gao *et al.* (2013) favoured EBV as a response variable instead of DRP. Hence, the choose of the optimal response variable and the construction of training sets are still open questions, especially for cow training sets and novel health traits.

Generally, for novel traits and multiple linear regression model applications via BLUP and Bayesian methods, GBV are expected to be biased, because the number of genotyped individuals (n) is much smaller than the number of SNP (p). Non-parametric methods might prevail over

parametrization, with regularization. Non-parametric methods such as support vector machines, neural networks and random forest (**RF**) were applied to estimate SNP effects in genomic predictions (e.g., Moser *et al.* 2009; Naderi *et al.* 2016). Random forest (Breiman, 2001) was one of the most popular non-parametric methods for genomic selection applications when the response trait was binary (Strobl *et al.*, 2009). Using SNP sequence data for genomic predictions in maize, Sarkar *et al.* (2015) identified superiority of RF over least absolute shrinkage selection operator (**LASSO**) and ridge regression (**RR**). Botta *et al.* (2014) described the properties of RF, suggesting such statistical technique also for genome wide association studies (**GWAS**). In farm animals, RF was used to identify significant SNP being associated with disease traits, and to map QTL on the genome (Minozzi *et al.* 2014). In humans, Goldstein *et al.* (2010) applied RF to large disease datasets, and identified genes being associated with multiple sclerosis.

Up to now, only a few studies used RF methodology to discover genetic variants spread over the genome with regard to disease resistance or disease susceptibility in farm animals, specifically in dairy cattle. Furthermore, the potential of RF for genomic predictions, especially in cows training sets, needs to be explored in detail. So far, only results from an RF simulation study (Naderi *et al*. 2016) gave recommendations for the optimal composition of cow training sets. In consequence, we used a large dataset of randomly selected cows (cows from commercial herds) with genotypes and phenotypes for health traits for the following objectives: i) to compare accuracies of GBV from RF and GBLUP for clinical mastitis, claw disorders, and female infertility, ii) to study the impact of disease incidences in the cow calibration group on accuracies of GBV, iii) to compare accuracies of GBV for either DRP and PCP as response variable, iv) to compare prediction accuracies from specific scenarios with ssGBLUP results, v) to screen the whole genome in order to identify SNP with large effects via classical GWAS and via RF, and vi) to infer overlaps between identified SNP with genes as reported in publicly available gene databases.

MATERIALS AND METHODS

Data

Phenotypes. Health traits were recorded in dairy cattle contract herds located in northeast Germany, in the federal states of Mecklenburg-Westpommerania and Berlin-Brandenburg. Dairy cattle farmers applied electronically disease trait recording systems. Basis for disease trait

recordings were the official ICAR guidelines (Stock et al., 2013). These guidelines follow a hierarchical recording system, i.e., considering overall disease categories as well as specific sub-categories and single diseases. The phenotype dataset included health traits from 80,741 first lactation Holstein cows kept in 58 large-scale herds. Herd size ranged from 68 to 5,668 cows, with an average herd size of 1,392 cows, considering the calving years 2007 to 2014. The time span between diagnosis date and calving date was calculated. Only disease entries during the quite sensitive period after calving were considered, i.e., the first 200 days in milk. Repeated measurements for the same disease were ignored. At least one entry for the same disease implied a score = 1 (sick); otherwise, the score = 0 (healthy) was assigned. For genomic predictions, 15 single health disorders were grouped according to the ICAR guidelines into the following three overall disease categories: Claw disorders, clinical mastitis and female infertility. Disease trait categorization was imperative, because of extremely low disease incidences for some specific diseases. In genetic-statistical models, low disease incidences might be associated with convergence problems. In contrast, for GWAS, only a very precise and specific phenotype is of interest. Hence, we considered the specific diseases laminitis and dermatitis digitalis from the overall claw disorder category, clinical mastitis, and endometritis from the overall infertility category. Disease trait descriptions, along with the disease incidences for the whole dataset and for genotyped cows, are given in Table 1.

Genotypes. Among the 80,741 cows with phenotypes, 6,744 cows were genotyped. 2,090 cows were genotyped with the *10K Illumina Bovine Eurogenomics* SNP chip, and 4,654 cows were genotyped with the *Illumina Bovine 50K SNP BeadChip*. The low density 10K genotypes were imputed to the 50K panel by the national genetic evaluation centre, following the same procedure as applied for national official German genomic evaluations. Editing of SNP genotypes (the original 50K and the imputed 50K) excluded 1,674 SNP due to a minor allele frequency (**MAF**) lower than 0.01, or due to a large deviation from Hardy–Weinberg equilibrium ($P < 0.000001$). Cows with a call rate lower than 0.95 for all loci were discarded. After SNP-data editing, the final dataset included 43,939 SNP from 6,737 cows.

Table 1: Overview of health disorders as used in the present study along with their disease incidences (in %) in the total data set (T) and in the dataset of genotyped cows (G)

Disorder category	Sub-category of health disorder	Descriptive of health disorder	Incidence	
			T	G
Claw Disorder	White line disease	Separation between sole and wall horn	31.9	29.0
	Laminitis[1]	Aseptic infection of the corium		
	Digital phlegmon	Digital inflammatory edema		
	Sole and Toe ulceration	Infection of the corium		
	Dermatitis digitalis[1]	Exudative infection of claw		
	Interdigital dermatitis	Infectious inflammation of the skin in the clef		
Clinical Mastitis[1]	Mastitis descriptive Mastitis – etiological	Obvious infection (pain, heat, redness) Inflammation of the udder	29.3	26.2
Female Infertility	Uterine infertility[1] (Endometritis I, II, III, IV)	Infertility related to changes in the uterus (endometrial bacterial infection)	27.0	24.3
	Ovarial infertility (Silent estrus, ovarian cysts and corpus luteum persistent)	Infertility related to changes in the ovaries		

[1]Disease traits used for GWAS

Response variables

De-regressed proof. DRP and PCP for the overall trait categories claw disorders, clinical mastitis and infertility were used as dependent variables in genomic predictions. For the calculation of DRP, first the cow EBV of corresponding traits were estimated applying the following generalized linear mixed model [1] with a logit link function:

$$\boldsymbol{\eta} = \mathbf{Xb} + \mathbf{Zu} \quad [1]$$

where $\boldsymbol{\eta}$ was a vector of logits for all cows, $\mathbf{X}$ was the incidence matrix for fixed effects, $\mathbf{b}$ was the vector for fixed effects including herd and year-season of calving, $\mathbf{Z}$ was the incidence matrix for random effects, $\mathbf{u}$ was the vector for additive genetic effects, with $\mathbf{u} \sim N(0, \mathbf{A}\sigma_a^2)$, where $\mathbf{A}$ was the pedigree based relationship matrix considering all animals from the pedigree data, and σ_a^2 was the

additive genetic variance. Afterwards, the equation by Garrick *et al.* (2009) was used to calculate DRP of cows:

$$DRP_i = \frac{EBV_i}{r_i^2}$$

where DRP_i was the DRP of cow *i*, EBV_i was the EBV for cow *i*, and r_i^2 was the reliability of the EBV for cow *i* calculated as follows (described by Mrode, 2005):

$$r_i^2 = 1 - \frac{(SEP_i)^2}{\sigma_a^2}$$

where SEP_i was the standard error of the prediction for cow *i*.

Pre-corrected phenotype. RF methodology does not account for the impact of fixed effects. This was the reason for applying a 2-step strategy. First, the phenotypes of cows for disease traits were pre-corrected for fixed effects using the following logit model [2] with the same fixed effects as specified for model [1]. Model [2] was:

$$\boldsymbol{\eta} = \mathbf{Xb} \quad \textbf{[2]}$$

Second, cows were classified as sick (code = 1) or healthy (code = 0) according to solutions for residuals from model [2] as follows: Cows were ranked according to solutions for residuals, and a threshold was defined based on the disease incidence. Cows above the threshold were considered as sick, and cows below the threshold were considered as healthy. For example for claw disorders, the first 1,958 of the ranked cows were defined as sick. It was our intention to study the impact of binary data on model evaluation criteria and prediction accuracies. Hence, this was the rationale for back transforming the solutions to binary data, implying that the PCP was binary.

Estimation of genomic breeding values. For the estimation of GBV, GBLUP and RF were applied to both response variables DRP and PCP. For GBLUP, the software package DMU (Madsen & Jensen 2013) was used. The following linear mixed model [3] was defined:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wg} + \mathbf{e} \quad \textbf{[3]}$$

where **y** was a vector of DRP and PCP for all cows in the training set, **X** was the incidence matrix for fixed effects, **b** was the vector of fixed effects including only the overall mean effect, **W** was the incidence matrix of random genomic effects, **g** was a vector of GBV for all genotyped individuals, and **e** was a vector of random residuals. It was assumed that $\mathbf{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$ where σ_g^2 was the

additive genetic variance, and **G** was a marker based genomic relationship matrix. The **G**-matrix was computed based on the algorithm as proposed by VanRaden (2008), using the **G**-matrix software (Su & Madsen 2013). For binary PCP, model [3] with a logit link function was used.

In addition, single step genomic BLUP (ssGBLUP) was applied to selected scenarios. Selected scenarios where those cow allocation schemes depicting largest prediction accuracies for GBLUP. For ssGBLUP application, the software package DMU (Madsen & Jensen 2013) was used. The following logit model [4] was defined:

$$\boldsymbol{\eta = Xb + Wu} \quad \textbf{[4]}$$

where $\boldsymbol{\eta}$ was a vector of logits for all cows (non-genotyped and genotyped cows), **X** was the incidence matrix for fixed effects, **b** was the vector for fixed effects including herd and year-season of calving, **W** was the incidence matrix of random genomic effects, and **u** was the vector of single-step genomic breeding values (ssGBV). It was assumed that $\mathbf{u} \sim N(0, \boldsymbol{H}\sigma_u^2)$, where σ_u^2 was the corresponding genetic variance, and **H** was the combined relationship matrix (Christensen *et al.* 2012; Legarra *et al.* 2009). ssGBLUP methodology considers information from non-genotyped and genotyped animals by blending the pedigree relationship matrix (**A**- matrix) and genomic relationship matrix (**G**- matrix) into an **H**- matrix. The inverse of **H** was constructed according to Christensen & Lund (2010) and Aguilar *et al.* (2010):

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

where $\mathbf{A^{-1}}$ was the inverse of the pedigree-based relationship matrix, $\mathbf{A_{22}^{-1}}$ was the inverse of the sub-matrix of the pedigree-based relationship matrix for genotyped animals, and $\mathbf{G}_{\boldsymbol{w}}$ was calculated as follow:

$$\mathbf{G}_{\boldsymbol{w}} = (0.95\times \mathbf{G} + 0.05 \times \mathbf{A_{22}})$$

where **G** was the genomic relationship matrix.

For the estimation of GBV via RF methodology, the java package RanFoG (Gonzalez-Recio & Forni 2011) was applied. Theoretical background in this regard is given by Naderi *et al.* (2016).

Scenarios for creating training and testing sets. First, all sick cows (i.e. 1,958 for claw disorders, 1,768 for clinical mastitis, and 1,645 for infertility) were allocated to the testing set. Vice versa, the healthy cows (i.e., 4,779 for claw disorders, 4,969 for clinical mastitis, and 5,092 for infertility)

were considered as training set. To allocate different numbers of sick cows to the training set, the number of sick cows in the training set gradually increased by moving 10% of the sick cows (selected randomly) from the testing set to the training set. The size of training and of testing sets were kept constant by moving the same number of healthy cows (selected randomly) from the training to the testing set. The cow allocation procedure was done stepwise in increments of 10% of the sick cows, until all of the sick cows were moved to the training set. Consequently, and in the final step, all cows in the testing set were healthy, and all sick cows were allocated to the training set. For all cow allocation scenarios, four replicates were carried out.

Assessing the prediction performance. GBV of cows in the testing set were predicted based on the realized genomic relationship matrix calculated from the 6,737 genotyped cows. For the response variable DRP, the average correlation coefficient from the four replicates between DRP and predicted GBV from cows in the testing set was the evaluation criterion "prediction accuracy of GBV" ($\mathbf{r_{GBV}}$). Furthermore, a corrected prediction accuracy according to Wellmann *et al.* (2013) ($\mathbf{r_{GBV\text{-}cor}}$) was calculated:

$$r_{GBV\text{-}cor} = r_{GBV} - a_1 (r_{EBV}^{TES} - 1)$$

where r_{GBV} was the correlation between GBV and DRP of one replicate, r_{EBV}^{TES} was the mean accuracy of EBV in the testing set of the corresponding replicate, and a_1 was an estimated regression coefficient based on the following equation:

$$r_{GBV} = a_0 + a_1 r_{EBV}^{TES} + a_2 r_{EBV}^{TRS} + e$$

where a_0, a_1 and a_2 were intercept and regression coefficients, respectively, r_{EBV}^{TRS} was the mean accuracy of EBV in the training set of a given replicate, and *e* was the residual component.

For PCP as response variable, prediction accuracy of GBV ($\mathbf{r_{GBV\text{-}PCP}}$) from the four replicates was calculated according to the following equation (as applied by Gerardo *et al.* 2016):

$$r_{GBV\text{-}PCP} = \frac{cor(GBV, y)}{\sqrt{h^2}}$$

where $r_{GBV\text{-PCP}}$ was the accuracy of GBV, y was the pre-corrected phenotype and h^2 was the heritability for the trait of interest.

Genome wide association study. SNP-effects from a GWAS for laminitis, dermatitis digitalis, clinical mastitis and endometritis were estimated in GCTA (Yang *et al*. 2011). We defined a linear mixed model, but for binary input data, the program transforms the estimates to a liability scale (Lee et al., 2011). The SNP model [5] was:

$$\mathbf{y = Xb + Wa + Zu + e \ [5]}$$

where **y** was the observation vector for health traits; **b** was the vector of the fixed effects (herd and year-season of calving), **X** was an incidence matrix for the fixed effects; **a** was the vector for the additive allele substitution effects of the candidate SNP to be tested for the association (a fixed regression); **W** was the design matrix for SNP genotypes coded as 0, 1 or 2; **u** was the vector for additive polygenic effects. It was assumed that $\mathbf{u} \sim N(0, \boldsymbol{G}\sigma_u^2)$, where σ_u^2 was the corresponding genetic variance and **G** was a marker based genomic relationship matrix, with the corresponding incidence matrix **Z**; and **e** was the vector of random residual effects. The genomic relationship matrix considered all SNP except the chromosome on which the candidate SNP is located. *P*-values $\leq 5 \times 10^{-5}$, and the corrected false discovery rate (**FDR**, Benjamini & Hochberg (1995)) < 10 % were used to identify significant associations between single SNP with disease traits. For the calculation of FDR, all *P*-values from the smallest to the largest were ranked. Afterwards, *P*-values were adjusted:

$$\text{FDR} = \frac{\text{Number of tests}}{\text{P rank}} \times \text{P}$$

where the number of tests is equal to number of SNP.

Association studies between single SNP with disease traits via RF were performed using the R package *Ranger* (Wright *et al*. 2017). Theoretical background addressing details for the calculation of SNP importance is given by Naderi *et al.* (2016). In brief, they specified that the RF algorithm utilizes two third of the data to construct a decision tree. Individuals not appearing in any sample are defined as "out of bag (**OOB**)". Based on OOB data, the prediction accuracy from the OOB sample of a given tree in relation to the same OOB sample, but with a permuted genotype for a given SNP, is measured by using the misclassification rate. The importance of a given SNP is defined as the percentage in prediction accuracy increase from OOB of the given tree in relation to the permuted OOB. The relative importance of a single SNP was defined as the ratio of the importance of a specific SNP in relation to the most important SNP. Hence, SNP values for the

relative importance were in the range from 0 to 1 (most important SNP: value of 1). SNP importance values from RF were correlated with corresponding SNP GWAS effects for specific chromosomes, and for specific chromosome segments.

Gene annotation and functional gene analysis. Annotated genes in a window of 500 kbp downstream and upstream of each significant SNP were identified from the Ensemble database (www.ensembl.org/biomart/martview). For the specification of gene functions, the GeneCards database was used (www.genecards.org). Furthermore, we studied the InnateDB database. InnateDB is a publicly available database including genes, proteins and signaling pathways, especially addressing immune response.

RESULTS AND DISCUSSION

Accuracy of genomic predictions

The response variable DRP. Prediction accuracies (r_{GBV}) from GBLUP and RF applications from four replicates with different proportions of sick cows in the training set are shown in Figure 1A for claw disorders, in Figure 1B for clinical mastitis, and in Figure 1C for infertility. GBLUP accuracies outperformed those from RF for all cow allocation schemes. Largest accuracies were observed when a larger number of sick cows were in the training set, but effects were only minor for GBLUP scenarios. However, for RF, increasing the number of sick cows in the training population was associated with a continuous increase in prediction accuracies. As one specific example, r_{GBV} for claw disorders from GBLUP and RF were 0.85 and 0.65, respectively, when only 10% of sick animals were allocated to the training set, while the highest accuracy was 0.86 for GBLUP and 0.73 for RF when 100% of sick animals were assigned to the training set (Figure 1A). For GBLUP, prediction accuracies for claw disorders and clinical mastitis were larger compared to those for infertility. Prediction accuracies from GBLUP ranged from 0.84 to 0.86 for claw disorders, from 0.83 to 0.86 for clinical mastitis, and from 0.79 to 0.82 for infertility. For RF, prediction accuracies for claw disorders were throughout higher (0.65 to 0.73) than those for the other two diseases, i.e. 0.64 to 0.71 for clinical mastitis, and 0.64 to 0.68 for infertility. Results indicate that both components disease incidences in training sets and disease trait heritabilities had impact on accuracies of genomic predictions. The heritability from model [1] was 0.069 for claw disorders, 0.056 for clinical mastitis and 0.051 for infertility. In addition, the highest disease

incidence was identified for claw disorders. The beneficial impact of increasing heritabilities on GBV accuracies for both methods GBLUP and RF is in agreement with results from simulation studies (e.g., Zhang *et al.* 2010; Naderi *et al.* 2016).

However, prediction accuracies from the present study were quite large, also when comparing to the results from previous studies. For example, Erbe *et al.* (2012) reported correlations between GBV and daughter yield deviations in the range from 0.56 to 0.58 for moderate heritability milk production traits, based on 2,257 genotyped Australian Holstein bulls. Kramer *et al.* (2014) estimated correlations between GBV and DRP in the range from 0.63 to 0.74 for low heritability functional traits. Data basis in this study was a cow calibration set including only 1,126 Brown Swiss dairy cows from Switzerland. Larger correlations in our study might be due to the substantially larger number of cows in the training set. Increasing the size of training populations is associated with an increase of related animals, leading to larger prediction accuracies due to a better exploitation of linkage disequilibrium (**LD**) (VanRaden, 2008; Habier *et al.* 2010).

Overestimated GBV might be another reason for large accuracies, because the complete pedigree was used to estimate breeding values. Utilization of the full pedigree implies that EBV of animals in the training and testing sets were from the same genetic evaluation, with possible impact on genomic prediction accuracies (Wellmann *et al.* 2013). Such impact was more obvious for low heritability traits, and especially when genotyped animals had only a few progeny (Amer & Banos 2010). One strategy to correct GBV accuracies was applied by Kramer *et al*. (2014). They divided the correlations between GBV and DRP by the EBV reliability. Nevertheless, in the present study and for most of the allocation schemes, those calculated accuracies were larger than one. This was mainly due to low EBV reliabilities for the low heritability traits. As a further possibility, the regression approach by Wellmann *et al.* (2013) was applied, with the regression coefficients $\hat{a}_0$ = 2.02, $\hat{a}_1$ = -0.39 and $\hat{a}_2$ = -2.17. Prediction accuracies from the "Wellmann-equation" ($r_{GBV\text{-}cor}$) are also depicted in Figure1A for claw disorders, in Figure 1B for clinical mastitis, and in Figure 1C for infertility. Those accuracies were in the range as indicated in previous studies (e.g. Su *et al*. 2014). However, the outperformance of GBLUP over RF preserved for all allocating schemes, and for all the three disease traits. For instance, using GBLUP, GBV accuracies for claw disorders ranged from 0.62 to 0.65, but were lower for RF in a range from 0.45 and 0.53. Again, accuracies from GBLUP and RF increased when adding more sick cows to the training population. Again, this effect was more pronounced for RF. In agreement with r_{GBV}, largest $r_{GBV\text{-}cor}$ were found for claw disorders (for

both methods GBLUP and RF), and $r_{GBV\text{-}cor}$ were lowest for infertility. Infertility was the trait with lowest reliabilities for pedigree based EBV, and the trait with lowest disease incidences, indicating impact of these parameters on $r_{GBV\text{-}cor}$. The reliability of EBV from the pedigree based relationship matrix (model [1]) was 0.47 for claw disorders, 0.43 for clinical mastitis, and 0.41for infertility. The effect of the EBV reliability on accuracies of genomic predictions was explained by, e.g., Hidalgo *et al.* (2015). They showed that the genomic prediction bias was reduced, and that genomic prediction accuracies increased, when removing less reliable EBV from the input dataset.

Using the regression approach for accuracy corrections, results from the current study for mastitis and infertility were in line with estimates by Su *et al*. (2014). Su *et al*. (2014) also used a quite large dataset of genotyped animals, with a reference set of ~4000 and a validation set of ~1150 animals. Gao *et al.* (2013) used a smaller dataset including ~3000 genotyped animals in the reference set and ~1000 animals in validation set. In consequence, their genomic prediction reliabilities were smaller, i.e., 0.29, 0.23 and 0.29 for claw disorders, mastitis and infertility, respectively.

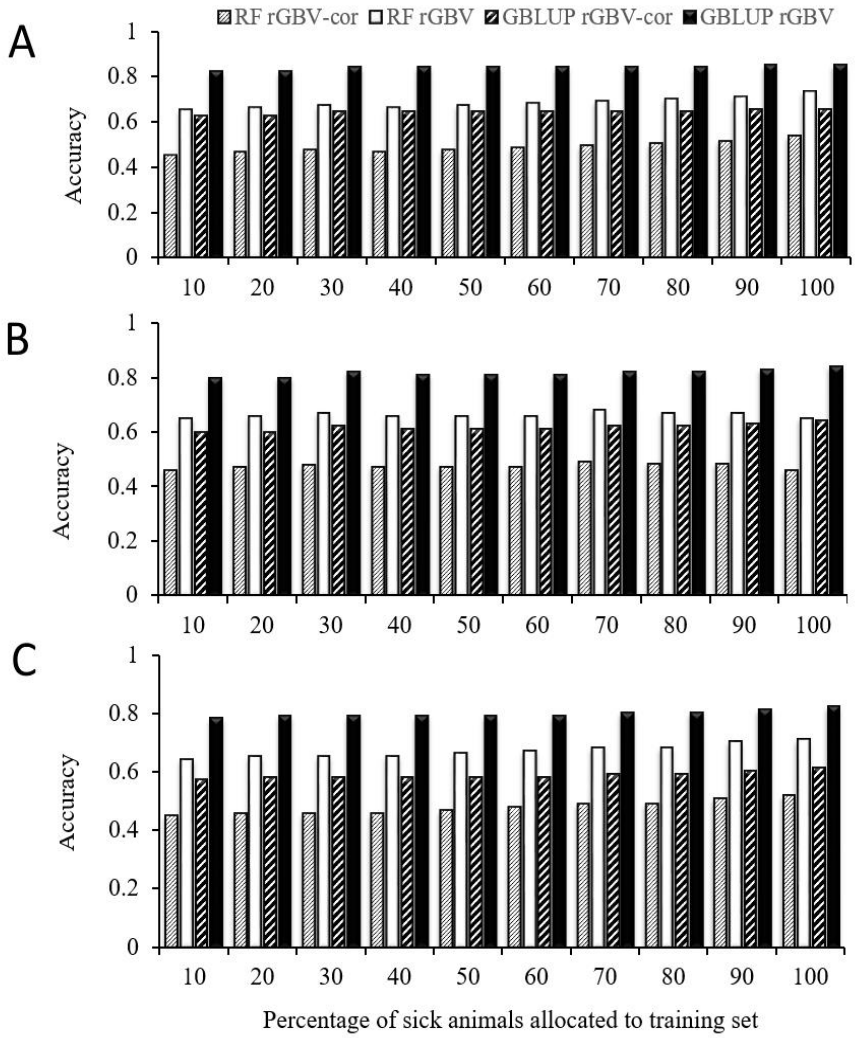

Figure 1: Correlation between de-regressed proofs and genomic breeding values (r_{GBV}), and the corrected prediction accuracy using the "Wellmann-equation" ($r_{GBV\text{-}cor}$) for claw disorders (**A**), for clinical mastitis (**B**) and for infertility (**C**) from random forest (RF) and genomic BLUP (GBLUP)

The response variable PCP. For PCP as response variable, calculated accuracies ($r_{GBV\text{-}PCP}$) (Figure 2A for claw disorders, Figure 2B for clinical mastitis, and Figure 2C for infertility) were lower than corresponding accuracies based on DRP. One explanation addresses the low heritability health traits in this study. Gerardo *et al.* (2016) also reported larger accuracies for the response variable EBV compared to pre-adjusted phenotypes for low heritability traits. Vice versa, they identified opposite results for high heritability traits. The main explanation for different accuracies when basing genomic predictions on either PCP or DRP addresses the variety of suggested specific equations (see materials and methods). For example for PCP, when dividing the correlation between GBV and

phenotypes by the square root of the heritability, a heritability estimate close to zero (as being the case for health traits) has substantial impact on the prediction accuracy increase. Furthermore, differences in "signal-to-noise ratios" (Daetwyler *et al.* 2013) of two distinct-quantity response variables PCP and DRP might contribute to differences inn realized accuracies.

For PCP as response variable, prediction accuracies ($r_{GBV\text{-}PCP}$) from GBLUP outperformed those from RF (plus 10 to 70% accuracy increase) for all disease traits and for all cow allocation schemes (Figure 2A for claw disorders, Figure 2B for clinical mastitis, and Figure 2C for infertility). For a larger percentage of sick cows allocated to the training set, better prediction accuracies were observed for RF. For example, for the "80% allocation scheme", the prediction accuracy for clinical mastitis was 0.39 from RF, and 0.37 from GBLUP. Accordingly, for infertility, accuracies were 0.42 (RF) and 0.39 (GBLUP), and 0.43 (RF) and 0.40 (GLUP) for claw disorders.

For PCP, an increase of the percentage of sick cows in the training set was associated with a substantial increase in prediction accuracies ($r_{GBV\text{-}PCP}$) for both methods RF and GBLUP (Figure 2A for claw disorders, Figure 2B for clinical mastitis, and Figure 2C for infertility). For DRP, only minor impact of disease incidence increases on prediction accuracies (r_{GBV}) was observed. Using PCP as response variable, $r_{GBV\text{-}PCP}$ from both methods was largest when disease incidences in training sets reflected the disease incidence in the whole population (Figure 2A for claw disorders, Figure 2B for clinical mastitis, and Figure 2C for infertility). For example for clinical mastitis, allocating 70% of the sick cows to the training set (implies a disease incidence of 24.8% in the training set) was associated with the best genomic prediction from GBLUP ($r_{GBV\text{-}PCP}$ = 0.39). The disease incidence for clinical mastitis considering genotyped cows was on a quite similar level (26.2%). One explanation for different reactions of DRP and PCP on the increased number of sick cows in training sets addresses the different distributions of response variables. For PCP as response variable, the distribution is binomial. For RF and a binomial response variable, Naderi *et al*. (2016) identified more informative trees with increasing disease incidences, leading to an increase in prediction accuracies.

Single step GBLUP methodology. Applying ssGBLUP for specific allocation schemes was associated with a further increase in prediction accuracies for all traits. The largest prediction accuracy for the "50% allocation scheme" and for claw disorders was 0.76 for ssGBLUP, and slightly smaller for GBLUP with 0.74 (Figure 2A). For clinical mastitis and the "70% allocation

scheme", prediction accuracies from ssGBLUP were considerably larger than from GBLUP (Figure 2B). For female infertility and the "60% allocation scheme", prediction accuracy differences between ssGBLUP and GBLUP were very small, i.e. 0.01 (Figure 2C). In agreement with our results, several previous studies reported the superiority of ssGBLUP over GBLUP methodology for datasets including a large number of animals with phenotypes, but a smaller number of animals with genotypes (Gao *et al.* 2012; Lourenco *et al.* 2015a, Ashraf *et al.* 2016). Aguilar *et al*. (2010) indicated that ssGBLUP is a simple and quick alternative to estimate GBV when both phenotypes and genotypes are jointly available. They argued that automatically defined weighting factors for the different information sources are the main advantage of ssGBLUP. In a similar context, Legarra *et al.* (2014) indicated the properly weighting of various information sources, in order to avoid double-counting of contributions due to genetic relationships and phenotypic records.

Genome wide associations for disease traits

Clinical mastitis. Applying MLM, we identified four SNP on chromosomes 8, 26, 15 and 6 which significantly contributed to clinical mastitis ($P < 5 \times 10^{-5}$ and FDR < 10%) (Figure 3). The chromosome number, name and position of SNP, FDR, -log10 *P*-values and annotated genes within 500 Kb up and downstream of the given SNP, are listed in Table 2. Interestingly, applying RF without using a genomic or additive-genetic relationship matrix, confirmed the highly significant SNP BTB-01737838 on chromosome 8. This SNP located at 81.63 cM was the most important SNP as detected by RF. The correlation coefficient between the SNP effects from GWAS with the RF importance criterion for single SNP, including all SNP from chromosome 8, was 0.52. RF identified additional four important SNP for clinical mastitis on chromosome 6 at 71.5 cM and 72.5 cM, and on chromosome 7 at 47.1 cM and at 63.8 cM (Figure 3). The correlation coefficient between the GWAS SNP effects with the RF importance criterion, including all SNP from chromosome 6, was 0.51, and 0.48 for chromosome 7.

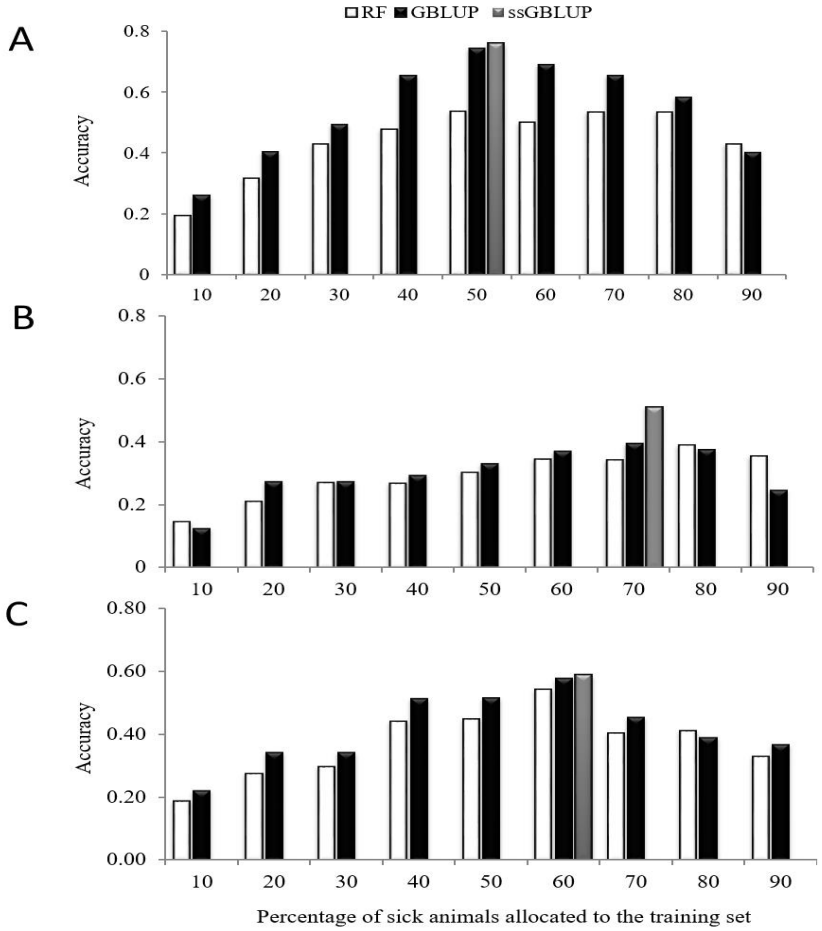

Figure 2: Prediction accuracy ($r_{GBV\text{-}PCP}$) for claw disorders (**A**), for clinical mastitis (**B**) and for infertility (**C**) from random forest (RF), genomic BLUP (GBLUP) and single step genomic BLUP (ssGBLUP)

A wide range of strong association signals for clinical mastitis was reported in previous studies. For example, Klungland *et al.* (2001) reported a QTL on chromosome 8 associated with clinical mastitis within the distance of 47.08-84.84 cM, being also the segment where the most significant SNP from the current study was located. On the data basis of 673 Holstein cows reared in tropical conditions, Iung *et al.* (2014) identified four SNP located on BTA 6, 8, 14 and 15 being associated with somatic cell score (SCS). Schnabel *et al.* (2005) found a QTL for SCS at 81.3 cM on chromosome 8, in close distance to the significant SNP from the current study.

The most important SNP detected by RF on chromosome 6 confirmed results by Klungland *et al.* (2001), who found a QTL affecting clinical mastitis within a segment of 35.39-70.74 cM on chromosome 6. In addition, Daetwyler *et al.* (2008) identified a QTL for somatic cell count on chromosome 6 at 72 cM. Sahana *et al.* (2014) reported a highly significant association signal for clinical mastitis on chromosome 6 at 88.97 cM in Holstein cattle. Sodeland *et al.* (2011) reported a QTL in association with clinical mastitis within a segment of 55.84-64.08 cM on chromosome 7, close to the important SNP ARS-BFGL-NGS-41589 as identified with RF at 63.84 cM.

According to the Ensemble database, the chromosome segment on chromosome 8, where the SNP BTB-01737838 is located, includes the *GAS1*-gene in a distance of 126 kb. *GAS1* contributes to abnormal gland morphology and physiology: Expression of *GAS1* caused the death of cells in the mammary gland, and prevented cell cycle progression (Jaggi *et al.* 1996). The chromosome segment for the SNP ARS-BFGL-NGS-63987 SNP on chromosome 6 (the most important SNP identified by RF) also includes glycerol-3-phosphate acyltransferase 3 (*GPAT3*). *GPAT3* produces a protein variant being involved in the synthesis of triacylglycerol in the bovine mammary gland (Bionaz & Loor 2008). Fatty acids affect inflammatory responses to common infectious diseases, such as mastitis and metritis.

The chromosome segment for BTA-36568-no-rs on chromosome 15 includes the *CYP2R1*-gene in a distance of 321kb. This gene catalyzes the synthesis of cholesterol, steroids and other lipids, and converts vitamin D into the active ligand for the vitamin D receptor (Breuer *et al*. 2013). An inherited mutation in the *CYP2R1* gene was associated with symptoms of vitamin D deficiency (Cheng *et al.* 2004). The relationship between vitamin D3 levels in the mammary glands and the strength of immune system mechanisms against udder infections, due to *Streptococcus*, was proved by Lippolis *et al.* (2011). Furthermore, according to InnateDB, the positional candidate genes for clinical mastitis, i.e., *GAS1*, *PDE3B*, *CYP2R1*, *INSC* are expressed in the fat pad of the mammary gland. Mammary fat pad supports the rudimentary structure of the bovine mammary gland (Sheffield 1988).

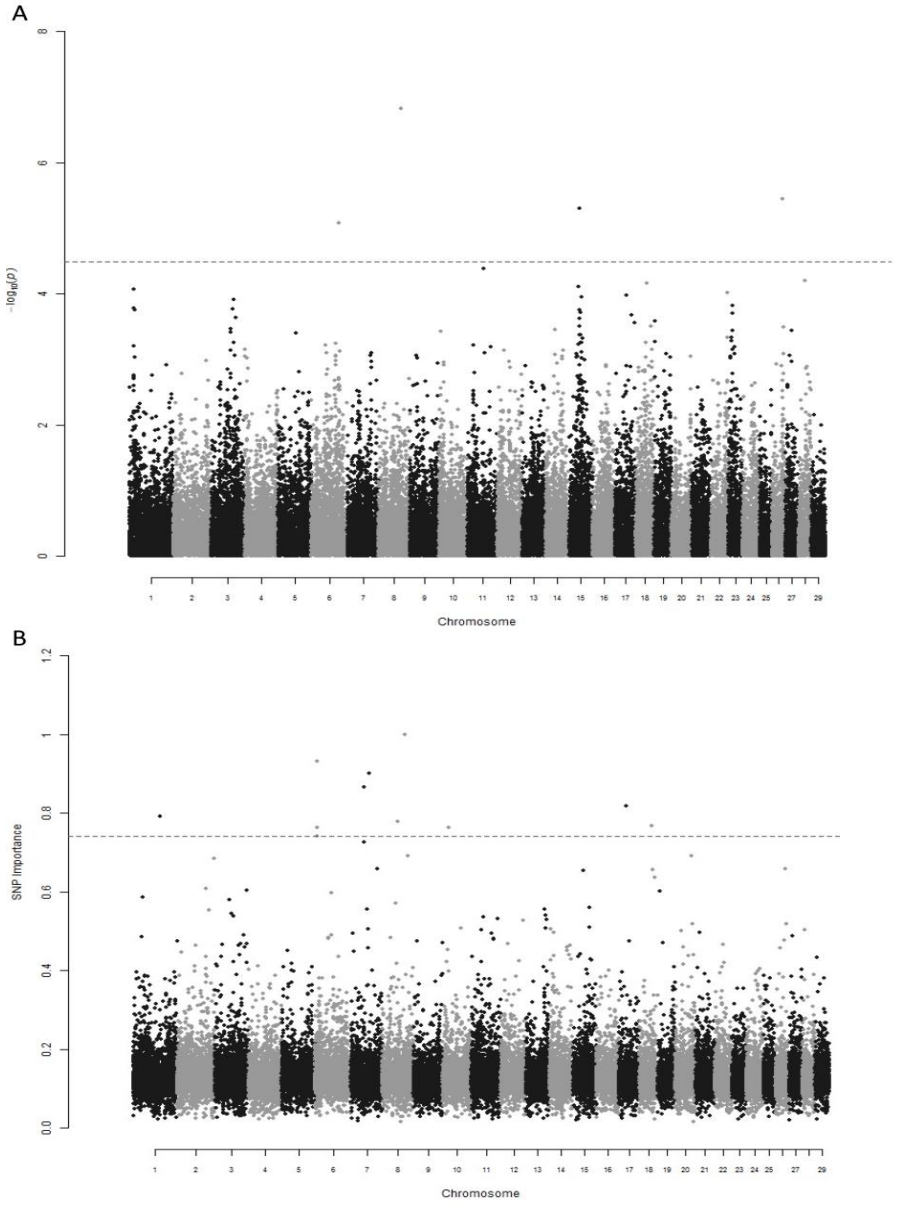

Figure 3: Manhattan plots for -log10 (p) from genome wide association studies (**A**) and SNP importance from RF (**B**) for clinical mastitis.

Claw disorders

Laminitis. Results from GWAS for laminitis are shown in Figure 4. Five significant SNP with a FDR lower than 10% and $P < 5 \times 10^{-5}$ were detected based on GWAS (Table 2). Significant SNP were located on chromosomes 7, 11, 5 and 4. ARS-BFGL-NGS-10231 and ARS-BFGL-NGS-114992 were two significant SNP on chromosome 7 at 63.2 cM and at 61.6 cM, respectively. BTB-

00466773 was a significant SNP located on chromosome 11 at 23.3 cM. Hapmap50590-BTA-120045 and Hapmap38318-BTA-45002 were two significant SNP on chromosome 5 and 4 at 79.6 cM and at 33.9 cM, respectively. RF identified the most important SNP for laminitis on chromosome 23 at 45.5 cM.

Interestingly, four of the significant SNP detected by GWAS were among the top ten SNP as identified via RF. The correlation coefficient between the SNP effects from GWAS with the RF importance criterion for single SNP, including all SNP from chromosome 7, was 0.53. The correlation between estimates from both methods including all SNP from all chromosomes was 0.47.

The positional candidate genes for laminitis within a 500 kbp window are listed in Table 2. So far, no apparent functions of most of these genes related to laminitis, were reported. Nevertheless, Serine Peptidase Inhibitor Kazal Type 5 (*SPINK5*) contributes to the inflammation of muscles and the skin. Speculating on the possible impact of claw disorders, the serine protease inhibitor is an important gene regulating the anti-inflammatory or antimicrobial protection of mucous epithelia, and protective barrier functions of the skin. The Solute Carrier Family 26 Member 2 (*SLC26A2*) gene was associated with severe disorders in arms and legs, and affected the development of bones. In addition, Wang (2010) identified Solute Carrier Family 26 Member 4 (*SLC26A4*) gene expressions in the laminar tissue of horses, with possible impact on hoof health. According to InnateDB, *HMGXB3* and *CSF1R* are potential positional candidate genes, which are intensively in the lymph gland (Breuer *et al.* 2013). In such physiological context, also the potential positional candidate genes *IPO8, KIAA1324L* and *GRM3* play a minor role. Shearer (1998) reported that lymph accumulation was associated with increased pressure, pain and foot tissue damage.

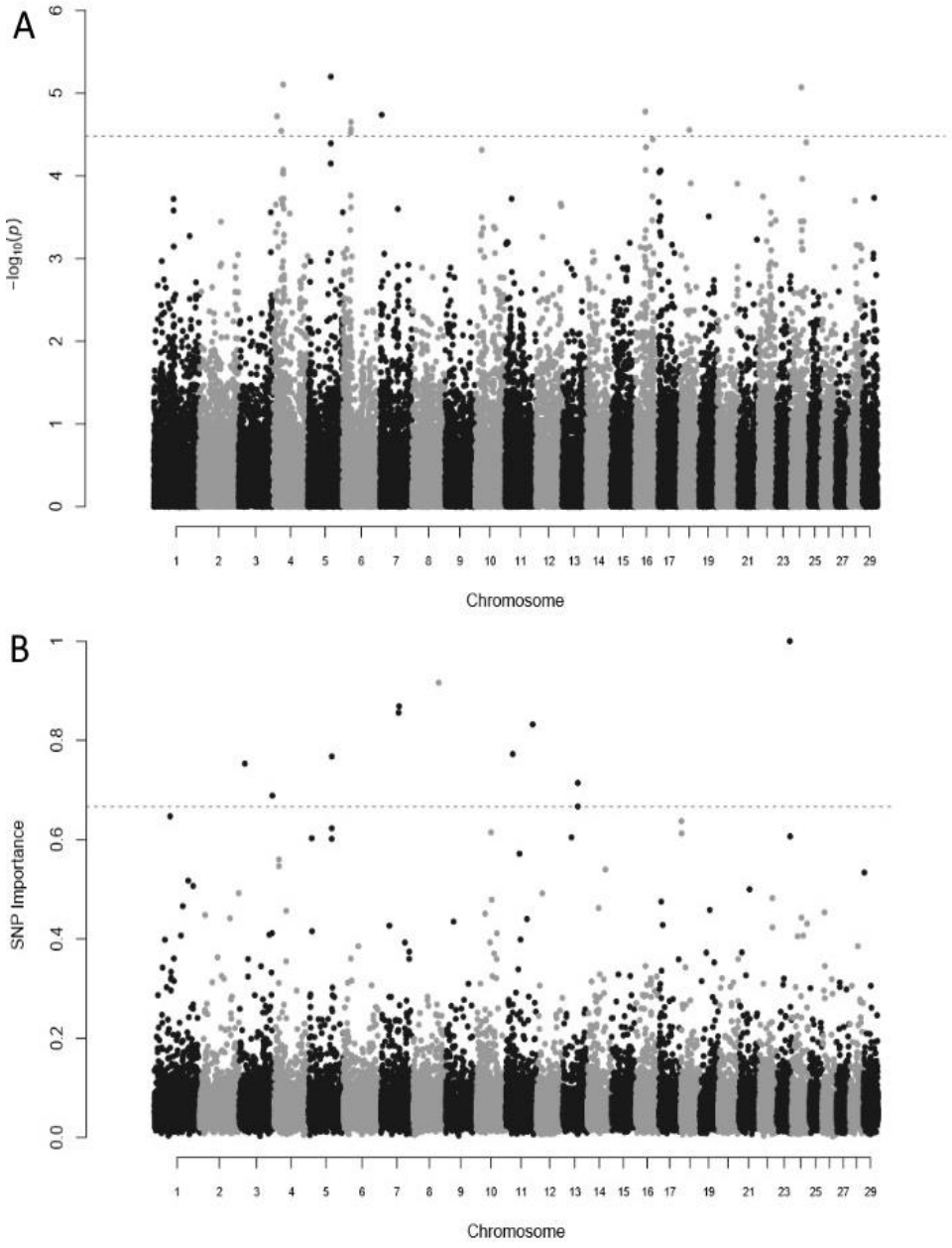

Figure 4: Manhattan plots for -log10 (p) from genome wide association studies (**A**) and SNP importance from RF (**B**) for laminitis.

Table 2: SNP associations for laminitis, dermatitis digitalis, clinical mastitis and endometritis with false discovery rate (FDR) ≤ 10% from GWAS, the most important SNP variable from RF (indicated with VIM = 1), and annotated genes within 500 Kb up and downstream of the given SNP

Disease	Chr	SNP	Position (bp)	FDR (%)	$-\log_{10}$ P / VIM	Annotated Gene
Laminitis	7	ARS-BFGL-NGS-10231*	63215249	0.76	6.76	*PPARGC1B, PDE6A, SLC26A2, HMGXB3, CSF1R*
	11	BTB-00466773*	23340880	0.41	6.72	-
	7	ARS-BFGL-NGS-114992*	61601418	0.61	6.38	*SPINK5, SPINK6, SPINK7*
	5	Hapmap50590-BTA-120045	79695650	6.9	5.19	*CAPRIN2, IPO8, TMTC1*
	4	Hapmap38318-BTA-45002	33987589	6.9	5.10	KIAA1324L, GRM3
	23	BTA-107511-no-rs**	45530763	-	1	*SMIM13, ELOVL2*
Dermatitis digitalis	3	Hapmap60335-rs29018229* **	41634074	1.5	6.46 / 1	*OLFM3*
	9	BTB-01594395*	65093167	1.2	6.26	*U6*
	9	BTB-01704243	64583261	1.5	5.99	*SYNCRIP, SNX14, SNORD50, U6*
Clinical mastitis	8	BTB-01737838* **	81638162	0.6	6.82 / 1	*GAS1, SNORA70*
	26	Hapmap50053-BTA-61516	38980475	7.9	5.44	*FAM204A*
	15	BTA-36568-no-rs	38099059	7.4	5.29	*INSC, CALCA, CYP2R1, PDE3B*
	6	ARS-BFGL-NGS-63987	10029854	9.1	5.08	*GPAT3, SNORA69, U6*
Endo-metritis	21	BTB-00803496	10171581	0.48	6.95	-
	1	BTA-28763-no-rs*	76331663	7.4	5.47	*FGF12, UTS2B, OSTN*
	2	Hapmap50978-BTA-48134**	81053048	-	1	*NABP1*

*identified SNP via both methods (GWAS and RF)

**the most important SNP identified via RF (VIM = 1)

Dermatitis Digitalis. Using GWAS, three significant SNP with FDR lower than 10% were identified (Figure 5).

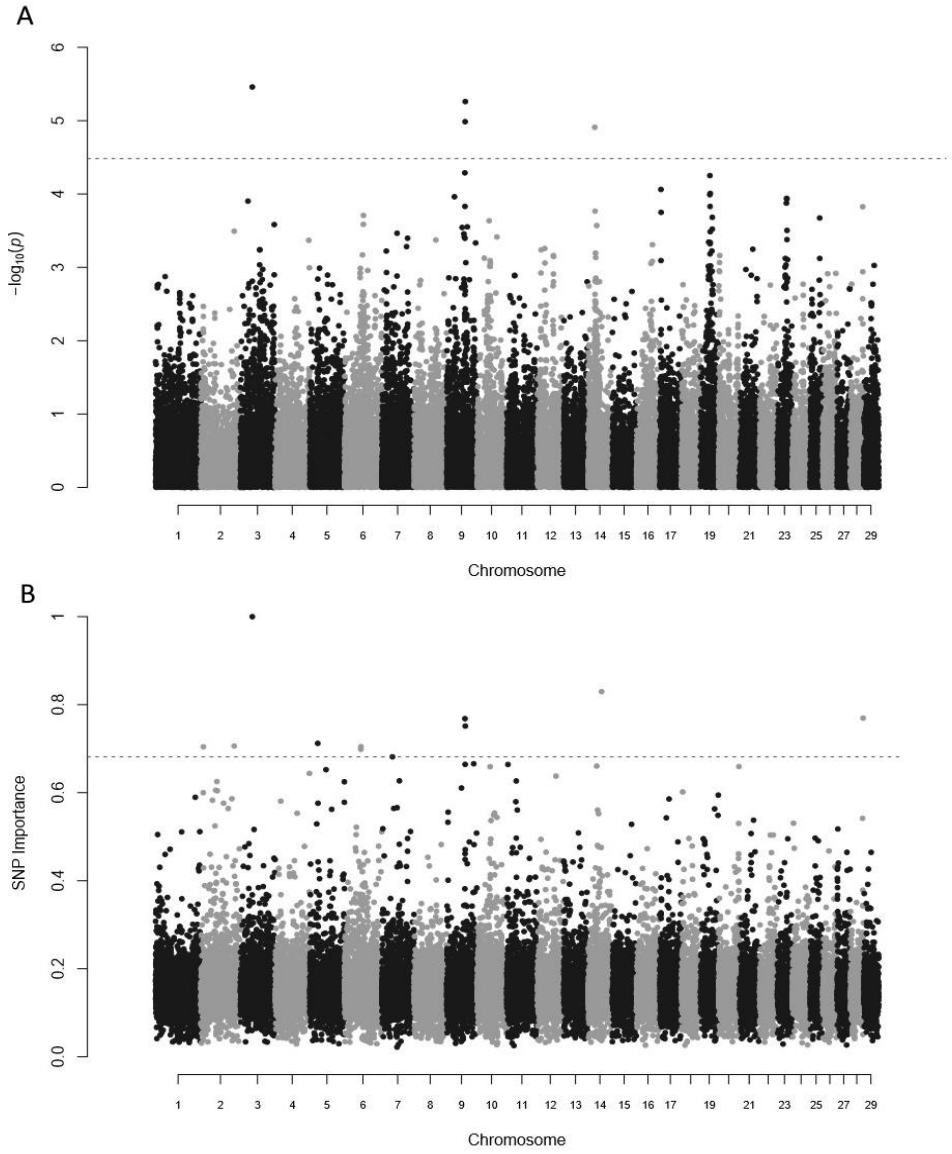

Figure 5: Manhattan plots for -log10 (p) from genome wide association studies (**A**) and SNP importance from RF (**B**) for dermatitis digitalis.

The chromosome number, name and position of SNP, FDR, -log10 P-values, and annotated genes within 500 Kb up and downstream of a given SNP are listed in Table 2. Hapmap60335-rs29018229, BTB-01594395 and BTB-01704243 are SNP located on chromosome 3 at 41.6 cM, and on chromosome 9 at 65.0 cM and 64.5 cM, respectively. The most important SNP identified by RF was

the most significant SNP from GWAS. In addition, BTB-01594395 on chromosome 9 at 65.0 cM was among the most important SNP as identified with RF (Figure 5). The correlation coefficient between SNP effects from GWAS with the RF importance criterion for single SNP, including all SNP from chromosome 3, was 0.42. The correlation was 0.47 for the SNP from chromosome 9.

To our knowledge, only a few GWAS addressed claw disorders in cattle, with a focus on laminitis and dermatitis digitalis. Van der Spek *et al.* (2015) identified three significant SNP for the claw disorder interdigital hyperplasia on chromosome 7, close to significant SNP from the current study. In total, Van der Spek *et al.* (2015) identified significant SNP on 20 different chromosomes, and they concluded that claw disorders are affected by many genes distributed across the whole genome, each of them explaining a small amount of the genetic variance. Using whole-genome sequence data, Wu *et al*. (2016) listed 49 significant SNP being associated with feet and leg disorders, located on chromosome 7 within 63 to 65 cM. Wu *et al.* (2013) found 12 SNP being significantly associated with feet and legs traits in Chinese Holstein cattle. Two of the identified SNP were located on chromosome 3 at 48.2 cM and 14.2 cM.

Endometritis. Results from GWAS for endometritis are shown in Figure 6. Using GWAS, two significant SNP with FDR lower than 10% were identified. BTB-00803496 and BTA-28763-no-rs were two significant SNP located on chromosome 21 and 1 at10.17 cM and at 76.3 cM, respectively. The most two important SNP identified via RF were located on chromosome 2 at 81.0 cM and at 80.9 cM. The most significant SNP from GWAS (BTB-00803496) on chromosome 21 at 10.17 cM was among the top ten SNP as detected via RF. The correlation coefficient between the SNP effects from GWAS with the RF importance criterion for single SNP, including all SNP from chromosome 21, was 0.57. The correlation for single SNP including all SNP from all chromosomes was 0.53.

Results from the current study were in line with Boichard *et al.* (2003). They detected a QTL associated with conception rate in the same segment where the significant SNP on chromosome 1 is located. Khatkar *et al.* (2014) focused on meta analyses considering 35 QTL and 23 GWAS studies

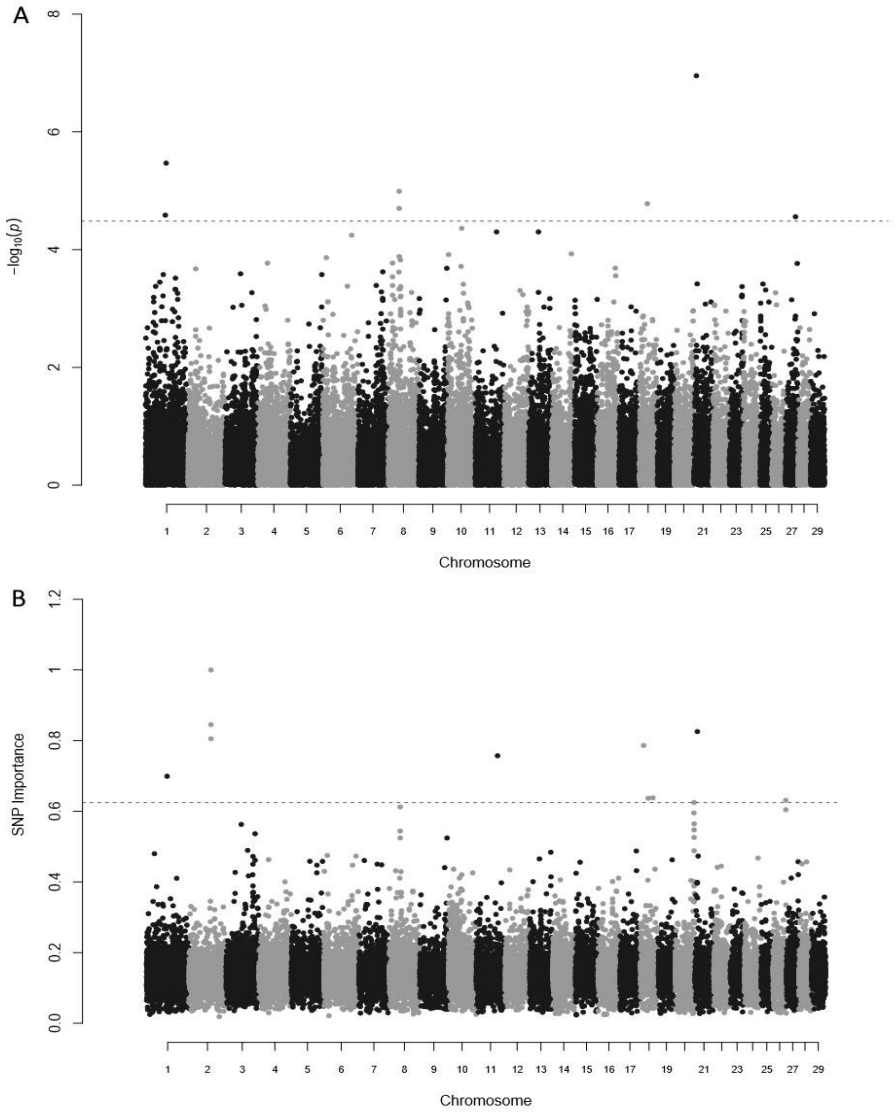

Figure 6: Manhattan plots for –log10 (p) from genome wide association studies (**A**) and SNP importance from RF (**B**) for endometritis

related to cattle fertility, based on more than 101,000 genotyped animals. They reported that most obvious peaks for QTL and SNP effects for female fertility were on chromosome 1. Fortes *et al.* (2013) reported that segments on chromosome 1 were associated with female reproduction traits. Nayeri *et al.* (2016) identified significant effects for SNP located on chromosome 21 on the days from calving to first service and on days open. Hawken *et al.* (2012) found 45 SNP located on chromosome 21 which were associated with female reproduction traits.

Potential positional candidate genes for endometritis within a 500 kbp window are listed in Table 2. However, direct functions of these genes with regard to infertility or endometritis have not yet been determined. One possible explanation for gene impact on endometritis addresses the *FGF* family member genes, which are associated with cell survival activities, and are involved in a variety of biological processes. These biological processes include embryonic development, cell growth, tissue repair, and tumor growth. Also Merhi *et al.* (2015) demonstrated the upregulation of FGF12 in cumulus granulosa cells, providing evidence of inflammation in obese women. Interestingly, according to InnateDB, the positional candidate gene *OSTN* is expressed in the intercaruncular uterus. The glandular intercaruncular regions preserves the uterus in a state of dormancy, and supports the growth of the feto-placental unit (Arosh *et al.* 2004). In addition, secretions from glands in the intercaruncular endometrium contribute to conception after insemination (Gray *et al.* 2001).

CONCLUSION

Prediction accuracies for low heritability health traits were larger when using DRP as response variable, compared to estimates from the PCP response. This was the case for both statistical methods RF and GBLUP. The cow training set composition was one substantial factor affecting prediction accuracies from RF and GBLUP. For PCP as response variable, optimal cow training sets had disease incidences, which were close to the whole population disease incidence. In most scenarios, GBLUP gave better prediction accuracies than RF. Only for PCP as response variable and a large number of sick cows allocated to the training set, RF performed better than GBLUP. With regard to the detection of significant SNP, there were strong overlaps using either RF or a GWAS based on a maximum likelihood approach. Studied gene functions and gene locations suggested *GAS1*, *GPAT3* and *CYP2R1* as potential positional candidate genes for clinical mastitis, *SPINK5* and *SLC26A2* as potential positional candidate genes for claw disorders, and *FGF12* as a potential positional candidate gene for endometritis.

ACKNOWLEDGEMENT

The authors gratefully acknowledge funding from the German Federal Ministry of Education and Research (BMBF) and from the Förderverein Bioökonomieforschung e.V. (FBV) / German Holstein Association (DHV) for the collaborative project "KMU-innovativ-10: Kuh-

L – cow calibration groups for the implementation of selection strategies based on high-density genotyping in dairy cattle", grant no. 031A416C

REFERENCES

Aguilar I., Misztal I., Johnson D. L., Legarra A., Tsuruta S. & Lawlor T. J. (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, **93**, 743–752.

Aguilar I., Misztal I., Tsuruta S., Wiggans G. R., & Lawlor T. J. (2011) Multiple trait genomic evaluation of conception rate in Holsteins. *Journal of Dairy Science*, **94**, 2621–2624.

Albrecht T., Wimmer V., Auinger H.J., Erbe M., Knaak C., Ouzunova M., Simianer H. & Schön C.C. (2011) Genome-based prediction of testcross values in maize. *Theoretical and applied genetics,* **123,** 339–350.

Amer P.R. & Banos G. (2010) Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. *Journal of dairy science*, **93,** 3320–3330.

Arosh J. A., Banu S. K., Chapdelaine P. & Fortier M. A. (2004) Temporal and tissue-specific expression of prostaglandin receptors EP2, EP3, EP4, FP, and cyclooxygenases 1 and 2 in uterus and fetal membranes during bovine pregnancy. Endocrinology, **145**, 407–417.

Ashraf B., Edriss V., Akdemir D., Autrique E., Bonnett D., Crossa J., Janss L., Singh R. & Jannink J. L. (2016) Genomic prediction using phenotypes from pedigree lines with no markers. *Crop Science*, **56**, 957–964.

Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,* **57,** 289–300.

Bionaz M. & Loor J.J. (2008) ACSL1, AGPAT6, FABP3, LPIN1, and SLC27A6 are the aost abundant isoforms in bovine mammary tissue and their expression is affected by stage of lactation. *The Journal of Nutrition,* **138,** 1019–1024.

Boichard D., Grohs C., Bourgeois F., Cerqueira F., Faugeras R., Neau A., Rupp R., Amigues Y., Boscher M.Y. & Leveziel H. (2003) Detection of genes inuencing economic traits in three French dairy cattle breeds. *Genetics selection evolution: GSE,* **35,** 77–101.

Botta V., Louppe G., Geurts P. & Wehenkel L. (2014) Exploiting SNP correlations within random forest for genome-wide association studies. *PloS one,* **9**, e93379.

Breiman L. (2001) Random forests. *Machine Learning,* **45,** 5–32.

Breuer K., Foroushani A. K., Laird M. R., Chen C., Sribnaia A., Lo R., Winsor G. L., Hancock R. E.W. Brinkman F. S. L. & Lynn D. J. (2013) InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation. *Nucleic Acids Research,* **41**, D1228.

Cheng J.B., Levine M.A., Bell N.H., Mangelsdorf D.J. & Russell D.W. (2004) Genetic evidence that the human CYP2R1 enzyme is a key vitamin D 25-hydroxylase. *Proceedings of the National Academy of Sciences of the United States of America,* **101,** 7711–7715.

Christensen, O. F. & Lund M. S. (2010) Genomic prediction when some animals are not genotyped. *Genetic Selection Evolution,* **42**, 2.

Christensen, O.F., Madsen P., Nielsen B., Ostersen T. & Su G. (2012) Single-step methods for genomic evaluation in pigs. *Animal,* **6**, 1565–1571.

Daetwyler H.D., Schenkel F.S., Sargolzaei M. & Robinson J.A.B. (2008) A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. *Journal of dairy science,* **91,** 3225–3236.

Daetwyler H.D., Calus M.P.L., Pong-Wong R., Los Campos G. de & Hickey J.M. (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics,* **193,** 347–365

de Los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D. & Calus M.P.L. (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics,* **193,** 327–345.

Ducrocq V. & Santus E. (2011) Moving away from progeny test schemes: consequences on conventional (inter)national evaluations. Interbull Bulletin 43 (http://www.interbull.org/images/stories/Ducrocq_copy.pdf).

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

Fernanda V.B., Neto J.B., Sargolzaei M., Cobuci J.A. & Schenkel F.S. (2011) Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC genetics,* **12,** 80.

Feucker W. & Staufenbiel R. (2003) Zentraler Diagnoseschlüssel. http://www.portal-rind.de/index.php?module=Downloads&func=prep_hand_out&lid=17. Accessed 16.03.2010.

Fortes M.R.S., Deatley K.L., Lehnert S.A., Burns B.M., Reverter A., Hawken R.J., Boe-Hansen G., Moore S.S. & Thomas M.G. (2013) Genomic regions associated with fertility traits in male and female cattle: advances from microsatellites to high-density chips and beyond. *Animal reproduction science,* **141,** 1–19.

Gao H., Christensen O. F., Madsen P., Nielsen U. S., Zhang Y., Lund M. S. & Su G. (2012) Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genetic Selection Evolution*, **6**, 44.

Gao H., Lund M.S., Zhang Y. & Su G. (2013) Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. *Journal of animal breeding and genetics.* **130,** 333–340.

Garrick D.J., Taylor J.F. & Fernando R.L. (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics selection evolution: GSE,* **41,** 55.

Gerardo A.F.J., Guilherme J.M.R., Valente B.D., Carvalheiro R., Fernando B., Garcia D.A., Gordo D.G.M., Espigolan R., Takada L., Tonussi R.L., de Andrade, W. B. F., Magalhaes A.F.B., Chardulo L.A.L., Tonhati H. & Albuquerque L.G. de (2016) Genomic prediction of breeding values for carcass traits in Nellore cattle. *Genetics selection evolution: GSE,* **48,** 7.

Gernand E., Rehbein P., Borstel U.U. von & Konig S. (2012) Incidences of and genetic parameters for mastitis, claw disorders, and common health traits recorded in dairy cattle contract herds. *Journal of dairy science,* **95,** 2144–2156.

Goldstein B.A., Hubbard A.E., Cutler, A. & Barcellos L.F. (2010) An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics,* **14**, 49.

Gonzalez-Recio O. & Forni S. (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics, selection evolution: GSE,* **43,** 7.

Gray C. A., Bartol F. F., Tarleton B. J., Wiley A. A., Johnson G. A., Bazer F.W. & Spencer T. E. (2001) Developmental biology of uterine glands. *Biology Reproduction*, **65**, 1311–1323.

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

Habier D., Tetens J., Seefried F.R., Lichtner P. & Thaller G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics, selection evolution: GSE,* **42,** 5.

Hawken R.J., Zhang Y.D., Fortes M.R.S., Collis E., Barris W.C., Corbet N.J., Williams P.J., Fordyce G., Holroyd R.G., Walkley J.R.W., Barendse W., Johnston D.J., Prayaga K.C., Tier B., Reverter A. & Lehnert S.A. (2012) Genome-wide association studies of female reproduction in tropically adapted beef cattle. *Journal of Animal Science,* **90,** 1398–1410.

Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K. & Goddard M.E. (2009) Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics selection evolution: GSE,* **41,** 51.

Hernandez J., Shearer J. K. & Webb D. W. (2002) Effect of lameness on milk yield in dairy cows. *Journal of the American Veterinary Medical Association*. **220**, 640–644.

Hidalgo A.M., Bastiaansen J.W.M., Lopes M.S., Veroneze R., Groenen M.A.M. & Koning D.J. de (2015) Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs. *Journal of Animal Science,* **93,** 3313–3321.

Hogeveen H., Huijps K. & Lam T. J. (2011) Economic aspects of mastitis: new developments. *New Zealand Veterinary Journal,* **59**, 16-23.

Iung L.H.S., Ramírez-Díaz J., Pertile, S. F. N., Petrini J., Salvian M., Rodriguez M.A.P., Lima R.R., Machado P.F., Coutinho L.L. & Mourão G.B. (2014) Genome-wide association for somatic cell score in Holstein cows raised in tropical conditions. Abstract 624 in Proc. 10th World Congr. Genet. Appl. Livest. Prod., Vancouver, BC, Canada.

Jaggi R., Marti A., Guo K., Feng Z. & Friis R.R. (1996) Regulation of a physiological apoptosis: Mouse mammary involution. *Journal of dairy science,* **79,** 1074–1084.

Khatkar M.S., Randhawa I. & Raadsma H.W. (2014) Meta-assembly of genomic regions and variants associated with female reproductive efficiency in cattle. *Livestock Science,* **166,** 144–157

Klungland H., Sabry A., Heringstad B., Olsen H.G., Gomez-Raya L., Våge D.I., Olsaker I., Ødegård J., Klemetsdal G., Schulman N., Vilkki J., Ruane J., Aasland M., Rønningen K. & Lien S. (2001) Quantitative trait loci affecting clinical mastitis and somatic cell count in dairy cattle. *Mammalian Genome,* **12,** 837–842.

Kramer M., Erbe M., Seefried F.R., Gredler B., Bapst B., Bieber A. & Simianer H. (2014) Accuracy of direct genomic values for functional traits in Brown Swiss cattle. *Journal of dairy science,* **97,** 1774–1781.

Legarra A., Aguilar I. & Misztal I. (2009) A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science,* **92**, 4656–4663.

Legarra, A., Chistensen O. F., Aguilar I. & Misztal I. (2014) Single step, a general approach for genomic selection. *Livestock Production Science*, **166**, 54–65.

Lee H.S., Wray N.R., Goddard M.E. & Visscher P.M. (2011) Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88, **294-305**.

Lippolis J.D., Reinhardt T.A., Sacco R.A., Nonnecke B.J. & Nelson C.D. (2011) Treatment of an intramammary bacterial infection with 25-hydroxyvitamin D3. *PloS one,* **6,** e25479.

Lourenco D. A. L., Tsuruta S., Fragomeni B. O., Masuda Y., Aguilar I., Legarra A., Bertrand J. K., Amen T. S., Wang L., Moser D. W.& Misztal I. (2015 a). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *Journal Animal Science,* **93**, 2653–2662.

Lourenco D. A. L., Fragomeni B. O., Tsuruta S., Aguilar I., Zumbach B., Hawken R. J., Legarra A. & Misztal I. (2015 b) Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genetics Selection Evolution: GSE*, **47**, 56.

Madsen P. & Jensen J. (2013): DMU – a package for analysing multivariate mixed models. Version 6, release 5.2. Aarhus University, Foulum, Denmark. Available from http://dmu.agrsci.dk.

Merhi Z., Polotsky A.J., Bradford A.P., Buyuk E., Chosich J., Phang T., Jindal S. & Santoro N. (2015) Adiposity alters genes important in inflammation and cell cycle division in human cumulus granulosa cell. *Reproductive sciences,* **22,** 1220–1228.

Minozzi G., Pedretti A., Biffani S., Nicolazzi1 E. L. & Stella A. (2014) Genome wide association analysis of the 16th QTL- MAS workshop dataset using the Random Forest machine learning approach. *BMC Proceedings*, **8** (Suppl 5):S4.

Misztal I., Aggrey S. E. & Muir W. M. (2013) Experiences with a single-step genome evaluation. *Poultry science*, **92**, 2530–2534

Moser G., Tier B., Crump R.E., Khatkar M.S. & Raadsma H.W. (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics selection evolution: GSE,* **41,** 56.

Mrode R. A. (2005) Linear models for the prediction of animal breeding values: 2rd edition, pp. 344, Cabi Publishing.

Naderi S., Yin T. & Konig S. (2016) Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. *Journal of dairy science,* **99,** 7261–7273.

Nayeri S., Sargolzaei M., Abo-Ismail M.K., May N., Miller S.P., Schenkel F., Moore S.S. & Stothard P. (2016) Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC genetics,* **17,** 75.

Oltenacu P. A., & Broom D. M. (2010) The impact of genetic selection for increased milk yield on the welfare of dairy cows. *Animal Welfare,* **19,** 39–49.

Ostersen T., Christensen O.F., Henryon M., Nielsen B., Su G. & Madsen P. (2011) Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genetics selection evolution: GSE,* **43,** 38.

Pszczola M., Strabel T., Mulder H.A. & Calus M.P.L. (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of dairy science,* **95,** 389–400.

Sahana G., Guldbrandtsen B., Thomsen B., Holm L.-E., Panitz F., Brondum R.F., Bendixen C. & Lund M.S. (2014) Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. *Journal of dairy science,* **97,** 7258–7275.

Sarkar R. K., Rao A. R., Meher P. K., Nepolean T. & Mohapatra T. (2015) Evaluation of random forest regression for prediction of breeding value from genomewide SNPs. *Journal of Genetics*, **94**, 187–192.

Schnabel R.D., Sonstegard T.S., Taylor J.F. & Ashwell M.S. (2005) Whole-genome scan to detect QTL for milk production, conformation, fertility and functional traits in two US Holstein families. *Animal genetics,* **36,** 408–416.

Shearer J. K. (1998) Lameness of dairy cattle: consequences and causes. *The Bovine Practitioner*, **32**, 79-85.

Sheffield L. G. (1988) Organization and growth of mammary epithelia in the mammary gland fat pad. *Journal of Dairy Science*, **71**, 2855-2874.

Sodeland M., Kent M.P., Olsen H.G., Opsal M.A., Svendsen M., Sehested E., Hayes B.J. & Lien S. (2011) Quantitative trait loci for clinical mastitis on chromosomes 2, 6, 14 and 20 in Norwegian Red cattle. *Animal genetics,* **42,** 457–465.

Stock K. F., Cole J., Pryce J., Gengler N., Bradley A., Andrews L., Heringstad B. & Egger-Danner C. (2013). Standardization of health data – ICAR guidelines including health key. *ICAR Technical Series*, **17**, 75–81.

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, **14**, 323–348.

Su G. & Madsen P. (2013) User's Guide for Gmatrix version 2, a program for computing genomic relationship matrix. Accessed Apr. 11, 2013. Available from- http://-www.-dmu.agrsci.-dk/-Gmatrix/Doc/.

Su G., Christensen O. F., Janss L. & Lund M. S. (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *Journal of dairy science*, **97**, 6547–6559

van der Spek D., van Arendonk J.A.M. & Bovenhuis H. (2015) Genome-wide association study for claw disorders and trimming status in dairy cattle. *Journal of dairy science,* **98,** 1286–1295.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *Journal of dairy science,* **91,** 4414–4423.

VanRaden P. M. (2012). Avoiding bias from genomic pre-selection in converting daughter information across countries. *Interbull Bulletin*, **45**, 1–5.

Wang J. (2010) Genome-wide transcriptome analysis of laminar tissue during the early stage of experimentally induced equine laminitis. *Dissertation, Texas A&M University*.

Wellmann R., Preuss S., Tholen E., Heinkel J., Wimmers K. & Bennewitz J. (2013) Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics selection evolution: GSE*, **45**, 28.

Wright M. N. & Ziegler A. (2017) Ranger: A Fast implementation of Random Forests for high dimensional data in C++ and R. *Journal of Statistical Software,* **77**, 1–17.

Wu X., Fang M., Liu L., Wang S., Liu J., Ding X., Zhang S., Zhang Q., Zhang Y., Qiao L., Lund M.S., Su G. & Sun D. (2013) Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. *BMC genomics,* **14,** 897.

Wu X., Guldbrandtsen B., Lund M. S. & Sahana G. (2016) Association analysis for feet and legs disorders with whole genome sequence variants in 3 dairy cattle breeds. *Journal of dairy science,* **99**, 7221–7231.

Yang J., Lee S. H., Goddard M. E. & Visscher P. M. (2011) GCTA: a tool for genome wide complex trait analysis. *The American Journal of Human Genetics*, **88**, 76–82.

Zhang Z., Liu J., Ding X., Bijma P., Koning D.J. de & Zhang Q. (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PloS one,* **5,** e12648.

4th Chapter

Assessing signatures of selection through variation in linkage disequilibrium within and between dual-purpose black and white (DSN) and German Holstein cattle populations

Naderi[1], S., Moradi[2], M. H., Farhadian[3], M., Yin[1], T., Jaeger[1], M., Scheper[1], C., Korkuc[4], P., Brockmann[4], G.A., König[1], S.

[1]Institute of Animal Breeding and Genetics, University of Gießen, Ludwigstr. 21b, Germany

[2]Department of Animal Sciences, Arak University, Shahid Beheshti Street, Arak, Iran

[3]Department of Animal Science, University of Tabriz, 29 Bahman Boulevard, Tabriz, Iran

[4] Albrecht Daniel Thaer- Institute for Agricultural and Horticultural Sciences, Humboldt Universität zu Berlin, Invalidenstr. 42, D-10115 Berlin, Germany

SUMMARY

Identification of genes or genomic regions that have been targeted of recent selection might contribute to a better understanding of adaptive evolution. Comparing distinct populations or sub-populations and inferring genomic regions with prominent genetic differentiation are the basic principles for the identification of selection signatures. The aim of this study was to infer adaptive genetic variation between the German Holstein and the DSN population as well as between sub-populations stratified according to geographical characteristics and disease incidences. We used cross-population extended haplotype homozygosity methodology (XP-EHH), which exploits linkage disequilibrium structures to reveal the most recent selection signatures. Furthermore, we calculated Wrights fixation index (F_{ST}). Analyses based on 4,654 genotyped Holstein cows and 261 genotyped DSN cows. In order to build up pronounced contrasts between populations and in order to save computation time, 2,076 high-yielding cows (average milk yield from the first three test-days) represented the German Holstein population. The geographical herd location was used as an environmental descriptor to create DSN subpopulations, i.e., the West-German DSN (172 cows) and the East-German DSN (89 cows) subpopulation. In addition, two groups of Holstein cows were created based on most extreme values for solutions of residual effects for dermatitis digitalis (DD), representing the most resistant and most susceptible cows for DD. These two German sub-populations were defined as healthy and sick (250 animal in each population), respectively. A threshold for the top 0.1 percentile of negative or positive XP-EHH scores were study in detail. Gene annotation based on the Ensembl database (www.ensembl.org/biomart/martview) and covered a window of 250 Kbp downstream and upstream of each core SNP corresponding to peaks of extreme XP-EHH scores. In addition, functional interactions among potential candidate genes were studied. The most outstanding XP-EHH score was on chromosome 12 (at the position of 77.34 Mb) for DSN and on chromosome 20 (at 36.29-38.42 Mb and at 69.43-69.66 Mb) for German Holsteins. The most extended selection signature for East-DSN was identified on chromosome 6 at 92.97 to 94.43 Mb. This genomic segment harbors known quantitative trait loci for several economically important milk and meat quality traits like milk kappa-casein percentage and marbling score. The average F_{ST} values were 0.068 between German Holsteins and DSN, 0.0085 between West-DSN and East-DSN sub-populations, and 0.010 between healthy and sick Holstein sub-populations. Gene annotation analyses for the selection signature regions revealed various potential candidate genes

associated with production traits, such as *CLU* and *WARS2*. Furthermore, several genes being associated with DD resistance were detected. In this regard, the most important genes were *FARS2, CCDC185, MIA3, TRIM27, LYRM4, CHDH, LAPTM5, HSCB, TRAF3IP3*, *TNFAIP3* and *MID2.* Among the annotated genes, *TRAF3IP3* is expressed in immune organs, promoting immune response. Our methodological approach, which based on populations or created sub-populations with different breeding history was suitable to understand principles of selection and adaptation in German Holstein cattle and their founder breed (DSN) in more detail.

Key words: Selection signatures, population stratification, genome annotations

INTRODUCTION

The present diversity of cattle breeds specialized for either milk or meat production, or selected for dual-purpose use, is the direct consequence of divergent artificial selection strategies. Divergent artificial selection contributes to selection signatures in the genome. Selection signatures are structural footprints at specific genome segments, controlling fitness or productivity. Due to linkage, those changes are detectable via loci directly influenced by selection, as well as in linked neutral loci (Kreitman, 2000).

Increasing availability of genomic information with knowledge on genomic variants in farm animals, as well as advances in statistical methods, enable the exploration of domestication processes. Domestication processes include studies on evolutionary history, and on biological differentiations between populations and breeds. Identified selection signatures allow insights into genomic regions altered by selection, contributing to a deeper understanding of the underlying biological and physiological mechanisms of artificial or natural selection in farm animals (Qanbari & Simianer, 2014). Selection causes not only the increase of allele frequencies of beneficial mutations, but also reduces the variation in regions linked to selected loci, so-called “selective sweeps” (Smith & Haigh, 1974). In addition, the extent of linkage disequilibrium (LD) among markers within an extended chromosomal segment indicates selection on annotated genes.

Several methods have been used to identify selection signatures based on LD structures: The extended haplotype homozygosity (EHH) (Sabeti *et al*. 2002), the integrated haplotype score (iHS) (Voight *et al*. 2006), and cross-population extended haplotype homozygosity (XP-EHH) (Sabeti *et al*. 2007). EHH is the probability that two randomly selected haplotypes carrying the same selected allele (core allele) are homozygous, for the entire interval from the core allele to a given locus

(Rothammer *et al*. 2013). The iHS methodology bases on the ratio of the integrated EHH-curves within a population (Voight *et al*. 2006). However, this method might lack power when aiming on the identification of selective sweeps, being the result from complete allele fixation (Rothammer *et al*. 2013). XP-EHH combines both methods EHH and iHS. The concept bases on artificially generated sub-populations, which can be used for the detection of most recent signatures of selection (Sabeti *et al*. 2007). The XP-EHH method identifies long haplotypes representing the most recent selection signatures, being indicators for recent breed differentiation and alterations on phenotypic scales. Applying XP-EHH, Chen *et al*. (2016) identified recent selection signatures in Chinese Holstein and Simmental cattle populations. Regions under positive selection pressure overlapped with a set of important genes being involved in biological processes. Lee *et al*. (2014) also used XP-EHH methodology, and they detected 250 genes in regions with "outlier SNP", including the alpha1casein (*CSN1S1*) and the beta-casein gene (*CSN2*). Both genes *CSN1S1* and *CSN2* have substantial impact on milk protein quality in Holstein cattle.

Assessing the variation of marker allele frequencies in different populations is a further tool to discover genome wide signatures (Holsinger & Weir, 2009). One specific strategy is to use a large number of SNPs across the genome and to compare specimen from different populations in order to identify regions with prominent genetic distinction (Gholami *et al*. 2014). In this regard, Wrights fixation index (F_{ST}) is one of the most popular methods. F_{ST} bases on single site differentiation for the detection of selection signatures, reflecting genetic differentiation between populations (Wright, 1949). F_{ST} was used for the detection of selection in several previous studies, e.g., Moradi *et al*. 2013 or Gholami *et al*. 2014.

The modern German Holstein breed originates from the local dual-purpose German black pied cattle breed (DSN). The DSN breed consequently followed a dual-purpose breeding goal for high meat and milk yield, and additionally on adaptation to grassland systems. In contrast, in German Holsteins, there was a strong selection focus on milk yield and on dairy character over decades by excluding meat traits (Brade & Brade, 2013; Jaeger *et al*. 2016). Selection intensities for functional traits in German Holsteins were quite low in breeding programs in the past (König *et al*. 2007). Selection on milk production traits in German Holsteins was significantly intensified during the last decades through the implementation of artificial insemination schemes since the 1960s (Skjervold & Langholz, 1964). This development was accompanied by increasing imports of specialized Holstein Friesian dairy lines from North America. A comparison of the modern German

Holstein population with one of its founder breeds, i.e., the DSN population, is a unique opportunity for the detection of recent selection signatures. In this specific case, we hypothesize the identification of selection signatures due to recent selection pressure on chromosomal segments being associated with milk production traits. Additionally, a further intra-breed comparison between the current West German-DSN and east German-DSN subpopulations will contribute to a deeper understanding of selection mechanisms. From a breeding history perspective, the local separation of Germany after the Second World War also separated the DSN population, preventing any exchange of animals until the reunification of Germany in 1990. During the separation, production system characteristics as well as breeding goal definitions were significantly different in both German states. Consequently, also against this background, the XP-EHH method might be appropriate in order to detect recent selection signatures.

Lack of precise health data hampered the inclusion of health traits into breeding goals (Gernand *et al*. 2013). Thus, natural selection was the only force determining adaptive evolution of health traits over decades. The implementation of direct electronic health recording systems combined with large-scale genotyping in cow training sets is a new basis for artificial health trait selection (Egger-Danner *et al*, 2014). Genomic regions as well as genes targeted by selection on disease resistance are functionally important, and gene polymorphisms directly related to phenotypic variation were detected (Nielsen *et al*. 2007). In consequence, stratification of German Holstein populations according to disease incidences allows detection of recent selection signatures for health and adaptation.

Localization of chromosomal segments via selection signatures is a first hint for functional gene identification of complex traits (Evangelou *et al*. 2014; Wang *et al*. 2010). Complex traits such as disease susceptibility mostly rely on cumulative gene effects, as well as on multiple gene interactions in functional pathways (Eleftherohorinou *et al*. 2009). The term “interacting” includes genes whose products are integrated in functional pathways, with possible impact on other gene expressions, and in causality, on the biological output. For instance, in a study addressing aspects of immune genetics, the pattern of gene variants within a pathway specified the intensity and nature of immune response of a host to specific pathogens (Hill, 2006).

In consequence, the aims of this study were i) to infer adaptive genetic variation between the German Holstein and the DSN populations, ii) to infer adaptive genetic variation between sub-

populations stratified according to geographical characteristics and disease incidences, and ii) to infer biological pathways of genes that are located in genomic regions of selection signatures.

MATERIALS AND METHODS

Data

Samples and Quality Control. 4,654 German Holstein dairy cows and 261 DSN cows were genotyped with the *Illumina Bovine 50K SNP BeadChip V2*. With regard to SNP quality control criteria, SNP marker with a minor allele frequency (MAF) lower than 0.01, and with a large deviation from Hardy–Weinberg equilibrium ($P < 0.000001$), were excluded. All SNP had a call rate larger than 95%. After SNP-data editing, the final dataset included 39,917 SNP. Furthermore, we excluded cows with a call rate of less than 0.95 for all loci. Finally, 4,654 German Holstein cows and 261 DSN cows remained for ongoing genomic analyses. The Software PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) was used for the SNP filtering processes.

Population stratifications

Holstein and DSN. From the total dataset of 4,654 genotyped first-lactation German Holstein cows, we considered a random sample of 2,076 cows from 28 different herds (because of computation time). For these 2,076 cows, the average milk yield from the first three test-days was 32.68 kg, while the average from all genotyped 4,654 Holstein cows was 31.05 kg. The 261 DSN cows from six different herds with an average milk yield of 24.46 kg represented the DSN population.

Healthy and sick Holstein cows. German Holstein cows were allocated to two different groups according to their susceptibility to / resistance against dermatitis digitalis (DD). In this regard, we focused on the first 200 days in milk, representing the quite sensitive period after calving. Repeated observations for DD from the same cow were ignored. At least one entry for DD implied a score = 1 (sick); otherwise, the score = 0 (healthy) was assigned. In a next step, binary DD phenotypes were pre-corrected using the following model:

$$\boldsymbol{\eta} = \mathbf{Xb} + \mathbf{e}$$

where $\boldsymbol{\eta}$ was a vector of logits for all cows, $\mathbf{X}$ was the incidence matrix for fixed effects; $\mathbf{b}$ was a vector for fixed effects including herd and year-season of calving, and $\mathbf{e}$ was a vector of random residual effects. Based on solutions for DD residuals, cows with most extreme values were allocated

either to a group A or to a group B (250 animals in each group). The 5% upper tail (= group A) included the DD resistant cows with residuals in the range from -0.88 to -0.31 (mean value: -0.44). The 5% lower tail (= group B) included the DD susceptible cows with residuals in the range from 0.75 to 0.99 (mean value: 0.83).

West-DSN and East-DSN. DSN sub-population creation based on the geographical herd location (i.e., former East versus former West Germany). The map indicates the location of DSN herds, along with the percentage of DSN cows within herds (Figure 1). The majority of West-DSN farms was located in north-west Germany (East Friesland area), and around Hannover and Bremen. These small-scale herds (average herd size: 49 cows) represented an intensive grazing system in coastal marshlands. The East-DSN sub-population was a random sample from one large-scale herd (herd size: 1712 cows), representing a strict indoor production system. The west-DSN sub-population included 172 cows, and 89 cows were allocated to the east-DSN sub-population.

Pedigree based genetic relationships between and within populations and sub-populations were calculated using the software package CFC (Sargolzaei *et al*. 2006). For further illustration of genomic composition and differentiation among population, a principle component analysis (PCA) was carried using the R package SNPRelate (Zheng *et al*. 2012).

Identification of selection signatures

As introduced by Sabeti *et al*. (2007), XP-EHH was calculated in order to identify recent positive selection signatures. XP-EHH bases on EHH values and evaluates LD decay across the genome. For a bi-allelic SNP with alleles A and a, EHH is defined as follows (Sabeti *et al*. 2002):

$$\mathrm{EHH} = \frac{\sum_{i=1}^{h_x} \binom{n_i}{2}}{\binom{n_a}{2}\binom{n_A}{2}}$$

where n_A and n_a are the number of haplotypes with alleles A and a, respectively, n_i is the count of the i^{th} haplotype within a sub-population, and h_x represents the number of distinct haplotypes in a genomic region up to a distance x from the core locus. In order to calculate XP-EHH for sub-populations 1 and 2, all SNPs located 1 Mb in both directions from a given core SNP were considered. Afterwards, EHH was integrated within these bounds (for the entire interval from the core SNP up to the distance x) considering both sub-populations. In a next step, unstandardized XP-EHH followed the principles as defined by Sabeti *et al*. (2007):

$$\mathrm{XP-EHH} = \log\left(\frac{\int \mathrm{D}\ \mathrm{EHH}_{\mathrm{pop1}}(x)dx}{\int \mathrm{D}\ \mathrm{EHH}_{\mathrm{pop2}}(x)dx}\right)$$

XP-EHH was standardized based on means and standard deviations. Extreme positive values reflect selection in sub-population 1, and the extreme negative values in sub-population 2.

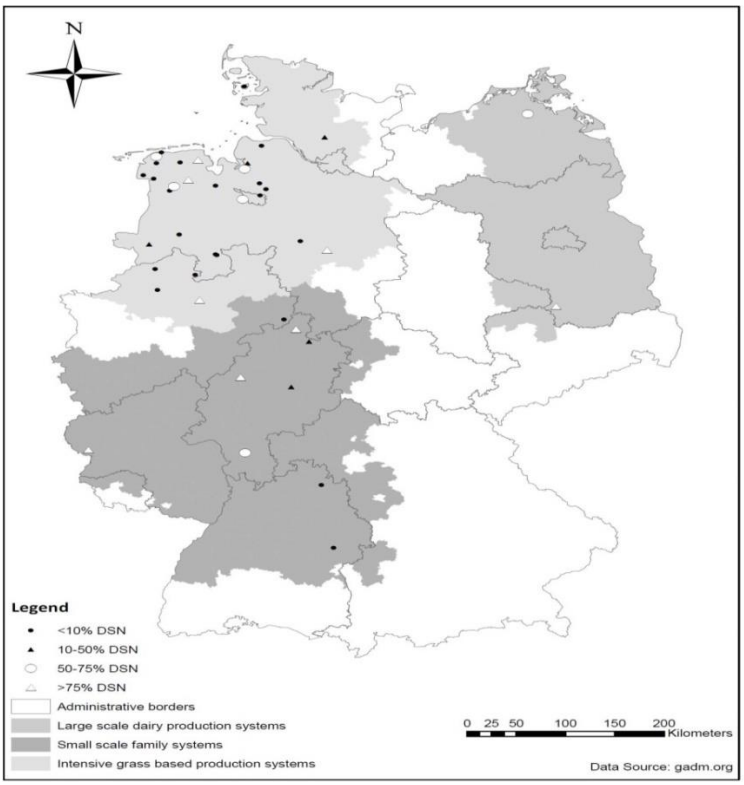

Figure 1: Distribution of DSN cattle herds across Germany. Different symbols indicate the DSN percentage per herd.

The software package FASTPHASE (Scheet & Stephens, 2006) was used to reconstruct haplotypes for each chromosome in both sub-population because phased haplotypes are required to calculate XP-EHH. Further selection signature analyses using fully phased haplotype data were carried out applying an XP-EHH software package at http://hgdp.uchicago.edu (coded by J. Pickrell). Calculation of XP-EHH bases on genomic distances between adjacent SNP. As introduced by MacEachern *et al*. (2009), F_{ST} was calculated as follows:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

where H_T and H_S are expected heterozygosities for the overall total population and for sub-populations, respectively.

Gene annotation and functional gene analysis

Genes were extracted for segments on the chromosome around the peaks of extreme XP-EHH values from Ensembl database (www.ensembl.org/biomart/martview). The "peak definition" corresponds to XP-EHH values beyond the upper and lower 1% of the observed genome-wide distribution of XP-EHH. A window of 250 kbp downstream and upstream of each core SNP as used by Maiorano *et al*. (2018) was considered for potential candidate gene identifications. Additionally, Cattle QTL database (https://www.animalgenome.org/cgi-bin-/QTLdb/BT/-search) was used to compare the detected regions in the current study with QTL regions previously identified.

Gene Network analysis. In addition, we studied functional interactions among proteins encoded by the candidate genes. In this regard, we used the STRING (https://string-db.org/) database which collected curated and experimental validated proteins interaction from published literature and protein-protein interactions network was constructed for each population (Szklarczyk *et al*. 2016). Constructed network clustered with the K means algorithm to define the functional modules. K-means is a popular clustering algorithm which is widely used in anomaly-based intrusion detection. It tries to classify a given data set into k (a predefined number) categories.

RESULTS AND DISCUSSION

Genetic relationships

The average relationship coefficient was 0.049 within the healthy German Holstein sub-population, 0.047 within the sick German Holstein sub-population, and 0.043 within the German Holstein population. With regard to dual-purpose DSN cows, the average relationship coefficient was 0.089 for the East-DSN, 0.028 for the West-DSN sub-population and 0.021 for the total DSN dataset. The average relationship coefficient between the sick and healthy German Holstein sub-population was 0.041, 0.0001 between the German Holstein and DSN population, and 0.038 between East-DSN and West-DSN. These results indicate that the German Holstein and the DSN breed can be clearly distinguished. Healthy and sick German Holstein, and East-DSN and West-DSN sub-populations, are genetically more similar, probably due to the more intense drift in the small DSN population, and shared founder effects. Furthermore and as shown in Figure 2 by using the first two principal components (PCA1 and PCA2) only two populations in this study (Holstein *vs.* DSN) were clearly distinguishable by two clusters. The Holstein populations (healthy Holstein and sick Holstein) and the DSN populations (West-DSN, East-DSN) are located at opposite sites. The healthy and sick

Holstein populations are located in the same group. Also, among West and East DSN populations are not a definite separation based on two first components (PCA1 and PCA2).

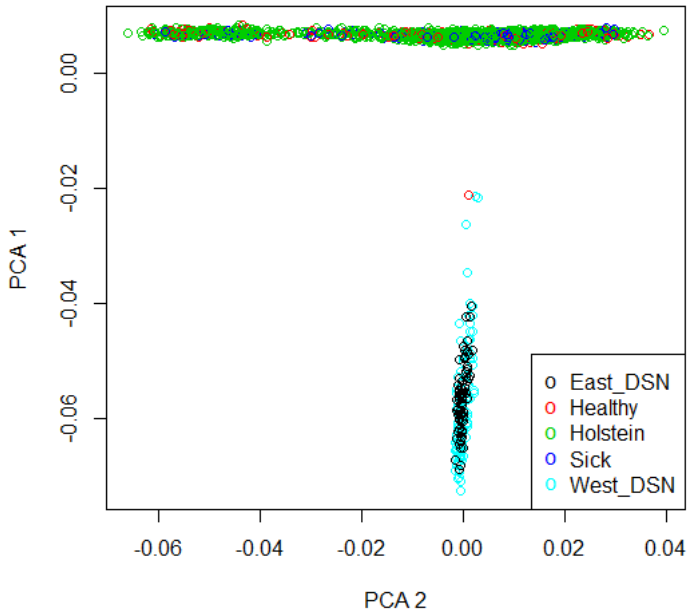

Figure 2: Principal components analysis between populations

Selection signatures

German Holstein versus DSN. XP-EHH was calculated for each SNP along the genome for the DSN and German Holstein population. The Manhattan plot of standardized XP-EHH scores across the genome is shown in Figure 3. Negative XP-EHH values reflect selection events in the DSN population. According to XP-EHH scores and a threshold for the top 0.1 percentile for negative XP-EHH scores (corresponds to XP-EHH < -2.9), we found evidence of selection across the genome on chromosomes 1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 14, 15, 18, and 23. An extremely negative value for XP-EHH (XP-EHH = -4.09) was identified on chromosome 12 at the position of 77.34 Mb. In addition, obviously negative XP-EHH scores were identified on chromosome 9 at positions from 13.63 to 13.94 Mb, and on chromosome 23 at positions from 5.21 to 5.52 Mb.

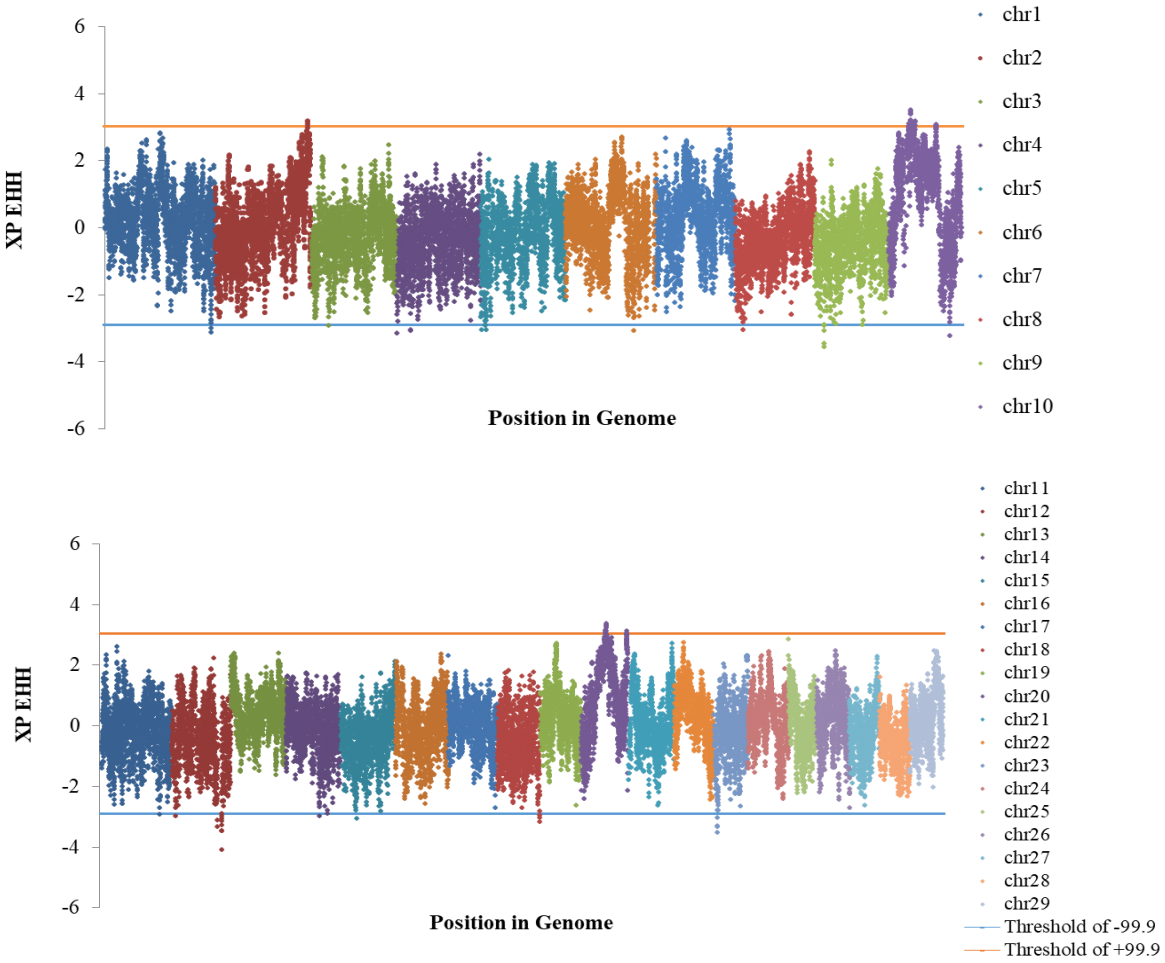

Figure 3: XP-EHH score for each SNP as a function of the chromosome position for the Holstein (positive values) and the DSN population (negative values).

The regions under positive selection and the annotated genes in a window of 250 Kb downstream and upstream of each core SNP are presented in Table 1 for DSN.

Table 1: The regions under positive selection in the DSN population and the annotated potential candidate genes within a window of 250 Kb downstream and upstream of each core SNP.

Chr.	Position	No. of SNP	Gene
1	150.61-151.12	2	*CHAF1B, CLDN14, SIM2, HLCS, HGNC, RIPPLY3*
3	24.28	1	*WARS2, TBX15*
4	0.25, 19.89-19.85	3	*THSD7A, TMEM106B, VSTM2A*
5	0.65, 6.49, 7.17	3	*TSPAN8, NAV3, ZDHHC17, CSRP2, E2F7*
6	97.58	1	*C4orf22, BMP3, PRKG2*
8	11.00	1	*SCARA5, PBK, ESCO2, CCDC25, SCARA3, CLU, TMEM215*
9	13.63-13.94	4	*CD109*
10	87.99	1	*C14orf1, TTLL5, TGFB3*
11	89.23	1	-
12	6.76, 70.06-70.09, 77.31-77.61	12	*LOC515333, UGGT2*
14	64.58, 53.02	2	*RRM2B, CSMD3*
15	25.18	1	*C15H11orf71, RBM7, REXO2, NXPE4, ZBTB16*
18	65.40-65-62	3	*ZNF211, ZSCAN4, LOC509810, LOC100124497, LOC104968476*
23	5.21-5.52	4	*GFRAL, HCRTR2, FAM83B, 5S_rRNA*

Network N1 shows physiological pathways of genes overlapping with signatures of positive selection (Figure 4). By using k-means clustering algorithm the network was divided in three clusters. In the first cluster (in green) *PBK*, *PCNA* and *ZDHHC17* were the hub genes of the network. The hub genes of second cluster (in red) and the third cluster (in blue) were *ACACB* and *PRKG2* genes respectively. Nevertheless, none of the genes involved in three clusters independently activated a pathway significantly. Associations of some genes in the network N1 with dairy cattle breeding goal traits have been identified in previous studies. For instance, *CLU* is one of genes of the network that reported by Li *et al*. (2016) as one of the most influential candidate genes affecting milk protein content. Wang *et al*. (2012) verified the effect of *CLU* on milk production traits in Chinese Holstein cows. *CLU* is induced in the mammary gland under stress, and plays an anti-inflammatory role (Guenette *et al*. 1994; Humphreys *et al*. 1999; Piantoni *et al*. 2010). Accordingly, Silkensen *et al*. (1994) found increased *CLU* expressions during stress periods, e.g., in sick animals.

Swanson *et al*. (2009) verified such increased *CLU* expressions during mastitis episodes of lactating dairy cows.

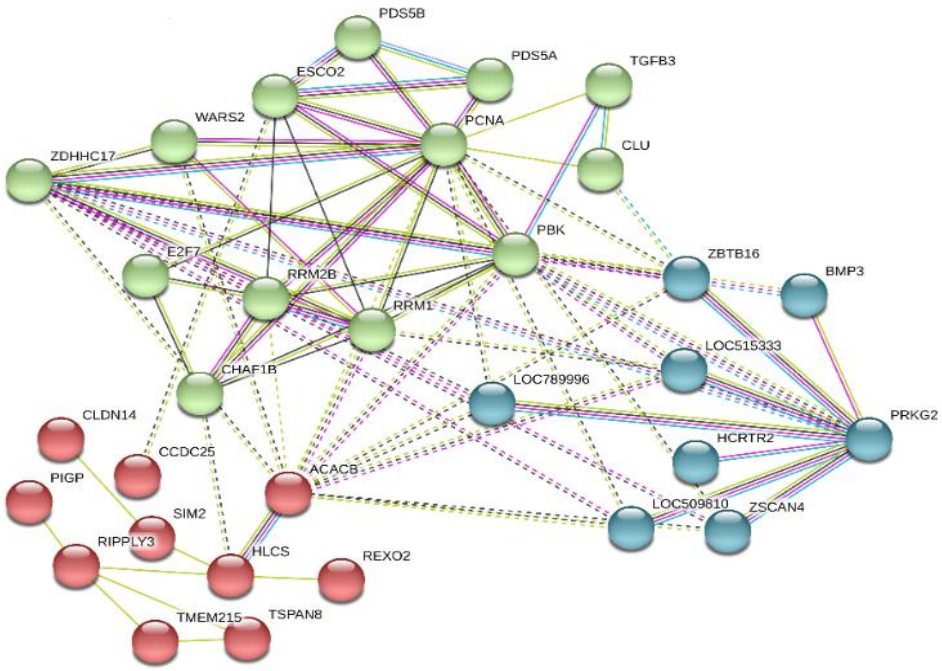

Figure 4: Network N1, interaction between the identified genes for DSN population.

Association of *PRKG2* (the hub gene of third cluster) with production and conformation traits was confirmed in previous studies. For example, Boegheim *et al.* (2017) reported that causative mutation in *PRKG2* is probably in linkage disequilibrium with several QTL associated with production traits in cattle. Genetic hitchhiking with the neighbouring gene variant leads to increase in homozygosity of that region. Furthermore and according to Koltes *et al*. (2009), in a certain period in history animals phenotypically were selected for a stocky and short-legged and mutation within the kinase domain of *PRKG2* could be the causative mutation of dwarfism in Angus cattle. *WARS2* is one of the genes in network N1, which is involved in regulation of subcutaneous fat mobilisation, and fat disposition in the human visceral adipose tissue (Schleinitz *et al*. 2014). Accordingly, Cesar *et al*. (2014) identified associations between specific fatty acid compositions and *WARS2* in Brazilian Nellore steers. *WARS2* encodes the essential enzyme to catalyse amino acylation of tRNA, and carries a signal peptide to regulate mitochondrial import (McKusick, 2007).

The regions under positive selection and the annotated genes in a window of 250 Kb downstream and upstream of each core SNP are presented in Table 2 for German Holsteins. The regions with the highest positive XP-EHH scores reflecting selection events in the Holstein population (Figure 3) are located on chromosome 10 (at 31.46 to 37.83 Mb and at 68.60 to 68.89 Mb), chromosome 20 (at 36.29 to 38.42 Mb and at 69.43 to 69.66 Mb), and on chromosome 2 (at 131.30 to 131.34 Mb). The positive selection signature in the two regions on chromosome 10 at 31.46 to 37.83 Mb and at 68.60 to 68.89 Mb encompasses several genes (i.e., *SNORA74*, *C15orf41*, *MEIS2*, *TMCO5A*, *snoU13*, *SPRED1*, *SNORA70*, and *PELI2*). Among these genes, Pellino E3 Ubiquitin Protein Ligase Family Member 2 (*PELI2*) is expressed in a broad range of tissues, including the mammary tissue. Gutiérrez-Gil *et al*. (2015) identified *PELI2* as a candidate gene within the core selective sweep regions of dairy cattle, regulating the toll-like receptor signalling pathway. This pathway contributes to immune response. Yang *et al*. (2014) reported similar *PELI2* pathway contributions in pigs.

Table 2: The regions under positive selection in the Holstein population and the annotated potential candidate genes within a window of 250 Kb downstream and upstream of each core SNP.

Chr.	Position	No. of SNP	Gene
10	31.46-37.83, 68.60- 68.89	24	*U2*, *SNORA74*, *C15orf41*, *MEIS2*, *U4*, *TMCO5A*, *snoU13*, *SPRED1*, *SNORA70*, *U7*, *PELI2*
20	36.29-38.42, 69.43-69.66	12	*NIPBL*, *bta-mir-2360*, *NUP155*, *WDR70*, *GDNF*, *U6*, *SNORA17*, *EGFLAM*, *IRX1*
2	131.30-131.34	3	*LDLRAD2*, *HSPG2*, *CDC42*, *WNT4*

Interestingly, the inferred high XP-EHH scores on chromosome 20 confirmed selection signatures identified in Israeli Holsteins (Glick *et al*. 2012), and in German Holstein bull dams intensively selected for production traits (Qanbari *et al*. 2011). Zhao *et al*. (2015) applied the iHS methodology in Holstein dairy cattle, and they also reported strong selection signatures on chromosome 20. The potential candidate genes in these regions are *NIPBL*, *bta-mir-2360*, *NUP155*, *WDR70*, *GDNF*, *SNORA17*, *EGFLAM* and *IRX1*. The gene *NIPBL* was the hub gene in network N2

(Figure 5) that was included in the first cluster of the network (in blue). However, the annotated genes in this cluster did not activate a pathway significantly but a genome-wide scan in Brazilian sheep breeds revealed *NIPBL* as a candidate gene involved in different biological functions, such as immunity, nervous system development and reproduction (Gouveia *et al*. 2017). Furthermore, *NIPBL* was one of the identified genes being associated with fat percentage and protein percentage in Chinese dairy cattle (Jiang *et al*. 2014). The extended selection signature on chromosome 2 spanned almost 0.021 Mb. The most outstanding genes from this chromosomal region were *LDLRAD2*, *HSPG2*, *CDC42*, and *WNT4*. *CDC42* is the hub gene of the second cluster of the network N2 (in green) is involved in activating of several pathways (i. e., MAPK signalling pathway, Jak-STAT signalling pathway and T cell receptor signalling pathway).Yamaji *et al*. (2013) confirmed the mandatory requirement of the JAK-STAT signalling pathway in development of mammary gland and lactation in mice. Furthermore, Arun *et al*. (2015) reported that JAK-STAT pathways play an important role in providing intracellular signals consequent in co-ordinate gene transcription in respond to a wide variety of hormones with specific function in development of mammary gland and lactation cycle. They concluded that JAK-STAT pathway genes are beneficial in expanding a model that estimated for significant variation in several important dairy production traits. MAPK signalling pathway by involving in cell proliferation, affects hyperplastic growth (Chang, 2007) and associated with residual feed intake trait (Rolf *et al*. 2012). The annotated genes involved in the third cluster of the network N2 that was shown in red (Figure 5) did not activate a pathway significantly. However, *HSPG2*, one of the genes involved in this cluster, is included in the inhibition of the matrix metalloproteinase pathway (Do *et al*. 2017). Key roles of matrix metalloproteinases include the regulation of mammary epithelial cell functions, cell proliferations, and cell differentiations (Uria *et al*. 1997).

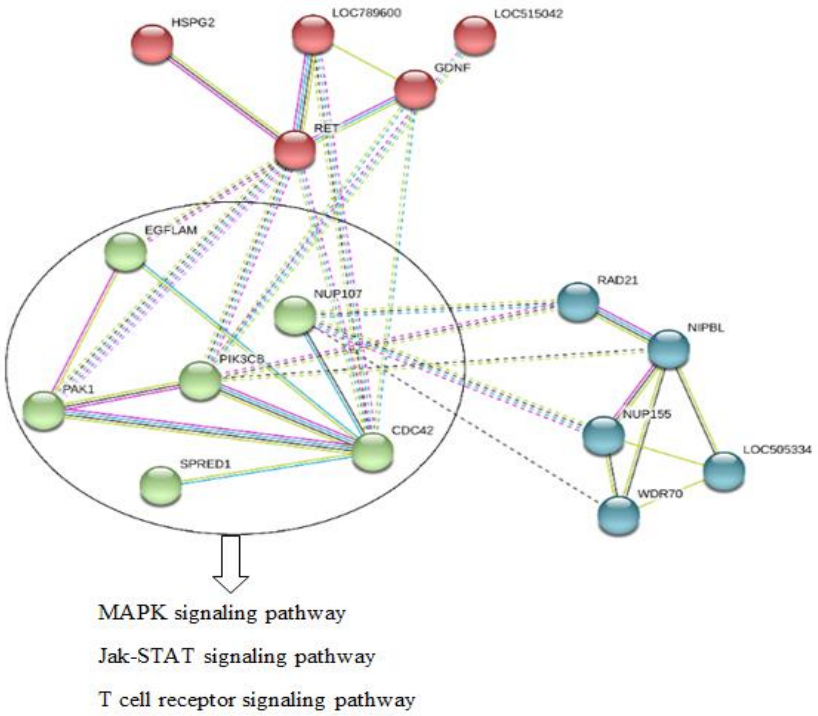

Figure 5: Network N2, interaction between the identified genes for Holstein population.

The main focus in this study was to detect recent signatures of selection in the DSN and Holstein populations. DSN is the founder breed of the modern German Holstein population, but with a breeding focus on different traits. In contrast to German Holsteins, DSN is a dual-purpose breed considering both trait categories meat and milk. The divergent breeding goals might contribute to pronounced genomic differentiation between both breeds. None of the identified regions under recent selection in the German Holstein population overlapped with identified regions in the DSN population. Hence, such findings refer to the fact that recent selection affected different loci in these two populations. On the other hand, the average F_{ST} between German Holsteins and DSN was 0.068 ± 0.051 (not shown), indicating a close relationship in the past. The F_{ST} and XP-EHH are two approaches reflecting population differentiation, but these measurements are complementary from a time scale perspective (Sabeti *et al*. 2007): XP-EHH mostly detects the recent positive selection signals, and F_{ST} selection signals from the past.

East-DSN versus West-DSN. XP-EHH values are plotted against genomic locations for East-DSN and West-DSN sub-populations (Figure 6). As indicated in Figure 6, a threshold for the top 0.1 percentile for negative or positive XP-EHH scores was defined. Signals beyond this threshold were

studied in detail, i.e., selection signals on different chromosomes (1, 4, 5, 6, 7, 8 10, 11, 12, 13, 14, 16, 17, 20 and 21). Negative XP-EHH scores reflect selection events in the East-DSN population. The most extended selection signature for East-DSN was identified on chromosome 6 at 92.97 to 94.43Mb, spanning a segment of almost 1.46 Mb. Interestingly, this region harbours quantitative trait loci (QTL) for several economically important traits, especially milk and meat quality traits. For instance, Buitenhuis *et al*. (2016) detected several significant QTLs associated with milk kappa-casein percentage in Danish Holsteins. In this chromosomal region, Cai *et al*. (2018) identified a QTL influencing milk fat yield in Nordic Holstein cattle. Surprisingly, Mateescu *et al.* (2017) detected two significant SNPs associated with marbling score in the same region in Angus cattle and Michenet *et al*. (2016) detected significant SNPs affecting body weight (weaning) in beef cattle. Furthermore, the extended selection signatures on chromosome 16 spanning a chromosomal segment from 4.93 to 5.25 Mb harbored several QTLs for body weight as well as carcass quality. For example, McClure *et al*. (2010) detected several QTLs that spanned this region associated to 12th rib fat thickness, marbling score and body weight (birth, weaning and mature) in Angus cattle. The regions under positive selection, and the annotated genes in a window of 250 Kb downstream and upstream of each core SNP, are listed in Table 3.

Table 3: The regions under positive selection in the East-DSN sub-population and the annotated potential candidate genes within a window of 250 Kb downstream and upstream of each core SNP.

Chr.	Region	No. of core SNP	Candidate Gene
6	92.97-94.43	28	*ART3, NUP54, SCARB2, STBD1, SHROOM3, SEPT11, CCNI, CCNG2, CXCL13, CNOT6L, MRPL1*
11	82.83	1	-
13	45.84, 34.16	2	*ZEB1*
15	41.78,		*GALNT18*
16	4.93, 5.25, 66.22	3	*C16H1orf116, YOD1, PFKFB2, C4BPA, SMG7, NCF2, ARPC5, APOBEC4*
21	25.09-25.56	2	*SH3GL3, RASGRF1, CTSH, MORF4L1*

Associations of some those genes with dairy cattle breeding goal traits have been detected in previous studies. For example, *YOD1* defined as a regulated gene in longissimus muscle of finishing

cattle by involving in protein processing in endoplasmic reticulum (Lee *et al*. 2017). Furthermore, *PFKFB2* was identified as a candidate genes related to lipid and phospholipid metabolism (Buchanan *et al*. 2016). In fact, *PFKFB2* plays an important role in synthesis of a regulatory molecule (fructose-2, 6-bisphosphate) which is involved in glycolysis (Hue and Rider, 1987). McClure *et al*. (2010) reported a QTL in this region associated with 12th rib fat thickness in Angus. Interactions between the annotated genes were illustrated in network N3 (Figure 7). As it is shown in the figure, the network was divided in three clusters by using k-means clustering algorithm.

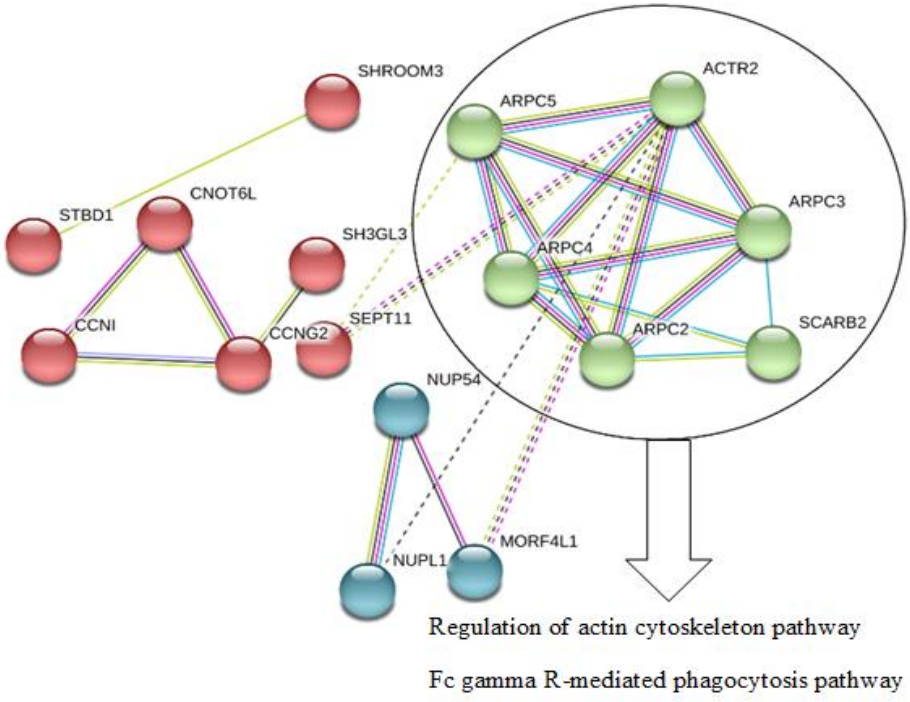

Figure 7: Network N3, interaction between the identified genes for East_DSN sub-population.

A threshold for the top 0.1 percentile for positive XP-EHH scores identified the signals of selection, being possible targets of recent selection in the West-DSN population (Figure 6).

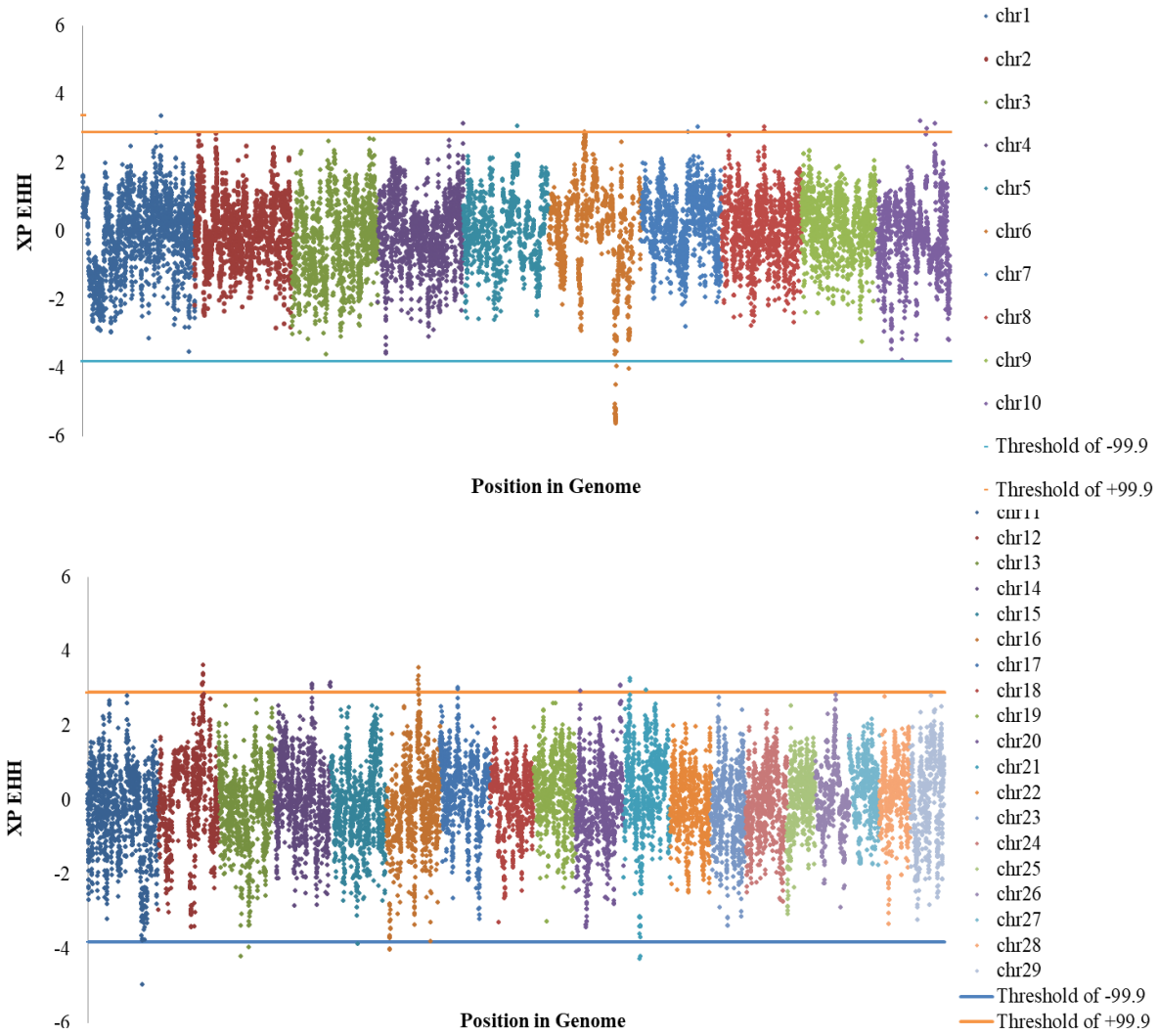

Figure 6: XP-EHH scores for each SNP for the West-DSN (positive values) and East-DSN sub-populations (negative values).

The extreme positive values of XP-EHH are located on different chromosomes (on chromosme1 (at 110.72 Mb), on chromosome 4 (at 120.64 Mb), on chromosome 5 (at 76.33 Mb), on chromosome 7 (at 80.27 Mb), on chromosome 8 (between 61.35 to 61.45 Mb), on chromosome 10 (between 62.40 to 82.57 Mb), on chromosome 12 (between 67.07 to 67.95 Mb), on chromosome 14 (in two regions at 57.63 to 57.87 and 84.55 to 84.61), on chromosome 16 (between 47.30 to 48.45), on chromosome 17 (between 26.39 to 26.41), on chromosome 20 (in two regions at 6.69 and at 67.01 to 67.49) and on chromosome 21 (in two regions at 9.74 to 9.78 Mb and 34.06 Mb)).

Table 4: The regions under positive selection in the West-DSN sub-population and the annotated potential candidate genes within a window of 250 Kb downstream and upstream of each core SNP

Chr.	Region	No. of core SNP	Candidate Gene
1	110.72	1	*RF00100*
4	120.64	1	*VIPR2*
5	76.33	1	*CYTH4, ELFN2, MFNG, USP18, ALG10*
7	80.27	1	*RFAM*
8	61.35-61.45	2	*MELK*
10	62.40, 71.69, 82.57	3	*DUT, RF00139, SLC12A1*
12	67.07-67.95	7	*GPC5*
14	57.63-57.87, 84.55-84.61	7	*TRHR, TMEM74, SNTB1*
16	47.30- 48.45	5	*DNAJC11, THAP3, PHF13*
17	26.39-26.41	2	*RF00003*
20	6.69, 67.01-67.49	3	*GFM2, NSA2, FAM169A,*
21	9.741-9.78, 34.06	3	*SIN3A,*

The extended selection signatures in the West-DSN sub-population that located on chromosome 8, 12, 16, 17 and 21 harbour several known QTLs mostly associated to body weight, carcass quality and milk composition. For instance, several QTL, associated with body weight, identified at chromosome 8 (Michenet *et al*. 2016), at chromosome 17 (MacNeil & Grosz, 2002) and at chromosome 21(McClure *et al.* 2010) which are in close distance with the region under positive selection in the current study. Accordingly, McClure *et al*. (2010) identified two QTL at chromosome 17 and 21 and MacNeil & Grosz (2002) detected a QTL at chromosome 16 associated with fat thickness at the 12th rib which strongly overlapped with the region under positive selection in the current study. In addition, Gutierrez-Gil *et al*. (2008) found a QTL on chromosome 16 at 22.5-49.5 Mb associated with juiciness and McClure *et al.* (2010) detected a QTL on chromosome 17 at 22.6-27.1 Mb associated with marbling score. Regarding to QTL associated with milk composition, Schopen *et al.* (2009) identified two QTL associated with milk alpha and kappa casein percentage respectively on chromosome 17 and 21 overlapped with identified signatures in the same regions of the current study. The genes of interest in these genomic regions are shown in Table 4, and their interactions are illustrated in network N4 (Figure 8).

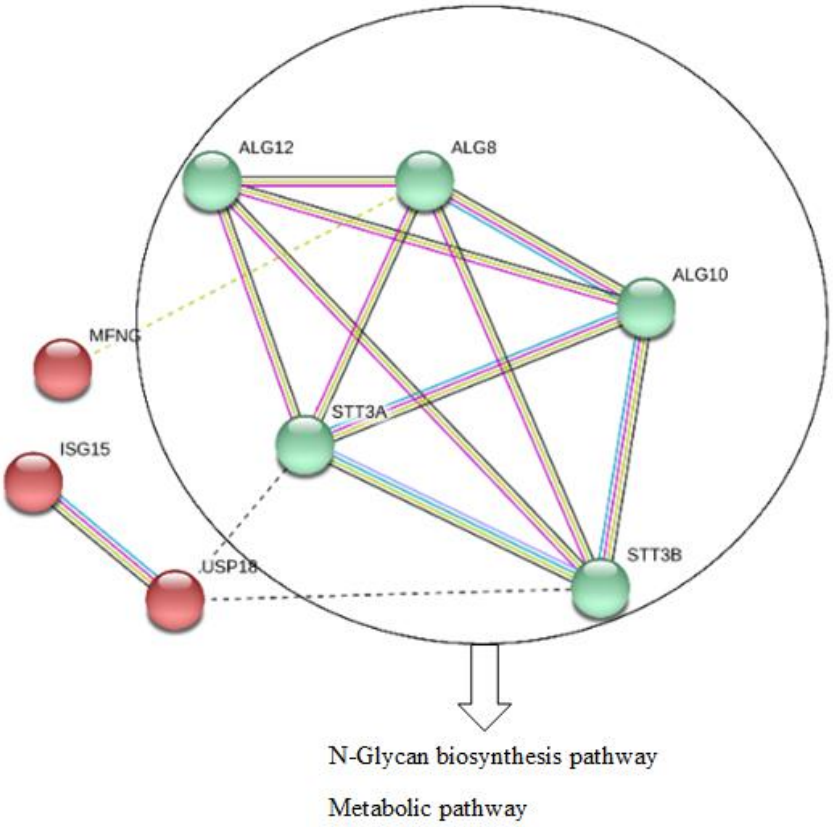

Figure 8: Network N4, interaction between the identified genes for West_DSN sub-population.

Due to low number of annotated genes, the network was divided in two clusters. The first cluster that has been illustrated in red did not activate any pathway and *USP18* was the hub gene of the network. *USP18* is involved in many biological pathways in various cell types and various immunological processes (Honke *et al*. 2016). Lindholm-Perry *et al.* (2016) detected that *USP18* gene expressed only in animals with low intake/high gain phenotypes and involving in proteolysis and hydrolase functions in beef cattle. Furthemore, Lee *et al.* (2015) detected *USP18* gene expressions differentially between low and high residual feed intake chicken population and involved in cell death and protein synthesis biological pathway. Furthermore, Magalhães *et al*. (2016) reported that *USP18* associated with marbling in Nellore Cattle. They reported influence of this gene in relation to carbohydrate and lipid metabolism. In fact interspersed fat deposition in muscle is consequent of lipid metabolism which derives from fat consumption and abundance

carbohydrate that reserve in the fat form. Two gene families of *ALG* and *STT* were involved in the second cluster of the network N4 that has been illustrated in green. Those genes activated the N-Glycan biosynthesis and metabolic pathway significantly. Palombo *et al*. (2018) confirmed that *ALG12* by involving in Glycan biosynthesis pathway considered as a best candidate gene associate with long-chain fatty acid trait (C18:1 trans-6–8) in Italian Holstein cows.

The identified regions under recent selection in the West-DSN sub-population was not overlapped with the detected regions in East-DSN sub-population and only chromosome 16 and 21 (but with different chromosomal segments) contributed to selection signatures in both West and East sub-populations.

In the current study, we focused on the identification of recent footprints being characteristics for the West-DSN and East-DSN sub-populations after the Second World War separation. During the separation, despite the close genetic relationships, the production environments as well as breeding goals in both German sub-populations were substantially different. Interestingly, the XP-EHH scores revealing recent selection signals were quite high, while the F_{ST} value between these two populations was very low (average of differentiation ~ 0.0085 ± 0.0024). The results underline the close relationship between these two populations in the past.

Sick versus healthy Holstein cows. XP-EHH scores for each SNP for healthy and sick Holstein cows are shown in Figure 9. The targets of recent selection are detected through a threshold of the top 0.1 percentile for negative or positive XP-EHH scores. The top 0.1 percentile for positive XP-EHH values reflects selection signals in the healthy sub-population. The most extended selection signature showing recent or still undergoing strong positive selection in the healthy sub-population was identified on chromosomes 16, 22 and 23. The positive selection signature extended 0.63 Mb on chromosome 16 at 26.99 to 27.62 Mb, on chromosome 22 at 46.85 to 49.25 Mb, and on chromosome 23 at 48.99 to 49.76 Mb (Figure 9). Smetko *et al*. (2015) found a pronounced population differentiation between local African breeds via F_{ST} values in the same region on chromosome 16. The population differentiation was also related to trypanosomosis susceptibility. Cole *et al*. (2011) identified a QTL associated with claw health indicator traits, i.e., foot angle rear leg quality, in this segment on chromosome 16. Supporting the health indicator aspect, Gernand *et al*. (2013) reported strong positive correlations between breeding values for foot angle and for rear leg side view with breeding values for DD. Kolbehdari *et al*. (2008) identified a significant SNP associated with bone quality, in close distance to our region of interest on chromosome 22. Gautier

et al. (2009) identified a region on chromosome 22 at 43.79 to 53.04 Mb, including a large number of SNP overlapping with relevant regions in the current study. Such overlaps might be due to the preservation of several haplotypes, which contain variants under selection in different populations.

The genomic regions with extreme peaks for XP-EHH for the healthy sub-population harboured the genes as listed in Table 5.

Table 5: The regions under positive selection in the healthy Holstein sub-population and the annotated potential candidate genes within a window of 250 Kb downstream and upstream of each core SNP

Chr.	Region	No. of core SNP	Candidate Gene
16	26.99-27.62	11	*TAF1A*, *MIA3*, *AIDA*, *BROX*, *AUH*, *TLR5*, *SUSD4*, *CCDC1*, *CAPN8*
22	46.85-49.25	22	*CACNA1D*, *CACNA2D3*, *LRTM1*, *SELENOK*, *ACTR8*, *IL17RB*, *CHDH*
23	48.99-49.76	6	*F13A1*, *NRN1*, *FARS2*, *LYRM4*, *PPP1R3G*, *RPP40*, *CDYL*

Interactions between identified genes are illustrated in network N5 (Figure 10). First cluster of the network (in red) was included 7 genes and *ACTR8* was identified as hub gene of the cluster. Nevertheless, those genes did not activate a pathway significantly. The second cluster of the network that was shown in green involved 6 genes and *FARS* family genes were the hub gene of the cluster. *FARS2* was detected by McCarthy *et al*. (2010) as one of the "down-regulated genes" in high-yielding Holstein dairy cows during the energy deficiency period in early lactation. We hypothesize strong associations between energy deficiency and the occurrence of claw disorders. In this regard, Collard *et al*. (2000) found a significant correlation between energy balance traits and laminitis. Boettcher *et al*. (1998) identified lameness as the most important disease for cows with a negative energy balance during the first 50 days in milk. With a focus on a more detailed disease description, lameness was mostly due to DD (Refaai *et al*. 2013). The annotation list is enriched with genes of biological interest, being involved in wound healing pathways. According to Shearer *et al*. (2015), these genes could be indirectly related to DD, because claw lesions in dairy cattle

follow a common wound healing process. Wound healing implies a rapid unrestricted re-epithelialization of the ulcer, but inflammatory cells might contribute to the non-healing nature of severe DD cases (Wilson-Welder *et al.* 2015).

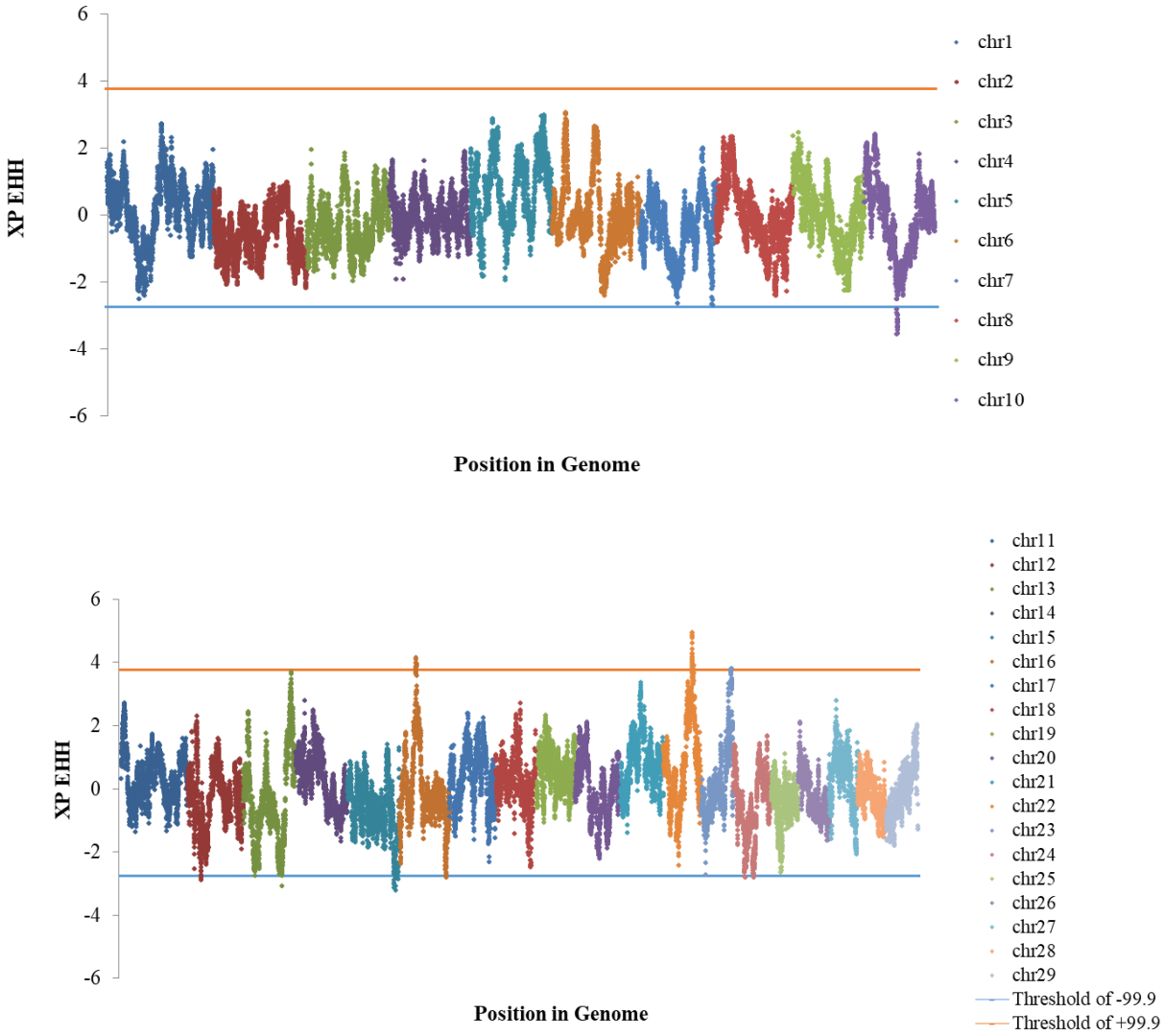

Figure 9: XP-EHH scores for each SNP for the healthy (positive values) and sick Holstein sub-populations (negative values).

The third cluster of the network N5 that was shown in blue involved 4 genes and *CACNA2D3* was the hub gene of the cluster. The annotated genes involving in this cluster activated

MAPK signalling and Oxytocin signalling pathway significantly. So far, no apparent functions of most of these genes as well as the pathways related to DD were reported.

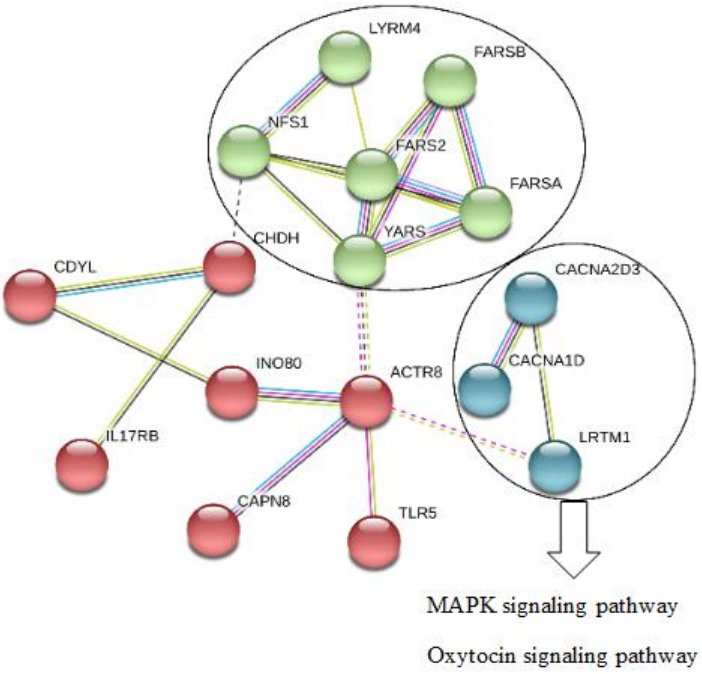

Figure 10: Network N5, interaction between the identified genes for Healthy population.

Nevertheless, the association of MAPK signalling pathway with residual feed intake trait in Angus cattle was confirmed by Rolf *et al*. (2012). A negative genetic correlation between feed efficiency and incidence of claw disorder was reported in Danish Red population (Wassmuth *et al*. 2010); where residual feed intake can be a useful indicators of feed efficiency in dairy breeding programmes (Manafiazar *et al*. 2016).

The detected selection signals in the sick German Holstein sub-population corresponds to extreme negative XP-EHH values (Figure 9). These negative values were detected on chromosome 10 (at 48.54 to 49.74 Mb), on chromosome 12 (at 22.21 to 23.06 Mb), on chromosome 13 (at 62.53 to 63.18 Mb), on chromosome 15 (at 76.86 to 82.76 Mb), on chromosome 16 (at 75.49 to 75.86 Mb), and on chromosome 24 (at 18.78 to 19.11 Mb). The most obvious effect was on chromosome 15, spanning a segment of almost 5.8 Mb, and harbouring the genes *SLC35C1*, *CRY2* and *MAPK8IP1*. The three genes *RORA*, *ICE2* and *ANXA2* are located in a 1.2 Mb segment on chromosome 10. The

detected region on chromosome 16 at 75.49 to 75.86 Mb encompasses several genes: *DIEXF, IRF6, C1orf74, TRAF3IP3, HSD11B1, G0S2, LAMB3* and *CAMK1G.*

The selection signature regions for the sick sub-population include several QTL, which are associated with claw disorders, and with feet and leg conformation traits. In this regard, in Holstein dairy cattle, van der Spek *et al*. (2015) and Wu *et al*. (2016) identified highly significant SNP on chromosome 10 associated with claw disorders, and with feet and legs disorders, respectively. Buitenhuis *et al*. (2007) detected a QTL on chromosome 15 at 78.0 Mb for bone quality, and in the same marker bracket a QTL for hock quality. A moderate genetic correlation between bone qualities with our target disease trait DD was reported in some previous studies (e.g., Onyiro *et al*. 2008).The regions under positive selection and the annotated genes in a window of 250 Kb downstream and upstream of each core SNP are presented in Table 6.

ICE2, TRAF3IP3, MAPRE1, COG6, ANXA2 are the annotated genes that mostly involved in the regulation of biological processes, in particular in the regulation of organelle and in cellular component organizations. Peng *et al.* (2015) found that *TRAF3IP3* is expressed in immune organs, promoting the immune response. Dong *et al.* (2017) recommended *MAPRE1* as a potential biomarker to infer pathological mechanism. The interactions between the annotated genes are shown in network N6 (Figure 11). The genes involved in the first two clusters of the network (in red and in blue) did not activate a pathway significantly but the third cluster (in green) which included 6 genes, activated circadian rhythm pathway significantly. *CRY1* was identified as hub gene of this cluster and regulates circadian rhythm by encoding flavin adenine dinucleotide-binding protein which is a key component of the circadian core oscillator complex (Wang *et al*. 2015). Interestingly, overexpression of *CRY1* is associated with lower blood glucose concentrations (Zhang *et al*. 2010) and animals with glucose concentrations out of the reference range show more claw disorder in first two month of lactation (Wilhelm *et al.* 2017).

None of the identified regions under recent selection in the sick German Holstein sub-population overlapped with identified regions in the healthy sub-population. Such finding refers to the fact that selection affected different loci in both sub-populations. Only chromosome 16 (but with different chromosomal segments) contributed to selection signatures in both healthy and sick sub-populations.

Table 6: The regions under positive selection in the sick Holstein sub-population and the annotated potential candidate genes within a window of 250 Kb downstream and upstream of each core SNP.

Chr.	Region	No. of core SNP	Candidate Gene
10	48.54-49.74	18	*RORA, ICE2, ANXA2*
12	22.21-23.06	3	*FOXO1, COG6, LHFP, SNORA48, NHLRC3, PROSER1, STOML3*
13	62.53-63.18	2	*KIF3B, ASXL1, NOL4L, NOL4L, COMMD7, DNMT3B, MAPRE1, SUN5, BPIFB2*
15	76.86-82.76	11	*SLC35C1, CRY2, MAPK8IP1, C11orf94, PEX16, LARGE2, PHF21A, CREB3L1, DGKZ*
16	75.49-75.86	2	*DIEXF, IRF6, C1orf74, TRAF3IP3, HSD11B1, G0S2, LAMB3,CAMK1G*
24	18.78-19.11, 33.30	4	*LAMA3, NKRD29, NPC1, C24H18orf8, IOK3, TMEM241*

CONCLUSION

Application of XP-EHH methodology in created diverse cattle sub-populations successfully identified several putative selection signature regions, harbouring genes or QTL being associated with disease and production traits. For instance, *CLU* was detected as a hub gene in the DSN population, and is one the most promising potential candidate genes affecting milk protein concentration. *USP18* was identified as a candidate gene in the west-DSN sub-population, and is associated with marbling score by involving in proteolysis and hydrolase functions in beef cattle as well as lipid and carbohydrate metabolism. Interestingly, *FARS2*, known as one of the most important “down-regulating genes” in high-yielding Holstein dairy cows with a negative energy status, was detected in the healthy German Holstein sub-population. The detected regions are worthy candidates for further investigations. Nevertheless, some of the detected selection signature regions were unrelated with genes that are relevant regarding adaptation to harsh environments, productivity or disease susceptibility.

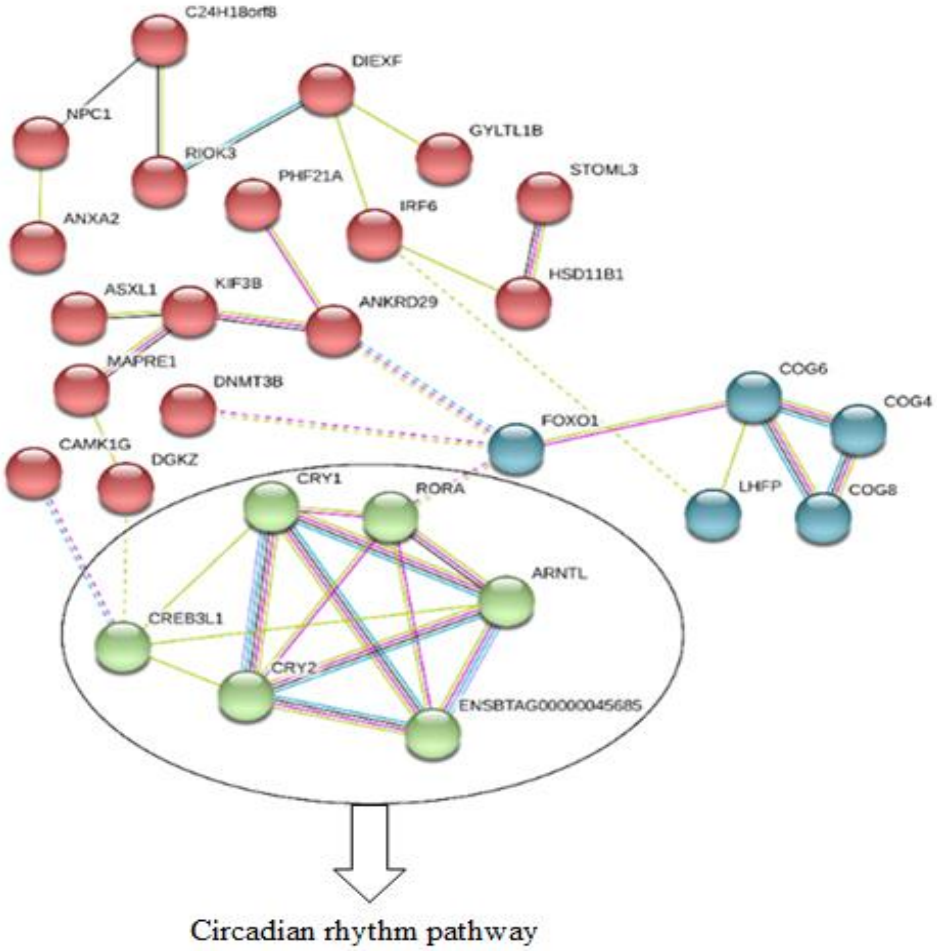

Figure 11: Network N6, interaction between the identified genes for Sick population.

ACKNOWLEDGEMENT

For the DSN-study, the authors acknowledge the financial support for this project provided by transnational funding bodies, being partners of the FP7 ERA-net project, CORE Organic Plus, and the cofund from the European Commission. Regarding the German Holstein genotypes,
the authors acknowledge funding from the German Federal Ministry of Education and Research (BMBF) and from the Förderverein Bioökonomieforschung e.V. (FBV) / German Holstein Association (DHV) for the collaborative project "KMU-innovativ-10: Kuh-L – cow calibration groups for the implementation of selection strategies based on high-density genotyping in dairy cattle", grant no. 031A416C.

REFERENCES

Arun S. J., Thomson P. C., Sheehy P. A., Khatkar M. S., Raadsma H. W. & Williamson P. (2015) Targeted analysis reveals an important role of JAK-STAT-SOCS genes for milk production traits in Australian dairy cattle. Front Genet, 6, 342-349.

Boegheim J. M., Leegwater P. A. J., van Lith H. A. & Back W. (2017) Invited review, Current insights into the molecular genetic basis of dwarfism in livestock. The Veterinary Journal, 224, 64-75.

Boettcher P. J., Dekkers J.C.M., Warnick L. D. & Wells S. J. (1998) Genetic Analysis of Clinical Lameness in Dairy Cattle. Journal of dairy science, 81, 1148–1156.

Brade V. W. & Brade E. (2013) Zuchtgeschichte deutschen Holsteinrinder. Berichte über Landwirtschaft - Zeitschrift für Agrarpolitik und Landwirtschaft. ISSN 2196-5099. Available at http://buel.bmel.de

Buchanan J. W., Reecy J. M., Garrick D. J., Duan Q., Beitz D. C., Koltes J. E., Saatchi M., Koesterke L. & Mateescu R. G. (2016) Deriving gene networks from SNP associated with triacylglycerol and phospholipid fatty acid fractions from ribeyes of angus cattle: Frontiers in genetics, 7, 116.

Buitenhuis A. J., Lund M. S., Thomasen J. R., Thomsen B., Nielsen V. H., Bendixen C. & Guldbrandtsen B. (2007) Detection of Quantitative Trait Loci Affecting Lameness and Leg Conformation Traits in Danish Holstein Cattle. Journal of dairy science, 90, 472–481.

Buitenhuis B., Poulsen N. A., Gebreyesus G. & Larsen L. B. (2016) Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle: BMC genetics,17, 114.

Cai Z., Guldbrandtsen B., Lund M. S. & Sahana G. (2018) Dissecting closely linked association signals in combination with the mammalian phenotype database can identify candidate genes in dairy cattle: BMC genetics, 19, 30.

Cesar A. S. M., Regitano L. C. A., Mourão G. B., Tullio R. R., Lanna D. P. D. & Nassu R. T. et al. (2014) Genome-wide association study for intramuscular fat deposition and composition in Nellore cattle. In BMC genetics, 15, 39.

Chang K. C. (2007) Key signalling factors and pathways in the molecular determination of skeletal muscle phenotype: Animal, 1, 681–98.

Chen M., Pan D., Ren H. Fu J., Li J., Su G., Wang A., Jiang L., Zhang Q. & Li J. F.. (2016) Identification of selective sweeps reveals divergent selection between Chinese Holstein and Simmental cattle populations. In Genetics, selection, evolution : GSE, 48,76.

Cole J. B., Wiggans G. R., Ma L., Sonstegard T. S., Lawlor T. J., Crooker B. A., van Tassell C. P., Yang J., Wang S., Matukumalli L. K. & Da Y. (2011) Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC genomics, 12, 408.

Collard B. L., Boettcher P. J., Dekkers J.C.M., Petitclerc D. & Schaeffer L. R. (2000) Relationships Between Energy Balance and Health Traits of Dairy Cattle in Early Lactation. Journal of dairy science, 83, 2683–2690.

Do D. N., Li R., Dudemaine P. L. & Ibeagha-Awemu E. M. (2017) MicroRNA roles in signalling during lactation. An insight from differential expression, time course and pathway analyses of deep sequence data. Scientific reports, 7, 44605.

Dong L. Y., Zhou W. Z., Ni J. W., Xiang W., Hu W. H., Yu C.& Li H. Y. (2017): Identifying the optimal gene and gene set in hepatocellular carcinoma based on differential expression and differential co-expression algorithm. Oncology reports, 37, 1066–1074.

Egger-Danner C., Schwarzenbacher H. & Willam A. (2014): Short communication: Genotyping of cows to speed up availability of genomic estimated breeding values for direct health traits in Austrian Fleckvieh (Simmental) cattle — Genetic and economic aspects. Journal of dairy science, 97, 4552–4556.

Eleftherohorinou H., Wright V., Hoggart C., Hartikainen A. L., Jarvelin M. R., Balding D., Coin L. &Levin M. (2009) Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PloS one, 4, e8068.

Evangelou M., Dudbridge F. & Wernisch L. (2014) Two novel pathway analysis methods based on a hierarchical model. Bioinformatics (Oxford, England), 30, 690–697.

Gautier M., Flori L., Riebler A., Jaffrézic F., Laloé D., Gut I., Moazami-Goudarzi K. & Foulley J. L. (2009) A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. BMC genomics, 10, 550.

Gernand E., Döhne D. A. & König S. (2013) Genetic background of claw disorders in the course of lactation and their relationships with type traits. Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie, 130, 435–444.

Gholami M., Erbe M., Gärke C., Preisinger R., Weigend A., Weigend S. & Simianer H. (2014) Population genomic analyses based on 1 million SNPs in commercial egg layers. PloS one, 9, e94509.

Glick, G.; Shirak, A.; Uliel, S.; Zeron, Y.; Ezra, E.; Seroussi, E. Ron M. & Weller J. I. (2012) Signatures of contemporary selection in the Israeli Holstein dairy cattle. Animal genetics, 43, Suppl 1, 45–55.

Gouveia J. J. S., Paiva S. R., McManus C. M., Caetano A. R., Kijas J. W., Facó O., Azevedo H. C., Araujo A. M., Souza C. J. H., Yamagishi M. E. B., Carneiro P. L. S., Lôbo R. N. B., de Oliveira S. M. P. & da Silva M. V. G. B. (2017) Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds. Livestock Science, 197, 36–45.

Guenette R. S., Corbeil H. B., Léger J., Wong K., Mézl V., Mooibroek M. & Tenniswood M. (1994) Induction of gene expression during involution of the lactating mammary gland of the rat. Journal of molecular endocrinology, 12, 47–60.

Gutiérrez-Gil B., Wiener P., Nute G. R., Burton D., Gill J. L., Wood J. D. & Williams J. L. (2008) Detection of quantitative trait loci for meat quality traits in cattle: Animal genetics, 39, 51-61.

Gutiérrez-Gil B., Arranz J. J. & Wiene, P. (2015) An interpretive review of selective sweep studies in Bos taurus cattle populations. Identification of unique and shared selection signals across breeds. Frontiers in genetics, 6, 167.

Hill A. V. S. (2006) Aspects of genetic susceptibility to human infectious diseases. Annual review of genetics, 40, 469–486.

Holsinger K. E., Weir B. S. (2009) Genetics in geographically structured populations. Defining, estimating and interpreting F(ST). Nature reviews. Genetics 10 (9), 639–650.

Honke N., Shaabani N., Zhang D. E., Hardt C. & Lang K. S. (2016) Multiple functions of USP18: Cell death & disease, 7, e2444.

Hue L. & Rider M. H. (1987). Role of fructose 2,6-bisphosphate in the control of glycolysis in mammalian tissues: Biochem. J. 245, 313–324.

Humphreys D. T., Carver J. A., Easterbrook-Smith S. B. & Wilson M. R. (1999) Clusterin has chaperone-like activity similar to that of small heat shock proteins. The Journal of biological chemistry, 274, 6875–6881.

Jaeger M., Brügemann K. & König S. (2016) Genetic relationships and trait comparison between and within selected lines of local dual-purpose cattle. 67th annual meeting of the European Association for Animal Production, Belfast

Jiang L., Liu X., Yang J., Wang H., Jiang J., Liu L., He S., Ding X., Liu J. & Zhang Q. (2014) Targeted resequencing of GWAS loci reveals novel genetic variants for milk production traits. BMC genomics, 15, 1105.

König S., Lessner S. & Simianer H. (2007). Application of controlling instruments for improvements in cow sire selection. Journal of Dairy Science, 90, 1967-1980.

Kolbehdari D., Wang Z., Grant J. R., Murdoch B., Prasad A., Xiu Z., Marques E., Stothard P., Moore S. S. (2008) A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. Journal of dairy science, 91, 2844–2856.

Koltes J. E., Mishra B. P., Kumar D., Kataria R. S., Totir L. R., Fernando R. L., Cobbold R., Steffen D., Coppieters W., Georges M., & Reecy J. M. (2009) A nonsense mutation in cGMP-dependent type IIprotein kinase (PRKG2) causes dwarfism in American Angus cattle. PNAS, 106, 19250-19255.

Kreitman M. (2000): Methods to detect selection in populations with applications to the human. In Annual review of genomics and human genetics, 1, 539–559.

Lee H. J., Kim J., Lee T., Son J. K., Yoon H. B., Baek K. S., Jeong J. Y., Cho Y. M., Lee K. T., Yang B. C., Lim H. J., Cho K., Kim T. H., Kwon E. G., Nam J., Kwak W., Cho S. & Kim H. (2014) Deciphering the genetic blueprint behind Holstein milk proteins and production. Genome biology and evolution, 6, 1366–1374.

Lee J., Karnuah A. B., Rekaya R., Anthony N. B. & Aggrey S. E. (2015) Transcriptomic analysis to elucidate the molecular mechanisms that underlie feed efficiency in meat-type chickens: Molecular genetics and genomics, 290, 1673–1682.

Lee S., Park S. J., Cheong J. K., Ko J. Y., Bong J. & Baik M. (2017) Identification of circulating miRNA involved in meat yield of Korean cattle: Cell biology international, 41, 761–768.

Li C., Cai W., Zhou C., Yin H., Zhang Z., Loor J. J., Sun D., Zhang Q., Liu J. & Zhang S. (2016) RNA-Seq reveals 10 novel promising candidate genes affecting milk protein concentration in the Chinese Holstein population. Scientific reports, 6, 26813.

Lindholm-Perry A. K., Butler A. R., Kern R. J., Hill R., Kuehn L. A., Wells J. E., Oliver W. T., Hales K. E.,Foote A. P. & Freetly H. C. (2016) Differential gene expression in the duodenum,

jejunum and ileum among crossbred beef steers with divergent gain and feed intake phenotypes: Animal genetics, 47, 408–427.

MacEachern S., Hayes B., McEwan J. & Goddard M. (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics, 10, 181.

MacNeil M. D. & Grosz M. D. (2002) Genome-wide scans for QTL affecting carcass traits in Hereford x composite double backcross populations: Journal of animal science, 80, 2316–2324.

Magalhães A. F. B.,de Camargo G. M. F., Fernandes G. A., Gordo D. G. M., Tonussi R. L., Costa R. B., Espigolan R., deO Silva R. M., Bresolin T., de Andrade W. B. F., Takada L., Feitosa F. L. B., Baldi F., Carvalheiro R., Chardulo L. A. L. & de Albuquerque L. G. (2016) Genome-wide association study of meat quality traits in nellore cattle: PloS one, 11, e0157845.

Maiorano A. M., Lourenco D. L., Tsuruta S., Ospina A. M. T., Stafuzza N. B., Masuda Y., Filho, A. E. V., Cyrillo, J. N. D. G., Curi R. A. & Silva, J. A. V. (2018) Assessing genetic architecture and signatures of selection of dual purpose Gir cattle populations using genomic information: PloS one, 13, e0200694.

Manafiazar G., Goonewardene L., Miglior F., Crews D. H., Basarab J. A., Okine E.& Wang Z. (2016) Genetic and phenotypic correlations among feed efficiency, production and selected conformation traits in dairy cows: Animal, 10, 381–389.

Mateescu R. G., Garrick D. J.& Reecy J. M. (2017) Network analysis aeveals putative genes affecting meat quality in angus cattle: Frontiers in genetics, 8, 171.

McCarthy S. D., Waters S. M., Kenny D. A., Diskin M. G., Fitzpatrick R., Patton J., Wathes D. C. & Morris D. G. (2010) Negative energy balance and hepatic gene expression patterns in high-yielding dairy cows during the early postpartum period. A global approach. Physiological genomics, 42A, 188–199.

McClure M. C., Morsci N. S., Schnabel R. D., Kim J. W., Yao P., Rolf M. M., McKay S. D., Gregg S. J., Chapple R. H., Northcutt S. L. & Taylor J. F. (2010) A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle: Animal genetics, 41, 597–607.

McKusick V. A. (2007) Mendelian Inheritance in Man and its online version, OMIM. American journal of human genetics, 80, 588–604.

Michenet A., Barbat M., Saintilan R., Venot E. & Phocas F. (2016) Detection of quantitative trait loci for maternal traits using high-density genotypes of Blonde d'Aquitaine: BMC genetics, 17, 88.

Moradi M. H., Nejati-Javaremi A., Moradi-Shahrbabak M., Dodds K. G. & McEwan J. C. (2013) Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC genetics, 13,10.

Nielsen R., Hellmann I., Hubisz M., Bustamante C. & Clark A. G. (2007) Recent and ongoing selection in the human genome. Nature reviews. Genetics, 8, 857–868.

Onyiro O. M., Andrews L. J. & Brotherstone S. (2008) Genetic parameters for digital dermatitis and correlations with locomotion, production, fertility traits, and longevity in Holstein-Friesian dairy cows. Journal of dairy science, 91, 4037–4046.

Palombo V., Milanesi M., Sgorlon S., Capomaccio S., Mele M.,Nicolazzi E., Ajmone-Marsan P., Pilla F., Stefanon B.,& D'Andrea M. (2018) Genome-wide association study of milk fatty acid composition in Italian Simmental and Italian Holstein cows using single nucleotide polymorphism arrays: Journal of dairy science, 101, 11004-11019.

Peng S., Wang K., Gu Y., Chen Y., Nan X., Xing J., Cui Q., Ge Q. & Zhao H. (2015) TRAF3IP3, a novel autophagy up-regulated gene, is involved in marginal zone B lymphocyte development and survival. Clinical and experimental immunology, 182, 57–68.

Piantoni P., Wang P., Drackley J. K., Hurley W. L. & Loor J. J.(2010) Expression of metabolic, tissue remodeling, oxidative stress, and inflammatory pathways in mammary tissue during involution in lactating dairy cows. Bioinformatics and Biology Insights, 4, 85–97

Qanbari S., Gianola D., Ben H., Flavio S., Steve M., Moore S., Thaller G.& Simianer H. (2011) Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC genomics, 12, p. 318.

Qanbari S., & Simianer H. (2014) Mapping signatures of positive selection in the genome of livestock. In Livestock Science, 166, 133–143.

Refaai W., Ducatelle R., Geldhof P., Mihi B., El-shair M. & Opsomer G. (2013) Digital dermatitis in cattle is associated with an excessive innate immune response triggered by the keratinocytes. BMC veterinary research, 9, 193.

Rolf M. M., Taylor J. F., Schnabel R. D., McKay S. D., McClure M. C., Northcutt S. L., Kerley M. S. & Weaber R. L. (2012) Genome-wide association analysis for feed efficiency in Angus cattle: Anim Genet, 43, 367–374.

Rothammer S., Seichter D., Förster M. & Medugorac I. (2013) A genome-wide scan for signatures of differential artificial selection in ten cattle breeds. BMC genomics ,14, 908.

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

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

Sargolzaei M., Iwaisaki H. & Colleau J. J. (2006) CFC - A Software Package for Pedigree Analysis and Monitoring Genetic Diversity. User's Manual. Release 1.0.

Scheet P. & Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data. Applications to inferring missing genotypes and haplotypic phase. American journal of human genetics, 78, 629–644.

Schleinitz, D., Böttche, Y., Blüher M. & Kovacs P. (2014) The genetics of fat distribution. Diabetologia, 57, 1276–1286.

Schopen G. C., Heck J. M., Bovenhuis H., Visker M. H., van Valenberg H. J. & van Arendonk J. A. (2009) Genetic parameters for major milk proteins in Dutch Holstein-Friesians: Journal of dairy science, 92, 1182–1191.

Shearer J. K., Plummer P. & Schleining J. (2015) Perspectives on the treatment of claw lesions in cattle. Veterinary Diagnostic and Production Animal Medicine,6, 273-292.

Silkensen J. R., Schwochau G. B. & Rosenberg M. E. (1994) The role of clusterin in tissue injury. Biochemistry and cell biology = Biochimie et biologie cellulaire,72, 483–488.

Smetko A., Soudre A., Silbermayr K., Müller S., Brem G., Hanotte O., Boettcher P. J., Stella A., Mészáros G., Wurzinger M., Curik I., Müller M., Burgstaller J. & Sölkner J. (2015) Trypanosomosis. Potential driver of selection in African cattle. Frontiers in genetics, 6, 137.

Smith J. M., & Haigh J. (1974) The hitch-hiking effect of a favourable gene. In Genet. Res. 23 (1), 23-45.

Swanson K. M., Stelwagen K., Dobson J., Henderson H. V., Davis S. R., Farr V. C. & Singh K. (2009) Transcriptome profiling of Streptococcus uberis-induced mastitis reveals fundamental differences between immune gene expression in the mammary gland and in a primary cell culture model. Journal of dairy science, 92, 117–129.

Szklarczyk D., Morris J. H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N. T., Roth A., Bork P., Jensen L. J. & von Mering C. (2016) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible: Nucleic acids research, 45,362–368.

Uría J. A., Stahle-Bäckdahl M., Seiki M., Fueyo A. & López-Otín C. (1997) Regulation of collagenase-3 expression in human breast carcinomas is mediated by stromal-epithelial cell interactions. Cancer Research, 57, 4882-4888.

van der Spek D., van Arendonk J. A. M. & Bovenhuis H. (2015) Genome-wide association study for claw disorders and trimming status in dairy cattle. Journal of dairy science, 98, 1286–1295.

Voight B. F., Kudaravalli S., Wen X. & Pritchard J. K. (2006) A map of recent positive selection in the human genome. PLoS biology, 4, e72.

Wang K., Li M. & Hakonarson H. (2010): Analysing biological pathways in genome-wide association studies. Nature reviews. Genetics, 11, 843–854.

Wang Z., Huang J., Zhong J. & Wang G. (2012) Molecular cloning, promoter analysis, SNP detection of Clusterin gene and their associations with mastitis in Chinese Holstein cows. Molecular biology reports, 39, 2439–2445.

Wang M., Zhou Z., Khan M. J., Gao J.,& Loor J. J. (2015) Clock circadian regulator (CLOCK) gene network expression patterns in bovine adipose, liver, and mammary gland at 3 time points during the transition from pregnancy into lactation: Journal of dairy science, 98, 1–12.

Wassmuth R., Madsen D., Jensen P., & Jensen P. (2000) Genetic parameters of disease incidence, fertility and milk yield of first parity cows and the relation to feed intake of growing bulls: Acta Agric Scand(A), 50, 93-102.

Wilhelm K., Wilhelm J.,& Fürll M. (2017) Claw disorders in dairy cattle - an unexpected association between energy metabolism and sole haemorrhages: Journal of Dairy Research, 84, 54–60

Wilson-Welder J. H., Alt D. P. & Nally J. E. (2015) Digital Dermatitis in Cattle. Current Bacterial and Immunological Findings. Animals, 5, 1114–1135.

Wright S. (1949) The Genetical Structure of Populations. Annals of Human Genetics, 15, 323–354.

Wu X., Guldbrandtsen B., Lund M. S., Sahana G. (2016) Association analysis for feet and legs disorders with whole-genome sequence variants in 3 dairy cattle breeds. Journal of dairy science, 99, 7221–7231.

Yamaji D., Kang K., Robinson G. W., & Hennighausen L. (2013) Sequential activation of genetic programs in mouse mammary epithelium during pregnancy depends on STAT5A/B concentration: Nucleic Acids Res, 41, 1622–1636.

Yang S., Li X., Li K., Fan B. & Tang Z. (2014) A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. BMC genetics,15, 7.

Zhang E. E., Liu Y., Dentin R., Pongsawakul P. Y., Liu A. C., Hirota T., Nusinow D. A., Sun X., Landais S., Kodama Y., Brenner D. A., Montminy M.,& Kay S. A. (2010) Cryptochrome mediates circadian regulation of cAMP signaling and hepatic gluconeogenesis: Nature Medicine, 16, 1152–1156.

Zhao F., McParland S., Kearney F., Du L. & Berry D. P. (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genetics, selection, evolution : GSE, 47, 49.

Zheng X., Levine D., Shen J., Gogarten S. M., Laurie C., & Weir B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics: 28, 3326-3328.

5th Chapter

General Discussion

Preface and Overview

We studied a variety of factors and parameters that affect the accuracy of genomic predictions. To do so, we used random forest methodology (RF) as well as genomic BLUP (GBLUP) method with strong focus on training set design. In addition, single-step genomic BLUP (ssGBLUP) was applied for selected scenarios to estimate genomic breeding values. In the following chapters, selection signature through variation in linkage disequilibrium (LD) within and between dual-purpose black and white (DSN) and Holstein cattle populations was identified.

First a stochastic simulation was applied in chapter 2 to investigate:

- The effect of composition of the training set (based on incidence of diseased cows in the training set);
- The impact of genomic architecture of traits (number of quantitative trait loci (QTL), heritability, LD structure and marker panel density);
- The impact of the model choice (by comparing the RF methodology with the GBLUP method);
- The potential of random forest methodology to detect locations of the most important single nucleotide polymorphism markers (SNP) and how the locations overlapped with the true QTL.

A mimic design of cow training and testing sets based on the incidence of diseased cows in the training set was used, being the basis for genomic predictions for disease traits with real data.

In chapter 3, a large dataset of commercial herds and randomly selected cows with phenotypes for health traits was genotyped to study:

- The potential of RF to estimate GEBV and compare it with GBLUP and ssGBLUP estimates for disease traits such as claw disorder, clinical mastitis and female infertility;
- The effect of cow disease incidences in the training set on accuracies of GEBV;
- The impact of using either de-regressed proof (DRP) or pre-corrected phenotype (PCP) as response variable on accuracies of GEBV;
- The identification of SNP with large effect by using RF and to compare with a classical genome wide association study (GWAS) approach
- The annotated genes in a close distance of significant SNPs

In chapter 4, the detection of selection signature was studied to:

- manifest adaptive genetic variation between the DSN and German Holstein populations demonstrate adaptive genetic variation between population strata according to disease incidences and geographical characteristics
- Infer biological pathways of the annotated genes (the genes that overlapped with selection signatures).

Impact of disease incidences in the cow training set on accuracies of GEBV

Design and optimization of the training set in genomic selection is one of the most important factors that affect the accuracy of GEBV. Cow training set as used in this study, especially for disease traits, represent a small sub population of selected commercial herds from eastern part of Germany (chapter 3). Revealing the impact of cow training set characteristics is very crucial when predicting the disease probability of a genotyped female calf or heifer from a different sub-population (i.e., genotyped female calves and heifers in small scale herds from west-Germany with significant differences for mean of disease incidence from large scale farms of east-Germany (Gernand et al., 2012)). On the other hand, allocating a determined budget for genotyping individuals, the first question that arises is how to choose the best animals to genotype for setting up a training set and to maximize the accuracy of GEBV. Hozé et al., (2014) showed that the choice of animals to be kept as training in small populations has a substantial impact on population structure and genomic predictions accuracy. Criteria to build an optimal training set have been highlighted in several studies (Saatchi et al., 2011.; Clark et al., 2012; Pszczola et al., 2012; Guo et al., 2014; Wang et al., 2017). The results of this study for either simulation scenarios or real genotyped data (chapter 2 and 3), showed that when using phenotype as response variable the increase of the number of sick cows in the training population was associated with a substantial rise in prediction accuracies for both methods GBLUP and RF. The highest prediction accuracies were achieved when the percentage of sick animals allocated to training population was the same as the disease incidences for both populations of training set and the whole population. For example, in chapter 2, allocating 3200 of sick animals (80%) as training population, implied a disease incidence of 20%, which is equal to the population disease incidence of 20%. Moreover, in chapter 3, for clinical mastitis, allocating 1240 sick cows (70% of sick cows) in the training set implied a disease incidence of 24.8%, which was close to the disease incidence of 26.2% in whole genotyped cows. It corresponded to the highest GEBV accuracy for GBLUP. One explanation for this result refers to the increase in the genetic

relationships between animals in the training and testing sets for higher numbers of sick cows in training sets. For example, in chapter 2, in the scenario S_IV (i. e. h^2= 0.3, No. of QTL= 725 and 50K SNP chip), when 10% of sick animals were allocated to the training set, the genetic relationship between the two sets was 0.0005. This value increased to 0.008 when 80% of sick animals were allocated into the training set. The impact of genetic relationships between training and testing sets on prediction accuracies have been verified in several studies (Wientjes et al., 2013; Habier et al., 2010). Clark et al. (2012) reported that distant related training and testing set was associated with decrease in genomic prediction accuracies. Habier et al. (2010) studied the effect of relationship on genomic prediction accuracies in German Holstein cattle by splitting animals into groups based on their maximum relationships. The accuracy was evaluated in each group and showed that the decrease in relationship was associated with decrease in the mean accuracy of each group.

Another reason for the realized result in our study might be due to increase of genetic variance in the training set, by allocating the intermediate percentages of sick animals into it. The significant impact of increases in genetic variances of training sets on genomic predictions accuracy was explored in previous studies (Daetwyler et al., 2008; Nirea et al., 2012). Pszczola et al. (2012) reported that optimal training set should have a loose family structure by including animals with low average relationship into the training population. They reported that related animals might slightly explain the same part of variation; hence, the theoretical maximum accuracy can be realized when the individuals in the training set are unrelated and with no identical-by-state alleles. These authors concluded that optimal design of the training set depends on the desired breeding strategy of the breed and might be different from one application to another. Lee et al. (2017) showed that an optimal design of a training set for a homogenous population (i. e. within the same breed) consist of both close and distant relatives and unrelated individuals, and to maximize genomic prediction accuracy a composite design is preferred. In fact, a composite design benefits from the effective information from relatives while it also takes advantages from information of unrelated individual.

Impact of genetic architecture of traits on accuracies of GEBV

To investigate the impact of genetic architecture of the trait on the genomic prediction accuracy in chapter 2, the number of QTL, the level of LD and the marker density were modified as well as heritability of the traits.

In chapter 3, accuracy of GEBV for different disease traits with different genetic architecture and

heritability was also evaluated. The results in chapters 2 and 3 revealed considerably the effect of trait genetic architecture on genomic prediction accuracy. Several studies have reported the impact of genetic architecture, different densities of markers and heritability on genomic prediction accuracy (Daetwyler et al., 2010; Zhang et al., 2011; Wang et al., 2017).

Goddard (2009) developed a deterministic method to predict the genomic predictions accuracy. The parameters included in the formula were the number of phenotypic records in training set (n), the length of the genome (l), the heritability of the trait (h^2) and the QTL distribution effects. Assuming a normal distribution for QTL effect, the genomic prediction accuracy can be predicted based on the following equation:

$$R=\sqrt{\left[\,1-\lambda/(2n\sqrt{a})*\ln\left(\left(1+a+2\sqrt{a}\right)/(1+a-2\sqrt{a})\right)\right]}$$

where $a=1+2\lambda/n$, and $\lambda=qk/h^2$, with $k=1/log(2N_e)$, where N_e is the effective population size. The parameter q=number of independent chromosome segments in the population. The value of q used here was $2N_el$, where l is the length of the genome in Morgans. By using the same number of phenotypic records and the same heritabilities of the traits, the deterministic prediction of accuracies depends on the effective population size and the length of the genome. Furthermore,(Daetwyler et al., 2010) showed that number of independent chromosome segments (M_e) considerably affected the genomic prediction accuracy for GBLUP and BayesB. Small effective population size is associated with low number of independent segments in the genome and in the consequent fewer markers are needed to capture the effects of all segments (Goddard et al., 2010). Hence, the effective population size affects the extent of LD in a population (Wientjes et al., 2013). Guo et al. (2014) found that genomic prediction accuracy was affected by genomic heritability of training and testing set. They reported an increase in genomic heritability of training and testing sets, associated with a large improvement in predictions of marker effects due to higher genetic variations. Villumsen et al. (2009) showed that by decreasing heritability of the trait, the genomic prediction accuracy decreased due to less information coming from each phenotypic observation available to estimate haplotype effect. The result of this study was in line with literature where the impact of genetic architecture as well as the trait heritability on genomic prediction accuracy was significant. In addition, the effect of marker panel density on genomic prediction accuracy was revealed as well.

Impact of the model choice on accuracies of GEBV

Several statistical models have been applied to predict genomic breeding values using genome-wide SNP markers. GBLUP is one of the most widely used methods which is a linear mixed model incorporating a genomic relationship matrix (GRM). Low computation time and being simple as traditional BLUP is the reasons for the popularity of this method (Su et al., 2014). Over-parameterization due to small number of rows in the mixed model equation (number of genotyped cows) in relation to number of column in the MME (number of genetic markers) is the main problem of using linear models in genomic prediction. In consequent, it associates with a probably biased and less accurate prediction. Application of RF is a likely alternative to overcome the over-parameterization problem for such circumstances. The result of this study in chapter 2 revealed outperformance of RF over GBLUP in some scenarios depending on genetic architecture (no. of QTL, level of LD) and on heritability. Moreover, in chapter 3 the result showed the outperformance of GBLUP over RF in low percentage of sick cows in training set but by increasing number of sick cows in training set RF performed better. Applying ssGBLUP in chapter 3 was associated with the outperformance of ssGBLUP over RF and GBLUP. Several studies reported the superiority of ssGBLUP over GBLUP method (Gao et al., 2012; Lourenco et al., 2015). Including the phenotype and pedigree information of large number of cows with phenotype but without genotype could be the probably reason for this outperformance (Ashraf et al., 2016).

Assessing predictive ability using receiver operator characteristic (ROC) curves

Area under the receiver operator characteristic curves (AUC) is a further assessment criterion for genomic predictions. Plotting true positive and false positive events for all successive thresholds is the principle of AUC calculation to evaluate the efficiency of classification models in clinical studies. Sparingly, AUC reflect the probability of correct assignment of animals in the class of healthy or in the class of diseased just based on their genotype (Wray et al., 2010). By using DRP as a response variable, the calculated AUCs from 4 replicate and all the allocation schemes for claw disorders, clinical mastitis and infertility are shown in Figures 1, 2 and 3, respectively. By allocating larger number of sick cows to the training set, the AUC increased. The AUC for 10% of allocated scheme was 0.59 from GBLUP and for claw disorder while the AUC for the same trait from GBLUP application and 80% of the allocated scheme was 0.61. The calculated AUC from the

RF method and 10% of the allocated scheme for clinical mastitis was 0.57 while for 90% of the allocated scheme was 0.60. By using DRP as a response variable, GBLUP ranked individuals more accurately than RF (higher accuracy for GBLUP) but the superiority of GBLUP to distinguish precisely the healthy from the sick cows (i.e. AUC) deteriorated and in some allocated schemes, RF performed better than GBLUP. For example, in 80% of the allocated scheme, the realized AUC for infertility from RF and GBLUP were 0.57 and 0.55, respectively. Moreover, the observed AUC for clinical mastitis and 90% of the allocated scheme for RF and GBLUP were 0.60 and 0.59, respectively. The allocated schemes associated with the highest AUC values reflect the disease incidences in the training sets, and as close as possible in the whole population. For example, assigning 80% of the sick cows into the training population represent a disease incidence of 32.5% for claw disorder. This number is close to the whole population disease incidence of 31.9%. An AUC of 0.60 implies the probability of precisely classifying the healthy and sick cows is 60% (based on genomic marker data). While, an AUC of 0.50 indicates that a classification of being healthy or sick is the same as a random guess. The calculated AUC values in this study were in the range found in literature. Tsairidou et al. (2014) reported an AUC value of 0.58 for 1151 animals and a heritability of 0.23 for bovine tuberculosis trait. Although the size of the training set in the current study was higher and expected higher AUC value, but the heritability of the traits in our study was lower than for tuberculosis trait. Gonzalez-Recio and Forni (2011) found the AUC values ranging from 0.58 to 0.70 for a trait with heritability of 0.25. Higher AUC values reported by Nguyen et al. (2015) (from 0.62 to 0.97) can be explained by the higher density (400K) SNP chip panel used by these authors. Hence, heritability, disease incidence, marker density and the size of training sets might be most important parameters affecting AUC.

Theoretical expectations and AUC_{max}

AUC_{max} (maximum AUC value) idea was first introduced by Wray et al. (2010). They expressed that AUC_{max} could be achieved if the classifier test was an ideal predictor of genetic risk. AUCmax is unique for each disease trait due to dependency of the AUC_{max} on the disease incidence (q) and the heritability of the trait based on the liability scale (h_l^2) (Wray et al., 2010). The online calculator developed by Wray et al. (2010) was used to calculate expected values for AUC_{max}. These values can be used as basis of comparison for the actual AUC values obtained in the present study. Maximum expected AUC from a genomic profile that explains all the genetic variance for the

disease prevalence of 0.20 was 0.66 for heritability of 0.1 and 0.77 for heritability of 0.3 (chapter 2). The maximum achievable AUC was 0.62 for claw disorder, 0.61 for clinical mastitis and 0.61 for female infertility (chapter 3). The calculated AUC values in the current study were associated with modest gain in the probability of correctly classified sick and healthy cows compared to random expectations. However, the AUC_{max} in the current study was not high enough to be an ideal predictor of genetic risk.

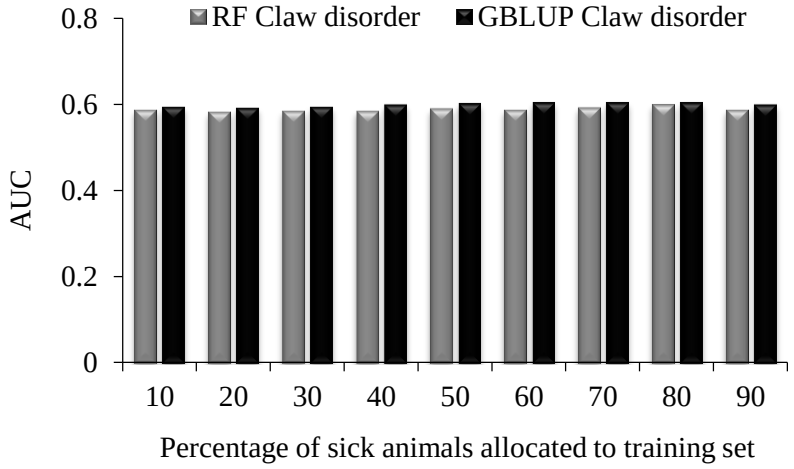

Figure 1: The area under the receiving operating characteristic curve (AUC) using de-regressed proof (DRP) as response variable for claw disorder.

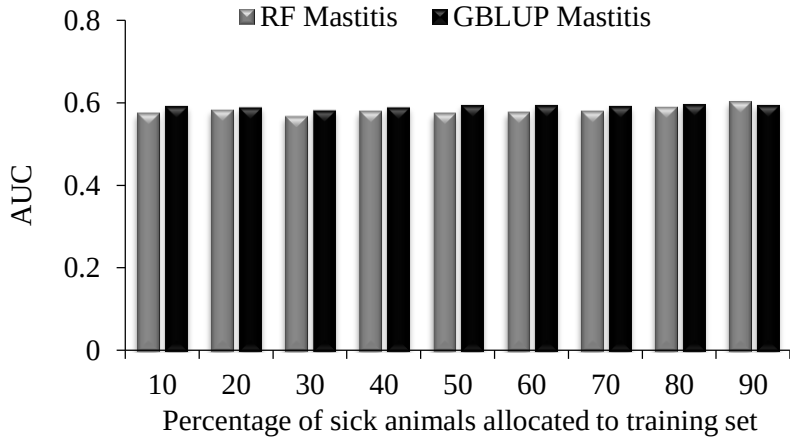

Figure 2: The area under the receiving operating characteristic curve (AUC) using de-regressed proof (DRP) as response variable for clinical mastitis.

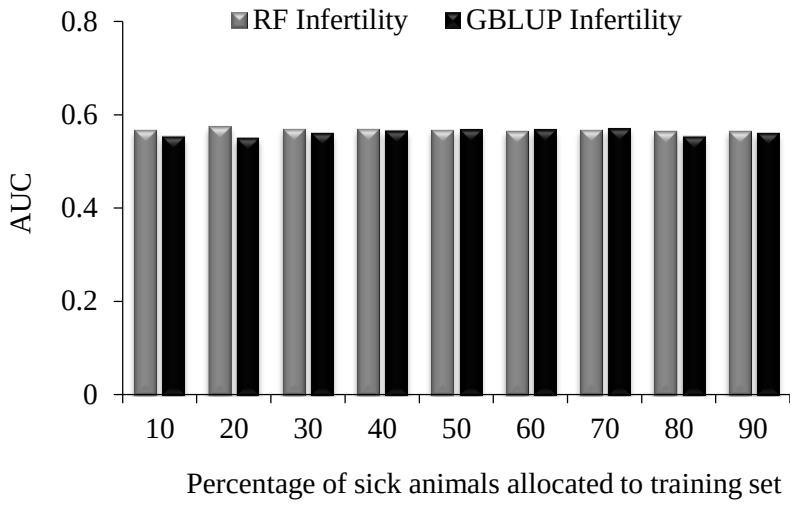

Figure 3: The area under the receiving operating characteristic curve (AUC) using de-regressed proof (DRP) as response variable for infertility.

Utilization SNP correlations from random forest for genome wide association studies

Result of this study verified that RF is a well-suited and promising alternative to classical approaches in terms of loci identification. In terms of detection of associated loci, RF methods have been able to retrieve the most effective QTL in simulation study (chapter 2) and recover most of the loci associated with given disease traits (chapter3) that have already been reported in the previous studies, thereby confirming its relevance in this context. For example, the most important SNP identified by RF on chromosome 6 confirmed results by Klungland et al. (2001), who found a QTL affecting clinical mastitis within a segment of 35.39-70.74 cM on chromosome 6. In addition, Daetwyler et al. (2008) detected a QTL for somatic cell count on chromosome 6 at 72 cM. The power of RF to recover the most effective QTLs (i.e. explaining a high part of the traits' phenotypic variation) increased by using a high marker density (chapter 2). The explanation refers to the fact that RF methodology is based on a random sampling of SNPs. Therefore, by using denser SNP chip panel, it might happen that those SNPs located in close distance to an effective QTL are sufficiently sampled and it means that the QTL signal is captured by SNP located in closer distance. In several studies, RF has been successfully applied to detect SNPs affecting susceptibility to disease (Strobl et al., 2007; Amaratunga et al., 2008; Meng et al., 2009). Botta et al. (2014) reported that RF, a tree-

based ensemble method, could be a robust analysis tool in the context of GWAS due to its intrinsic multivariate and nonlinear attributes.

Although RF reveals the most predictive SNPs based on provided variable importance measures; the real importance values are hard to interpret due to dependency of them on the signal in the data and on the parameters of the algorithm (Genuer et al., 2008). Usually, the top ranked SNPs based on decreasing importance values are declared as important and number of selected SNPs is often arbitrary (Szymczak et al., 2016). However, Breiman and Cutler suggested a classical statistical test by estimating *z-scores* and estimation of asymptotic *p-values* but number of trees which is a tuning parameter in RF affects the power of this test. Several simulation studies showed that when the sample size is small and the effect of the SNPs on the given trait is low, a large number of the SNPs with highest variable importance values are not causally associated to the trait (Kim et al., 2009; Kim et al., 2011). However, in this study, we followed a novel variable selection approach called recurrent relative variable importance measure (r2VIM) proposed by Szymczak et al. (2016). In this method several runs of RF performed, each resulting in importance values measured relative to the minimal importance value and SNPs with large relative variable importance of all the runs are considered as most important SNP.

Detection of selection signature

In this thesis, two methods based on inter-population statistics for recent breed differentiation and alterations on phenotypic scales were used to detect selection signature: F_{ST} and XP-EHH. Yi et al. (2010) has shown that for recently diverged populations, the methods based on inter-population statistics are more robust to identify selection signatures than intra-population. Therefore, the XP-EHH method that identifies long haplotypes representing the most recent selection signature was discussed in chapter 4. It was demonstrated that the XP-EHH method has substantial power in identification of recent selection signature of divergent populations stratified based on geographical characteristics and disease incidences. Using this method detected relevant genes to selection strategy of the populations and disease resistance including *BCL10*, *TRAF3IP3* and *FARS2*. In addition, the outlier approach was used in this study to identify the genomic regions under selection. This method is an effective and widely used approach for the detection of genomic regions under selection regardless of the phenotypes (Narum and Hess, 2011). Applying this approach, the most extreme pattern of variation (typically 1% of top or bottom of a statistic test) will be detected as the

targets of selection. Nevertheless, Akey (2009) confirmed that a detected signal applying outlier approach is not necessarily reflects the regions being under selection. Qanbari and Simianer (2014) argued that the difficulties of deriving a null distribution for a statistical test to reveal candidate region significantly. Therefore, in most of the studies the outlier approach has been applied avoiding specifying a statistical test (Akey et al., 2002).

Signatures of positive selection revealed by F_{ST}

Assessing the variation of marker allele frequencies of different populations is another tool to discover genome wide signature (Holsinger and Weir 2009). One strategy at this point is to use large number of SNPs information in the genome and compare samples from different populations to identify regions of genome with prominent genetic distinction (Gholami, 2014). The statistic F_{ST}, is widely used for the detection of natural selection signature (Porto-Neto et al., 2013) or of the effect of artificial selection on domesticated animals (Yang et al., 2014) and for the genetic differentiation between several populations at any locus (Moradi et al., 2012).

Holstein and DSN

The windowed F_{ST} against position in genome was plotted and shown in Figure 4. The average of differentiation between Holstein and DSN breeds was 0.068 ± 0.051. Considering the top 0.1 percentile of F_{ST}, thirty-nine signature signals were detected. The windowed F_{ST} in detected regions was higher than 0.33 and there was a tendency among the outlier SNPs to be clustered. The detected signals were located on chromosomes 1, 2, 3, 4, 6, 7, 9, 10, 16, 17, 18, 20, 21, 24 and 26. Among the different chromosomes the sharp F_{ST} peak clearly observed on chromosome 1 in the region of 76.62 to 76.73 Mb corresponded to the highest value of F_{ST} (0.46). Moreover, regions of 19.33 to 19.61 Mb on chromosome 4, and 47.61 to 47.80 Mb on chromosome 7 included the highest number of signals. Chromosome 10 with 4 sharp F_{ST} peaks at different regions might be also a target of recent selection.

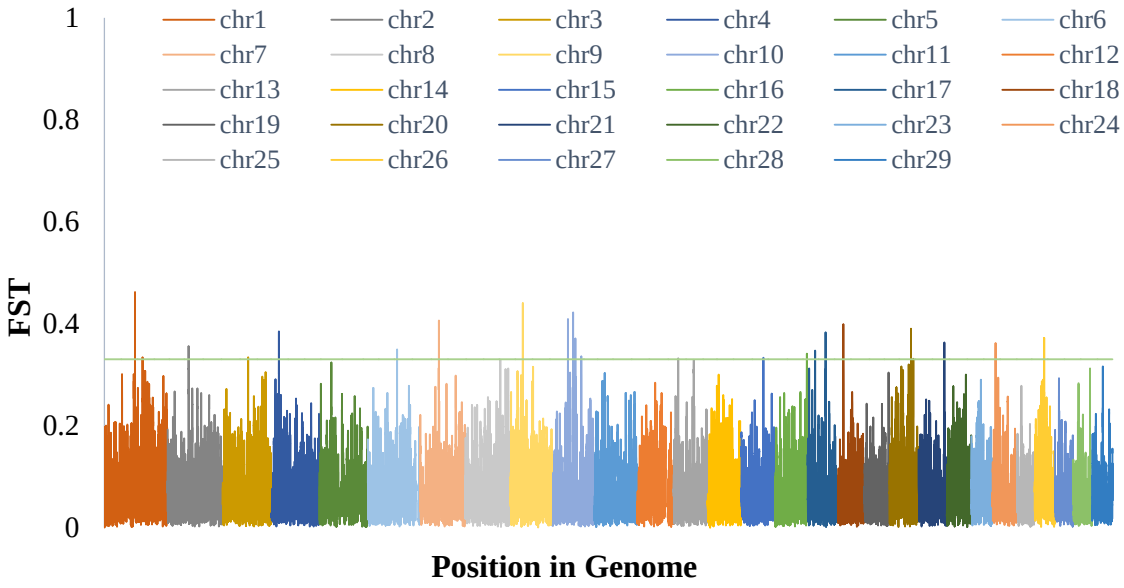

Figure 4. F_{ST} score as a function of chromosome position for Holstein and DSN population.

East-DSN and West-DSN

The plotted windowed F_{ST} against position in genome for East-DSN and West-DSN was shown in Figure 5. There was a very low average of differentiation between West-DSN and East-DSN (0.0046 ± 0.0036). Considering the top 0.1 percentile of F_{ST}, the regions with highest F_{ST} values located on different chromosomes (1, 2, 3, 4, 6, 8, 11, 13, 14, and 17). As shown in this figure, the outlier SNPs were mostly located on chromosomes 1, 11 and 13. Specifically, the evidence of selection was detected in three regions with windowed F_{ST} value > 0.10 on chromosomes 1 at 136.90 Mb, on chromosome 11 at 24.12 Mb and on chromosome 13 at 75.24 Mb respectively.

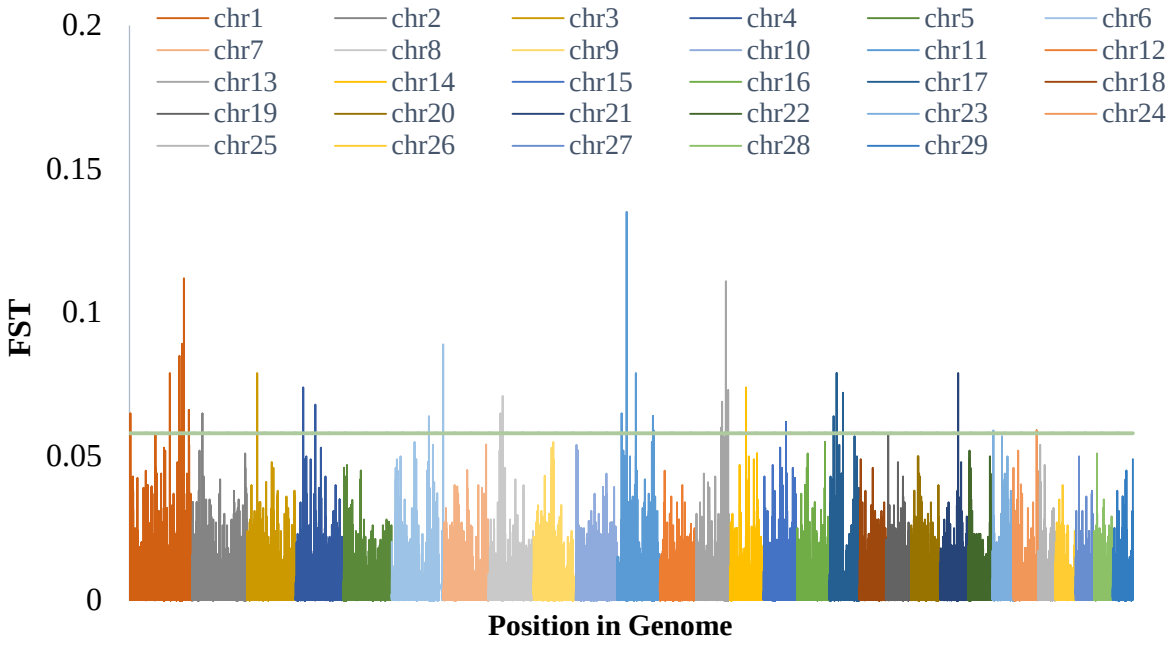

Figure 5. F_{ST} score as a function of chromosome position for East-DSN and West-DSN population.

Sick and healthy Holstein cows

The plotted windowed F_{ST} for healthy and sick Holstein populations against position in genome was shown in Figure 6. There was a low average of differentiation between healthy and sick (0.010± 0.0082). As shown in the figure, most of the windows above the top 0.1 percentile of F_{ST}, were located on chromosomes 4, 10, 12, 14 and 21 with averaged F_{ST} higher than 0.061. The outlier windows on chromosome 21 at 57.94 to 58.47 Mb corresponded to highest F_{ST} value. Detected signals of selection on chromosome 10 located at two regions of 59.15 to 59.47 Mb and 100.88 to 101.35 Mb. In addition, the regions across the position 61.81 to 62.18 Mb on chromosome 4 reflected the high values of F_{ST} ranging from 0.062 to 0.068.

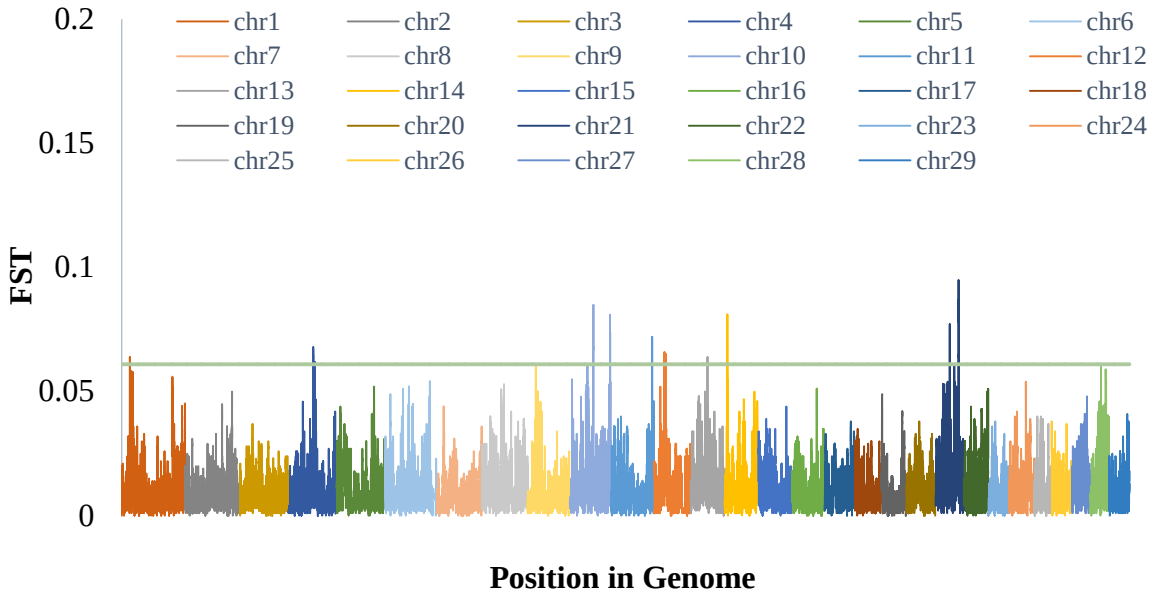

Figure 6. F_{ST} score as a function of chromosome position for healthy and sick Holstein population.

The overlapped signals between the F_{ST} and XP-EHH test was very small and only one region on chromosome 10 were detected through both methods for sick and healthy populations. This result is attributed to the different features of the methods. The F_{ST} and XP-EHH are two method representing population differentiation and antonymous in time scale. F_{ST} detects highly differentiated alleles frequencies between two population, where selection in one region induce a larger frequency differences in comparison to neutrally evolving alleles and reveal the selection signatures that might have been occurred in very long time in the past. The XP-EHH approach identifies selected alleles that have risen to near fixation in one but not another population in the recent generations (Sabeti et al., 2007).

CONCLUSION

Based on the result of this thesis, it can be concluded that the composition of training set is one the most crucial parameters affecting predictions accuracies for both methods GBLUP and RF. The highest prediction accuracy will be achieved when disease incidences in training set is as close as possible to the population disease incidence. Furthermore, also this study revealed increasing prediction accuracy when increasing genetic relationships between training and testing sets. Moreover stronger impact of genetic architecture (number of QTL, level of LD) and of heritabilities

on accuracies of GEBV was identified when applying RF compared to GBLUP. Generally, RF method was more precise than GBLUP to differentiate between healthy and sick animals (higher AUC), especially when increasing the marker density. RF was successfully applied whole-genome screenings to identify important SNP in close distance to a QTL or candidate gene. Using DRP as response variable instead of PCP to prediction genomic values for low heritability health traits corresponded to larger accuracy for both methods of RF and GBLUP. In term of detection of significant SNP, there were strong overlaps using either RF or a GWAS using mixed linear model. Study gene functions revealed that *GAS1, PDE3B, CYP2R1, INSC* (highly expressed in the fat pad of mammary gland) might be potential candidate genes for clinical mastitis, *HMGXB3, CSF1R* (highly expressed in the lymph and body cavities) potential candidate genes for claw disorders, and *OSTN* (expressed in uterus intercaruncular) a potential candidate gene for endometritis. Moreover, application of XP-EHH methodology successfully identified several putative selection signature regions, harbouring genes or QTL being associated with disease and production traits. The detected regions are worthy candidates for further investigation.

REFERENCES

Akey, J.M.; Zhang, G.; Zhang, K.; Jin, L.; Shriver, M. D. (2002). Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Res. 12, 1805–1814.

Akey, J.M., (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res. 19, 711–722.

Amaratunga, D.; Cabrera, J.; Lee, Y. S. (2008). Enriched random forests. Bioinformatics (Oxford, England) 24, 2010–2014.

Ashra,; B.; Edriss, V.; Akdemir, D.; Autrique, E.; Bonnett, D.; Crossa, J.; Janss, L.; Singh, R.; Jannink J. L. (2016). Genomic prediction using phenotypes from pedigree lines with no markers. Crop Sci. 56, 957–964.

Botta, V.; Louppe, G.; Geurts, P.; Wehenkel, L. (2014). Exploiting SNP correlations within random forest for genome-wide association studies. PloS one 9, e93379.

Breiman, L.; Cutle,r A. Random forests [Internet]. Available from: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

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

Daetwyler, H. D.; Pong-Wong, R.; Villanueva, B.; Woolliams, J. A. (2010a). The impact of genetic architecture on genome-wide evaluation methods. Genetics 185, 1021–1031

Daetwyler, H. D.; Villanueva, B.; Woolliams, J. A. (2008), Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one 3, e3395.

Gao, H.; Christensen, O. F.; Madsen, P.; Nielsen, U. S.; Zhang, Y.; Lund, M. S.; Su, G. (2012). Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genetics selection evolution, GSE,6, 44.

Genuer, R.; Poggi, J. M.; Tuleau, C. (2008). Random Forests: Some methodological insights. Research report INRIA Saclay, RR-6729. Available from: http://hal.inria.fr/inria-00340725/fr/

Gernand, E.; Rehbein, P.; von Borstel, U. U.; König, S. (2012). Incidences of and genetic parameters for mastitis, claw disorders, and common health traits recorded in dairy cattle contract herds. J. Dairy Science, 95, 2144–2156.

Gholami, M. (2014). Selection signature detection in a diverse set of chicken breeds. Dissertation, Faculty of Agricultural Sciences, Georg-August-University Göttingen, Germany.

Goddard, M. E.; Hayes, B. J.; Meuwissen, T. H. E. (2010). Genomic selection in livestock populations. Genetics Res. 92, 413–421.

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

González-Recio, O.; Forni, F. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics selection evolution, GSE 43, 7.

Guo, Z.; Tucker, D. M.; Basten, C. J.; Gandhi, H.; Ersoz, E.; Guo, B.; Xu, Z.; Wang, D.; Gay, G. (2014). The impact of population structure on genomic prediction in stratified populations. Theoretical and applied genetics. 127, 749–762.

Habier, D.; Tetens, J.; Seefried, F. R.; Lichtner, P.; Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics selection evolution, GSE, 42, 5.

Holsinger, K. E.; Weir, B. S. (2009). Genetics in geographically structured populations. Defining, estimating and interpreting F(ST). Nature reviews. Genetics 10, 639–650.

Hozé, C.; Fritz, S.; Phocas, F.; Boichard, D.; Ducrocq, V.; Croiseau, P. (2014). Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Science, 97, 3918–3929.

Kim, Y.; Li, Q.; Cropp, C. D.; Sung, H.; Cai, J.; Simpson, C. L. (2011). Performance of random forests and logic regression methods using mini-exome sequence data. BMC Proc. 5, Suppl 9, S104.

Kim, Y.; Wojciechowski, R.; Sung, H.; Mathias, R. A.; Wang, L.; Klein, A. P.; Lenroot, R. K.; Malley, J.; Bailey-Wilson, J. E. (2009). Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. BMC Proc. 3 (Suppl 7), S64.

Klungland, H.; Sabry, A.; Heringstad, B.; Olsen, H.G.; Gomez-Raya, L.; Våge, D.I.; Olsaker, I.; Ødegård, J.; Klemetsdal, G.; Schulman, N.; Vilkk,i J.; Ruane, J.; Aasland, M.; Rønningen, K.; Lien, S. (2001). Quantitative trait loci affecting clinical mastitis and somatic cell count in dairy cattle. Mammalian Genome, 12, 837–842.

Lee, S. H.; Clark, S.; van der Werf, J. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. bioRxiv, 119164

Lourenco, D. A. L.; Tsuruta, S.; Fragomeni, B. O.; Masuda, Y.; Aguilar, I.; Legarra ,A.; Bertrand, J. K.; Amen, T. S.; Wang, L.; Moser, D. W.; Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93, 2653–2662.

Meng, Y. A.; Yu, Y.; Cupples, L. A.; Farrer, L. A.; Lunetta, K. L. (2009). Performance of random forest when SNPs are in linkage disequilibrium. BMC bioinformatics 10, 78.

Moradi, M. H.; Nejati-Javaremi, A.; Moradi-Shahrbabak, M.; Dodds, K. G.; McEwan, J. C. (2012). Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC genetics 13, 10.

Narum, S. R.; Hess, J. E. (2011). Comparison of F(ST) outlier tests for SNP loci under selection. Molecular ecology resources 11, 184–194.

Nguyen, T. T.; Huang, J. Z.; Wu, Q.; Nguyen, T.; Junjie, M. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics 16, S5

Nirea, K. G.; Sonesson, A. K.; Woolliams, J. A.; Meuwissen, T. H. E. (2012). Strategies for implementing genomic selection in family-based aquaculture breeding schemes. Double haploid sib test populations. Genetics selection evolution, GSE 44, 30.

Porto-Neto, L. R; Lee, S. H.; Lee, H. K.; Gondro, C. (2013). Detection of signatures of selection using Fst. Methods Mol. Biol. 1019, 423–436.

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

Qanbari, S.; Simianer, H., 2014. Mapping signatures of positive selection in the genome of livestock. Livest. Sci. 166, 133–143.

Saatchi, M.; McClure, M. C.; McKay, S. D.; Rolf, M. M.; Kim, J. W.; Decker, J. E.; Taxis, T. M.; Chapple, R. H.; Ramey, H. R.; Northcutt, S. L.; Bauck, S.; Woodward, B.; Dekkers, J. C. M.; Fernando, R. L.; Schnabel, R. D.; Garrick D. J.; Taylor, J. F.. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics selection evolution, GSE 43, 40.

Sabeti, P. C.; Varilly, P.; Fry, B.; Lohmueller, J.; Hostetter, E.; Cotsapas, C. et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913–918.

Strobl, C.; Boulesteix, A. L.; Zeileis, A.chim; Hothorn, Torsten (2007), Bias in random forest variable importance measures. Illustrations, sources and a solution. BMC bioinformatics 8, 25. 25.

Su, G.; Christensen, O. F.; Janss, L.; Lund, M. S. (2014). Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Science 97, 6547–6559.

Szymczak, S.; Holzinger, E.; Dasgupta, A.; Malley, J. D.; Molloy, A. M.; Mills, J. L.; Brody, C. L.; Stambolian, D.; Bailey-Wilson, J. E.. (2016). r2VIM. A new variable selection method for random forests in genome-wide association studies. BioData mining 9, 7.

Tsairidou, S.; Woolliams, J. A.; Allen, A. R.; Skuce, R. A.; McBride, S. H.; Wright, D. M.; Bermingham, M. L.; Pong-Wong, R.; Matika, O.; McDowell, S. W. J.; Glass, E. J.; Bishop, S. C. (2014). Genomic prediction for tuberculosis resistance in dairy cattle. PloS one 9, e96728.

Villumsen, T. M.; Janss, L.; Lund, M. S. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie 126, 3–13.

Wang, Q.; Yu, Y.; Yuan, J.; Zhang, X.; Huang, H.; Li, F.; Xiang, J. (2017). Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei. BMC genetics 18,45.

Wientjes, Y. C.; Veerkamp, R. F.; Calus, M. P. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193, 621–631.

Wray, N. R.; Yang, J.; Goddard, M. E.; Visscher, P. M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS genetics 6, e1000864.

Yang, S.; Li, X.; Li, K.; Fan, B.; Tang, Z.(2014). A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. BMC Genetic. 15, 7.

Yi, X.; Liang, Y.; Huerta-Sanchez, E.; Jin, X.; Cuo, Z. X. P.; Pool, J. E. et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science (New York, N.Y.) 329, 75–78.

Zhang, Z.; Ding, X.; Liu, J.; Zhang, Q.; de Koning, D . J. (2011). Accuracy of genomic prediction using low-density marker panels. J. Dairy Science 94, 3642–3650.